T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANA BİLİM DALI
İSTATİSTİK BİLİM DALI

VERİ MADENCİLİĞİ TEKNİKLERİ İLE MOBİL TELEKOM SEKTÖRÜNDE MÜŞTERİLERİN KREDİ SKORLAMASINA İLİŞKİN İSTATİSTİKSEL BİR ANALİZ

Yüksek Lisans Tezi

KUBİLAY KARAKUŞ

İstanbul, 2009

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANA BİLİM DALI
İSTATİSTİK BİLİM DALI

VERİ MADENCİLİĞİ TEKNİKLERİ İLE MOBİL TELEKOM SEKTÖRÜNDE MÜŞTERİLERİN KREDİ SKORLAMASINA İLİŞKİN İSTATİSTİKSEL BİR ANALİZ

Yüksek Lisans Tezi

KUBİLAY KARAKUŞ

Danışman: Yrd.Doç.Dr. İ. ESEN YILDIRIM

İstanbul, 2009

Marmara Üniversitesi
Sosyal Bilimler Enstitüsü Müdürlüğü

Tez Onay Belgesi

EKONOMETRİ Anabilim Dalı İSTATİSTİK Bilim Dalı Yüksek Lisans öğrencisi KUBİLAY KARAKUŞ nın VERİ MADENCİLİĞİ TEKNİKLERİ İLE MOBİL TELEKOM SEKTÖRÜNDE MÜŞTERİLERİN KREDİ SKORLAMASINA İLİŞKİN İSTATİSTİKSEL BİR ANALİZ adlı tez çalışması ,Enstitümüz Yönetim Kurulunun 16.07.2009 tarih ve 2009-12/34 sayılı kararıyla oluşturulan jüri tarafından oybirliği/oyçokluğu ile Yüksek Lisans Tezi olarak kabul edilmiştir.

	Öğretim Üyesi Adı Soyadı	İmzası
Tez Savunma Tarihi	: 28.7.2009	
1) Tez Danışmanı :	**YRD. DOÇ.DR. ESEN YILDIRIM(ZEREN)**	.....
2) Jüri Üyesi	: **DOÇ. DR. DİLEK ALTAŞ**	.....
3) Jüri Üyesi	: **YRD. DOÇ.DR. İBRAHİM EDİN**	.....

ÖZET

Risk yönetimi; riski öngörerek zarar oluşmadan gerekli aksiyonların alındığı sürekli güncellenen bir süreçtir. Bu süreci oluşturabilmek ve yönetebilmek için, geleceği öngörme becerilerine ve araçlarına sahip prediktif bir kurum olmak gerektir. Rekabetin artması ile bir yandan yeni abone almanın zorlaşması, ekonomik çalkantıların meydana gelmesi ile de artan şüpheli alacak tutarı arasındaki optimum dengenin yakalanması günümüzde mobil operatörlerin yönetmeleri gereken kredi risk sorunlarının en önemlilerinden bir tanesidir.

Veri Madenciliği altyapısı üzerinde kurulan Davranışsal Kredi Skorlama sistemi ile şüpheli alacaklara neden olacak abonelerin önceden tahmin edilerek zamanında müdahale ile zararın minimize edilmesi amaçlanmaktadır. Geçmişte oluşan şüpheli alacakların yapısı, veri madenciliği uygulamaları kullanılarak analiz edilip modellenmiştir. Özellikle ödeme performansını yansıtan değişkenlerin yoğun olduğu veri setinde farklı Veri Madenciliği algoritmaları denenmiştir. Lojistik Regresyon, Karar Ağaçları ve Yapay Sinir Ağ modelleri söz konusu veri setine uygulanmıştır. Birbirlerine yakın ve tutarlı sonuçlar vermesine rağmen Lojistik Regresyon modeli bu yöntemler içerisinde en başarılısı olmuştur.

Anahtar kelimeler: Davranışsal Kredi Skorlama, Veri Madenciliği, Lojistik Regresyon, Yapay Sinir Ağları, Karar Ağaçları

SUMMARY

Risk management is continuously updated process in which required actions are taken before existence of loss by using predictive analytics. Creating and managing this process is necessarily required to be a predictive enterprise that has prediction ability and tools. As a result of aggressive competition, acquiring new subscribers is getting more difficult and as a result of economical fluctuation and ambiguity bad debt amount is increasing in these days. Carrying out balance between these two issues is one of the most important credit risk problems in mobile telecom operators.

It is aimed to minimize bad debt loss by predicting risky subscribers formerly by using behavioral credit scoring based on data mining methods and infrastructure. Structure of formerly occurred bad debtors were analyzed and modeled by using data mining techniques. Various data mining algorithms applied on the data set in which there are particular variables that reflects payment performance. Logistic regression, decision tree and neural network were applied to data set. Although they generated consistent and correspondent results, logistic regression was found as the most successful method that performed well on the data set.

Key words: Behavioral Credit Scoring, Data Mining, Logistic Regression, Neural Networks, Decision Trees

ÖNSÖZ

Çok büyük bilişim altyapılarına sahip firmalarda operasyonel işlemler sonucunda birbirinden farklı alanlarda üretilip depolanan dağınık veri parçalarının bir araya getirilerek, kullanılabilir bilgiye dönüştürülmesi gerçek bir işletmecilik başarısıdır. Bu başarı ancak Veri Madenciliği teknikleri ile sağlanabilir. Veri Madenciliğinin önemi gittikçe daha fazla anlaşılmaktadır ve yeni alanlarda denenmektedir. Veri Madenciliğinin iddialı olduğu alanlardan bir tanesi de Kredi Skorlama konusudur. Bu çalışma, ekonominin en dinamik sektörü olan mobil telekomünikasyon alanında yapılmış ve farklı disiplinlerin bir arada nasıl etkin bir şekilde kullanılabileceğine dair önemli bir uygulama aktarılmıştır.

Tezimin hazırlanması sürecinde bana değerli eleştirileri ile yol gösteren tez danışmanım Yrd. Doç. Dr. İ. Esen Yıldırım'a, tezimin her aşamasında fikirleri ile katkıda bulunan değerli çalışma arkadaşım Hürcan Coşkun'a, veri konusundaki engin bilgi birikimini paylaşan Hakan Sarıbıyık'a, modelleme konusundaki deneyimiyle beni destekleyen Selim Deliloğlu'na çok teşekkür ederim. Ayrıca bana karşı anlayış destekleri için yöneticilerime, kardeşlerim Muradiye ve Kenan'a ve tabi ki anneme sonsuz teşekkürlerimi sunuyorum.

İstanbul, 2009 Kubilay KARAKUŞ

İÇİNDEKİLER

SAYFA

TABLO LİSTESİ

ŞEKİL LİSTESİ

KISALTMALAR

ABK : Anahtar Başarı Kriterleri *(Key Performance Indicators - KPIs)*

ASCII : American Standard Code for Information Interchange

BDDK : Bankacılık Düzenleme ve Denetleme Kurumu

BKS : Bilgi Keşif Süreci *(Knowledge Discovery Process- KDP)*

BSM : Başvuru Skorlama Modeli *(Application Scoring Model – ASM)*

BPN : Back Propagation Neural Network

BT : Bilgi Teknolojileri *(Information Technology – IT)*

BTN : Bellek Temelli Nedenleme *(Memory Based Reasoning - MBR)*

CART : Classification and Regression Tree

CBR : Case Based Reasoning

CDR : Call Detail Records

CHAID: Chi-Square Automatic Interaction Detector

ÇDURE: Çok Değişkenli Uyarlamalı Regresyon Eğrileri *(Multivariate Adaptive Regression Splines – MARS)*

ÇKKD : Çok Kriterli Karar Destek *(Multicriteria Decision Aid - MCDA)*

DSGY : Doğru Sınıflandırılmış Gözlemlerin Yüzdesi *(Percentage Correctly Classified Cases – PCC)*

DSM : Davranışsal Skorlama Modeli *(Behavioural Scoring Model – BSM)*

DVM : Destek Vektör Makineleri *(Support Vector Machine – SVM)*

EAS-VMS: Endüstriler Arası Standart Veri Madenciliği Süreci *(CRoss-Industry Standard Process for Data Mining - CRISP-DM)*

EKFK : Eşit Kredi Fırsatı Kanunu *(Equal Credit Opportunity Act – ECOA)*

EKK : En Küçük Kareler

FMS : Fraud Management System

GP : Genetik Programlama

GSM : Global System for Mobile Communications

HTML : Hypertext Markup Language

ISAS : Internet Sahtekârlık Alarm Sistemi

İZ : İş Zekâsı *(Business Intelligence - BI)*

KA : Karar Ağaçları *(Decision Tree – DT)*

KL : Kredi Limiti

KDS : Karar Destek Sistemi *(Decision Support System - DSS)*

KKB : Kredi Kayıt Bürosu

KMS : Kısa Mesaj Servisi *(Short Message Service – SMS)*

KRS : Kredi Referans Sistemi

LDA : Lineer Diskriminant Analizi

LKS : Limit Kontrol Sistemi

LP : Lineer Programlama

LR : Lojistik Regresyon

MERNİS: Merkezi Nüfus İdaresi Sistemi

ML : Maksimum Likelihood

MLP : Multi-Layer Perception

NB : Naive Bayes

NBD : Net Bugünkü Değer *(Net Present Value – NPV)*

OLAP : Online Analytical Processing

OLTP : Online Transaction Processing

ÖİG : Özel İlgili Grubu *(Special Interest Group - SIG)*

QDA : Quadratik Disktriminant Analizi

RBF : Radial Basis Function

RDMBS: Relational Database Management Systems

ROC : Receiver Operating Characteristic

SOM : Self-Organizing Map

SYS : Sesli Yanıt Sistemi *(Interactive Voice Response – IVR)*

TANs : Tree Augmented Naive Bayes Classifiers

TBA : Temel Bileşenler Analizi *(Principal Component Analysis - PCA)*

TCKN : Türkiye Cumhuriyeti Kimlik Numarası

TCMB : Türkiye Cumhuriyeti Merkez Bankası

TK : Telekomünikasyon Kurumu

UTADIS: UTilités Additives DIScriminantes

VBK : Veriden Bilgi Keşfi *(Knowledge Discovery from Data-KDD)*

VM : Veri Madenciliği (*Data Mining - DM*)

VTYS : Veri Tabanı Yönetim Sistemleri *(Database management systems - DBMSs)*

WWW : World Wide Web

YSA : Yapay Sinir Ağı *(Artificial Neural Network - ANN)*

YSD : Yapısal Sorgulama Dili *(Structured Query Language-SQL)*

YZ : Yapay Zekâ (Artificial Intelligence - AI)

1. GİRİŞ

Karmaşıklaşan ve belirsizleşen bir dünyada geleceği öngörmek önemini gittikçe daha da artırmaktadır. Bunu sağlayabilmek için prediktif bir kurum olmak gerekir. Çünkü prediktif kurumlar geleceği öngörme ve yönetme becerilerine ve ilgili araçlara sahiptir. Bu kurumların kullandığı temel yöntem "iş zekâsı" dır. İş zekâsı, mevcut iş koşullarını doğru anlamak ve organizasyon tarafından üretilen tüm iş verilerini analiz ederek anlamlı bilgiye dayalı iş kararları almaktır. İş zekâsı sonuçlarının kullanılabileceği, uygulamaya yönelik bir hedefi vardır.

İş zekâsı, geçmişe bakarak bugünü anlayarak yönetmek, geleceği öngörerek kazanmaktır. Firmaları stratejik olarak ön plana çıkaracak pek çok bilginin doğru değerlendirilebilmesi, bu kavram ile anlatılır. İş zekâsı kavramının en önemli bileşenlerinden bir tanesi Veri madenciliğidir. MIT (Massachusetts Institute of Technology) Technology Review'e göre veri madenciliği dünyayı değiştirecek on gelişen teknolojiden bir tanesidir.

1980'li yıllardan itibaren finansal sermayenin küresel işbirliğini artırması ile kredi sektöründe rekabetin artması, internet ve mobil bankacılık uygulamalarının gelişmesi, pek çok farklı yeni kredi ürününün geliştirilmesi, hızla gelişen kredi pazarlama faaliyetleri, gelişmiş risk yönetimi tekniklerine dolayısıyla kredi skorlamaya ihtiyacı artırmıştır. Gelişen kredi piyasalarında karlılığı artırmak ve maliyetleri düşürmek için kredi riskinin doğru olarak ölçülebilmesine olanak sağlayan yöntemler de gün geçtikçe daha önemli hale gelmektedir.

Kredi risk değerlemesi, finansal analiz alanında çok önemli bir yere sahiptir. Risk değerlemesi konusunda, sektör içinde ve akademik alanda yapılan pek çok araştırma mevcuttur. Kredi skorlama modelleri olarak da anılan pek çok önemli teknik, özellikle banka verileri kullanılarak geliştirilmiş ve gerçek hayata ait uygulamalarda başarılı sonuçlar elde edilmiştir. Kredi skorlaması, sadece başlangıçta kredi verme aşamasında veya bir defaya mahsus olarak kullanılmayıp, müşterinin hayat eğrisinin tüm aşamalarında anahtar bir role sahiptir.

Kredi skor modellerinin geliştirilmesi, kredi sektöründeki rekabetin artması ve şüpheli alacak sorunları nedeniyle çok önemli bir konu haline gelmiştir. Bundan dolayı kredi skorlama modellerinin çoğu, model sonuçlarının doğruluğunu artırmak için istatistiğin çeşitli alanlarında yaygın bir şekilde çalışılmaktadır. Kredi skorlama modellerini kullanmanın

avantajları; kredi analizlerinin maliyetini azaltmak, daha hızlı kredi kararı almak, kredi tahsilâtlarını garantiye almak ve olası riskleri azaltmak olarak ifade edilebilir.

Kredi skorlama projesi aslında tam bir entegrasyon prosesidir. Çünkü başta kredi büroları olmak üzere farklı dış kaynaklardan, daha sonra da içerideki farklı kaynaklardan veriler alınarak işlenir ve hazırlanan sonuçlar da kullanılmak üzere farklı sistemlere gönderilir. Abone değerleme konusunda Kredi Risk bölümünün yanı sıra Müşteri Hizmetleri, Pazarlama ve Satış bölümleri, hatta Hukuk bölümlerinin yakın işbirliği içerisinde çalışması gerekir.

Kredi skorlaması ile abonenin kredi riskini tespit edecek yeni teknikler geliştirilerek firmanın şüpheli alacaklarının azaltılması sağlanırken, yeni satışlardan kaynaklanan geliri de maksimize edilir. Bu iki temel amacın yanı sıra, riski yüksek abonenin elenmesi ve risksiz abonenin şebekeye dâhil edilmesi ile Müşteri Hizmetleri tarafından gerçek aboneye daha fazla hizmet sunulur, ayrıca aynı network yatırımı ile daha kaliteli iletişim hizmeti sağlanır.

Hemen tüm sektörlerde olduğu gibi telekomünikasyon sektöründe de abonenin anlaşılması için toplanan veriler çok büyük miktarlara ulaşmıştır. Çoğunluğu operasyonel sistemler tarafından toplanan bu veriler, pazarlama ve satış faaliyetleri kadar risk alanında da karar verme amacıyla kullanılmaktadır. Sektördeki zorlu rekabet şartlarından dolayı yeni abone kazanmanın ve mevcut müşteriyi elde tutmanın önemi artsa da, ödenmeyen faturaların neden olduğu zararın finansal boyutları da ekonomik çalkantıların olduğu dönemlerde çok önemli boyutlara ulaşmaktadır. Faturalı mobil telekomünikasyon abonelerinin fatura ödeme kabiliyetlerinin sürekli takip edildiği etkin sistemlere sahip olmak, tüm operatörler için kesin bir zorunluluktur.

Mobil telekomünikasyon abonelerinin davranışsal skorlama modelinin oluşturulmasının bazı zorlukları mevcuttur. Bunlardan birincisi, verinin kendisi ile ilgilidir. Mobil abonelerin ürettiği veriler çok boyutludur; aktivasyon anında alınan ve değişmeyen özelliklerin yanı sıra, aylık fatura hareketlerine, günlük işlem ve anlık konuşma verilerine sahiptir. Bu veriler de özellikleri gereği farklı sistemler tarafından oluşturulur, farklı formatlarda ve veri yapılarında depolanırlar. İkinci olarak, kredi değerlendirme altyapısının her aboneliği ayrı olarak mı yoksa diğer faturalı ve ön ödemeli hatları ile birlikte mi ele alacağına karar verilmesi gerekir. Her iki seçimin de birbirlerine göre avantaj ve dezavantajları mevcuttur.

Bu çalışmada, veri madenciliği altyapısı irdelenerek kredi skorlama sürecinde nasıl etkin olarak kullanılabileceği mobil telekomünikasyon sektörüne ait bir uygulama veri seti üzerinden aktarılmaya çalışılmıştır.

Çalışma toplam yedi alt bölümden oluşmaktadır. Genel çerçeveyi aktaran "Giriş" bölümünden sonra aktarılan ikinci bölümde, tezin bütünlüğünü sağlamak açısından tüm üretim süreçlerinin ihtiyaç duyduğu veri konusu, "Veri Ambarı" alt başlığında yer almaktadır. Verilerin kullanıma uygun şekillerde nasıl depolandığı ve teknolojide gelinen son durumda nasıl yönetildiği konuları bu bölümde sunulacaktır.

Üçüncü bölümde, 1980'lerden sonra tüm iş dünyasında kendini giderek daha fazla kabul ettiren "iş zekâsı" kavramının en önemli bileşeni olan veri madenciliği (VM) kavramı aktarılmaktadır. Veri madenciliğinin geçirdiği süreçler kısaca anlatıldıktan sonra, özellikle VM metodolojisi ele alınmış ve VM tekniklerinden bahsedilerek bu alandaki yeni trendlerden bahsedilecektir.

Dördüncü bölümde, Kredi Skorlamanın temel kavramları, başlıca uygulama alanları, aşamaları, geliştirilme süreçleri ve kullanılabileceği diğer alanlar ele alınmıştır. Kredi skorlamanın ürünü olarak nitelendirilebilecek "skor kart" ın nasıl geliştirileceği özet olarak aktarılmıştır.

Beşinci bölümde ise, bankacılık sektörü başta olmak üzere telekomünikasyon ve perakendecilik sektöründe gerçekleştirilen uygulamalar aktarılmaktadır. Burada özellikle, çalışmanın da konusu olan bireysel abonelerin incelendiği çeşitli örnekler de yer almaktadır.

Çalışmanın altıncı bölümünde, Türkiye'de mobil telekomünikasyon sektöründe faaliyet gösteren bir firmaya ait belirli bir abone kitlesi için elde edilen örnek veri seti kullanılarak davranış skorlama modelleri geliştirilmiştir. Geliştirilen modeller Clementine 12.0 ve SAS Data Miner 4.5 programlarında hazırlanmıştır. Bu amaçla oluşturulan veri setinde yer alan 10.000 bireysel aboneye ait demografik özellikler, ödeme performansı, işlem çeşitlerini içeren pek çok veri kullanılmıştır. Örnek veri seti üzerinde kurulan farklı modellere ait sonuçlar daha sonra uygun yöntemler ile karşılaştırılarak elde edilen sonuçlar yorumlanmıştır.

Sonuç bölümünde ise, araştırma problemine ne kadar uygun çözümler getirildiği tespit edilecek ve hazırlanan modellerin sonuçları yorumlanarak genel olarak değerlendirilecektir. Bunun yanı sıra, kredi skorlama konusunda tüm sektörün önündeki engeller ve bazı çözüm yolları aktarılmıştır.

2.VERİ TABANLARI

Karşı karşıya kalınan problemlerin karmaşıklığının sürekli artması ve değişimin hayatın her alanında hızlanması, bu aşamaların her birinde karar vericinin işini hayli zorlaştırmıştır. Günümüzde teknolojideki baş döndürücü ilerlemeler sonucu, bilgi saklama yöntem ve araçlarının ucuzlaması ve yaygınlaşması, toplanan bilgi miktarının önceki zamanlarla karşılaştırılamayacak derecede artmasına yol açmıştır.

Tarih boyunca karar verici pozisyonunda olan her kişi için 'karar verme işlemi' sekiz temel aşamadan oluşmuştur[1]:

1. Problemin tanımlanması
2. Amacın belirlenmesi
3. İlgili verinin toplanması
4. Yapılabilir alternatiflerin tanımlanması
5. Alternatiflerin muhakeme edilmesi için gerekli kriterlerin belirlenmesi
6. Karşılıklı ilişkilerin modellenmesi
7. Alternatiflerin sonuçlarının tahmin edilmesi
8. En iyi alternatifin seçilmesi

2010'larda dünya üzerindeki bilgi tabanının her 11 saatte iki katına çıkacağı tahmin edilmektedir[2]. Bilgi dağları olarak da nitelenen bu bilgi yığınları içerisinden istediğimiz bilgilere zamanında ulaşmak mevcut durumda bile kolay olmadığı gibi giderek daha da zorlaşacaktır.

Geçmişe baktığımızda kullanılan verinin hemen tamamı araştırma amaçlı olarak toplanmış veya üretilmiştir. Günümüzde ise iş dünyasında çok yüksek miktarlı operasyonel veri mevcuttur. Bu operasyonel veri aslında ilk başta analizlerde kullanılmak amacıyla oluşturulmamıştır, bu nedenle yapılan işlem fırsat oluştuğunda kullanmak üzere verilerin toplanması olarak tanımlanabilir. Bu durum, değişkenlerin belirli problemlere cevap vermek

[1] Newman D. G. **Engineering Economic Analysis**, 2nd Edition, San Jose : Engineering Press, Inc. http://www.msstate.edu/dept/AgEdExp/4163/decision/decision.htm (24.02.2009)
[2] IBM Global Technology Services, **The toxic terabyte: How data-dumping threatens business efficiency**, London: 2006. s. 2.

amacıyla kontrol edilip değiştirilebildiği deneysel veri mantığının karşıtı bir durumu temsil etmektedir[3].

Tablo 1 : Deneysel Veri Üretimi ile Fırsata Dayalı Veri Üretiminin Karşılaştırılması

	Deneysel	Fırsata Dayalı
Amaç	Araştırma	Operasyonel
Değer	Bilimsel	Ticari
Üretim	Aktif olarak kontrol edilebilir	Pasif bir şekilde izlenebilir
Boyut	Küçük	Çok büyük
Hijyen	Temiz	Kirli
Durum	Statik	Dinamik

Bilime ve bilgiye dayalı bir sonuca ulaşabilmek için veri temeldir ve ancak veri mevcut ise analizden bahsedilebilir. Kullanılabilir ve sistematik veri yığınlarını oluşturmak ve istenen veriye erişirken kullanılan yöntem ve araçlar bu bölümün ana konusunu oluşturmaktadır. Bu bölümde, veritabanları ve veri ambarları hakkında özet ve bütünsel bir bakış ortaya konulacaktır.

2.1.Veri

Bilgiye ulaşma olarak isimlendirilen işlemler bütününün ilk adımı veri toplama ile başlar. Verinin tanımı konusunda internetin kütüphanesi olarak nitelendirilebilecek wikipedia'da bulunan tanım şu şekildedir[4]: "Veri; bir tecrübe, gözlem veya deneyin veya bir bilgisayar sistemindeki işlemlerin veya bir önermeler kümesinin sonuçları olarak toplanan gerçeklerin bir koleksiyonunu belirtir. Bu koleksiyonun içerisinde değişkenlerin bir kümesinin ölçümleri ve gözlemleri olarak sayılar, kelimeler veya görüntüleri içerebilir. Veri, enformasyonun ve bilginin türetildiği en düşük seviyedeki soyutlama olarak görülür."

Veri kelimesi Latince "verilen şey" anlamına gelen "datum" un çoğuludur[5]. Bilgisayar ortamında, görüntü ve seslerin de sayılarla ifade edilebildiğini dikkate alırsak, veriyi basitçe sayı ve kelimelerden oluşan bir yığın olarak ele alabiliriz. Bu yığının içerisinden istediğimiz veriyi nasıl bulacağımız sorusunun cevabı, veriyi belli bir düzen içinde bir araya getirmek ve

[3] SAS Institute Inc. **Appliying Data Mining Techniques Using SAS Enterprise Miner Course Notes.** Cary, NC : 2005, s. 3.
[4] http://en.wikipedia.org/wiki/Data (23.02.2009)
[5] Wilkinson, L., **The Grammar of Graphics (Statistics and Computing).** Second Edition, New York: Springer, 2005, s.41.

ulaşmanın etkin yöntemini ortaya koymaktır. Bu düzenin bilgisayar bilimleri alanındaki adı Veritabanı Yönetim Sistemleridir[6].

2.2.Veritabanı Yönetim Sistemleri

Verinin depolanması olarak adlandırılan ve verinin değişik kaynaklardan alınarak, bilgisayarda manyetik bir ortamda kalıcı olarak saklanması işlemi dosyaların yardımı ile olur. Dosyalar, belli bir ismi olan, manyetik ortamda büyüklüğü, içindeki veri miktarı ile doğru orantılı olarak artan, tekst veya binary olarak adlandırılan iki farklı formatta olabilen yapılardır. Tekst formatı önceden belirlenmiş karakter setlerini içerir. Bunların en çok bilineni ASCII karakter setidir. Binary yani ikilik formatı ise tamamen 1 ve 0'lardan oluşur[7].

Bu dosyaların sayısının artması, bu dosyalardan istenen bilgilerin alınmasını zorlaştıran önemli bir faktördür. Bilgisayarların yaygınlaşmaya başladığı ilk yıllarda bu problemin çözümü için, dosya işleme dilleri olarak nitelendirilen bilgisayar dilleri kullanılmıştır. Bu dillerin en yaygınlarından birisi AWK'dır ve günümüzde halen kullanılmaktadır[8]. Gelişen ihtiyaçlar sonucunda ASCII dosyalar ile veri saklama ve AWK gibi diller ile bu dosyaları işlemek yerine, veriyi saklama ve veriye ulaşmak için özel yapılar ve araçlar sağlayan Veritabanı ismi verilen yazılımlara doğru bir yönelim olmuştur.

Bu konuda, literatürde oldukça çok sayıda kaynakta bahsedildiği üzere, bilgisayarlar bit denen 1 ve 0'lardan oluşan veriler üzerinde işlemler yapma kabiliyetine sahip makinelerdir. Dolayısı ile veri depolama gerçekte bu bazda yapılır. Yani herhangi bir görüntü ya da ses içeren veriler öncelikle 1 ve 0'lardan oluşan dosyalar haline getirilirler. Bu işleme sayısallaştırma denir. Bunun dışında günlük hayatta karşılaştığımız sayı ve kelimeler de 1 ve 0'lara çevrilir. Kelimeler ASCII olarak nitelendirilen standart sayılara karşılık gelen bir kodlama ile bilgisayarda depolanır[9]. Dolayısı ile her şey 1 ve 0'lardan oluşan bir dosya yapısı şeklinde manyetik ortama aktarılır. Bu dosyaların içeriğine 'ham veri' ismi verilir, yani bir bütünlük oluşturmalarını sağlayacak mantıksal bir yapıya sahip değillerdir.

---

[6] Ramakrishnan, R. and Johannes Gehrke, **Database Management Systems**, Third Edition, New York: The McGraw-Hill Companies, 2003, s.4.
[7] Hyde R., **The Art of Assembly Language Programming.** San Francisco: No Starch Press, 2003, s.11.
[8] Close, Diane B., Arnold D. Robbins, Paul H. Rubin, Richard Stallman, Piet van Oostrum, **The AWK Manual,** Edition 10, 1995. s. 1.
[9] Kodlamalar için bknz http://www.asciitable.com (14.02.2009)

Veritabanlarında, verilerin birbiri ile olan anlamsal, yapısal ya da kullanıma dair ilişkileri önemli rol oynar. Veritabanı yönetim sistemleri adını alan bu yapıların üç ana probleme çözüm sunan yazılım araçları olduğunu düşünebiliriz[10].

a) Verinin depolanması

b) İstenen verilere hızlı ve en az maliyetle ulaşılması

c) Verilerin güvenliğinin sağlanması

Ham verilerin işlenmesi için belirli sistematik yapı geliştirilmesi ve depolama esnasında sürekli yapılan işlemlerin hızlı ve en az zahmetle yapılmasının sağlanması çok önemlidir. Aynı zamanda istenen verilere hızlı ve en az maliyetle nasıl ulaşılacağı da çok önemlidir. Bunun için indeksleme, sıralama, rastgele hale getirme (hashing) gibi ileri teknikler kullanılır[11].

Hangi veriye kimlerin erişebileceğinin kontrolü de verilere ulaşmanın kolaylaştırıldığı bir ortamda çok daha önem kazanmaktadır. Bu nedenle bilgiye erişim ve bilgi güvenliği günümüzün en büyük güvenlik sorunlarından birisi haline gelmiştir.

Bütün bu özellikleri bir arada sunan bir yapıyı her bilgisayar kullanıcısının baştan kurması tabii ki düşünülemez. Bu nedenle, bu üç temel probleme çözümler üreten yazılımlar geliştirilmiştir. Bu yazılımlara Veritabanı Yönetim Sistemleri (VTYS) denir.

Veritabanı yönetim sistemlerinin üstünlükleri aşağıdaki gibi özetlenebilir[12] :

- Verinin tekrarlanmasını engeller (aynı verinin birden çok uygulama tarafından ortak kullanılacağı şekilde bütünleşik bir yapı mevcuttur)
- Verinin tutarlı olmasını sağlar (verinin doğruluğunu sağlayacak uygun filtrelere sahiptir)
- Aynı andaki erişimlerde tutarsızlıkların ortaya çıkmasını engeller (veritabanı uygulamalarında veritabanı nesneleri farklı uygulamalar tarafından paylaşılmasına rağmen yapılan işlemlerde tutarsızlık olmasını engeller)
- Verinin güvenliğini sağlar (farklı erişim yetkilendirmeleri tanımlayarak, tüm kullanıcılar için sadece ilgili alanlara erişmeleri sağlanır)

[10] Ramakrishnan, R. and Johannes Gehrke, s.4.
[11] Hand D., Heikki Mannila, and Padhraic Smyth. **Principles of Data Mining.** Massachusetts: A Bradford Book The MIT Pres, 2001, s. 205.
[12] Özkan, Yalçın. **Veri Madenciliği Yöntemleri.** İstanbul: Papatya Yayıncılık Eğitim. 2008, s.15-16.

Veritabanı sistemlerinde veriyi oluşturan her bir birimin diğerleri ile olan ilişkileri mantıksal olarak dikkate alınır ve veri birbirleri ile ilişkilendirilir. Bu ilişki şekli, eldeki verinin içeriğine bağlı olarak değişik Veri Modellerine ihtiyaç duyulmasına yol açmıştır. Veri modelleri verilerin arasındaki mantıksal ilişkiyi tanımlamada hangi kriteri temel aldığımızı ortaya koyar.

Seçilen veri modeline göre oluşturulan veritabanları çeşitlilik gösterir. Bunların en genel olarak gruplaması şu şekilde yapılabilir[13]:

i. **Nesne-yönelimli veritabanları;** Nesneye yönelik programlama paradigmasından hareketle oluşturulan bir veri modeline dayanır. Bu veri modelleri kendi metodlarını ve diğer nesne denen yapılarla mesajlaşmalarını yönetecek yapıları barındırır.
ii. **Nesne ilişkili veritabanları;** Nesne ilişkili veri modelini alır. Veri tipleri; kompleks ağaçlar, grafikler, listeler ve hiyerarşiler olan ilişkisel veritabanlarıdır. Bu veritabanlarında sorgulama yapmak için özel diller gerekmektedir.
iii. **İlişkisel veritabanları;** Veriyi tablolar olarak ele alan ve bu tabloların içindeki verilerin diğer tablolardaki veriler ile olan ilişkilerinin yine tablolarda tutulduğu Veri Tabanı Yönetim Sistemleri (VTYS)'dir. En yaygın olan yapıdır. İlişkilerin kurulma şekline göre alt gruplara ayrılabilir; yıldız yapılı (Star schema), kartanesi yapılı (snowflake schema) yapılar gibi. En yaygın VTYS'leri; Oracle, Sybase, Access, Mysql, Informix, SQL Server dir.

   İlişkisel veritabanlarında veriyi sorgulamak için özel diller kullanılır. Bu dillerden en yaygın olanı Yapısal Sorgulama Dili (YSD) *(Structured Query Language-SQL)* dir. Bu dil 30 kadar temel komutu içeren deklaratif yapıda bir dildir. Kullanıcı istediği bilgileri ve ilişkilerini belli kurallar çerçevesinde tanımlar ve YSD sorgu işleyicisi bu sorgudan yola çıkarak veriyi getirir. YSD üç temel işlem yapmaya izin verir.

   a) Veriyi değiştirme (veriyi getirme ve değiştirme imkânı verir)

   b) Veriyi tanımlama (yapıları tanımlamayı, tabloları yaratmayı, ilişkileri kurmayı sağlar)

   c) Veri kontrolü (veritabanını kullanan kullanıcıların hak ve yetkilerini belirlemeyi sağlar)

[13] Han, J. and Micheline Kamber, **Data Mining: Concepts and Techniques.** Second edition, San Francisco: Morgan Kaufmann Publications, 2006, s.12.

iv. **İşlem bazlı veritabanları;** Her bir işlemin ayırt edici bir tanımlayıcısının olduğu kayıtlı dosyalardan oluşan veritabanlarıdır. Alışveriş işlemlerinin kaydedildiği sistemlerde kullanılabilir.
v. **Uzaysal (spatial) veritabanları;** Bunlar iki nokta arasındaki mesafelerin ölçülebildiği görüntü ve haritaların tutulduğu VTYS'dir. İki format yaygın olarak kullanılır, raster formatı ve vektör formatı.
vi. **Zamansal veritabanları;** Bunlar zamanın her bir veri için önemli olduğu VTYS'dir. Güncel veriden daha çok verilerin zaman içindeki değişimi önemlidir. Borsa, hava durumu verileri bu tür verilerdir.
vii. **Metin veritabanları;** Uzun ya da kısa cümleler ve kelimelerden oluşan verilerin tutulduğu VTYS'dir. Verilerin yapısız ya da yapılı olmaları mümkündür. Örneğin; ilaçların veya yemeklerin nasıl yapıldığının tutulduğu yapılar.
viii. **Çoklu medya veritabanları;** Görüntü ve sesin saklanması ve sorgulanmasını kolaylaştıran VTYS'dir. Örneğin bir filmdeki belirli bir sahneyi hızlı bir şekilde bulmamızı sağlayan bir yapı sunar.
ix. **WWW;** World Wide Web, dünya geneline yayılan bilgi ağı olarak nitelendirebileceğimiz bu yapılarda hiperlinkler ve HTML (Hypertext Markup Language) temel araçlardır. Google, yahoo gibi arama motorları günümüzde herkesin kullandığı VTYS sorgulama araçları olmuştur.

2.3. Veri Ambarı

Veritabanları, konu odaklı operasyonel sistemlerdir. Bu sistemlerin yukarıda belirttiğimiz, karar verme pozisyonunda olan kişilere karar vermede yardımcı olması pratik olarak çok zordur. Bu nedenle bir kurumdaki birçok operasyonel veritabanının konsolide bir şekilde tutarlı, zaman boyutunu içerecek şekilde bütünleştirilmesi gerekliliği ortaya çıkmıştır. Bu yapı karmaşık sorgulara cevap verecek şekilde bir optimizasyona sahip olmalıdır. Bu yapılara Veri Ambarı denir.

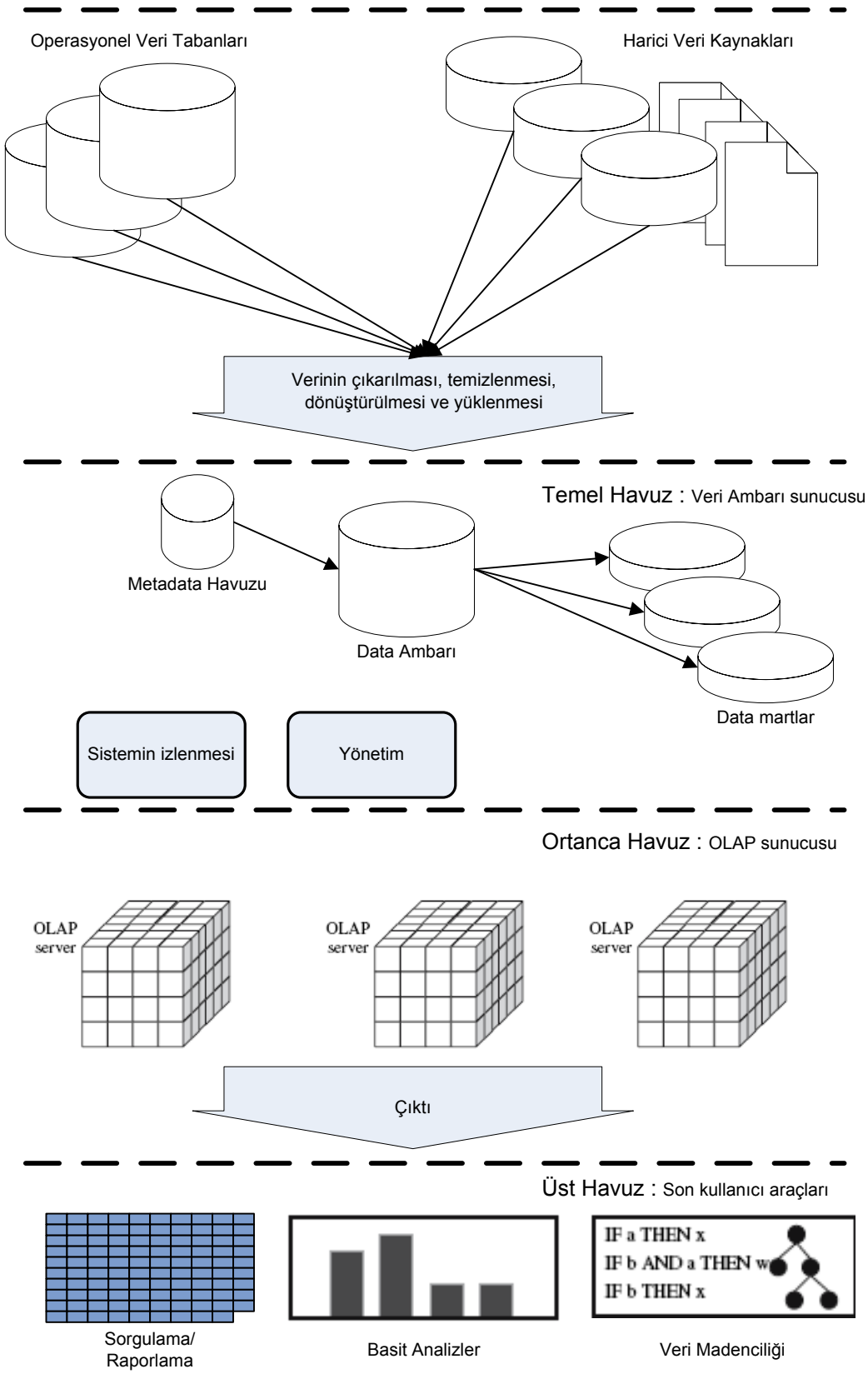
Veri ambarı, birbirleriyle ilişkili farklı veri tabanlarından toplanan verinin konsolide edilmiş halde, düzenli olarak depolandığı yerdir. Veri ambarı, etkin iş kararlarının zamanında alınmasını sağlamak için gerekli olan stratejik bilgiye hızlı ve güvenilir bir şekilde erişmeye imkân verir. Veri ambarının oluşturulduğu kurumda tüm kullanıcılar tek bir kaynaktan veri

aldığı için veri farklılığından kaynaklanan çelişkiler yaşanmaz. Ayrıca ulaşılan verinin doğruluğunu garanti ederek verilecek kararlara güven duyulmasını sağlar.

Veri ambarı, farklı kaynaklar tarafından toplanan verilerin beraber analiz edilmesi amacıyla oluşturulur. Burada yer alan verilerin çoğunluğu operasyonel sistemlerden gelir. Bilişim teknolojilerinin gelişmesiyle ve karar almada bilginin öneminin kavranmasıyla birlikte, tüm özel sektör ve kamu kuruluşları eskiye göre çok daha yüksek miktarda veri üretip bunları depolamaya başlamıştır. Hâlihazırda sayısal ortamda yapılan tüm işlemler kayıt olarak oluşturulup, kullanılmak amacıyla saklanmaktadır. Bankalarda, hastanelerde, telekomünikasyon sektöründe, süpermarketlerde, belediye ve tapu gibi kamu kurumlarında gerçekleşen her işlem kayıt altına alınmakta, depolama ve veri işleme maliyetlerinin de düşmesiyle birlikte farklı amaçlarla kullanılmak üzere veri olarak saklanmaktadırlar.

Veri ambarı oluşturulurken mimari yapı, farklı bölümlerin talep ve gereksinimlerine göre şekillenir. Çünkü amaç iş sorunlarına yanıt arayan son kullanıcıların ihtiyacı olan veriyi hızlı, doğru ve tam olarak üretmektir. Veri madenciliği için Veri ambarı zorunlu değilse de Veri madenciliği uygulamalarının etkinliğini arttıracak özelliklere sahiptir.

Datamart; veri ambarının belirli bir konuya göre düzenlenmesi ile ortaya çıkan alt birimidir. Daha küçük boyutlara sahiptir, belirli kullanıcılar tarafından ulaşılır ve kurumun özel ihtiyaçlarına cevap vermesi için tasarlanırlar. Bu çalışmada kullanılan veri seti de bir datamart olarak geliştirilmiştir. Metadata ise; veri setlerinin içerisinde bulunan tüm alanlar için VTYS'nin ihtiyaç duyduğu tanımlamaların (veri tipi, verinin bulunacağı alanın boyutu vb.) yapıldığı ve bu bilgilerin tutulduğu yerdir.



Şekil 1 : Üç Aşamalı Veri Ambarı Mimarisi
Kaynak: Cios Krzysztof J. ve diğerleri, Data Mining: A Knowledge Discovery Approach. New York: Springer, 2007, s. 110.

2.4. Veri Tabanı Yönetim Sistemlerinde Optimizasyon ve İşleyiş

VTYS tasarlanırken veritabanını oluşturacak verilerin kaydedilme, okunma, değiştirilme işlemlerinin amacı ve yöntemi konusunda bazı kararlar alınması gerekmektedir. Bu kararlar veritabanının işleyişinin ihtiyaçları karşılaması açısından çok önemlidir[14].

Bir kurumdaki günlük verilerin işlendiği veritabanlarının çalışmasında; kayıt okuma, ekleme, güncelleme ve silme gibi işlemler çokça yapılır. Bu nedenle, bu tür bir ihtiyaç için oluşturulmasına karar verilen veritabanı; hızlı okuma, yazma, güncelleme, değiştirme, değişikliği geri alma, çok kullanıcının kullanımına cevap verebilme gibi ihtiyaçlara cevap verebilmelidir. Bu şekilde optimize edilmiş veritabanları OLTP (Online Transaction Processing) sistemler olarak isimlendirilir. Bir anlamda işlem tabanlı bir optimizasyon vardır[15].

Bunun yerine tarihsel sorgulamaların yapılmak istendiği, analiz ve raporlama ihtiyaçlarına yönelik, kompleks sorgulara cevap verecek şekilde, sadece okuma bazlı işlemler yapmak için kurulmak istenen bir veritabanında optimizasyon ve yapı farklı olmak durumundadır. Bu tip optimizasyon ve yapıları barındıran veritabanlarına OLAP (Online Analytical processing) sistemler denmektedir[16] [17].

[14] Cios Krzysztof J. ve diğerleri. **Data Mining: A Knowledge Discovery Approach.** New York: Springer, 2007, s.107.
[15] Cios Krzysztof J. ve diğerleri, s.107.
[16] Cios Krzysztof J. ve diğerleri. s.93.
[17] Sumathi S. and S.N. Sivanandam, **Introduction to Data Mining and its Applications.** Berlin: Springer, 2006, s.70-72.

Tablo 2 : OLAP ve OLTP Sistemlerinin Farklı Kriterlere Göre Karşılaştırılması

	OLTP	OLAP
Kullanıcılar	Veri girmekle görevli memurlar, Bilgi Teknolojileri (BT) çalışanları, yüzlerce hatta binlerce kişi olabilir. Örnek; nüfus memurlukları	Karar destek uzmanları, birkaç kişi, ya da onlarca kişi. Şirket ya da kurumların analiz ve raporlama bölümlerinde çalışan BT uzmanları örnek olarak verilebilir.
Veri	Güncel, detaylı, basit ilişkili	Tarihsel, çok boyutlu, entegre, özetlenmiş
Boyut	MB'dan GM'a kadar	GB'dan TB'a kadar
Erişilen kayıt sayısı	Onlarca	Milyonlarca
Amaç	Müşteri tabanlı, günlük, istenen işleme yöneliktir. Örneğin; nüfus müdürlüğüne gelen bir vatandaşın nüfusa bağlı olduğu yerin değişmesi talebi sonucu nüfus bilgilerinde yapılacak değişiklikler.	Pazara yönelik, karar destek amaçlı, konuya odaklı. Örnek; seçimde oy verme yetisine ve özelliklerine sahip olan seçmenlerin belirlenmesi ve bu seçmenlerin yaş dağılımlarının ve eğitimlerinin vb. özelliklerinin olduğu bir analiz raporu.
Temel özellikler (erişim tipi, çalışma birimi, veri erişimi örüntüsü, veri tabanı dizaynı)	Okuma, yazma, değiştirme. Basit, işlem bazlı, sık sık, ER diyagramları	Sadece okuma, karar destek, kompleks sorgular, Ad hoc analiz, yıldız şema

Kaynak: Cios Krzysztof J. ve diğerleri, Data Mining: A Knowledge Discovery Approach. New York: Springer, 2007, s. 107.

Günümüz veri ambarı teknolojilerinde OLAP tekniğinin önemli bir yeri vardır. Kullanıcıların sorduğu karmaşık iş sorgularına hızlı bir şekilde yanıt alınması OLAP teknik ve yöntemleri kullanılmaksızın imkânsızdır. Bu nedenle veri madenciliği çalışmalarında OLAP özelliği olan veri ambarları önemli avantajlar sunmaktadır.

OLAP teknikleri beş ana komuta dayanır[18].

1) ROLL UP: Bir veri küpünde boyutlar üzerinde gruplama işlemi yapılması anlamına gelir. YSD'deki "group by" komutunun kullanılmasına eşdeğerdir. Günlük bir tablodan aylık bazdaki satışların raporlanması bu tür bir işlemdir.

2) DRILL DOWN: ROLL UP ın tersidir. Aylık satışlardan günlük satışlara geçiş örnek verilebilir.

3) SLICE: Veri küpündeki bir kesiti almayı sağlar. Örneğin; Mayıs ayındaki satışlar içinde İstanbul'daki satışları almak.

4) DICE: Veri küpündeki sadece bir hücreyi alma işlemidir. Veri küpünden her bir boyutu belirlenmiş bir veriyi alma işlemidir.

5) PIVOT: Veri küpünün boyutlarına bakışımızı değiştirmek için kullanılır. Örneğin; zaman-ürün boyutlarından bakarken zaman-yer boyutuna geçmek.

[18] Cios Krzysztof J. ve diğerleri, s. 116.

Son zamanlarda, OLAP araçlarına olan ilginin artması İş Zekâsı (IZ) sektörünün önde gelen firmalarının göreceli küçük IZ firmalarını satın alması ile kendini göstermektedir. Bu listede son satın almalar aşağıdaki tabloda yer almaktadır[19]. (Detaylı bilgi için aynı adrese bakılabilir, konuyla ilgili satın almaların uzun bir listesi vardır.)

Tablo 3 : Önemli İş Zekası Sistemleri Sağlayıcıları

Firma	Ürün	Satın alan Firma
Cognos	Cognos 8, Planning, Controller, TM1	IBM
Business Objects	BusinessObjects, Crystal, Cartesis	SAP
Hyperion Solutions	Essbase, Hyperion Planning, HFM, former Brio	Oracle
Pilot Software	PilotWorks	SAP
ProClarity Corporation	ProClarity	Microsoft
Siebel	Siebel Analytics	Oracle

[19] Pendse, N., Business Application Research Center, 2007. http://www.olapreport.com/consolidations.htm (10.05.2009)

3-VERİ MADENCİLİĞİ

Bu bölümde, kullanım alanı hızla genişletmekte olan Veri madenciliği kavramı; kullanım alanları, temel teknikleri ve metodolojisi ile birlikte anlatılacaktır.

3.1. Bilgi Keşif Süreci (BKS)

Bilgi keşif süreci (BKS), tam tanımıyla veri tabanlarında bilgi keşif süreci olarak da adlandırılır. Bazı uygulama alanlarında yeni bilgiyi aramak, veri içindeki geçerli, farklı, potansiyel olarak kullanılabilir ve nihai olarak anlaşılır kalıpları belirlemenin önemli bir süreci olarak tanımlanır. Verinin başlıca kaynağı olarak veritabanı vurgulanmasına rağmen, süreç veritabanında olmayan verilere de genelleştirilir. Bir adımı Veri Madenciliği (VM) olmak üzere birçok adımı içerir. Kapsamında bulunan her bir adım, belirli bir keşif görevini tamamlamayı amaçlar ve her biri keşif yönteminin uygulanması ile başarılı olarak sonuçlandırılır. Bilgi keşfi, bilgi çıkarma sürecinin tamamı ile ilgilenir; verinin nasıl tutulduğu ve veriye nasıl erişildiğini, büyük veri setlerini analiz etmek için etkin ve ölçeklenebilir algoritmaların nasıl kullanıldığını; sonuçların nasıl görselleştirileceği ve yorumlanacağını, insan ve makine arasındaki ilişkinin nasıl modelleneceği ve destekleneceğini içerir. Ayrıca uygulama alanının öğrenilmesi ve analizi için destek verilmesi ile ilgilenir[20].

İnsanoğlunun giriştiği hemen her alanda çok büyük veri tabanlarının hızlı bir şekilde oluşması, veriyi faydalı, iş odaklı bilgiye dönüştürmek için yeni güçlü araçlara olan büyük bir talebi yaratmıştır. Bu ihtiyacı karşılamak için yapılan çalışmalarda, araştırmacılar makine öğrenmesi, gizli örüntü (pattern) tanıma, istatistik veri analizi, veri görüntüleme, sinir ağları gibi fikir ve yöntemleri keşfetmişlerdir. Bu çalışmalar sıklıkla veri madenciliği ve bilgi keşfi olarak adlandırılan yeni bir araştırma alanının ortaya çıkmasına neden olmuştur[21]. Veri tabanlarında bilgi keşfi; veri içindeki geçerli, yeni potansiyel olarak faydalı ve nihai olarak anlaşılır örüntüleri tanımlamanın önemli bir sürecidir[22].

"Veri Madenciliği - Data Mining" ve "Veriden Bilgi Keşfi (VBK)- Knowledge Discovery from Data (KDD)" terimlerinin tam anlamı hakkında bir karışıklık vardır. VBK, veriden bilgi

[20] Cios Krzysztof J. ve diğerleri, s. 10.
[21] Sumathi S. and S.N. Sivanandam, s. 2.
[22] Sumathi S. and S.N. Sivanandam, s. 187.

çıkarmanın tüm sürecini betimlemek için 1995'te ortaya atılmıştır. Bu bağlamda, bilgi veri elemanları arasındaki ilişkiler ve örüntü anlamına gelmektedir. "Veri madenciliği" ise yalnızca VBK sürecinin keşif aşaması için kullanılmalıdır[23].

Bilgi keşif sürecini ortak bir çerçevede şekillendirmek için, bir süreç modeli bulunmalıdır. Model, kurumların bilgi keşif sürecini daha iyi anlamalarına yardımcı olmakta ve projenin planlanması ve uygulamasında takip edilmek üzere bir yol haritası sağlamaktadır. Bunun geri dönüşü de zaman ve maliyet kazancı, bu tip projelerin sonuçlarının daha iyi anlaşılması ve kabul edilmesi gibi faydalar ile olmaktadır. Bu tip projelerin önemli ve çoklu adımları, sonuçların kısmi incelenmesini, olası birkaç tekrarı ve veri sahipleri ile yapılacak etkileşimleri içermektedir. Bilgi keşif sürecini standartlaştırılmış süreç modeli olarak yapılandırmak için birkaç neden vardır[24]:

1. Son ürün, verinin kullanıcısı/sahibi için kullanılabilir olmalıdır,
2. İyi tanımlanmış bir bilgi keşif süreci modeli mantıklı, uyumlu, iyi düşünülmüş bir yapı ve yaklaşıma sahip olmalıdır ki bilgi keşif sürecinin arkasındaki mekanizmayı, ihtiyacı ve değeri anlamakta zorluk çekebilecek karar vericilere sunulabilsin,
3. Bilgi keşif projeleri, sağlam bir çerçeve üzerine kurulu önemli bir proje yönetimi gerektirir,
4. Bilgi keşfi, diğer mühendislik disiplinlerinin mevcut kurulmuş modellerini takip etmelidir,
5. Bilgi keşif sürecinin standartlaştırılması için onaylanmış, geniş kapsamlı bir ihtiyaç vardır.

BKS modeli kurmak için ilk teşebbüs akademik camiadan gelmiştir. 1990'ların ortalarında VM şekillenmeye başlarken, karmaşık bilgi keşif dünyasında VM araçları kullanan kullanıcılara rehberlik etmesi için araştırmacılar çok adımlı süreçleri tasarlamaya başladılar. Asıl önem verilen nokta, isteğe bağlı olmak üzere BKS'ni uygularken ardışık aktiviteleri sağlamaktı. Süreçle ilgili 1996'da Fayyad tarafından dokuz adımlı bir model ve 1998'de de Anand ve Buchner tarafından sekiz adımlı bir model geliştirilmiştir[25].

[23] Sumathi S. and S.N. Sivanandam, s. 3.
[24] Cios Krzysztof J. ve diğerleri, s. 9.
[25] Cios Krzysztof J. ve diğerleri, s. 11.

Endüstriyel modeller, kolayca akademik çabaları takip etmiştir. Birkaç farklı yaklaşım kabul edilmiştir. Bunlar yoğun endüstri tecrübesi olan bireylerin teklif ettiği modellerden, büyük endüstri konsorsiyumlarının teklif ettiği modellere kadar değişmektedir. İki temsili endüstri modeli Cabena (IBM desteği ile) tarafından sunulmuş Beş-Adım Modeli ve Avrupalı büyük endüstri konsorsiyumu tarafından geliştirilmiş altı adımlı CRIS-DM'dir. Akademik ve endüstri modellerinin geliştirilmesi, her iki bakış açısını içeren hibrid modellerin geliştirilmesine de öncülük etmiştir[26].

BKS, her adımı sırasıyla gerçekleştirilen çok adımlı bir sürece sahiptir. Bir sonraki adımın başlayabilmesi için bir önceki adımın başarılı bir şekilde tamamlanmış olması gerekir, çünkü bir önceki adımın sonuçlarını girdi olarak kullanır. Sürekli tekrarlanan bir yapıya sahiptir, çünkü süreçte pek çok geribildirim döngüsü ve tekrar olup, her biri revizyon sürecini tetikler[27].

Aşağıda BKS ile ilgili akademik alanda geliştirilen, endüstriyel firmalar tarafından geliştirilip kullanılan ve her ikisinin karşımı olarak nitelendirilebilecek beş farklı modelin karşılaştırması yer almaktadır[28]:

[26] Cios Krzysztof J. ve diğerleri, s. 12-14.
[27] Cios Krzysztof J. ve diğerleri, s. 20.
[28] Cios Krzysztof J. ve diğerleri, s. 17-18.

Tablo 4 : Bilgi Keşif Süreci ile İlgili Standartlaştırılmış Süreç Modelleri

Model	**Fayyad ve diğerleri.** *(1)*	**Anand & Buchner** *(2)*	**Cios ve diğerleri.** *(3)*	**Cabena ve diğerleri.** *(4)*	**CRISP-DM** *(5)*
Orijin / Adım sayısı	Akademik / 9	Akademik / 8	Hibrid Akademik/ Endüstriyel / 6	Endüstriyel / 5	Endüstriyel / 6
Adımlar	1.Uygulama alanını anlamak ve geliştirmek 2.Hedef veri seti yaratma 3.Veri temizleme ve önişleme 4.Veri İndirgeme ve tahmin 5.Veri madenciliği yöntemini seçmek 6.Veri madenciliği algoritmasını seçmek 7.Veri Madenciliği 8.Bulunan desenlerin yorumlanması 9.Keşfedilen bilginin konsolide edilmesi	1.İnsan kaynaklarının tanımlaması 2.Problemin ayrıntılarıyla tanımlanması 3.Verinin incelenmesi 4. İlgi alanı bilgisi edinme 5.Data önişleme 6.Metodolojinin belirlenmesi 7. Desenlerin keşfedilmesi 8. İşleme sonrası bilgi	1.Problemin anlaşılması 2.Verinin anlaşılması 3.Verinin hazırlanması 4.Veri madenciliği 5.Keşfedilen bilginin değerlendirilmesi 6.Keşfedilen bilginin kullanılması	1. İş hedeflerinin belirlenmesi 2.Verinin hazırlanması 3.Veri madenciliği 4. Sonuçların analiz edilmesi 5.Bilginin sindirilmesi	1.İş sorusunu anlama 2.Verinin anlaşılması 3.Verinin hazırlanması 4.Modelleme 5.Değerlendirme 6.Uygulamaya geçirme

Not (1) : En popüler ve alıntı yapılmış model; veri analizine yönelik detaylı tasvir sağlar fakat iş bakış açısı eksiktir. Bu süreç tekrarlamalıdır. Bu modelin yazarları herhangi iki adım arasında bazı döngülerin genellikle uygulandığını açıklamışlardır fakat detaylı bilgi vermemişlerdir. Bu model veri analizi ile ilgili detaylı bilgi vermektedir fakat iş bakış açısına yönelik tanımlar eksiktir. Bu model sonraki modeller için bir köşe taşı olmuştur.
Not (2) : İlk süreçler ile ilgili detaylı döküm sağlamaktadır; eksik adım bilgi keşif uygulaması ve proje dokümantasyonu ile ilgilidir
Not (3) : Hem akademik hem de endüstri modellerinden oluşturulmuştur ve iteratif durumları vurgular; birkaç geri besleme düğümünü tespit eder ve tanımlar. Model sürecin iteratif durumları üzerinde durur, önceki modellerin kullanıcılarının tecrübelerinden şekillenir.
Not (4) : İş odaklıdır ve veri madenciliği uzmanı olmayan kişiler tarafından kolayca anlaşılabilir; model tanımları veri madenciliği jargonu dışındaki kavramları kullanır.
Not (5) : Kolay anlaşılır bir dil kullanır; iyi bir dokümantasyonu vardır; tüm gerekli detayları sunacak şekilde tüm adımlar alt adımlara bölünmüştür. Model kolay anlaşılır bir dil ve iyi dokümantasyon ile nitelendirilmektedir. Tüm gerekli detayları sunacak şekilde tüm adımlar alt adımlara bölünmüştür. Ayrıca adımlar arası düğümleri ile sürecin güçlü tekrarlanabilir doğasını kabul etmektedir. Genel olarak, çok başarılı ve yaygın olarak kullanılan bir modeldir; çünkü pratik, endüstriyel ve gerçek dünya bilgisinin keşif deneyimine dayandırılmıştır.

3.2. Veri Madenciliği

Veri madenciliği; bilgi madenciliği, bilgi keşfi, bilgi hasadı, bilgi çıkarımı, veri arkeolojisi, desen (pattern) işleme, desen analizi gibi isimlerle de anılmaktadır. Veri madenciliğinin amacı veriyi uygulanabilir stratejik kararlar almak için yararlı bilgiye dönüştürmektir.

Veri madenciliği oldukça yeni ve gelecekte daha çok kullanılacağı düşünülen bir teknolojidir. İstatistiksel analiz, makine öğrenmesi, yapay zekâ ve veri görselleştirme tekniklerini kullanarak veri ambarlarında saklı büyük miktardaki veriden anlamlı yeni korelâsyonlar, örüntüler ve trendlerin keşfedilmesinin bir süreci olarak tanımlanabilir[29]. Veri madenciliği, madencilerin topraktan değerli madenleri alıp çıkardığı gibi büyük miktardaki veriden öz bilginin çekilip çıkarılmasını ifade eder[30]. Veri madenciliğinin amacı, çoğunlukla incelenmemiş büyük veri yığınları içerisinden bir anlam ifade edecek bilgiyi çıkarmaktır[31].

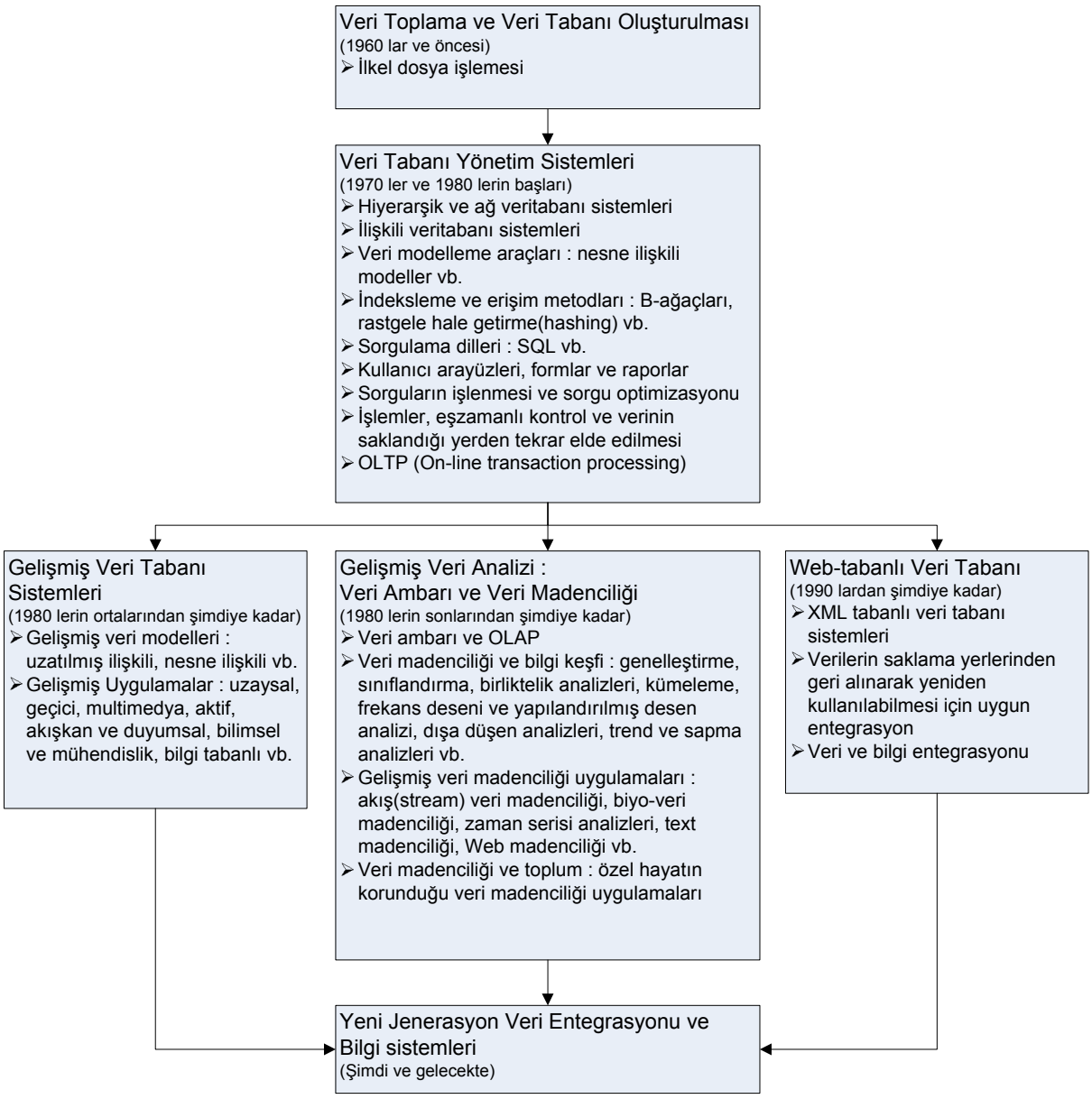
Veri madenciliği ve bilgi keşfi alanında devam eden dikkate değer büyüme, aşağıda belirtilen faktörlerin kesişmesiyle ivmelenmiştir[32]:

- Veri toplamadaki hızlı büyüme; süper market tarayıcılarının yaygınlaşması bunlara örnek verilebilir.
- Verinin veri ambarlarında saklanması; böylece bütün kurum, güvenilir ve güncel veriye kolayca erişebilmektedir
- Web üzerinden veya intranet üzerinden veriye kolayca ulaşabilme imkânı
- Küreselleşen ekonomide pazar payını artırmak için artan rekabet baskısı
- Ticari kullanıma hazır veri madenciliği yazılımlarının geliştirilmesi
- Depolama kapasitesinde ve hesaplama gücündeki muazzam büyüme

Veri madenciliği ile istatistik arasında iki temel farklılık mevcuttur: Birincisi istatistikte çalışmalar genellikle bir hipotez ile başlarken, Veri madenciliği hipoteze gerek duymaz. İkincisi de istatistiksel analizler niceliksel ve niteliksel verileri kullanırken, Veri madenciliği yöntemlerinde metin, ses, görüntü verileri gibi farklı tiplerde veriler kullanılabilir.

[29] Sumathi S. and S.N. Sivanandam, s. 8.
[30] Han, J. and Micheline Kamber, s. 5.
[31] Cios Krzysztof J. ve diğerleri, s. 3.
[32] Larose Daniel T., **Discovering Knowledge in Data an Introduction to Data Mining.** New Jersey: John Wiley & Sons, 2005. s. 4.

Aşağıdaki tablodan da görüldüğü gibi, 1960'larda kurumlar tarafından sistematik olarak veri toplanmaya başlandı. 1980'lerde veri sistemlerini yönetmekte karşılaşılan zorluklar sonucunda ilişkili veri tabanları (RDMBS) sistemi ortaya çıktı. 1990'larda veri ambarı teknolojisi ve OLAP küpleri yaygın olarak kullanılmaya başlandı.



Şekil 2 : Veri Tabanı Sistemlerinin Teknolojilerinin Değerlendirilmesi
Kaynak: Han, J. and Micheline Kamber, Data Mining: Concepts and Techniques. Second edition, San Francisco: Morgan Kaufmann Publications, 2006. s. 2.

Veri toplama, depolama ve imaj işleme, dijital sinyal işleme, metin işleme ve heterojen verinin çeşitli formatlarını işleme gibi uygulamalardaki büyük veri miktarlarının transfer edilme tekniklerinde muazzam gelişmeler oldu. Bununla birlikte, depolanan veri miktarındaki çok büyük artış, bu veriyi işleyecek daha iyi, daha hızlı ve ucuz yöntemlere büyük talep

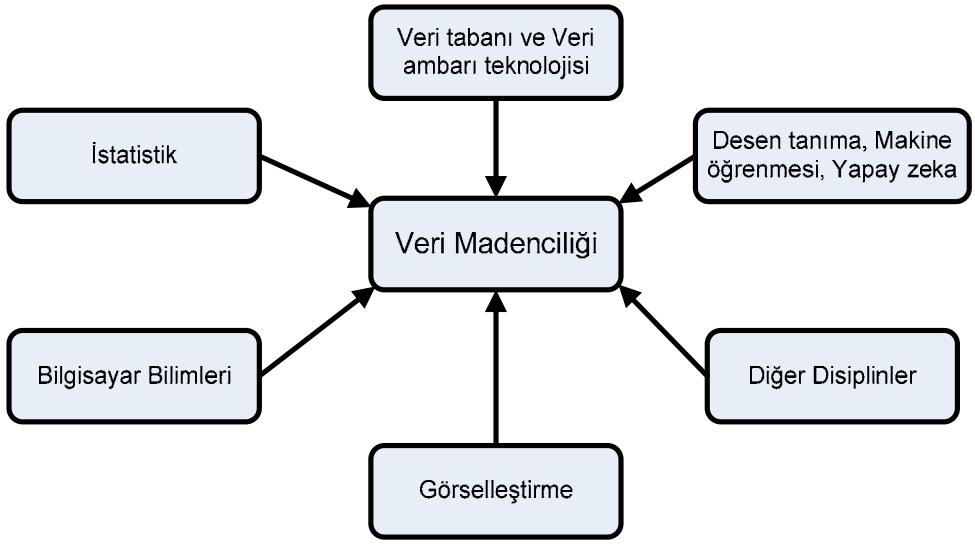
oluşmasına neden oldu. Diğer bir deyişle, dünyadaki bütün verinin, veriyi etkin ve verimli işleyecek, ondan anlamlı bilgi çıkaracak araçlar olmadan hiçbir değeri yoktur. U. Fayyad, H. Mannila, G. Piatetsky-Shapiro, G. Djorgovski, W. Frawley, P. Smith gibi öncüler bu acil ihtiyacı gördüler ve veri madenciliği alanı doğdu[33].

Çok hızlı şekilde artan ve müthiş boyutlara ulaşan veriler, toplanıp uygun şekillerde depolandıktan sonra, karar verme amaçlı kullanılabilmeleri için güçlü veri analiz araçlarına ihtiyaç duyulmuştur. Bu aşamada geliştirilen VM algoritmaları ile verideki gizli örüntülerin keşfi, düşük maliyetli ve hızlı bir şekilde gerçekleştirilebildi. VM ile elde edilen bulgular kullanılarak hem bilimsel alanda hem de ekonominin tüm alanlarındaki karmaşık iş problemlerine anlamlı çözümler üretildi.

İş problemlerinin çözümlenmesi için iş bilgisine sahip kullanıcılar, başarı için temel unsurdur ve ilgili iş problemlerini otomatik olarak çözümleyecek bir veri madenciliği aracı mevcut değildir. Projenin her aşamasında uzman analist müdahalesi gerekli olup, ayrıca model sonuçlarının yorumlanarak uygulamaya dönük çözümlerin üretilmesi de ancak iş bilgisi ve teknik bilgi birikimine sahip iş kullanıcısı tarafından gerçekleştirilebilir.

Veri madenciliğinin çıkış noktası istatistik bilimi olmasına karşın, pek çok farklı disiplindeki teknolojik gelişmenin bir sonucu olarak ortaya çıkmıştır. Veri madenciliği disiplinler arası bir alandır, veri tabanı sistemleri, istatistik, makine, görselleştirme ve enformasyon bilimi gibi disiplinlerin kesiştiği noktadır. Bundan başka, kullanılan veri madenciliği yaklaşımına bağlı olarak; sinir ağları, bulanık ve/veya yaklaşımlı küme teorisi, bilgi gösterimi, tümevarımsal mantık programlama veya yüksek performanslı bilgi işlem gibi diğer disiplinlerden teknikler de uygulanabilir[34].

[33] Cios Krzysztof J. ve diğerleri, s. 5.
[34] Han, J. and Micheline Kamber, s. 29.



Şekil 3 : Farklı Kilit Teknolojilerin Kesişimi Olarak Veri Madenciliği
Kaynak: Han, J. and Micheline Kamber, Data Mining: Concepts and Techniques. Second edition, San Francisco: Morgan Kaufmann Publications, 2006. s. 29.

Tipik bir Veri madenciliği mimarisi aşağıda sıralanan temel bileşenlere sahiptir[35]:

- Veritabanı, veri ambarı, WWW (World Wide Web) veya başka bir bilgi ambarı: İşlemek üzere kullanılan veri kaynağıdır. Veri temizleme ve veri entegrasyon teknikleri burada gerçekleştirilir.
- Veritabanı veya veri ambar sunucusu: Bu kısım, kullanıcının veri madenciliği taleplerini karşılamak üzere, ilgili verinin kaynağından çekilmesinden sorumludur.
- Bilgi Tabanı: Araştırma veya sonuçlanan kalıpların ilgi çekiciliğinin değerlendirilmesine yol göstermek için kullanılan ilgi alanı bilgisidir.
- Veri madenciliği motoru: Bu veri madenciliği sistemi için gereklidir ve ideal olarak görevler için tanımlama, birliktelik ve korelasyon analizi, sınıflandırma, öngörü, kümeleme analizi, dışa düşen analizi ve değişim analizi gibi fonksiyonel modüllerin bir kümesini içerir.
- Desen değerlendirme modülü: Bu bileşen tip olarak ilgi çekicilik ölçütlerini kullanır ve ilgi çekici model kalıplarını araştırmaya odaklanma amacıyla veri madenciliği modelleri ile etkileşimde bulunur.
- Kullanıcı ara yüzü: Bu modül kullanıcılar ve veri madenciliği sistemleri arasında bağlantı sağlar, kullanıcı veri madenciliği sorgusu veya görevi tanımlayarak sistem ile etkileşimde bulunur, araştırmaya odaklanmaya yardımcı olacak bilgi sağlar ve orta düzey veri madenciliği sonuçlarına dayalı keşfe yönelik veri madenciliğini icra eder.

[35] Han, J. and Micheline Kamber, s. 7.

Birbirinden pek çok farkı olan VM çözümlerinden en uygun olanı seçmek her zaman kolay olmayabilir. Bu çözümler içerisinden hangisinin kullanılacağına karar verirken aşağıdaki kriterlere dikkate edilmelidir:

- Farklı veri kaynaklarına erişebilmesi,
- Esnek ve kurumsal bir platforma sahip olması,
- Veri hazırlama, analiz ve sonuçların kolay bir ara yüzde uygulanmasına yönelik kullanım kolaylığı sağlaması,
- Hem veri hazırlama hem de modelleme özelliğini barındıran entegre bir sisteme sahip olması, ihtiyaç duyulan sistemin tüm bileşenleri için komple çözüm üretmesi,
- Kullanımı esnasında mümkün olduğu ölçüde az IT kaynağına ihtiyaç duyması,
- Modellemeye yönelik çok çeşitli algoritmalar içeren çözüm ve genel kabul görmüş veri madenciliği metodolojisi sunması,

Veri madenciliği, prensip olarak her türlü veri tipine veya veri havuzuna uygulanabilir. Ancak verilerin depolanma şekline göre uygulanan VM yöntemleri değişir.

Veri madenciliğinin tüm iş problemlerine sihirli çözümler getirdiğini düşünmek bu konuda yapılan en büyük yanılgıdır. VM, iş problemlerinin çözümünde fayda sağlayacak kaliteli bilginin veriden yola çıkılarak elde edilmesidir.

Rekabetin hızla arttığı günümüzde kurumlar, müşterileri daha iyi anlamak ve karar optimizasyonunu sağlamak için çok büyük miktarlarda veri toplamaya başlamıştır. Ancak birçok yerde veri toplama, araç olmaktan çok amaca dönüşmüş ve veri çöplükleri oluşmuştur. Sürdürülebilir bir başarıyı yakalamak için veri yığınları içerisindeki nitelikli bilgiye ulaşmak ve geleceği öngörmek için kurulan sistemler vazgeçilmez bir faktör haline geldiler. Bundan dolayı veri madenciliğine dayalı analitik bir sistem, pek çok durumda artık zorunlu olarak kullanılması gereken bir araç haline gelmiştir.

Günümüzün pazarlama stratejilerinde müşterilerin anlaşılması çok önemli bir gerekliliktir. Yeni müşterilerin başarılı bir şekilde firmaya kazandırılması ve mevcut müşterilerden özellikle yüksek değerliliğe sahip olanların elde tutulması, bu şekilde mümkün olmaktadır. Bu amaca ulaşabilmek için pek çok firma büyük hacimli verileri uygun şekilde toplama,

düzenleme, saklama, analiz ederek, sonuçlara uygun aksiyon alma sürecini etkin ve doğru şekilde uygulamaya çalışır.

3.3. Veri Madenciliği Metodolojisi

Veriden kullanılabilir bilgiye ulaşmak için, veri analizinde kullanılan pek çok algoritmayı bilmek yeterli olmayıp, tüm süreci anlamak ve sürece hâkim olmak gerekmektedir. Başarılı bir Veri madenciliği projesi, ancak yeterliliği kanıtlanmış bir süreci takip ederek gerçekleştirilebilir.

Veriden Bilgi Keşfi (VBK) (Knowledge Discovery from Data - KDD) konusunda önemli araştırmaları olan Piatetsky-Shapiro tarafından editörlüğü yapılan, veri madenciliği ile ilgili kitaplar, farklı ürünler, eğitimler, araştırmalar, iş ilanları ve haberler konusunda sürekli güncellenen önemli bir kaynak konumundaki KDnuggets (www.kdnuggets.com) tarafından 2002 ve 2004'de gerçekleştirilen, Veri madenciliğinde kullanılan metodoloji seçimine yönelik araştırma sonuçları yer almaktadır. Veri madenciliği projeleri yapan kuruluşlara 'Veri madenciliği için kullanılan ana metodoloji nedir?' sorusu sorulmuştur. Alınan yanıtların karşılaştırmalı sonuçları ekteki tabloda yer almakta olup, CRISP-DM metodolojisi açık ara ile en fazla tercih edilen metot olmuştur.[36]

Tablo 5 : Veri Madenciliği Metodolojisi Kullanımı Anket Sonuçları

Veri Madenciliği Metodolojileri	2002 Yılı Araştırma Sonuçları		2004 Yılı Araştırma Sonuçları	
	Adet	Oran	Adet	Oran
CRISP-DM	96	51%	72	42%
SEMMA	22	12%	17	10%
Kendi Yöntemin	43	23%	48	28%
Organizasyonuma ait yöntem	13	7%	11	6%
Diğer	8	4%	10	6%
Hiçbiri	7	4%	12	7%
Toplam	189	100%	170	100%

Uygulamada ikinci en yüksek sırada kullanılan yöntem olan SEMMA, SAS firması tarafından geliştirilmiştir. SEMMA kelimesi VM sürecini ifade eden şu beş aşamanın ifadelerinin kısaltmasıdır: **S**ample (Örnekleme), **E**xplore (Keşfetmek), **M**odify (Değiştirmek ve Düzeltmek), **M**odel (Modelleme), **A**sses (Değerleme).

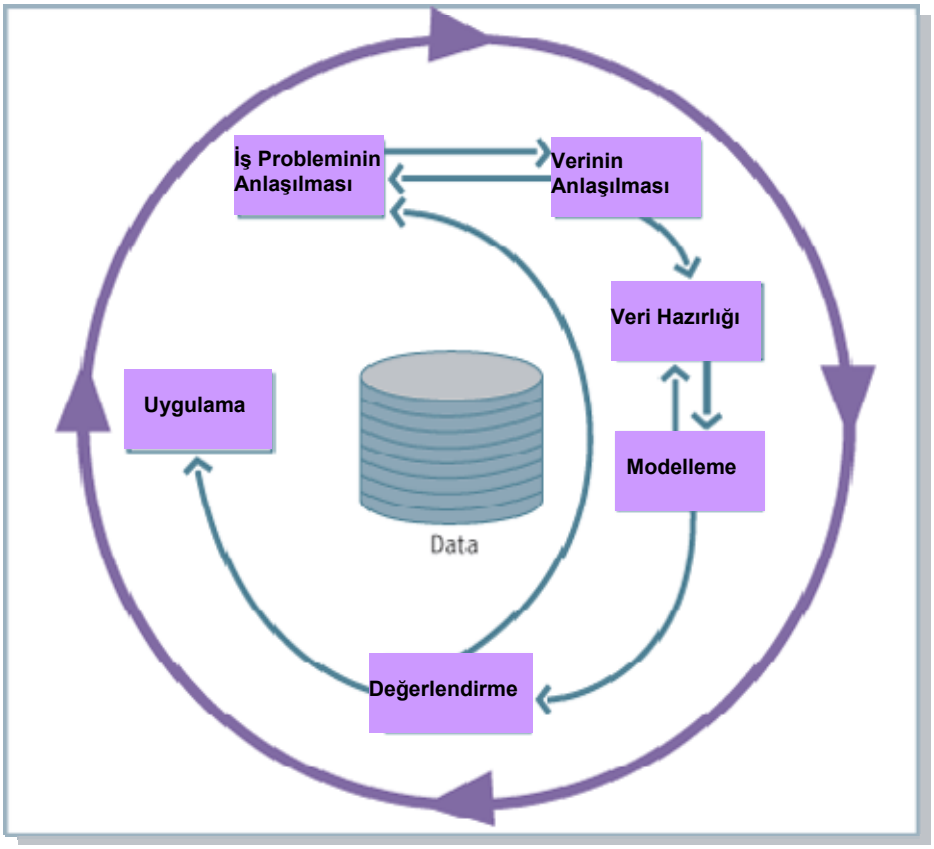
[36] http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm (30.06.2009)

CRISP-DM (CRoss-Industry Standard Process for Data Mining) 1990'ların sonlarına doğru dört Avrupalı firma tarafından oluşturuldu. Bu firmalar şunlardır; Integral Solutions Ltd. (ticari veri madencilik çözümleri üreticisi, daha sonradan SPSS olmuştur), NCR (veritabanı sağlayıcısı), DaimlerChrysler (otomobil üreticisi), ve OHRA (sigorta firması). Daha sonra endüstrinin lider modeli olarak kabul görmüştür ve hâlihazırda en çok uygulanan yöntemdir.

CRISP-DM; ne sadece teorik, teknik prensipler ile çalışan akademik usul üzerine kurulmuştur ne de seçkin gurulardan oluşan bir komite tarafından kapalı kapılar ardında yaratılmıştır. Birbirinden tamamıyla farklı bu iki yöntem de geçmişte denenmiş metodolojiler geliştirmek amacıyla kullanılmışlardır, fakat nadiren pratik, başarılı ve yaygın olarak uyarlanmış standartlara yol açmışlardır. CRISP-DM başarılı olmuştur, çünkü emin bir şekilde pratik, insanların veri madenciliği projelerini yaptığı gerçek dünya tecrübesine dayandırılmıştır[37].

CRISP-DM modelinin geliştirilmesi güçlü endüstriyel desteği arkasına almıştır. Ayrıca Avrupa Komisyonu tarafından finanse edilen ESPRIT programı tarafından da desteklenmiştir. CRISP_DM Özel İlgili Grubu (ÖİG) (Special Interest Group - SIG) geliştirilmiş süreç modelini desteklemek amacıyla oluşturulmuştur. ÖİG, VM uygulamalarından bağımsız bir endüstri standardı oluşturmak için uzman uygulamacılardan, VM sağlayıcılarından veya yönetim danışmanlarından oluşan bir gruptur. 1999'un ortasında CRISP-DM Version1.0 olarak yayımlandı. Hâlihazırda da CRISP-DM metodolojisi güncellenmesi amacıyla ÖİG düzenli çalışma grupları ve faaliyetlerde bulunmaktadır. Güncel durumda, ÖİG 300 VM kullanıcısı, VM araç veya hizmet sağlayıcısını içermektedir[38].

[37] CRoss Industry Standard Process for Data Mining, The CRISP-DM consortium August 2000, http://www.crisp-dm.org/Overview/index.htm (18.04.2009)
[38] http://www.crisp-dm.org/SIG/index.htm (18.04.2009)



Şekil 4: CRISP-DM Sürecinin Aşamaları

3.3.1. İş Probleminin Tanımlanması

Sürecin ilk adımı; iş probleminin, iş ihtiyaçlarının veya ulaşılmak istenen iş hedefinin iş perspektifi ile anlaşılarak net tanımın yapılmasıdır. Daha sonra bu bilgi, VM problemi olarak revize edilir ve başlangıç proje planı oluşturulur. Her iş problemi mutlaka VM problemi olarak ifade edilebilmelidir. Ancak bu şekilde bir VM projesi olarak hayata geçecek bir süreç oluşturulabilir ve bu aşamada başlangıç proje planı çıkarılabilir.

İş ihtiyaçlarının anlaşılması, ancak iş perspektifi ile konuya yaklaşılmasını gerektirir. İş hedeflerinin belirlenmesi, geçmişin anlaşılmasını ve mevcut durumun değerlendirilmesini kapsar. Bunun için eldeki kaynaklar, kısıtlar ve ihtiyaçlar arasında bir kıyaslama yapılır. Mevcut riskler ve fırsatlar gözden geçirilir.

Her iş probleminin mutlaka bir finansal sonucu mevcut olup, projenin sonuçlarının firmanın maliyetlerinin azalmasına veya karlılıkta meydana getireceği artışa göre tahmini getirisi de hesaplanmalıdır.

3.3.2. Verinin Tanımlanması

Verinin anlaşılması; öncelikle başlangıç verisinin toplanması adımı ile başlar. Bu aşamada ilgili tüm veri kaynakları belirlenir. İş problemi ile ilgili olduğu düşünülen tüm veriler ilk aşamada bir araya toplanır. Veri kaynaklarında belirtilen veriye erişimin sağlanıp sağlanamayacağı test edilir ve verinin düzenli olarak alınarak kullanılması aşamasında karşılaşılabilecek sorunlar tespit edilmeye çalışılır.

Daha sonra toplanan verinin tanımlanması adımına geçilir. Bu, veride yer alan değişkenlerin, değerlerin ne anlama geldiklerinin araştırılması ve belirlenmesi anlamına gelir. Verilerin tutulduğu tabloların yapısı, büyüklüğü, kullanışlı olup olmaması ve veritabanları sistemindeki ilişkileri gibi konular incelenir. Analizi yapacak ekibin tamamının aynı veriden aynı şeyi anlaması ve veri hakkında net olması gerekmektedir. Bundan dolayı veriyi tanıyan uzmanlardan mutlaka yardım alınmalıdır.

Verinin ne anlama geldiğinin tam olarak anlaşılmasından sonra verilerin incelenme safhasına geçilir. Burada genel inceleme, sorgulama, temel istatistiklerine, değişkenlerin doluluk oranlarına, ölçü birimlerine vb. bakılarak raporlanır. Bu esnada verinin kalitesi ölçülerek bir sonraki veri hazırlama aşamasına girdi niteliğinde bilgiler sağlanır.

Ayrıca veri setleri içerisinde gizli bilgiler içerdiğinden şüphelenilen önemli alt veri grupları da iş bilgisine sahip kullanıcıların yardımıyla tespit edilmeye çalışılır. Çünkü buradan elde edilecek bilgiler modelleme aşamasında çok önemli katkılar sağlayabilir.

Veri madenciliği modellemeleri için gerekli olan veri, kurum içinden olduğu gibi kurum dışından da elde edilebilir. Kurum içi veri kaynakları; abonelik veritabanı, abone işlemleri veri tabanı gibi kaynaklardır. Kurum dışı veri kaynakları ise kamuya ait İstatistik Kurumu tarafından yayımlanan genel demografik ve ekonomik veritabanları veya bankalar arası Kredi Kayıt Bürosu (KKB) gibi belirli bir sektöre ait veritabanı olabilir.

3.3.3. Verinin Hazırlanması

Veri hazırlığı, verinin değişik veri kaynaklarından derlenerek VM modellemesinde kullanılacak final veri setine dönüştürmek için gereken tüm aktiviteleri kapsamaktadır. Bu

aşama; tablo oluşturma, kayıt ve değişken seçimi, veri temizliği, yeni değişkenlerin oluşturulması ve verilerin dönüştürülmesi işlemlerini kapsamaktadır. Veri hazırlama aşaması, projeden projeye değişmekle birlikte tüm proje zamanının en büyük kısmını oluşturur.

Veri hazırlama aşamasında kullanılacak veri setleri belirlenir ve verinin hazırlanmasındaki her aşamada kalite kontrolü yapılır. Çünkü VM sonuçlarının güvenilir ve tutarlı olması verinin kalitesine doğrudan bağlıdır. Üstelik veri kalitesine önem verilmeyen hemen her durumda VM modelleri yetersiz sonuçlar üretir.

Veri kalitesinin sağlanması da bu aşamada yapılır. Verinin doğrulanması, verinin tam olup olmadığı, içerdiği hatalar anlaşılmaya çalışılır. Veri kalitesini artırmaya yönelik pek çok yöntem mevcuttur. Örneğin, kayıp verilerin uygun değerlerle doldurulması gibi yöntemlerle veri kalitesi arttırılır.

Veri temizliği, en önemlileri aşağıdaki aktarılan problemlerin giderilmesi işlemine denir:

- Mükerrer kayıtlar
- Yanlış veya tutarsız veri
- Yazım hataları
- Güncellenmemiş değerler
- Format farklılıklarından kaynaklanan tutarsızlıklar

Eksik verilerin analizi de bu aşamada yapılarak modellemeye hangi şekilde alınması gerektiğine karar verilir. Örneğin bir abonenin hiç harcama tutarının olmaması o abonenin daha yeni abone olduğunu gösterirken, yaş değişkeninin boş olması anlamsız kabul edilip uygun şekilde doldurulmalı veya bu değişkenin analizden çıkarılmasına karar verilmelidir.

Mevcut alanlardan yeni alanların türetilmesi gibi veri yapılandırma işlemleri de burada gerçekleştirilir. Bu işlem, modellerle daha kesin sonuçlar üretebilmek için yeni değişkenlerin oluşturulması için yapılır. Oluşturulan bu değişkenlerin modellerde daha anlamlı katsayılara sahip olacağı varsayımına dayanılarak iş bilgisinin de kullanılmasını gerektirmektedir.

Modellemeye girecek veri setinin oluşturulması sürecinde, farklı tabloların birleştirilmesi işlemi de yeni değişken oluşturmak kadar sık yapılan bir işlemdir. Data entegrasyonunun sağlanmasından sonra datalara uygun formatların atanması işlemi de burada yapılır.

3.3.4. Modelleme

Bu aşamada değişik modelleme teknikleri seçilerek hazırlanan veri setine uygulanır. Belirli VM problemleri için uygun birçok modelleme tekniği mevcuttur. Her modelleme yönteminin sahip olduğu birbirinden farklı parametreler kullanılarak modele girecek değişkenler ve bu değişkenlerin optimum katsayıları bulunmaya çalışılır.

Kurulan VM probleminin çözümüne uygun modelleme teknik veya teknikleri belirlenir. Mevcut problemin ve kullanılan değişkenlerin, uygulanacak modellerin varsayımlarını yerine getirip getirmediği araştırılır. Farklı modelleme yöntemleri mutlaka denenmelidir.

VM modelleme sürecinde test tasarımı da bu aşamada yapılır. Veri test ve öğrenim (train) olarak ikiye ayrılır. Bazı durumlarda bu ayrım; öğrenim (train), doğrulama (validation) ve test olmak üzere 3 alt grup da olabilir. Test verisi tüm verinin % 5 ile %50'si arasında değişebilir veya hiç eklenmeyebilir. Bu oranlar, VM uygulamasının geliştirildiği koşulları, uygulamacının deneyimleri gibi koşullara bağlı olarak değiştiğinden bu konuda net bir oran vermek mümkün değildir. Ancak örnek vermek gerekirse KXEN ürününde veriseti için başta belirlenen (varsayılan - default) oran %75 öğrenim, %25 doğrulama olarak iki kısma ayrılmaktadır, ancak gerekirse bu oranlar değiştirilebilmektedir.

Modelleme aşaması sırasında sık sık veri hazırlama aşamasına geri dönüşler yapılır. Çünkü modelleme, sürekli yenilenen bir deneme yanılma süreci olduğundan yeni bir değişkenin gerekli görülmesi, hedef değişken tanımının değiştirilmesi gibi ihtiyaçlar ortaya çıkabilir. Mevcut veriler kullanılarak açıklama gücü çok yüksek yeni değişkenler oluşturulabilir. Daha önce mevcut olmayan yeni bir veri kaynağına ulaşılmış olabilir. Ayrıca, bazı tekniklerin kullanılabilmesi için değişkenlerin uygun formatlara sahip olması gerektiğinde format değişikliği gerekebilir.

Modellemede başarı açısından değişken adedinin çokluğu bir avantajdır. Ancak değişkenlerin gereğinden fazla olması – bazı durumlarda bu sayı binlerle ifade edilmektedir - modelleme

sürecini zahmetli ve içinden çıkılmaz bir hale de sokabilir. Bu nedenle modeller üzerinde çalışılmaya başlanmadan önce değişkenler üzerinde kapsamlı çalışmalar sonucu değişken seçimi işlemi yapılabilir. Değişken seçimi işlemleri de yine Karar Ağaçları ve Lojistik Regresyon gibi algoritmaların yardımıyla yapılmaktadır. Bu seçim, modelleme sürecinin daha hızlı ve etkin yapılmasını sağlar.

Küçük örneklemler üzerinde yapılan pek çok istatistiksel model için modelleme öncesinde verilerin ilgili modele uygunluğu model varsayımlarına göre test edilir. VM veri setlerinde ise bu varsayımlara uygunluk kontrolü ya daha geniş bir aralıkta yapılır ya da hiç dikkate alınmaz.

Genel olarak aralarında yüksek korelasyon bulunan değişkenlerden bazıları seçilerek modelleme sürecine alınır, geri kalanı elenir. Ayrıca modeller kurulurken iş bilgisi ön plana çıkar. Modellerde ön sıralarda çıkmasına rağmen bazı değişkenlerin elenmesi, önemli katkı yapmadığı görülse bile bazı değişkenlerin mutlaka modellerde yer almasına bu aşamada karar verilir. Bu aşamanın sonunda, üretilen her model hem kendi içinde katsayıların tutarlılığı ve sonuçların beklentileri karşılama düzeyi açısından değerlendirilir, hem de aynı soruna yanıt veren diğer modeller ile belirli ölçütlere göre kıyaslanarak değerleme yapılır.

Modelin kullanıcılar tarafından kolay anlaşılması, gerçek hayata uygulanabilir olması ve kullanışlı olması istenir. Modelin yeni verilerle veya test verisi ile geçerliliğini kanıtlaması da gereklidir. Bunlardan daha önemlisi de bir amaca hizmet edecek bilgi sağlaması istenir. Bu bilgi, bir pazarlama kampanyası için abone segmentasyonunun yapılması da olabilir, düzenli bir tahmin prosesinin uygulamaya alınması da olabilir veya projeye başlanmadan belirlenmiş bir varsayımın olumlanması veya olumsuzlanması da olabilir.

Modeller, geliştirildikleri anın koşullarının gelecekte de devam edeceği varsayımına dayanarak kurulurlar. Ancak genel ekonomik dengelerin değişmesi, kullanım kalıplarının zaman içinde farklılaşması gibi nedenlerle dönemsel olarak yenilenmelidirler.

3.3.5. Değerlendirme

Veri analizi açısından veya açıklayıcı değişkenler ile hedef değişken arasındaki ilişkiler açısından değerlendirildiğinde çok güzel sonuçlar üreten, yüksek kaliteye sahip bir veya daha

fazla model bir önceki aşamada hazırlanıp bu aşamada girdi olarak kullanılır. Burada model, iş hedefleri perspektifinden incelenerek değerlendirilir ve iş hedeflerini ne ölçüde karşıladığı ölçülür.

Modellerin değerlendirilmesi, iş amaçlarını gerçekleştirmesi açısından önceden belirlenen belirli başarı kriterlerine göre yapılır. Bu kriterler; sahte başvuruların %50 sinden fazlasının tespit edilmesi, kampanya geri dönüş oranının genel kampanya geri dönüş oranlarının 3 katından fazla olması vb. gibi olabilir.

Burada yapılan değerlendirme sonucunda; ya önceki adıma geri dönülerek daha iyi sonuçlar üreten bir model araştırılır ya da bir sonraki adıma geçilir. En iyi model bile yetersiz ise sebepleri araştırılır. Kullanılan veri, kullanılan hedef değişkenin tanımı, yanlış modelleme tekniği, ilgisiz değişkenler, yanlış veri gibi akla gelebilecek her neden hesaba katılarak tüm süreç gözden geçirilir.

3.3.6. Uygulama ve İzleme

Veri setini iyi tanımlayan, istatistiksel olarak ve kuramsal olarak anlamlı bir modelin ortaya çıkarılması bir veri madenciliği projesinin sonu değildir. Veriden keşfedilen bu bilginin iş amaçlarına göre organize edilerek kullanıma aktarılması gerekmektedir. Kullanım yerine göre bu aşama, belirli bir raporun üretilmesi kadar basit olabileceği gibi, sürekli yenilenen ve mevcut sistemlerle entegre çalışan karmaşık bir yapıya da sahip olabilir.

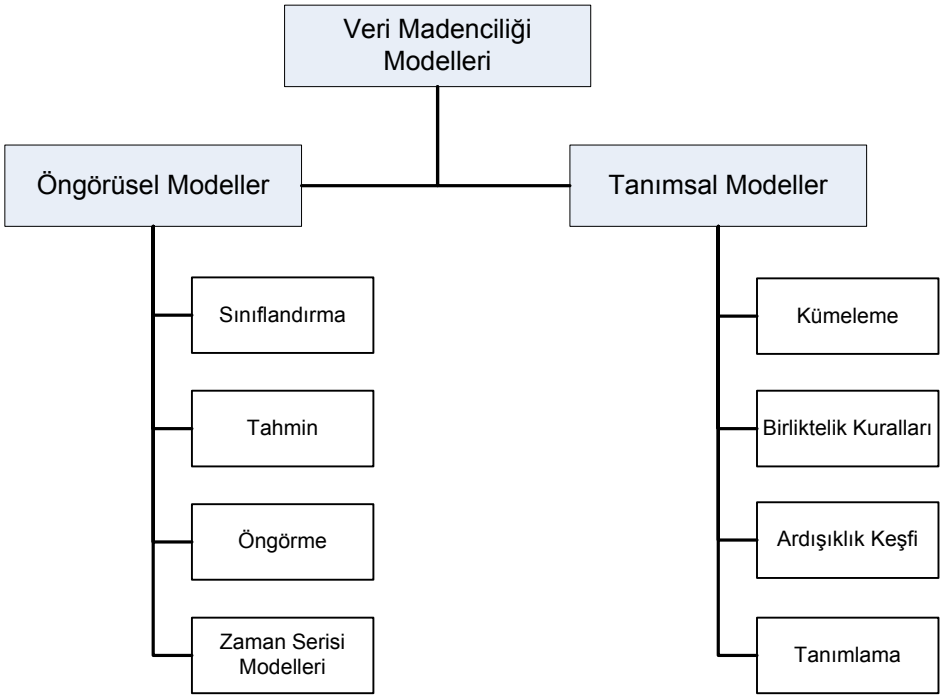
Uygulama canlı sisteme alındıktan sonra sürecin izlenmesi ve bakımının yapılması gerekmektedir. Canlı sistemlerin izlenmesi günlük ve aylık gibi periyodik raporlamalarla mümkün olabilir. Sistemin genel yapısında bir bozulma meydana gelirse, bu raporlardan ortaya çıkar.

Proje sonunda tüm aşamaların değerlendirilmesinin yapıldığı bir sonuç raporu yazılması uygun olur. Ancak bundan da önemlisi, proje süresince yapılan ve uygulamaya alınan tüm altyapının detaylı akış şemalarını içeren iyi bir dokümantasyonun hazırlanması, sistemin bakım aşaması için hayati öneme sahiptir.

Belirli bir hedef veya iş problemi belirlendiğinde mutlaka anahtar başarı kriterleri (Key Performance Index) de belirlenmelidir. Proje uygulamaya alındıktan sonra model sonuçları belirlenen bu kriterlere göre raporlanarak takip edilmelidir. Ancak bu takdirde proje sonucunun gerçek hayattaki başarısından veya başarısızlığından bahsetmemiz mümkün olabilir.

3.4. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller öngörüsel ve tanımsal olmak üzere iki alt gruba ayrılır. Öngörüsel modeller, sonuçları bilinen veriler kullanılarak geliştirilir ve sonuçları bilinmeyen aynı özelliklere sahip veri setine uygulanarak sonuç değerleri tahmin edilir. Tanımsal modellerde ise veri setindeki gizli örüntüler ortaya çıkarılarak karar vermede kullanılır. Tanımsal yöntemler daha çok veriyi anlamaya, tanımaya, keşfetmeye yönelik olarak kullanılır ve daha sonra uygulanacak yöntemler için fikir vermeyi amaçlar. Öngörüsel yöntemler ise veriden bilgi ve sonuç çıkarmaya yönelik olarak kullanılmaktadır



Şekil 5 : Veri Madenciliği Modelleri

3.4.1. Sınıflandırma (Classification)

Sınıflandırma; makine öğrenmesindeki, veri madenciliğindeki ve örüntü tanımadaki ana işlerden birisidir. Etiketlenmiş hedeflerin sınıflandırılması ile ilgilenir. Sınıflandırma için kullanılacak bilgi; sınıflandırma kuralları, diskriminant fonksiyonları ve karar ağaçları gibi farklı şekillerde ifade edilebilir[39]. Sınıflandırma, yeni sunulan bir hedefin özelliklerinin incelenmesi ve daha önce tanımlanmış sınıflardan birisine atanmasından oluşmaktadır. Sınıflandırma işi sınıfların iyi anlatılmış tanımı ve daha önce sınıflandırılmış örneklerin bir eğitim seti ile karakterize edilir. Asıl önemli iş, sınıflandırılmamış veriyi sınıflandırmak için uygulanabilecek bir model geliştirilmesidir[40].

Sınıflandırmada, hedef bir kategorik değişken vardır. Veri madenciliği modeli kayıtların geniş bir veri kümesini inceler, her bir kayıt hedef değişken ve girdi veya tahmin edici değişkenlerin bilgisini içerir. Ticari ve araştırma alanındaki sınıflandırma işlerine aşağıdaki örnekler verilebilir[41]:

- Belli bir kredi kartı işleminin hilekârlık amaçlı olup olmadığının tespit edilmesi
- Özel ihtiyaçlar ile ilgili olarak yeni bir öğrencinin belirli bir alana yerleştirilmesi
- Bir emlak kredi başvurusunun iyi veya kötü kredi riskine sahip olduğunun belirlenmesi
- Mevcut belirli bir hastalığın teşhisi
- Bir vasiyetin gerçekten ölmüş kişi tarafından mı, yoksa hilekârlık amacıyla başka biri tarafından mı imzalandığının tespit edilmesi
- Belirli bir finansal veya kişisel davranışın olası bir terörist tehdidi işaret edip etmediğinin belirlenmesi

3.4.2. Tahmin (Estimation)

Tahmin ile sınıflandırma bir istisna dışında benzerdir; tahminde hedef değişken kategorik değil sayısaldır. Modeller, tahmin ediciler gibi hedef değişkenin değerini sağlayan "eksiksiz" kayıtlar kullanılarak geliştirilir. Sonra, yeni gözlemler için, tahmin edicilerin değerine dayalı olarak hedef değişken tahminleri yapılır. Örneğin, hastane kayıtlarında okunan sistolik kan

[39] Yao, J. T., Y. Y. Yao and Y. Zhao, "Foundations of Classification". In Lin Tsau Y., Setsuo Ohsuga, Churn-Jung Liau, Xiaohua Hu (Eds.), **Foundations and Novel Approaches in Data Mining.** Berlin Heidelberg: Springer, 2006. s. 75.
[40] Berry, Michael J.A. and Gordon S. Linoff, s. 9.
[41] Larose Daniel T., 2005. s. 14.

basıncını, hastanın yaşı, cinsiyeti, vücut ağırlık indeksi ve kan sodyum seviyesine bağlı olarak tahmin etmek isteyebiliriz. Sistolik kan basıncı ve eğitim setindeki tahmin edici değişkenler arasındaki ilişki bir tahmin modeli sağlayacaktır. Daha sonra bu modeli yeni vakalara uygulayabiliriz[42].

Pratik hayatta tahmin, çoğunlukla bir sınıflandırma görevini yerine getirmek için yapılır. Bunun yanında, tahmin yaklaşımı her kaydın gerçekleşme olasılığına göre sıralanabilmesi nedeniyle önemli bir avantaja sahiptir. Regresyon modelleri ve sinir ağları tahmin amaçlı kullanılmak için çok uygun modellerdir. Bunların yanında da sağ kalım analizi, müşterinin alışverişi durdurması gibi bir olayda zamanın tahminini amaçlayan işler için çok uygundur[43].

3.4.3. Öngörme (Prediction)

Birçok veri madenciliği problemi gelecek çıktıları tahmin edebilecek modeller, denklemler veya makine öğrenmelerinin yapımına bağlıdır. Birçok uygulamada sadece gelecekte doğru tahminlere sahip olmakla ilgilenmeyiz, aynı zamanda bir gözlem ve çıktıları arasındaki niteliklerin ilişkilerini de öğrenmeyi isteriz. Amacımız yalnızca tahmin veya altında yatan mekanizmayı anlamak olsa da, iyi yapılandırılmış ve iyi tahmin edilmiş öngörü modeli sürecin birinci adımıdır[44].

Öngörü, sınıflandırma ve tahmine benzerdir, ancak öngörü için bir istisna vardır, çünkü sonuçlar gelecekte yatmaktadır. Sınıflandırma ve tahmin için kullanılan yöntem ve tekniklerden herhangi birisi uygun şartlar altında ayrıca öngörü için de kullanılabilir. Bunlar nokta tahmini ve güven aralığı tahminlerinin geleneksel istatistik yöntemlerini, basit doğrusal regresyon ve korelâsyonu, çoklu regresyonu, sinir ağlarını, karar ağaçlarını ve k-en yakın komşuluk yöntemlerini içermektedir[45].

[42] Larose Daniel T., 2005. s. 12.
[43] Berry, Michael J.A. and Gordon S. Linoff, s. 9-10.
[44] Ridgeway G. "Strategies and Methods for Prediction". In Ye N. (Ed.). **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. s. 160.
[45] Larose Daniel T., 2005. s. 13.

3.4.4. Zaman Serisi Analizleri (Time Series Analysis)

Bir zaman serisi eşit veya değişen zaman aralıklarında kaydedilmiş veri değerlerinin bir dizisidir. Zaman serileri verisi genellikle her bir gözleme ait değerlerin kaydedildiği zamanı belirten zaman göstergelerini içerir. Zaman serileri verisi genellikle bir veri akışı şeklinde aktarılır, örneğin veri değerlerinin sürekli akışı gibi. Zaman serileri analizi ve veri madenciliği yöntemleri amaç ve kullandığı algoritmalar bakımından farklıdırlar. Mevcut yöntemlerin çoğu; trend analizi ve öngörü, benzerlik araştırması ve ilişki analizi gibi kategorilerden bir tanesinin içerisinde yer alır. Zaman serileri ve veri madenciliği; finans, biyomedikal ve meteoroloji gibi birçok alanda ortak uygulamaya sahiptir, çünkü zaman serileri verisi farklı alanlardaki çeşitli kaynaklar tarafından oluşturulabilir[46].

Zaman serileri verisinin yaygınlığı ve önemi bu verilerin analizi üzerinde yapılan birçok araştırmayı ateşlemiştir. Zaman serileri üzerine istatistik literatürü çok büyüktür ve esas olarak örüntü belirleme ve trend analizi, mevsimsellik ve tahmin gibi problemler üzerine odaklanılmıştır. Bununla birlikte, istatistikçiler zaman serileri benzerliği ve indeksleme problemleri için uygun olacak yöntemler üzerinde çalışmamışlardır; bununla ilişkili problemlerin çoğu bilgisayar bilimi topluluğu tarafından çözülmüştür ve bu sorulara cevap vermek için veri madenciliği araştırmaları önerilmiştir[47].

3.4.5. Kümeleme (Clustering)

Kümeleme, verinin toplanmasındaki ayrımı keşfetme yapısı ile ilgilidir. Kümeleme işi hem kavramsal olarak hem de hesaplama ile ilgili olarak zorlayıcıdır. Kümeleme işleminden, isminin de ifade ettiği gibi, üzerinde düşünülen veri setindeki bireysel veri noktaları arasındaki benzerlikleri veya farklılıkları (noktalar arasındaki uzaklıklar gibi) inceleyerek kendiliğinden yapıyı keşfetme yeteneğinde olan uygun danışmansız bir algoritma beklenir. Bu yüksek derecede sezgisel ve çekici rehber, aldatıcı derecede basit gibi görünmektedir. Çünkü yapılan işlem temelde "iki veri noktasının her biri eğer birbirine 'yakın'sa kümele ve yeni şekillenen kümeler ve kalan veri noktaları arasındaki mesafeleri inceleyerek aynısını yapmaya devam et" şeklinde ifade edilebilir. Ancak kümelerin oluşumu için kullanılabilecek farklı

[46] Mehmet Sayal, "Time Series Analysis and Mining Techniques", In John Wang (Ed.). **Encyclopedia of Data Warehousing and Mining.** Hershey: Idea Group Inc, 2006. s. 1120.
[47] Das G. and Dimitrios Gunopulos, "Time Series Similarity and Indexing", In Ye N. (Ed.) **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. s. 280.

stratejilerin sayısı çok fazladır ve çok sayıda yaklaşım veri içindeki elemanlar arasındaki "benzerlik" in ne anlama geldiğini belirlemeye çalışmaktadır[48].

Kümeleme; kayıtları, gözlemleri veya vakaları benzer amaçların sınıflarını oluşturacak şekilde gruplama ile ilgilidir. Bir küme, birbirine benzeyen ama diğer kümelerdeki kayıtlara benzemeyen kayıtların derlemesidir. Kümeleme, kümelemek için bir hedef değişken olmadığından dolayı sınıflandırmadan farklıdır. Kümeleme işi bir hedef değişkenin değerini sınıflandırmaya, tahmin etmeye veya öngörmeye çalışmaz. Bunun yerine kümeleme algoritması, bütün veri setini alt homojen gruplar veya kümeler oluşturacak şekilde bölümlemeye çalışır; bunu yaparken de küme içindeki kayıtların benzerliğini en yükseğe ve küme dışındaki kayıtlar ile olan benzerliği en düşüğe indirmeye çalışır[49].

3.4.6. Birliktelik Kuralları (Association Rules)

Birliktelik kuralları madenciliği, kümelemenin haricindeki başka bir diğer anahtar danışmansız veri madenciliği yöntemidir; veri parçalarının büyük veri setleri içerisindeki ilgi çekici birliktelikleri (ilişkiler, bağımlılıklar) bulur. Veri parçaları, bir dış süreç tarafından oluşturulabilecek veya ilişkisel veri tabanları veya veri ambarlarından çekilecek işlemler şeklinde depolanır. Birliktelik kuralları algoritmalarının iyi ölçeklenebilir özelliklerinin olmasından ve biriktirilen datanın sürekli büyümesinden dolayı, birliktelik kuralları veriden bilgiyi çıkarmak için gerekli temel bir veri madenciliği aracıdır. İlgi çekici ilişkilerin keşfedilmesi, karar vermek için iş birimleri tarafından çoğunlukla kullanılan bir bilgi kaynağını sağlamaktadır[50]. İş dünyasında en yaygın olan, benzerlik analizi veya pazar sepet analizi olarak bilinen, birliktelik analizi iki veya daha fazla özellik arasındaki ilişkiyi ölçmek için kuralları açığa çıkarmaya çalışır[51].

Birliktelik kural keşfi, sınıflandırma kural çıkarımının birçok özelliğini paylaşır. Her ikisi de veri seti içindeki düzenliliği karakterize eden kuralları kullanır. Bununla birlikte, bu iki model amaçları itibariyle farklılık gösterir. Sınıflandırma amaçlı kural çıkarımı, öngörüde bulunmak için bir yetenek elde etmeye odaklanırken; birliktelik kural keşfi, bir şeyin iç yüzünü kavramaya yönelik bilgi sağlamaya odaklanır. Özellikle, veri elemanları arasındaki

[48] Cios Krzysztof J. ve diğerleri, s. 257.
[49] Larose Daniel T., s. 16.
[50] Cios Krzysztof J. ve diğerleri, s. 289.
[51] Larose Daniel T., s. 17.

beklenmeyen ilişkileri keşfetmeye ve karakterize etmeye odaklanır. Sınıflandırma kural çıkarımı tipik olarak, eğitim verisinin müşterek olarak çoğunluğunu kapsayan az sayıdaki kuralı bulmak için bulgusal (heuristic) araştırmayı kullanır. Birliktelik kural keşfi tipik olarak tüm eğitim verisini kapsamayı dikkate almaksızın çok sayıdaki kuralı bulmak için tam araştırmayı kullanır. Küçük kural setlerini keşfetmeye odaklanmaktan dolayı, sınıflandırma kural çıkarım sistemleri sıklıkla benzer performanslı alternatif kurallar arasından seçim yaparlar. Buna karşın, birliktelik kural keşfi sistemleri kullanıcı tanımlı kısıtları karşılayan tüm kuralları getirir ve kullanıcının hangi belirli kuralın daha büyük değere sahip olduğunu tespit etmesine izin verir[52].

Birlikte gerçekleşme olasılığı yüksek olaylar arasındaki ilişkiyi tespit eder. Pazar sepet analizi, çapraz satış, bilgisayar ürünleri satan bir firmanın ilgili ürünlerin web sayfalarına yaptığı yönlendirme, bu yöntemle yapılan uygulamalara örnek olarak verilebilir.

3.4.7. Ardışıklık Keşfi (Sequence Discovery)

Birliktelik kuralları aynı anda meydana gelen durumları bulur; örneğin, "hangi parçalar verilen bir zamanda birlikte alınmış?" gibi. Bir sonraki doğal soru olayların sırası ve ne anlama geldiği ile ilgilenir. Ardışıklık kuralları kurma süreci, birliktelik kuralları kurma süreci ile benzerdir. Ardışıklık keşfini yapabilmek için bir zaman serisi parçaları sıralı bir dizi haline getirilir. Bu sadece sıralı bir işlemler dizisinden farklıdır. Genel olarak, zaman serileri müşteri hakkındaki bilgiyi tespit etmeyi içerir, çünkü bu bilgi seri içindeki farklı işlemleri birbirine bağlamak için kullanılır[53].

Ardışıklık keşfi, birbirini takip eden olayların sırasını tespit eden kuralların bulunması sürecidir. Bu analiz sadece yukarıda ifade edildiği gibi sadece pazarlama fonksiyonlarına yardımcı olmak amacı ile kullanılmaz. Örneğin, Hilekârlık yapmak isteyen kişiler tarafından sıklıkla uygulanan bir olaylar dizisini tespit ederek ödememe riskini belirlemek için de kullanılabilir.

[52] Webb, Geoffrey I. "Association Rules", In Ye N. (Ed.). **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. s. 26.
[53] Berry, Michael J.A. and Gordon S. Linoff, s. 318.

3.4.8. Tanımlama (Description)

Bazen veri madenciliğinin amacı basitçe insanlar, ürünler veya ilk noktada veri üreten süreçler hakkındaki anlayışımızı artıracak şekilde karmaşık veritabanlarında neler olup bittiğini betimlemektir. Bir davranışın yeterli derecede iyi bir betimlemesi, çoğunlukla bir açıklama da ortaya koyacaktır. En azından, iyi bir betimleme, bir yorum ararken nereden başlanacağı hakkında fikir verir[54].

Araştırmacılar ve analistler basitçe verinin içinde saklı örüntü ve eğilimleri tanımlamak için yollar bulmaya çalışmaktadırlar. Örneğin, bir anket işten çıkarılanların başkanlık seçiminde mevcut görevdeki başkanı destekleme olasılığının düşük olmasının altında yatan nedenleri açığa çıkarabilir. Örüntü ve eğilimlerin betimlemeleri, çoğunlukla bu örüntü ve eğilimler için olası açıklamaları da ortaya koyar. Yüksek kaliteli betimleme çoğunlukla keşifsel veri analizi tarafından yerine getirilebilir ve bunlar genellikle örüntülerin ve eğilimlerin araştırılmasındaki veri keşfinde kullanılan grafiksel yöntemlerdir[55].

3.5. Veri Madenciliğinin Uygulama Alanları ve Örnekleri

Veri madenciliği disiplininin bu hızlı gelişiminin arkasında iş problemlerinde ortaya konan olumlu sonuçlar birinci derecede rol oynamıştır. VM, bankacılık, sigorta, finansal ürünler piyasası, hızlı tüketim, mühendislik ve tıp gibi çok farklı sektörlerde olduğu gibi Gümrük Müsteşarlığı, Emniyet, Genelkurmay, Devlet Planlama Teşkilatı gibi pek çok kamu kuruluşu tarafından da kullanılmaktadır. Aşağıda VM başlıca Pazarlama Yönetimi ve Risk Yönetimi gibi alanlara göre uygulama konuları yer almaktadır:

- Müşteri segmentasyonu (Customer Segmentation)
- Pazar sepet analizi (Market Basket Analysis)
- Müşteri ilişkileri yönetimi (Customer Relationship Management)
- Müşterilerin satın alma alışkanlıklarını belirlenmesi
- Hedef müşteri kitlesinin belirlenmesi
- Ürün yönetimi

[54] Berry, Michael J.A. and Gordon S. Linoff, **Data Mining Techniques for Marketing, Sales, and Customer Relationship Management.** Indianapolis: Wiley Publishing, 2004., s. 12.
[55] Larose Daniel T., 2005. s. 11.

- Kampanya yönetimi (Campaign Management)
- Fiyatlandırma
- Müşteri sadakat analizi
- Mevcut müşterilerin elde tutulması (Retention)
- Sadakat Analizi, sistemden ayrılacak müşterilerin öngörülmesi (Churn)
- Yeni müşteri kazanımı
- Müşteri ömür değeri
- Satış tahmini
- Gelir tahmini
- Çapraz satış (Cross Sell)
- Dikey satış (Up-sell)
- Kredi skorlama
- Kredi kart sahtekârlığı
- Kara para aklama
- Sigorta sahtekârlığı
- Telefon kullanım sahtekârlığı
- İnternet işlemleri sahtekârlığı

Asıl kullanım alanları yukarıda yer alan başlıklar olsa da pek çok başka alanda veri madenciliği uygulamaları kullanılmaktadır. Finans, hızlı tüketim, telekomünikasyon gibi çok miktarda verinin yoğun olarak kullanıldığı sektörlerde kullanılmakla birlikte hemen tüm sektörler içerisinde kendisine uygun bir kullanım alanı mutlaka bulmaktadır.

Telekomünikasyon sektöründe veri madenciliği yapmak pazarın anlaşılması ve yüksek rekabet şartlarına uyum sağlayabilmek için şarttır. VM başlıca; müşteri segmentasyonu, kampanya yönetimi, sistemden ayrılacak müşterilerin öngörülmesi (churn analysis), sahtekâr abonelerin tespit edilmesi, abone kullanım alışkanlıklarının belirlenmesi, kaynakların en iyi şekilde planlanması ve servis kalitesinin artması ve tabi ki kredi skorlama amacı ile yapılmaktadır. VM kullanılarak kredi skorlama ile her müşteriye en uygun kredi limitinin atanması, erken uyarı sistemi oluşturarak riskli abonelere uygun aksiyonların zamanında alınması, düzenlenecek cihaz kampanyalarından maksimum getirinin elde edilmesi mümkün olabilmektedir.

29 NBA takımının 16'sı IBM tarafından geliştirilen Advanced Scout isimli veri madenciliği programını kullanmaktadır. Bu yazılım, Amerikan Ulusal Basketbol Birliğinin basketbol oyunu içerisinde "olay" olarak tanımladığı basket, şut, rebound, top çalma gibi verileri kullanarak takım koçlarının gözden kaçırabileceği gizli desenlerin ortaya çıkarılmasını sağlamaktadır. Buradan elde edilen bilgilere dayanarak, takım çalıştırıcıları eksikliği görülen konularda strateji geliştirebilmektedirler[56].

Medya alanında da ilginç veri madenciliği çalışmaları mevcuttur: 1998 ile 2002 yılları arasındaki 834 filme ait yedi adet bağımsız değişken kullanılarak, filmlerin gişe başarıları yapay zekâ modelleri ile tahmin edilebilmiştir[57]. İtalya'da televizyon izleyicisi için; rekabetçi başka kanallar mevcut iken, seçili kanalda yayımlanacak hangi programın maksimum izleyiciyi ekran başına çekeceği ile ilgili tahminler üretilmiştir[58].

100'den fazla farklı kaynaktan alınan 49.652 kayıt kullanarak Bellek Temelli Nedenleme (BTN) (Memory Based Reasoning - MBR) yeni vakalara sınıflandırma kodları atamıştır ve yazarlardan birisi tarafından yürütülmüş bir çalışmaya dayandırılmıştır. Bu vakadan çıkarılan sonuçlar göstermiştir ki MBR, yüzlerce kategori ve kullanımı zor serbest-metin verisi içeren bir problem üzerinde insanlar gibi işlem yapabilmektedir. Gerçek editörler tarafından yapılan doğru sınıflama oranı %88 iken model %80'lik başarı oranını yakalamıştır. Bu çalışma, biraz daha fazla deneme ve çabayla MBR tarafından üretilen sonuçların, önemli bir eğitime sahip ve zaman alıcı bir iş yapan gerçek editörler ile bahsedilen noktada rekabet edebilecek düzeye gelebileceğini göstermiştir[59].

Ergenekon Davası olarak isimlendirilen davanın 2.455 sayfalık ilk iddianamesinde ismi geçen kişilerin sosyal diyagramını gösteren çalışma, Ergenekon.tc sergisi ile yayımlanmıştır. Ergenekon.tc projesini oluşturan bilgisayar programlarından biri iddianamenin tümünü analiz ederek isimlerin metin içindeki yakınlıklarına göre ilişkilerini çıkarmakta; diğeri ilişkilerden oluşan ağı göstermektedir. Ortaya çıkan haritada isimlerin büyüklüğü tekrarları, isimler arası

[56] Larose Daniel T., s. 3.
[57] Olson, David L. and Dursun Delen, **Advanced Data Mining Techniques.** Berlin: Springer, 2008. s. 164.
[58] Giudici, P., **Applied Data Mining: Statistical Methods for Business and Industry**. West Sussex: John Wiley & Sons Ltd., 2003. s. 323.
[59] Berry, Michael J.A. and Gordon S. Linoff, s. 265.

göreceli yakınlıklar ilişkilerin yoğunluklarını, koyu bölgeler bağlılık sıklığından oluşan merkezleri gösterir[60].

Şekil 6 : Ergenekon Davası İlk İddianamesinin Diyagramı
Kaynak: Ergenekon.tc çalışması (Burak Arıkan)

KDnuggets tarafından yıllık olarak yapılan VM uygulamaları konusundaki araştırmanın özet sonuçları aşağıda Tablo 6'da yer almaktadır. Bu araştırmaya göre, müşteri analizi ve bankacılık en yüksek oranda VM uygulanan alanlar olurken, ilgili üç yılın ortalamalarına bakıldığında VM uygulamalarının yaklaşık % 15'i de kredi skorlama alanında gerçekleştirilmektedir.

[60] Ergenekon.tc - Burak Arıkan, http://ergenekon.tc/ (25.05.2009)

Tablo 6 : Veri Madenciliğinin Kullanıldığı Sektörler Araştırma Sonuçları

Endüstri / Alan	2006		2007		2008	
	Adet (n)	Yüzde (%)	Adet (n)	Yüzde (%)	Adet (n)	Yüzde (%)
Müşteri Hizmetleri Yönetimi	43	40.2%	36	26.1%	41	36.9%
Bankacılık	1	0.9%	33	23.9%	34	30.6%
Doğrudan pazarlama / Fon toplama	22	20.6%	28	20.3%	15	13.5%
Sahtekarlık tespiti	24	22.4%	26	18.8%	21	18.9%
Bilim	12	11.2%	26	18.8%	11	9.9%
Telekomünikasyon	14	13.1%	21	15.2%	13	11.7%
Kredi skorlama	**21**	**19.6%**	**19**	**13.8%**	**14**	**12.6%**
Bioteknoloji / Genomics	17	15.9%	16	11.6%	12	10.8%
Web kullanım madenciliği	12	11.2%	14	10.1%	8	7.2%
Perakendecilk	11	10.3%	14	10.1%	13	11.7%
Tıp / Eczacılık	8	7.5%	13	9.4%	8	7.2%
Sigortacılık	12	11.2%	12	8.7%	11	9.9%
Hükümet / Askeriye	7	6.5%	10	7.2%	4	3.6%
Sağlık hizmetleri / İnsan kaynakları	5	4.7%	10	7.2%	10	9.0%
Web içerik madenciliği / Araştırma	15	14.0%	9	6.5%	6	5.4%
Üretim	7	6.5%	9	6.5%	9	8.1%
e-ticaret	6	5.6%	8	5.8%	8	7.2%
Eğlence / Müzik	2	1.9%	6	4.3%	3	2.7%
Güvenlik / Anti-terörizm	5	4.7%	5	3.6%	6	5.4%
Yatırım / Hisse senedi	11	10.3%	4	2.9%	14	12.6%
Seyahat / Turizm	5	4.7%	3	2.2%	3	2.7%
Toplu / İstenmeyen mail	2	1.9%	3	2.2%	3	2.7%
Sosyal politika araştırmaları	0	0.0%	5	3.6%	8	7.2%
Finans	0	0.0%	10	7.2%	18	16.2%
Reklamcılık	0	0.0%	0	0.0%	13	11.7%
Sosyal Ağ	0	0.0%	0	0.0%	2	1.8%
Diğer	15	14.0%	18	13.0%	14	12.6%

Kaynak: http://www.kdnuggets.com/polls/2008/data-mining-applications.htm (26.06.2009)
Not: 2006'da 111, 2007'de 138, 2008'de de 107 araştırmacı tarafından cevaplanmış olup birden fazla cevap alınmıştır

3.6. Veri Madenciliğinde Yeni Gelişen Alanlar

Verinin, veri madenciliği işlerinin ve veri madenciliği yaklaşımlarının çeşitliliği, veri madenciliğinde birçok zorlayıcı araştırma konusunu ortaya çıkarmaktadır. Aşağıdaki maddeler bu zorlayıcı konuları yansıtan veri madenciliği eğilimlerinden bazılarını tanımlamaktadır[61]:

- Metin madenciliği
- Web madenciliği
- Görselleştirme
- Grafik madenciliği, bağlantı analizi ve sosyal ağ analizi

[61] Han, J. and Micheline Kamber, s. 681-683.

- Ölçeklenebilir ve etkileşimli veri madenciliği yöntemleri
- Çoklu ilişkili ve çoklu veritabanlı veri madenciliği
- Başvuru inceleme
- Veri madenciliği ile veri tabanı sistemleri, veri ambarı sistemleri veya web veri tabanı sistemlerinin entegrasyonu
- Veri madenciliği dilinin standartlaştırılması
- Gerçek zamanlı veri madenciliği
- Veri madenciliğinde özel yaşamın korunması ve bilgi güvenliği

Bu gelişmelerin pek çoğu, uzaya ait veya coğrafi veri (spatiotemporal), multimedya verisi, text verisi ve web verisi gibi kompleks veri tipleri için uygun madencilik yöntemleri geliştirilmesini kapsamaktadır. Bunların yanısıra uygulamada karşılaşılan zorlukları aşmak için geliştirilmesi gereken alanlar da yukarıdaki listede yer almışlardır. Bunların içerisinde, günümüzde yoğun olarak başvurulan bir kısım VM alanı detaylı olarak aşağıda tanıtılmıştır.

Veri madenciliğinin önceki çalışmalarının çoğu ilişkisel, işlemsel veya veri ambarında bulunan veri gibi yapısal veriye odaklanmıştır. Bununla birlikte, gerçekte mevcut bilginin önemli bir kısmı metin veritabanlarında (veya doküman veri tabanları) tutulmaktadır. Yeni makaleler, araştırma yayınları, kitaplar, dijital kütüphaneler, e-posta mesajları veya web sayfaları gibi çok çeşitli kaynaklar, çok geniş bir doküman koleksiyonunu içerirler[62]. Metin madenciliği ile ele alınan en popüler problemlerden birisi doküman sınıflandırmasıdır. Doküman sınıflandırma, içeriklerine dayalı olarak önceden tanımlanmış kategorilere göre dokümanların sınıflandırılmasını amaçlar. Metin madenciliği ile ele alınan diğer önemli problemler; içeriğine bağlı olarak doküman araştırmayı, otomatik doküman özetlemeyi, otomatik doküman gruplamayı ve doküman hiyerarşisi oluşturmayı, doküman yazarlığının tespitini, doküman intihalinin tespitini, başlık tespiti ve takibini, bilgi çıkarmayı, yardımlı metin analizini ve kullanıcı profili çıkarmayı içerir[63].

Çoğu metin veritabanlarında tutulan veri yarı yapılandırılmış veridir; ne tamamen yapısızdır ne de tamamen yapılandırılmıştır. Geleneksel bilgi çıkarma teknikleri, hızlı bir şekilde artan büyük miktardaki metin verisi için yetersiz kalmaktadır. Bilginin elde edilmesi, uzun

[62] Han, J. and Micheline Kamber, s. 614.
[63] Mladenić D. "Text Mining-Machine Learning on Documents". In Wang J.(Ed.) **Encyclopedia of Data Warehousing and Mining.** Hershey: Idea Group Inc, 2006. s. 1109.

yıllardan beri veri tabanları ile paralel gelişen bir alandır. Yapısal verinin sorgulanması ve işlenmesi sürecine odaklanan veri tabanı sistemlerinin alanından farklı olarak bilginin bulunup elde edilmesi, çok miktardaki metin tabanlı dokümanlardan bilginin elde edilmesi ve organizasyonu ile ilgilenir. Tipik bir bilginin bulunup elde edilmesi problemi, kullanıcının sorgusuna bağlı olarak, sıklıkla ihtiyaç duyulan bilgiyi tarif eden anahtar kelimelerdir ve bir doküman koleksiyonu içerisindeki ilgili dokümanları belirler[64].

Web madenciliği üç sınıfa ayrılabilir: 1) Web yapı madenciliği; 2) Web içerik madenciliği; 3) Web kullanım madenciliği. Web yapı madenciliğinin amacı, Web sitesi ve Web sayfalarının birbirleri ile bağlantısına bakarak bilgi üretmektir. Web içerik madenciliği Web dokümanlarından yararlı ve gerekli bilgiyi elde etmek için kullanılır. Web kullanım madenciliği veri madenciliği teknikleri kullanılarak, Web kullanıcılarının İnternetteki davranışları için örüntü oluşturmak biçiminde tanımlanabilir[65].

Grafikler ve görselleştirme, özellikle büyük miktardaki verinin yönetilmesini, işlenmesini ve analiz edilmesini gerektiren uygulamalar için, veri inceleme ve anlama içinde temel noktadır. Veri madenciliği amaçlı grafik çizme tekniklerinin faydaları; görsel inceleme vasıtasıyla analizini, örüntü ve korelâsyonların keşfini, çıkarımını ve özetlemeyi içerir. Grafik çiziminden kaynaklanan teknolojik çözümlerden faydalanan farklı alanlar arasında başlıcaları şunlardır: Internet programlama, sosyal bilimler, yazılım mühendisliği, bilgi sistemleri, ulusal güvenlik, web araştırma ve sayısal biyoloji[66].

[64] Han, J. and Micheline Kamber, s. 615.
[65] Gündüz Ş. ve Eşref Adalı, "Web kullanıcılarının davranışları için örüntü bulma ve modelleme", **İTÜ Dergisi/d Mühendislik**, Cilt:3, Sayı:6, Aralık 2004,16-17.
[66] Didimo W. and Giuseppe Liotta, "Graph Visualization and Data Mining", In Cook, D.J. and Holder L.B. (Ed.) **Mining Graph Data**. New Jersey: John Wiley & Sons, 2007. s. 36.

4.KREDİ SKORLAMASI

4.1. Kredi Skorlaması

Türk Dil Kurumu sözlüğüne göre Kredi, 'ödünç alınan veya verilen mal veya parayı' ifade etmekte olup, aynı zamanda 'borç ödemede güvenilir olma durumu' için de kullanılmaktadır. Kredi kavramı biraz daha yakından incelendiğinde; geri alınmak üzere verildiği, süresiz veya belirli bir süreye sahip olduğu, belirli bir faiz veya gelir karşılığı kullandırıldığı vb. gibi pek çok detaya sahiptir. En genel olarak, müşterinin krediyi geri ödememesi olasılığını (probability of default) hesaplamaya kredi skorlama denir.

Kredi verilmesi kararını etkileyen en önemli nokta, kredinin tamamının ve faizlerinin geri alınıp alınamayacağıdır. Bu alanda, özellikle yığın ile ilgili alınacak kararlarda, kredi skoru kullanılır. Kredi skorlaması, abonelere ait geçmiş veriyi kullanarak, aynı bireylerin benzer davranacakları varsayımı altında, geleceği tahmin etmeye çalışan tekniklere denir. Yeni kredi başvurularının değerlendirmesi veya mevcut kredi limitinin belirlenmesi gibi kararların alınmasında kredi veren yetkiliye yol gösterir.

Skor kart ise, bireyleri iyi veya kötü ödeme gibi bazı istenen niceliklere dayalı olarak derecelendirmek için kullanılan matematiksel bir araçtır. Bu yüzden etkin bir skor kart, iyi ve kötü performans gösteren hesapları birbirinden ayrıştırabilir. Aynı şekilde skor kartlar, kredinin geri ödeme olasılığını tahmin edecek bir gösterge sağlarlar[67].

Kredi skorlamada kullanılan veriler; abone tarafından beyan edilen veriler, abone tarafından sağlanan dokümanlar vasıtasıyla elde edilen veriler, firma içerisinde daha önceden geçmişi varsa buradan elde edilen veriler ve dış kaynaklardan sağlanan verilerdir. Kullanılan verilerin bir kısmı doğrulanmış bir kısmı da doğrulanma ihtimali olmayan veriler olabilir.

Skor kartlar ham veriden bilginin elde edilmesi amacıyla kullanılırlar, ancak asıl faydası skorlama sistemi etkin olarak kullanıldığında ortaya çıkar. Veri, yalnızca karar vermeye yardımcı olacak skorlama sistemini besler. Skorlama sistemi de şirket stratejileri ile uyumlu olarak hayata geçirildiğinde anlamlıdır. Skorlama sisteminin, şüpheli alacakları azaltmak,

[67] Kindred D., "What is Scorecard?". In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s. 7-8.

pazar payını büyütmek, otomasyonu artırmak, hilekârlık ile mücadele etmek gibi, nasıl kullanılacağına karar verildiği zaman işin stratejisi ve amaçları dikkate alınmalıdır, şirketin ana stratejileri ile uyumlu olmalıdır.

Geleneksel kredi değerleme yöntemi, her ne kadar belirli veriler ve politikalar mevcut olsa da kredi analistinin geçmiş tecrübesine, dolayısı ile objektifliğe açık bir süreçtir. Kredi skorlama, bir çıktıyı öngörmek için sayısal bir formül kullanarak bilginin spesifik parçalarına puan atamaktır. Bugün Kredi Skorlama Modelleri sadece kredibilitenin tahmin edilmesinde değil aynı zamanda potansiyel batma riski, gelir tahmini, yeni aktivasyon, kullanım, karlılık, tahsilât, müşteri kaybı, sahtekârlık, sigorta kayıpları ve küçük işletmelerin kredi risklerinin belirlenmesinde de kullanılmaktadır[68]. Bunların haricinde, aboneye göre kredi şartlarının belirlenmesi, riske göre kredi fiyatlaması, borçlunun limit artış, azalışı ve kredinin yenilenmesi gibi konularda da kredi skorlaması kullanılabilir. Günümüzde ise amaç fonksiyonu, müşterilerin borç ödememelerini minimize etmenin yerine şirketlerin bu tür müşterilerden daha fazla nasıl gelir elde edebilecekleri üzerine kurulmaya başlanmıştır.

Skor kartın işleyişi şu şekilde gerçekleşir: Değerlendirilmek istenen her başvuruya veya hesaba ait özellikler aldıkları değere karşılık gelen puana, toplama veya çıkarma yoluyla sahip olur. Skor kartların kullanımı basit, anlaşılması kolaydır. Her hesabın aldığı skor, adayın gelecekteki davranışlarını tahmin etmede kullanılır. Alınan skorlar, adayların sıralanmasında ve böylece belirli gruplara atanarak en uygun aksiyonun alınmasında kullanılabilir.

Aşağıdaki Tablo 7, altı değişkene sahip bir skor kart örneğidir. Burada yer alan değişken adedi sınırlı tutulmuştur, gerçek uygulamalarda bu sayı daha fazladır. Belirli bir gözleme ait veriler de buradaki tabloda taralı alan ile işaretlenmiştir. Bu duruma göre örnek gözlemin alacağı toplam skor değeri 440 (74+79+77+85+50+75) olacaktır. Alınan bu skora karşılık gelen odds oranına göre en uygun davranış firma tarafından uygulanacaktır[*].

[68] Chandler G. "Generic and customized scoring models: A comparison". In Elizabeth Mays (Ed.). **Credit Scoring for Risk Managers.** Thomson, South-Western, 2004. s. 15.
[*] Odds oranı ile ilgili detaylı bilgi altıncı bölümde verilmiştir.

Tablo 7 : Skor Kart Örneği ve Bir Vaka Hesaplaması

Son 6 ay içinde toplam bilançonun yüzdesi olarak toplam ödeme tutarı	0-3 % 60	4-8 % 74	9-12 % 81	13-35 % 90	36-100 % 95
Borcun vadesinden sonra ay olarak geçen süre	0 - 3 21	4 - 5 54	6 - 9 67	10 + 79	Never 90
Bilançonun yüzdesi olarak ilgili dönemde yapılan satın almalar	1 – 19 60	20 – 49 67	50 – 89 71	90 – 99 77	100 + 71
Son 6 ay için ortalama bilanço	<$250 62	$250-499 75	$500-3499 85	$3500+ 49	
Bilançonun yüzdesi olarak nakit ön ödemeler	0 85	1 – 19 68	20 – 49 50	50 + 43	
En yüksek bilançonun yüzdesi olarak güncel bilanço	1 – 39 95	40 – 69 87	70 – 79 80	80 – 89 75	90 – 100 71

Kaynak: Thomas Lyn C., Edelman David B. and Crook Jonathan N. (Ed.) Readings in Credit Scoring. New York: Oxford University Press, 2004. s. 40.

Odds oranı, iyi abonelerin kötü abonelere oranıdır. Aşağıdaki tabloda da görüldüğü gibi Odds oranı skor aralığı ile doğru orantılı olarak artar. Skor aralığına göre örnekteki gözlem sayısının dağılımı yaklaşık uniformdur. Odds oranı skor sonuçlarını yorumlamak için çok önemli bir ölçüt olup, sadece Odds oranına dayanılarak bile skor sonuçlarına göre abonelere uygulanabilecek farklı aksiyonlar belirlenebilir. Buna göre dağılımdaki risk seviyeleri belirlendikten sonra bu seviye aralıkları için uygun aksiyonlar belirlenir. Aşağıda yer alan Tablo 8'de örnek skor sonuçlarına ve Odds oranının dağılımına göre varsayılan bir başvuru skorlama ve davranışsal skorlama aksiyonları da ayrı sütunlarda aktarılmıştır.

Tablo 8 : Skor Sonuçlarına Göre Hesaplanan Odds Oranı ve Alınabilecek Aksiyon Tipleri

Skor Aralığı	İyi adet (n)	Kötü adet (n)	Odds Oranı (İyi adet /Kötü adet)	Toplam içindeki Oranı	Başvuru Skorlama Aksiyonu	Davranış Skorlama Aksiyonu
0 - 80	400	400	1	11.7%	Başvuruyu Reddet	Aboneyi Engelle / Yasal takip için harekete geç
81 - 120	500	250	2	11.0%		
121 - 160	400	100	4	7.3%	Tekrar değerlendir/ Manuel incelemeye gönder / Kısıtlı şartlar ile içeri al	
161 - 200	520	65	8	8.6%		Aboneyi riskleri hakkında bilgilendir / İdari takibe geç
201 - 240	600	60	10	9.7%		
241 - 260	525	35	15	8.2%	Başvuruyu Kabul et	
261 - 300	450	18	25	6.8%		Risksiz abone
301 - 340	600	15	40	9.0%		
341 - 400	880	11	80	13.0%		
401 - 440	1.000	10	100	14.8%		
Toplam	**5.875**	**964**	**6.1**	**100.0%**		

Kredi skorlamanın firmalara pek çok faydası mevcuttur:

1. Kararların otomasyonu ile borç verme sürecinin etkinliğini artırır ve zaman tasarrufuna neden olur,
2. Tüm başvurulara aynı standartlar ile yaklaşıldığından sübjektifliği ortadan kaldırır,
3. Mevcut abonelerin daha yakından izlenmesine imkân vererek muhtemel riskleri azaltır,
4. Nakit akışının geliştirilmesini sağlar,
5. Tahsilâtı güvence altına alır ve tahsilâtın hızlandırılmasını sağlar,
6. Kredi kararları üzerindeki kontrolü sağlayarak üst yönetimin karar etkinliğini artırır,
7. Geliri artırır, şüpheli alacak miktarını düşürür.

Kredi skorlama modelleri ile adayların finansal yükümlülüklerini yerine getirme kabiliyetleri değerlendirilerek, adaylar kredi limiti ataması yapıldıktan sonra veya doğrudan iyi veya kötü kredibiliteye sahip olarak nitelendirilirler. Bundan dolayı kredi skorlama problemleri, aslında yaygın olarak uygulanan sınıflandırma problemlerinin kapsamı içindedir.

Kredi skorlama, kredi başvurusunda bulunan gözlemlere ait ana kütleyi birbirinden önemli derecede farklı risk karakteristiklerine sahip spesifik alt kümelere ayrıştırıp tanımlamaktır. Bu konuda, Fisher tarafından (1936) tanıtılan gruplar arasında ayrıştırmaya dayalı fikirler ile başlayarak, istatistik ve yöneylem araştırması yöntemleri kullanılarak birçok farklı yaklaşım geliştirilmiştir. İstatistik araçları; doğrusal diskriminant analizi ve lojistik regresyon gibi ayırma analizleri ile sınıflandırma ve karar ağaçları gibi tekrarlanan bölümlere dayalı ayırma algoritmalarını içermektedir. Yöneylem araştırması yöntemleri, öncelikle lineer programlama gibi matematiksel programlama yöntemlerini içerirler. Ayrıca son zamanlarda bazı yeni parametrik olmayan ve yapay zekâ yaklaşımları geliştirilmiştir. Bu yöntemler her yerde görülen sinir ağlarını, uzman sistemleri, genetik algoritmaları ve en yakın komşuluk yöntemlerini içerirler[69].

Kredi skorlama ticari bankacılığın temel bir yetkinliği olarak kabul edilir. Tüm ticari banka yönetimleri, skorlama modeli ne kadar iyiyse, kötü seçim riskinin o kadar düşük ve bankaya olan katma değerinin de o kadar yüksek olduğunu kabul etmektedir. Kredi veren kuruluş derecelendirme sistemini iyileştirerek, doğru skorlamanın yan etkisi olarak portföy azalışlarını ortadan kaldırıp kredi portföyünü önemli oranda artırabilir. Bu şekilde doymuş kredi

[69] Bugera, V., Hiroshi Konno, and Stanislav Ursayev. "Credit Cards Scoring with Quadratic Utility Functions". **Journal Of Multi-Criteria Decision Analysis**. Vol.11, 2002. s. 198.

pazarında bile artış meydana gelir. Bu etkiden dolayı, gelişmekte olan ve büyüyen kredi pazarında, kâr ve pazar payındaki artış daha fazla olacaktır[70].

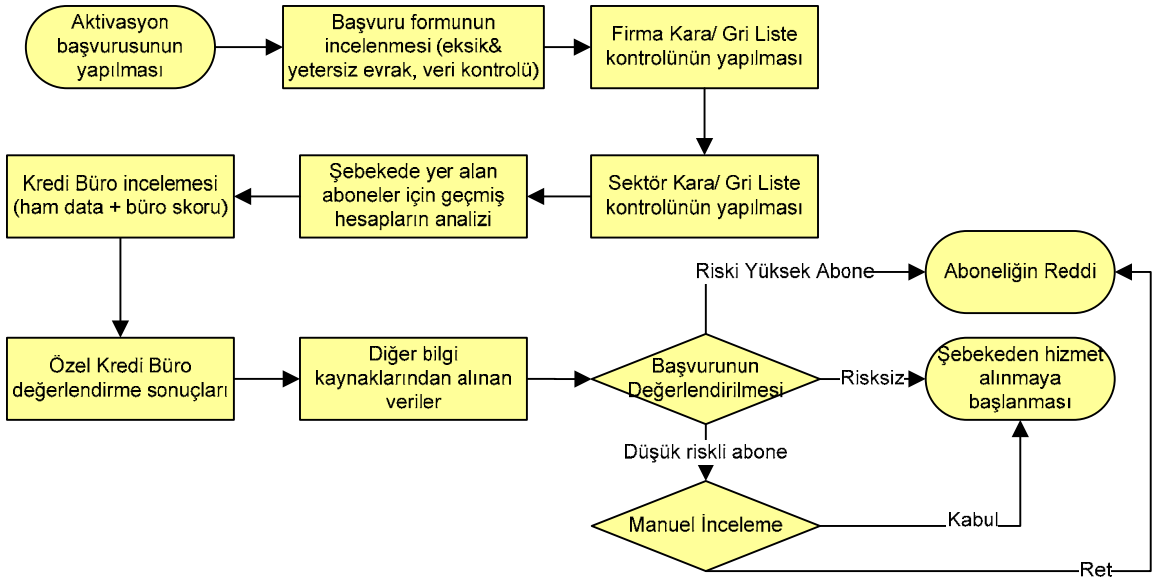
Basel II komitesi önerileri altında [Basel Committee on Banking Supervision. International Convergence of Capital Management and Capital standards: A Revised Framework. Bank for International Settlements Press & Communications, Basel, Switzerland, June 2004], sermaye paylaşımının etkinliğini artırmak amacıyla karmaşık kredi skorlama modellerini kullanmak bankalar için denetleyici kurumun bir zorunluluğu haline gelmektedir[71]. Buna göre, kredi skorlama sistemleri yasal zemine de dayanarak daha da ilerleyecek ve gelişecektir.

Müşterinin yaşam sürecinin tüm aşamalarında farklı amaçlarla Veri madenciliği uygulamaları veya skorlama yapılabilir. Bu sürecin çok önemli bir kısmını oluşturan Kredi skorlama işlemi de temel olarak iki alt başlığa ayrılır; Başvuru skorlaması ve Davranışsal skorlama

4.2. Başvuru Skorlama Modelleri

Başvuru Skorlama Modellerinde (BSM) amaç kredi almak için başvuru yapan adayı iyi ve kötü risk gruplarından bir tanesine atamaktır, yani kredinin verilip verilmeyeceğini belirler. Burada başlıca başvuru sırasında alınan demografik bilgiler, finansal veriler ve dışsal kaynaklardan alınan jenerik skor sonuçları birlikte değerlendirilir. Mobil operatörden hat almak isteyen bir abone için yapılacak BSM'de; abone tarafından tercih edilen aktivasyon kanalı, aktivasyon esnasında tercih ettiği ödeme tipi, ilgili operatörün veya sektörün kara/ gri listesinde yer alıp almaması, adres bilgisi başta olmak üzere verdiği bilgilerinin doğrulanıp doğrulanamaması, kredi bürosu tarafından verilen puan gibi pek çok veri kullanılmaktadır.

[70] Blöchlinger, A. and Markus Leippold. "Economic benefit of powerful credit scoring". **Journal of Banking & Finance.** Vol.30, 2006, s. 872.
[71] Laha, A. "Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring". **Advanced Engineering Informatics** Vol.21, 2007. s. 281.



Şekil 7 : Mobil Telekomünikasyon Sektöründe Uygulanabilecek Örnek Bir Başvuru Skorlama Süreci

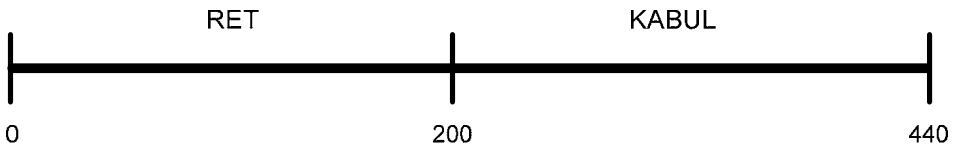
Başvuru Skorlama Modellerinin başlıca kullanılma amaçları aşağıdaki gibidir[72]:

1) Kabul edilen müracaatlar arasındaki şüpheli alacak yüzdesini azaltmak

2) Mevcut durumda reddedilen başvurular içerisinde şüpheli alacaklar ile sonuçlanmayacak olanları da kabul ederek satışların değerini ve hacmini iyileştirmek

3) Kredi onay yetkililerinin iş yüklerini azaltmak

BSM, aboneleri iyi ve kötü olmak üzere belirlerken belirli bir sınır veya eşik değerinin (cut-off point) belirlenmesi gerekir. Skor karta tam bir güven mevcutsa veya hata yapmanın maliyeti çok yüksek değilse uygulamada belirli bir nokta, kredi kabul/ reddi için sınır değeri olarak belirlenir.

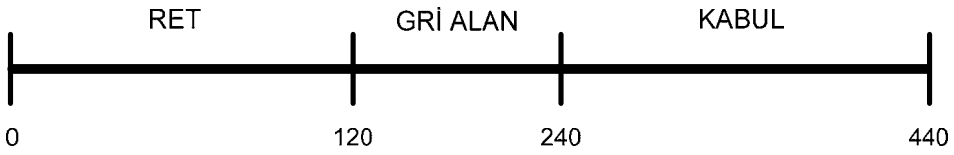


Şekil 8 : Tek Eşik Değerinin Kullanılması

BSM bir kısım aboneyi net olarak iyi veya yüksek riskli olarak belirleyemeyebilir. Çünkü pratik hayatta kredibilite olarak tanımlanan kavram sürekli bir değişkendir. Bu gibi durumlarda modeller; kabul, ret ve gri alan olarak üç gruplu bir sonuç üretebilir. Bu durumda

[72] Sarantopoulos, G. "Data Mining In Retail Credit". **Operational Research. An International Journal.** Vol.3, No.2, 2003. s. 111.

da iki eşik değeri belirlenir. Aşağıdaki örnekte alt eşik skoru 120, üst eşik skoru da 240 olarak belirlenmiştir.



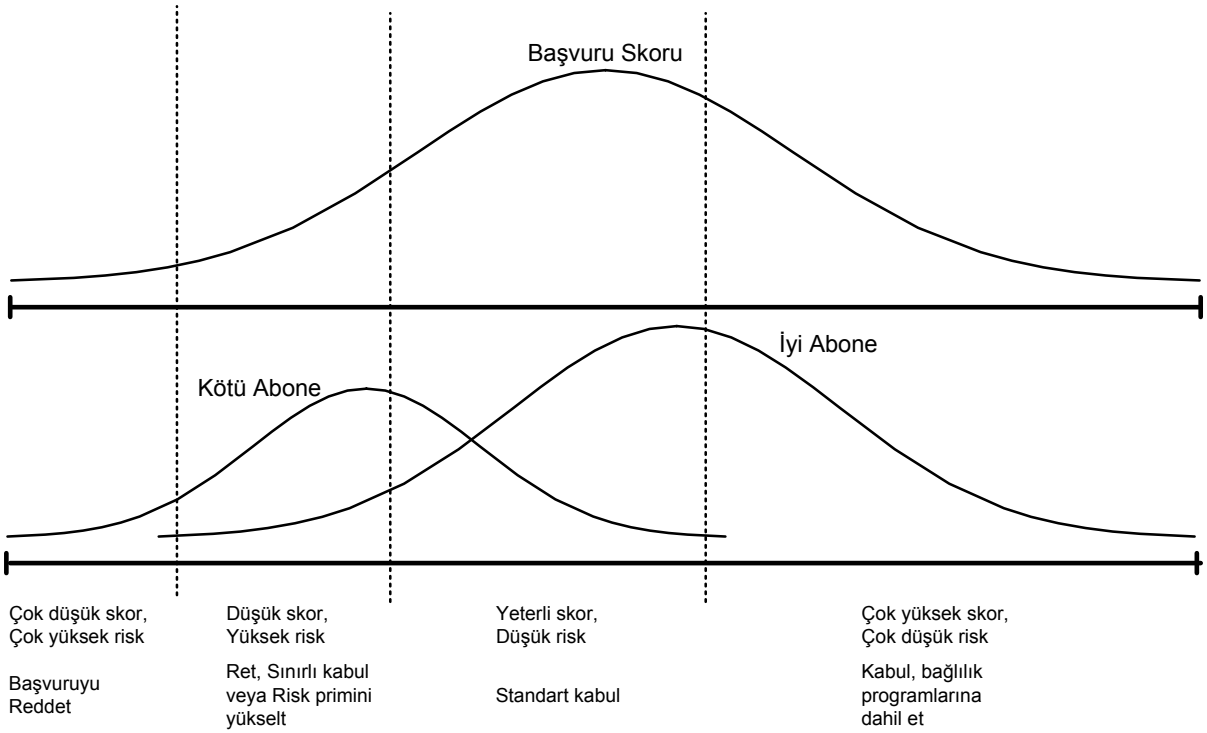
Şekil 9 : İki Eşik Değerinin Kullanılması

Gri alan olarak adlandırılan bölüme düşen başvurular, Kredi Değerleme Analistleri tarafından manüel olarak değerlendirilirler. Rengi tam olarak belli olmayan ve sınır noktasına yakın yere düşen bu aboneler için yeni bir skorlama modeli çalıştırılırsa buna İki Aşamalı Skorlama Yöntemi adı verilir. Bir diğer yol da bu abonelerden teminat veya depozit alınmasıdır. Gri alana düşen abonelere nasıl yaklaşılacağı; bu başvuruları değerlendirmek için kullanılacak işgücüne, ayrılacak kaynağa ve sektörün rekabet koşulları gibi firmanın kredi stratejilerine bağlı olarak belirlenir. Ancak bu gruba giren abonelerin oranı, daha gelişmiş modeller ile düşürülmelidir.

Kredi talebini reddetmenin maliyeti, başlıca Gelir Kaybı maliyeti iken, başvuruyu kabul etme maliyeti; Yatırım maliyeti, Operasyon maliyeti, Tahsilât maliyeti ve Ödenmeme maliyetlerinden oluşur. Yukarıda belirlenen eşik değeri, aynı zamanda belirli bir risk algılamasına göre bu maliyetlerin dengede olduğu noktadır. Ancak, reddedilen aboneden kaynaklanan memnuniyetsizlik gibi ölçülemeyen başka maliyetler de buraya dahil edilebilmelidir.

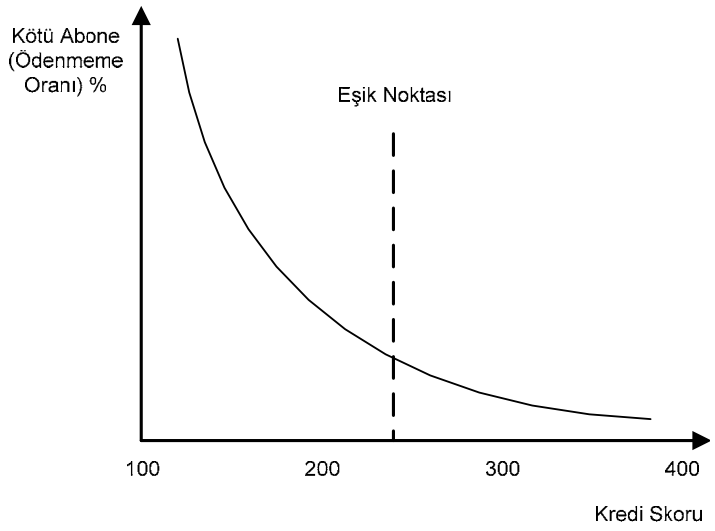
Kredi veren kuruluşların kredi politikaları nedeniyle, her durumda skor sonuçlarına dayanılarak karar alınmayabilir. Manüel analiz sonucunda veya müşteriye özel nedenlerden dolayı sistemin atadığı skor sonucuna karşıt yönde kararlar alınabilir. Buna karar ezme (override) denir.

Aşağıdaki Şekil 10, başvuru skorlama sonuçları skalası üzerinde iyi ve kötü abonelerin nasıl dağıldığını göstermektedir. İyi ve kötü abone gruplarının dağılımlarının üst üste çakışmasının önüne geçilememektedir. Bu nedenle mevcut veri ile elde edilebilecek dağılımların iyi tespit edilerek buna uygun aksiyonun iş kullanıcılarına ait olması gerekmektedir.



Şekil 10 : Başvuru Skoru ile Başvuran Abonelerin Risk Gruplarına Göre Dağılımı

Kredi kararının verilmesinde kullanılan sınır noktası; kabul edilen başvuru oranı (pazar payını etkiler) ve kötü başvuru oranı (şüpheli alacak miktarını etkiler) arasındaki denge gözetilerek tespit edilir. Zaten kurulan model sonuçları tarafından her bir skor aralığı için üretilen; kazanılacak gelir, oluşacak zarar ve odds oranı (iyi aboneler / kötü aboneler) gibi kriterler kullanılarak uygulanacak kredi stratejilerine uygun bir nokta sınır değeri olarak belirlenir. Sınır noktasının belirlenmesinde kullanılabilecek bir diğer yaklaşım da marjinal gelir ve marjinal zarar analizi yapmaktır. Eşik değeri, kötü abonelik veya ödenmeme oranı ile ters orantılı olarak hareket eden skor değerleri arasındaki kriterlere göre optimum noktada seçilir.



Şekil 11 : Riskli Abonelik ve Kredi Skoru İlişkisi

Başvuru skor kartlarının parametreleri genellikle reddedilen başvuruları dışarıda bırakan bir örneklem kullanılarak tahmin edilir. Bundan dolayı tüm başvurulara uygulandığı zaman modelin yanlı olduğu gösterilebilir. Crook, J. ve Banasic, çalışmalarında normal olarak reddedilecek aboneleri de içeren nadir bir örneklem kullanmışlardır. Ret kararının en doğru uygulamasının tüm başvuruları içeren popülasyon için hesaplanan potansiyel iyi-kötü oranının doğru tahminine bağlı olduğu görülmüştür. Reddedilen başvuruların hariç tutulması ve sadece kabul edilen başvurulara dayalı bir skor kartın oluşturulması amacıyla çok geniş bir katsayılar kümesi üzerinde çok detaylı bir şekilde çalışmanın bulgular üzerinde çok fazla potansiyel fayda sağlamadığı da görülmüştür[73]. Sadece kabul edilen başvurulara dayalı bir sınıflandırma kuralı kurmak sapmalı tahminlere yol açabilir. Özellikle, reddedilen başvurunun temerrüt riskini değerlendirmek için bu kuralı kullanmayı düşünen birisi çok dikkatli olmalıdır. Bu kredi skorlama literatüründe kısaca 'ret kararı problemi' olarak adlandırılmaktadır[74].

4.3. Davranışsal Skorlama Modelleri

Davranışsal Skorlama Modeli (DSM) mevcut müşteriler için kullanılır. Başvuru skorlama modelleri için elde edilen verilere ek olarak, borçlunun mevcut ödeme performansını gösteren değişkenler kullanılır. Borçlunun ödemelerini zamanında yapıp yapmadığı, ödeme yöntemleri gibi abonenin borç ödeme konusundaki kanıtlanmış istekliliğini ve ödeme gücünü ölçen bilgiler abonenin gelecekteki davranışlarının tahmin edilmesinde kullanılır. Telekomünikasyon sektöründe DSM, başlıca mevcut abonelere kredi limiti ataması yapmak ve cihaz kampanyası gibi kampanyalar düzenlemek amacıyla kullanılır.

Davranışsal skorlama, bir müşterinin memnuniyet verici bir durumda kalma veya bu duruma gelme olasılığını belirleyen bir araçtır. Davranışsal skorlama kredi yöneticilerine bir müşterinin gelecekte memnuniyet verici bir durumda olma olasılığı için sayısal bir değer belirlemesine imkân verir. Ancak davranış skorları hangi müşterilerin memnuniyet verici bir durumda olacağı veya hangilerinin olmayacağının kesinliği ile bir tutulmamalıdır, çünkü insan davranışının karmaşıklığı ve fiziksel dünyanın doğası herhangi bir alanda kesin tahmini

[73] Crook, J. and John Banasik. "Does reject inference really improves the performance of application scoring models?" **Journal of Banking & Finance**. Vol.28, 2004, s. 857–873.
[74] Feelders, A.J. "Credit Scoring and Reject Inference With Mixture Models". **International Journal of Intelligent Systems in Accounting, Finance & Management**. Vol.9, 2000, s. 1.

imkânsızlaştırmaktadır. Davranışsal skorlama sadece memnuniyet verici bir performansın olasılığını belirler[75].

Firmalar, kendi aboneleri için alınacak aksiyonların acil olup olmamasına ve önemine göre kullandıkları Davranışsal Skorlama Sisteminin çalışma düzenini belirlerler. En sık rastlanan durum, Davranışsal Skorlama Modellerinin ayda bir defa çalıştırılmasıdır. Ancak bu sistem, güçlü IT altyapısı ve desteği ile ihtiyaçlar doğrultusunda günlük yapıya dönüştürülebilir.

Davranışsal Skorlama Modelleri başlıca aşağıdaki durumlarda kullanılırlar:

1. Abone riskinin anlık olarak takip edilmesinde
2. Abonenin ödemelerini zamanında yapıp yapmayacağının tahmininde
3. Aboneye yeni bir finansman kaynağının yaratılıp yaratılmamasının kararının verilmesinde
4. Mevcut borcun yeniden finansman kararının verilmesinde
5. Vadesi gecikmiş tahsilâtların yapılmasında
6. Mevcut aboneye çapraz satış modellerinde

Davranışsal skorun skor kart ile ölçülmesi konusunda uzmanların iki temel eleştirisi mevcuttur. Bunlardan birincisi değerlendirilen abonelere olasılık değerinin atanamaması, ikincisi de müşteriyi doğru olarak sınıflandırabilmek için yeteri kadar sayıda faturanın mevcut olması gerektiğidir. Bu sebeplerden dolayı veri madenciliği teknikleri daha yaratıcı ve güvenilir sonuçlar üretmektedir[76].

Bu durumun daha iyileştirilmiş hali olaya dayalı skorlamadır. Olaya dayalı skorlama, müşterinin son ödeme tarihinden ziyade bir olaya dayalı olarak yeniden skorlama kabiliyetidir. İdeal olarak, en uygun aksiyon üzerinde bir karar verildiği zaman, performans skoru yeniden değerlendirilmelidir. Olaya dayalı skorlama için uygun örnekler şunlardır; karşılıksız çekten sonraki tahsilât süreci, kredi limiti artış için talep ve kredi limitini aşan resmi bir poliçe[77].

---

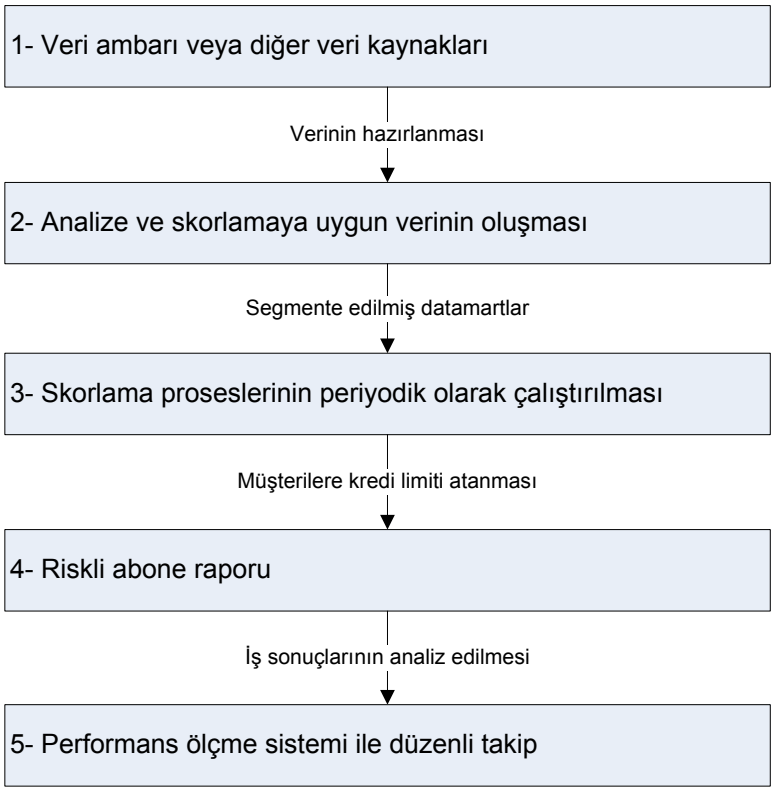
[75] Hopper M.A. and E.M. Lewis, "Behaviour Scoring and Adaptive Control Systems". In Thomas Lyn C., David B. Edelman, and Jonathan N. Crook (Ed.). **Readings in Credit Scoring.** New York: Oxford University Press, 2004. s. 39.

[76] Madeira, S. C., Oliveira, A. L. and Conceição C. S., "A Data Mining Approach to Credit Risk Evaluation and Behaviour Scoring". Fernando Moura Pires, Salvador Abreu (Eds.): **EPIA 2003 LNAI Lecture Notes in Artificial Intelligence No 2902**, Springer-Verlag Berlin Heidelberg, 2003, s. 184.

[77] Betts J., "Credit Scoring Systems". In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s. 31.

DSM, mevcut aboneler için uygulandığından abone hakkındaki pek çok ve aboneye özel veriye anlık olarak ulaşılabildiğinden BSM'ne göre çok daha güçlüdür ve daha yüksek seviyede geleceği öngörme kabiliyetine sahiptir. Bunun yanında DSM sonuçları, mevcut abonelerin yeni kredi başvurularında da BSM sonuçları ile birlikte kullanılabilir. Bu geliştirme ile kredi başvurularında iyi ve kötü başvurular birbirlerinden daha yüksek doğrulukla ayrıştırılır ve müşteri portföyü daha güvenilir olarak geliştirilebilir.

Aşağıda mobil telekomünikasyon sektöründe uygulanabilecek DSM uygulamasının süreçleri kısaca aktarılmıştır. Birinci aşamada hangi veri kaynaklarından ne tür verilerin alınması gerektiği kararlaştırılır. İkinci aşamada, veriler belirlenmiş anlamlara göre uygun formatta düzenlenerek analiz ve modellemeye hazır hale getirilirler. Daha sonra, belirlenmiş skorlama adımı çalıştırılır. Skorlamadan önce yapılacak segmentasyon bu sürecin verimliliğini artırır. Skor modelleri amaçlara uygun olarak belirlenen periyotlarda veri setine uygulanır. Bu aşamadan sonra da çıkan riskli abone raporunda yer alan abonelere otomatik olarak veya manüel inceleme sonrasında risk azaltıcı önlemler uygulanır. Son aşamada da tüm sürecin, modellerin ve müdahalelerin, performansı uygun kriterler ile ölçülerek raporlanır, böylece sistem düzenli olarak izlenir.



Şekil 12 : Mobil Telekomünikasyon Sektöründe Uygulanabilecek Örnek Bir Davranışsal Skorlama Süreci

İş dünyasında davranışsal skorlama modelinin en başarılı uygulaması, veritabanlarının içinde şekillendirilmektedir. Bu; müşterinin geçmişinin analiz edildiği, mevcut müşteri tercihleri arasından benzer davranış kalıplarının araştırıldığı ve bu kalıpların mevcut veya gelecekteki müşterilerin bir hedef kitlesi için kullanıldığı bir yaklaşımdır. Yapılan çalışmalar; hangi hedef müşteri gruplarının daha fazla harcaması için özendirileceği, ne kadar kredi limitinin atanacağı, belirli bir müşteri grubuna yeni ürün promosyonu yapılıp yapılmayacağı ve eğer geri ödeme kapasitesi kötüye dönerse şüpheli alacağın nasıl kurtarılacağı gibi kararları içerebilir. Bundan dolayı, bir davranışsal skorlama modeli bilgi güdümlü bir pazarlama sürecidir ve pazarlamacılara; pazarlama program ve stratejilerini geliştirme, test etme, uygulama, ölçme ve uygun bir şekilde kişiselleştirme imkânı vermektedir[78].

Davranışsal skorlama modeli abonenin yaşam eğrisinin her aşamasında kullanılabilir. Bankacılıkta kredi taksitleri düzenli olarak ödenmediği zaman abone önce İdari Takip aşamasına alınır. Bu aşamadan sonra belirli bir süre daha ödeme gerçekleşmezse veya abone iflası gibi belirli şartlar ortaya çıktığında da Yasal Takip süreçleri işletilir. Aynı durum telekomünikasyon sektörü için de geçerlidir. Faturasını zamanında ödemeyen bir aboneye sırasıyla SMS ile hatırlatma yapılır, hat aramaya kısıtlanır, borç bilgilendirme mektupları gönderimi yapılır, sesli yanıt sistemi (Interactive Voice Response- IVR) üzerinden veya canlı analist vasıtasıyla aranarak borç bilgisi verilir, hat aranmaya kısıtlanır, mahkeme kanalıyla tahsilât yapmak için avukatlık bürolarına iletilir. Tüm bu aşamaların sonunda hat iptal edilerek abonelik yaşam döngüsü sonlandırılır.

Mobil telekomünikasyon sektöründe riskli abonelerin tespit edilmesi amacıyla yapılan bu Davranışsal Skorlama çalışmasında Veri madenciliği kullanılmasının başlıca nedenleri aşağıda aktarılmıştır:

- Daha iyi ve daha nitelikli kararlar verebilmek için büyük veri yığınlarındaki gizli kalıpları ve ilişkileri keşfetmenin en gelişmiş ve hızlı yöntemlerini sunar
- Veri madenciliği başarı için gerekli; Bilgisayar Bilimleri, İstatistik ve Yapay Zeka ile gelişmiş veri toplama ve veri yönetimi gibi kilit teknolojilerin optimum bileşimidir
- Güçlü bir proje metodolojisi sunar
- Geleceği öngörmemizi sağlayacak ve hızlı aksiyon almamızı sağlayacak araçlar sunar
- Değişime hızlı uyumu sağlayacak esnek geliştirilebilir yöntemler sunar

[78] Hsieh, Nan-C. "An integrated data mining and behavioral scoring model for analyzing bank customers". **Expert Systems with Applications.** Vol.27, 2004, s. 625.

- İyi karar vermeye yardımcı olacak doğru veri altyapısının üzerine kurulmuş güçlü ve çok çeşitli analitik yöntemler sağlar

4.4. Skorkart ve Kredi Skorlama Altyapısının Gelişimi

Kredi skorlaması kavramının başlangıcı olarak David Durand'ın 1941 yılında National Bureau of Economic Research dergisinde yayımladığı çalışma gösterilir. Burada 37 farklı firma tarafından verilen 7.200 iyi ve kötü sonuçlanmış kredi işlemi analiz edilmiştir. Ki-kare analizi ile iyi ve kötü kredilerin belirlenmesinde hangi değişkenlerin anlamlı olduğu belirlenerek, geliştirilen Etkinlik İndeksi ile riski düşük olan başvuruların diğerlerinden etkin olarak ayrıştırılabildiği ortaya konulmuştur. Daha sonra da Diskriminant Fonksiyonu ile kredi Skorlama Modelleri geliştirilmiştir. HFC (Household Finance Corporation) Durand'ın geliştirdiği yöntemi uygulamaya geçirmiştir[79].

İlk defa 1940'larda Chicago'da kurulan Hausehold Finans ve Speigal Mail Order isimli iki firma kredi kayıplarını modellemeyi denemişlerdir. İlk skor kartlar kâğıt üzerinde manüel olarak karar vermede kullanılan çizelgeler halinde hazırlanmıştır. Manüel olarak uygulandıklarından beri, kolaylık için skor kartlar, eklendiğinde toplam skoru veren tam sayıları içeren nokta skorları ile tasarlanmıştır. Şartlı olasılıklar, toplam olasılığı bulmak için genellikle birbirleri ile çarpılırlar. 1950'lerin sonunda, iki matematikçi olan Bill Fair ve Earl Isaac olasılıkları ham nokta skorlarına dönüştürmek için doğal logaritmayı kullanmıştır. Bu yolla Amerikan Yatırım Kurumu (AIG) için ilk başarılı modelleri geliştirmişlerdir. Skor kart çalışmalarına Ford Kredi 1960'ların ortasında başladı ve birkaç eğitilmiş sigortacı skorlama kavramının temelini atmıştır. 1990'larda yeni kanallar ve çağrı merkezleri ile birlikte gelen değişim hızı, bankalar ve yapı kurumlarının skorlamayı kolayca kabul etmelerini sağlamıştır[80].

1990'lı yıllarda kullanılan kredi-skorlama sistemlerinin çoğu zarar olarak gösterilmiş müşterilerin incelenmesi ile geliştirilen skor kartlara dayandırılmıştır. Seçilen değişkenler ve atanan skorlar temel olarak yargısaldı. Fakat bunların kullanılması en azından kredi verme

[79] Johnson R.W. "Legal, Social, and Economic Issues in Implementing Scoring in the United States". In Thomas Lyn C., David B. Edelman, and Jonathan N. Crook (Ed.). **Readings in Credit Scoring.** New York: Oxford University Press, 2004. s. 5-6.
[80] Bailey M., "An Introduction to the Principles", In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s. 2-3.

sürecine biraz tekdüzelik ve öngörüsellik getirdi. Hızla genişleyen kredi portföylerinin çağında, sıklıkla acemi kredi analistlerinin yargılarından daha iyi sonuçlar ürettiler. Böylece ekonomik baskılar ve bilgisayar teknolojisi 1960'ların sonunda bir araya geldi ve 1970'lerin başında kredi skorlama sistemleri, geliştirilen deneysel değerleme yöntemlerine dayandırıldı ve yavaş yavaş kabul gördü[81]. Skor kart uygulaması, öncelikle finans kuruluşlarında ortaya çıkıp gelişmiştir. Mevcut durumda da en önemli müşterileri banka ve diğer finans kuruluşlarıdır. Ancak günümüzde skor kart sonuçları doğrudan pazarlama teknikleri için hedef müşteri kitlesinin tespitinde bile kullanılabilmektedir.

Kredi skorlaması, riskin ölçülmesini sağlayacak bir sıralama mekanizması kurar. Dolayısı ile birinci prensip, kredi başvurusu yapanların risklerine göre sıralanmasıdır. Daha yüksek skora sahip olanın şüpheli alacak oluşturma riskinin daha düşük olduğu ileri sürülür. İkinci prensip, geçmiş geleceği belirler. Diğer bir deyişle skor modelleri geçmiş veriye dayanarak kurulur ve aynı ürün için başvuran yeni adayların kredi başvurularının kararlarının verilmesinde kullanılır. Üçüncü prensip, bireysel başvuruların genel karakteristikleri genel başvuru kredi değerliliğinin öngörülmesinde kullanılır. Bir diğer ifadeyle, kredi değerleme kararları tutarlı olmalıdır[82].

Özellikle düşük miktarda ve yüksek adette kredi veren kuruluşlar kredi verme kararlarını kullanırken skor modellerinden faydalanırlar. Risk algılamasına göre de başvurunun kabul edilmesinde belirli bir sınır veya eşik değerini kabul ederler.

Kredi skorlama modellerinin başlangıçta kredi kartlarındaki kullanımından başarılı sonuçlar alınmasının ardından bireysel krediler, ev kredileri ve küçük işletmelere yönelik krediler gibi başka finansal araçlar için de kullanılmaya başlanmıştır. Daha sonraki yıllarda, reklam kampanyası veya cihaz kampanyası gibi doğrudan pazarlama alanında da skor kartlarının kullanılması ile kampanya geri dönüşlerinin arttığı gözlenmiştir. Ancak buradaki asıl başarı belki de skor kartın içeriği kadar, bu skor kartları geliştirmek için kullanılan veri madenciliği tekniklerinde aranmalıdır.

[81] Johnson R.W. , s. 6.
[82] Bailey M., s. 3-5.

4.5. Genel Amaçlı ve Özelleştirilmiş Skorlama Modelleri

Müşterilerine kredi veren veya müşterilerini kredilendirmeye ihtiyaç duyan hemen tüm firmalar geçmiş müşteri verilerini kullanarak kendi amaçlarına uygun Kredi Skorlama modelleri oluşturarak bunları uygulamaya alırlar. Ancak bunların yanında, farklı kredi sağlayan firmaların veya veri kaynaklarının (banka, sigorta, telekomünikasyon, perakendecilik, hükümet vb) verilerinin birlikte kullanılabildiği ve sonuçlarının belirli bir meblağ karşılığında isteyen tüm kurumlarca kullanılabildiği Jenerik Skorlama modelleri de mevcuttur. EK-1 de farklı skorlama firmaları tarafından üretilen ürünlere ait bilgiler mevcuttur.

Kredi veren firmalar, sadece kendi verilerinden oluşturduğu kredi skor modellerini kullanabileceği gibi; hazır olarak sunulan, ancak belirli bir maliyeti olan ve doğruluk oranı genelde daha düşük olan jenerik modelleri veya her ikisinin ortak bir bileşimini karar vermek amacıyla kullanabilirler. Bu karar, kredinin tipine (kurumsal/ bireysel), büyüklüğüne, içinde bulunulan ekonomik şartlara veya üst yönetimin belirlediği stratejiler gibi ölçütlere göre değişebilir. Çünkü her iki yöntemin de birbirlerine göre; geliştirme, uygulama, sistemin yönetimi ve yasal konularda avantaj ve dezavantajları mevcuttur.

Genel amaçlı (jenerik) skorlama sistemlerinin özelleştirilmiş (customized) sistemlere göre pek çok doğal avantajı vardır. Bunlar aşağıdaki gibi ifade edilebilir[83]:

- Tüm kredi verenler için kullanışlıdır, hatta küçük olanlar veya küçük-hacimli ürünleri olanlar bile kullanabilir. Kullanıcılar için geliştirme esnekliği sorunu ortadan kalkar.
- Kredi verene ait popülasyon grubunun tarihsel tecrübesi, kredi ürünleri ve coğrafik alanları ile sınırlı değildir
- Kullanıma hazırdır, geliştirme zamanı veya maliyeti yoktur.
- Kullanıcının bilgisine ve skorlama tecrübesine daha az bağlıdır
- Uygulaması kolaydır – skorlar genellikle başkaları tarafından üretilir
- Az sayıdaki karar için daha az maliyetlidir
- Kredi bürosu bilgisi detaylı bir şekilde kullanılabilir
- Kredi bürosu bilgisinin kullanımında çok ekonomiktir
- İflas gibi bazı kesin sonuçların tahmininde daha iyidir

[83] Chandler G., s. 43.

- Danışman ağı tarafından desteklenir
- Kredi bürosu tarafından korunduğundan güvenilirdir

Birçok avantajının yanında, genel amaçlı skorlama sistemlerinin özelleştirilebilir sistemler ile karşılaştırıldığında bazı dezavantajları da vardır. Bunlar aşağıdaki gibi belirtilebilir[84]:

- Potansiyel olarak doğruluğu daha azdır, çünkü kredi verenin kendi tecrübesi, ürünü ve müşterilerine dayandırılmamıştır
- Rakipler tarafından da kullanabilir
- Yüksek işlem hacmine sahip kullanımlarda daha pahalıdır
- Fikri mülkiyet hakkına tabidir – skorlama sisteminin detayları genellikle gizlidir
- Zamana bağlı tahmin sistemlerinde ve performans izleme amaçlı kullanımlarda daha zordur
- Ters aksiyon kodlarının tanımında ve prosedürlerin seçiminde katıdır.

Jenerik skorlamanın rolü artmaya devam edecektir ve uygulamada yeni tipleri ve modelleri ortaya çıkacaktır. Jenerik modellerden çıkarılan standartlaştırılmış risk ölçümleri kredi derecelendirme alanında önemli bir etkiye sahip olacaktır. Özelleştirilmiş skorlama sistemlerinin hala bir yeri olacaktır, ancak bununla birlikte, genelleştirilmiş modellemenin dışında nadiren kullanılacaklardır[85]. Bu görüşün geçerli olabilmesi için skorkart üreten firmaların, kendi müşterileri tarafından elde edilen verilerin çoğuna ulaşabildiğini ve skorlamaya konu olan bireylerin veya kurumların tek bir kimlik ile tanımlanmasının gerektiği varsayılmalıdır.

Genel olarak firmanın kendi verilerini kullanarak ürettiği skorlama modellerinin sonuçları, doğruluk oranları açısından jenerik olanlara göre daha yüksektir. Ancak jenerik modeller, değerlendirme açısından kendini ispatlamış daha uygun standart ölçümler sağlarlar.

4.6. Kurum İçi Geliştirilen (In House) Skorlama Modeli

Modern organizasyonlar gün geçtikçe, çoğunluğu kendi sistemlerinden olmak üzere daha fazla veriye ulaşma imkânını bulmaktadırlar. Bu özelliklerinden dolayı da firmaların müşterileri hakkında daha hızlı ve doğru karar almaları beklenir. Toplanan verinin

[84] Chandler G., s. 43.
[85] Chandler G., s. 48.

uygulanabilir kredi kararına dönüşmesinde kullanılan en önemli araç ise skorlama modelleridir. Aşağıda temel adımları aktarılan abone hayat eğrisinin her evresinde skorlama modelleri anahtar rolde kullanılabilir. İlk aşamada abone kitlesi çok fazladır ve her aşamada mevcut abone sayısı giderek azalır, buna karşılık elde kalan aboneler hakkında sahip olunan veri miktarı da artar.

Şekil 13 : Abone Yaşam Süreci ve Veri Madenciliği Uygulamaları

Başvuru ve kayıt safhalarında abone hakkında elde edilebilecek veriler; demografik veri, varsa aynı firmadaki daha önceki kredi sürecinde oluşturduğu veriler, herhangi başka bir kredi kurumu ile kredi ilişkisi olduysa kredi büro sonuçları, kredi başvurusunda bulunduğu kanal tercihi ve adres bilgisidir. Kayıt aşamasını geçip kabul edildikten sonra abonenin ödeme performansı, kullanım yapısı ve tercihleri ile ilgili veriler sistemden otomatik olarak toplanır. Tahsilât aşamasında analiz veri setine ek olarak abonenin verdiği taahhütlere uyup uymadığı ve geciktirdiği borç yapısı verileri eklenir. Borcun şüpheli alacak konumuna düştüğü son aşamada ise aboneye karşı alınan ek aksiyonların sonucunda abonenin verdiği tepkiler yeni girdiler olarak skorlamada kullanılır.

Bir kurum davranışsal skorlama modellerini abonenin kredi yaşam döngüsü boyunca yapabilir. Bu modeller, borçlunun ödeme performansını ve kullanım aktivitesini analiz ederek, ödememe riskinin tespit edilmesi, kredi limitinin ayarlanması, yeni ürün ve hizmetlerin teklif edilmesi, hilekârlığın tespit edilmesi, portföy riskinin tahmin edilmesi, geri alınabilir borç miktarının ve vadesinde ödenmeyen tutarın tahsil edilme yöntemlerinin hesaplanması amacıyla kullanılır[86].

Kurum içinde geliştirilen skor kartlarının faydaları sıklıkla aşağıdaki gibi aktarılır[87]:

- Skor kartların geliştirilmesi ve bakımında maliyet tasarrufu
- Skor kart geliştirme süreci üzerinde artan kontrol

[86] Trinkle, Brad S. and Baldwin, Amelia A. "Interpretable Credit Model Development via Artificial Neural Networks". **Intelligent Systems in Accounting, Finance and Management.** Vol.15, 2007, s. 125.
[87] Dekker B. "In-house Scorecard Development", In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s. 156.

- Müşteri yönetimi ve karar stratejileri ile çalışan analistin değerinin ve tecrübesinin artması
- Yeni teknolojiler ve veri madenciliği ile kuvvetlendirilen analistler sayesinde yenilikçiliği teşvik etme
- Yeni modellerin entegrasyonunda zamandan kazanma
- En iyi uygulamanın geliştirilmesi ve uygulanmasında tutarlılık
- 'Kara Kutu' anlayışının kaldırılması ve güvenin artırılması
- Belirli iş problemlerine odaklanma ve çözme

'Daha güçlü skor kartların' bu şekilde ortaya çıkmayacağı da düşünülebilir. Skor kartın gücü teknoloji ve kullanılan tekniklerden gelir. Kurum içi bir takım, bu bakış açısında bir danışman ekipten daha iyi olmayacaktır ve iç takımlar kendi uygulamalarında çok dar görüşlü ve at gözlüklü olma riski taşıyabilirler. Buna rağmen, kaliteli veri iyi bir skor kart için esastır ve bir kurum içi takımın veriyi daha iyi değerlendireceği ve bundan dolayı optimal skor kart geliştirme olasılığının daha fazla olduğu tartışılabilir[88].

Skorlama modelleri, uygun segmentlere ayrıştırılmış veri setleri üzerinden geliştirilmelidir. Segmentasyon belirli modeller vasıtasıyla yapılabildiği gibi, uzman görüşleri doğrultusunda da oluşturulabilir. Aşağıda bu alanda en çok kullanılan segmentasyon tanımları verilmiştir. Amaca uygun olmak üzere aşağıdaki bir veya birden fazla kıstasa göre oluşturulacak segmentler uygulamada çoğu zaman yeterli olmaktadır:

- Kurumsal başvuru / Şahıs başvurusu gibi, başvuranın belirli bir özelliğine göre belirlenebilir.
- Daha önceden firma ile iş ilişkisinin var olup olmamasına göre; Yeni müşteri / Eski müşteri şeklinde de ayırt etmek de mümkündür.
- Başvuran hakkında yeterli veri bulunup bulunmamasına bağlı olarak; Kalın dosya / İnce dosyalı müşteriler.
- Başvuru için kullanılan verilerin güvenilirliğine dair şüphe mevcutsa; Temiz dosya / Kirli dosya skorlama modelleri şeklinde de segmentasyona da gidilebilir.

[88] Dekker B. "In-house Scorecard Development", In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s.156.

Skorlama sistemleri, skorlama politikası ve kapsamı ile yargısal analizleri bile içeren bir değerlendirme sisteminin herhangi bir bileşeni, baştanbaşa bütün değerlendirme sistemine uyacak şekilde tasarlanmalı ve uygulanmalıdır. Bileşenlerin koordinasyonu çok kritik bir husustur. Sıralı değerlendirme çoğu kez ekonomik ve kolaydır, fakat matris yaklaşımı çok daha doğru olabilir. Tüm değerlendirme sisteminin sürecinin ve bileşenlerinin sıkı bir şekilde izlenmesi, sistemin uygun yönetilmesi açısından oldukça kritiktir[89].

4.7. Kredi Kayıt Bürosu ve Kanuni Düzenlemeler

Kredi bürolarında bulunan veriler, banka, finans kurumları, sigortalar, tahsilât büroları ve mahkemeler gibi pek çok kaynaktan elde edilmektedir. Bu kaynaklar tarafından gerçekleştirilen veri aktarımı ya on-line ya da periyodik şekilde olmaktadır. Kredi büroları, elde ettikleri bu verileri kullanarak modeller oluşturur, bunların doğruluğunu garanti eder, düzenli olarak günceller ve sonuçlarını takip ederler. Farklı ülkelerde özel ve/ veya kamu kurumu niteliğindeki kredi bürolarının yer aldığı birbirinden çok farklı uygulamalar mevcuttur.

Türkiye'de faaliyet gösteren Kredi Kayıt Bürosu (KKB) A.Ş. 5411 sayılı Bankacılık Kanununda öngörüldüğü üzere (md.73/4) kredi kuruluşları (mevduat bankaları ile katılım bankaları) ile finansal kuruluşlar (kredi kuruluşları dışında kalan ve sigortacılık, bireysel emeklilik veya sermaye piyasası faaliyetlerinde bulunmak veya Bankacılık Kanununda yer alan faaliyetlerden en az birini yürütmek üzere kurulan kuruluşlar ile kalkınma ve yatırım bankaları ve finansal holding şirketleri) tarafından kurulan şirketler vasıtasıyla yapacakları her tür bilgi ve belge alışverişini sağlamak üzere 11 Nisan 1995'te kurulmuş bir şirket olup; üyesi olan kuruluşlar da müşterilerine ait kredi bilgilerini bu Kanunun aynı maddesi uyarınca birbirleriyle paylaşmaktadırlar. KKB A.Ş. bu amacı doğrultusunda geliştirdiği sistemi Nisan 1999'da sektörün hizmetine sunmuş olup; bugün itibariyle 38 üye kuruma hizmet vermektedir. Kredi Referans Sistemi (KRS), bireysel kredi ürün müşterilerine ait detaylı bilginin, kredi kararı aşamasında risk faktörünün belirlenmesine ve dolayısıyla da riskin minimize edilebilmesine olanak sağlamak üzere çağdaş yöntemler kullanılarak paylaşımına olanak sağlayan bir "bilgi paylaşım sistemidir". KKB A.Ş. negatif ve pozitif nitelikli tüm bireysel kredi ürün bilgilerini üyelerinin paylaşımına sunmaktadır. KKB, Türkiye'de kurulan ilk ve tek kredi bürosudur. Bunun haricinde, Türkiye Cumhuriyeti Merkez Bankasında

[89] Chandler G., s. 48.

(TCMB) tutulan negatif kayıtlar veritabanı da kredi değerlemede bankalar tarafından kullanılabilmektedir.

KKB skor sonuçları, on-line ve gerçek zamanlı (real-time) olarak oluşturulur. Türkiye Bankalar Birliğinin talebi üzerine KKB tarafından Limit Kontrol Sistemi (LKS) ve Internet Sahtekârlık Alarm Sistemi (ISAS) geliştirilmiştir. Bu şekilde, sahte doküman, bilgi, bildirim ve başvuruları üyeleri arasında paylaşarak kötü niyetli potansiyel riskleri ortadan kaldırmaya yardımcı olmaktadır. Sisteme dahil edilen tüm gözlemlerin eksik verileri tamamlandıktan ve veriler doğrulandıktan sonra skorlama amacıyla kullanılmaktadır. Sadece 2007 yılı içerisinde KKB'den yapılan sorgulama adedi 70 milyonun üzerindedir[90]. Bu veri tabanı, KKB üyesi bankalar tarafından sürekli güncellenmesi, içerideki verilerin pek çok farklı sorgulamaya uygun şekilde düzenlenmesi, yıllardan beri amacına uygun çalışması ve verilerin zenginleştirilerek etkin şekilde yönetilmesi açısından Türkiye'deki en önemli veri kaynaklarından bir tanesidir.

Gelişmekte olan veya geçiş konumundaki pek çok ülkede kredi acenteleri ve kredi bürolarının mevcut olmaması, borç almak için başvuruda bulunan kişilerin kredi davranışlarını gösteren ilgili verilerin mevcut olmamasına da neden olmaktadır[91]. Kredi bürolarından elde edilen veriler ve skor sonuçları, kredi değerlemede her zaman için önemli girdilerdir. Ancak kredi bürolarında hem pozitif hem de negatif verilerin tutulması ve değerlendirmede kullanılması en uygunudur.

Kredi bürosundan yararlanmayan kurumlar yeni kredi taleplerini değerlendirmeye alırken müşteri beyanı ile elde edilen verileri ve abone kendi veritabanında mevcut ise buradaki geçmiş verileri, yeni abone ise sadece müşteri beyanına dayalı verileri kullanarak kredi skoru üretir. Kredi bürolarına üye kurumların pek çoğu, kendi üyeleri hakkında kredibilite ile ilgili karar alacakları zaman büro sonuçlarını kullanmayabilirler. Bunun haricinde, orta ve büyük ölçekli kredi talebinde bulunulduğu durumlarda, kurum kendi istihbarat kaynaklarını kullanarak bir sonuç elde etmeyi tercih edebilir.

Ortak bir veritabanına ulaşılmasının faydası, özellikle başvuru skorlama sürecinde hayati bir öneme sahiptir. Hatta müşteri tarafından verilen bilgiler ve haricen elde edilen veriler çoğu

[90] Kredi Kayıt Bürosu, **Annual Report 2007**, İstanbul: 2008.
[91] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan. "Consumer credit scoring models with limited data". **Expert Systems with Applications**. Vol.36, 2009, s. 4742.

durumda müşteri hakkında olumlu ya da olumsuz kanıya varmaya yetecek kadar olmayabilir. Mutlaka bir karar vermek gerektiği durumda ise hata oranı çok yüksek değerlendirme sonuçları ile ortaya çıkar.

Kredi Büro skorunu kullanmanın başlıca nedenleri olarak şunlar sayılabilir:

- Abone hakkında pek çok veri kullanılarak oluşturulmaktadır
- Sadece bu konuda çalışan uzmanlar tarafından oluşturulmaktadır
- Tahmin gücü belirli bir güvenilirlik düzeyinin üstünde sonuçlar verir
- Uzun zamandan beri kullanıldığından güvenilirliği ispatlanmıştır
- Ürettiği skorlar için standardizasyon mevcuttur
- Kanunlara uygun sonuçlar üretilir (İstenmeyen değişkenler dahil edilmez, yasal olarak kredi açılamayacak kişilere skor atanmaz vb.)

Yurtdışındaki kredi büroları, Türkiye'dekinden farklı olarak sadece bankacılık sistemi tarafından toplanan ödeme performans verilerine değil, aynı zamanda perakendecilik sistemlerinin veritabanlarına, hukuki olayların sonuçlandırıldığı davaların veritabanlarına, sektörler arası kara/ gri listelere, devletin elindeki gayrimenkul veritabanları gibi daha fazla veriye ulaşabilmektedir.

ABD kongresinin, kredi verenlerin kötü risklileri ve iyileri sınıflandırmasını engellemeyi amaçlamadığı ile ilgili güçlü deliller vardır. Amerikan kongresi, kredi skorlama ve yargısal sistemlerinin içine yasaklanmış değişkenlerin dahil edilmesini engelleyerek, bu değişkenlerden bir veya daha fazlasına sahip tüketicilere farklı davranılmasının engellendiğine inanmaktadır. Ancak kredi değerliliğinin temelindeki ayrıma müdahale edilmemiştir. Eşit Kredi Fırsatı Kanununun (ECOA-Equal Credit Opportunity Act) 1972'de kongreden geçmesinden kısa bir süre sonra, izin verilen birçok değişkenin yasaklanan değişkenler ile yüksek ilişkili olduğu için ayrımın hala var olduğu ile ilgili kaygılar ortaya çıkmıştır. Mevcut kanunu desteklemek için, kongre sonraki değişiklikleri 1976'da onaylamış ve ırk, renk, etnik köken, din gibi özelliklerin, kamu sosyal yardım menfaatlerinin edinilmesi ve tüketiciyi koruma sözleşmesi altındaki hakkın iyi niyetli uygulamasını garanti altına almıştır[92].

[92] Johnson R.W., s. 10-11.

Hâlihazırda Türkiye'de faaliyet gösteren telekomünikasyon firmaları KKB veritabanlarını kullanamamakta, sistem tarafından üretilen sonuçlar paylaşılmamaktadır. Pek çok sigorta firmasında görüldüğü gibi, telekomünikasyon sektöründe de aktivasyon süreçlerinde elde edilen verilerin eğitim seviyesi ve tecrübesi çok farklı kişiler tarafından alınması ve verinin elde edilmesinde yapılan kontrol seviyesinin çok yükseltilememesi nedeniyle özellikle demografik verilerin kalitesi düşük olmaktadır. Dolayısıyla aboneler hakkında elde edilen ilk verilere dayanarak başvuru skorlaması yapmak çok güvenilir olmamaktadır.

Ülkemizdeki terör olayları başta olmak üzere tüm devlet kurumları tarafından ihtiyaç duyulan "hattın gerçek sahipliğinin belirlenebilmesi ve kullanıcının tespit edilebilmesi" için mobil operatörlere getirilen zorunluluk ile 01.06.2008 tarihinden itibaren MERNİS (Merkezi Nüfus İdaresi Sistemi) üzerinden Türkiye Cumhuriyeti Kimlik Numara (TCKN) doğrulaması yapılmadan yeni hat aktivasyonu yapılmamaktadır. TCKN doğrulaması yapıldıktan sonra aktivasyonun yapılması ile pek çok sahtekârlık amacıyla yapılan başvuru da engellenmiştir.

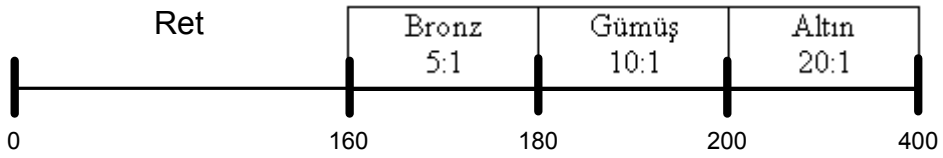
4.8. Kredilendirme Stratejileri ve Kredi Sektöründeki Gelişim

Kredi kararının verilmesinde skor kart veya skorlama sonuçlarının tek başına yeterli olduğu düşünülmemelidir. Skorlama, borç (kredi) veren firmanın kredi verme konusundaki kabul/ ret kararını vermede kullandığı Kredi Politikasının bir alt bileşenidir. Kredi Politikasının diğer kuralları aşağıdaki gibi olabilir[93]:

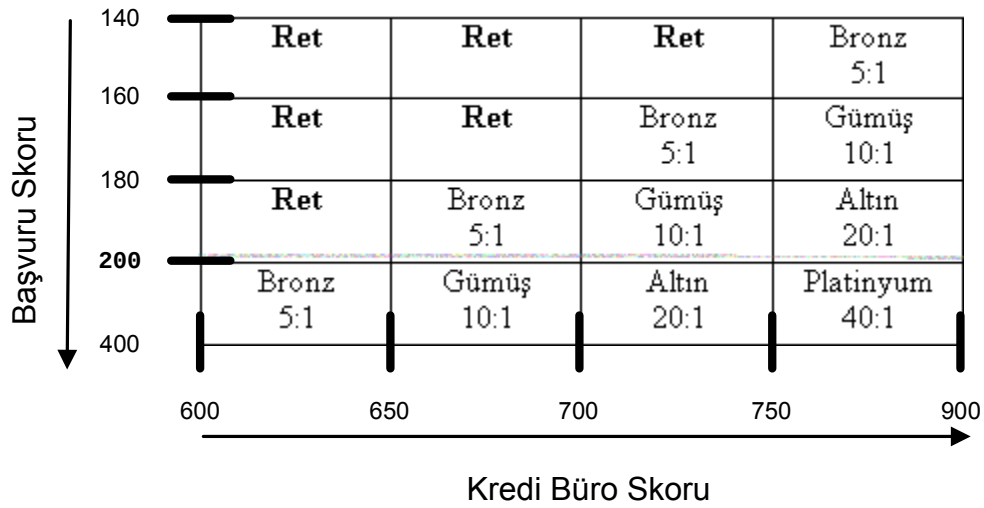
- Başvuruyu yapan 18 yaşından gün almış olmalı
- Yeterli düzeyde geliri olmalı (Öğrenci, işsiz, sigortasız çalışanlar gibi kesimler hariç tutulabilir)
- Daha önce sahtekârlık nedeni ile herhangi bir yerde kaydı olmamalı[**]
- Bilgilerin doğru olduğunun beyan edilmesi
- Belirli bir ikametgâhı olmalı
- Başvuru formundaki tüm bilgiler eksiksiz olmalı[***]
- Birden fazla başvurusu olmamalı
- Daha önce birden fazla defa kredi reddi yapılmamış olmalı vb.

[93] Kindred D., s. 12.
[**] Firmanın kara listesinde bulunması veya sektör kara listelerinin paylaşımı ile eski abonelerin ve gerçekleşmiş olayların tespit edilmesi amacıyla kullanılır.
[***] Türkiye'de bu amaç için; aboneden kendi adına düzenlenmiş elektrik, su, doğalgaz vb. faturası istenir. Adres doğrulama sistemine geçilmesi ile bunun merkezi olarak yapılması sistemin etkinliğini artıracaktır.

Uygulanacak kredi stratejisinde yine skora göre sıralama ve uygun gruplama ile abonelerin homojen hale getirilmesi işlemi ilk aşamada yapılmalıdır. Burada da verilen skor sonucu tek başına değerlendirmede kullanılabileceği gibi, hem firma içinde üretilen kredi skoru hem de kredi bürolarından elde edilen skor ortaklaşa kullanılarak da karar verilebilir. Başvuru skorlaması detayında benzer kullanımdan bahsedilmiştir. Buradaki farklılık, hem verilen skorlar hem de odds oranları açısından homojen gruplara ayrıştırılan abonelere verilecek kredi türü ve miktarında bir farklılaştırmanın yapılmasıdır. Aşağıda tek boyutlu ve iki boyutlu olarak aktarılabilecek iki farklı örnek strateji yer almaktadır[94].



Şekil 14 : Tek Boyutlu Eşik Değeri Stratejisi – Risk Profili
Kaynak: Dekker B. "Lending Strategies". In Murray Bailey (Ed.). Credit Scoring, The Principles and Practicalities. Second Edition. Bristol: White Box Publishing, 2004. s.78.



Şekil 15 : İki Boyutlu Eşik Değeri Stratejisi – Risk Profili
Kaynak: Dekker B. "Lending Strategies". In Murray Bailey (Ed.). Credit Scoring, The Principles and Practicalities. Second Edition. Bristol: White Box Publishing, 2004. s.79.

Kredi pazarının doygun olup olmaması da önemli bir kıstas olup, gelişmiş piyasalarda karlılığı artırmak, maliyetleri düşürmek ve özellikle abonelerin elde tutulmasına çalışılır.

[94] Dekker B. "Lending Strategies". In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s. 78-79.

Gelişen piyasalarda ise belirli bir risk düzeyi ile pazardan mümkün olan en fazla payın alınması uygun bir yaklaşımdır. Bundan dolayı Türkiye'de rahatlıkla kredi verilen bir kişiyle göreli olarak aynı şartlara sahip başka birine, gelişmiş piyasalarda faaliyet gösteren bir kurum kredi açmayabilir.

İnternet ve mobil uygulamalarının yayılmasıyla kredi sektöründeki uygulamalar önemli ölçüde etkilenmişlerdir. İnternet yeni pazarlama fırsatları ve yeni iş modellerinin uygulamaya girmesine imkân tanımış, bununla birlikte yeni süreçlerin ve risklerin oluşmasına neden olmuştur. Artık günümüzde kredi talepleri veya yeni hat aktivasyonları internet üzerinden yapılabilmektedir.

Kredi talep edenler açısından da çok farklı alternatifler ortaya çıkmış ve tüketiciler, hakları ve imkânları hakkında daha bilinçli hareket etmeye başlamışlardır. Tüketiciler birlikte çalıştıkları kurumlardan artık daha az bürokrasi, daha esnek ve hızlı bir yaklaşım tarzı beklemektedirler. Yüksek rekabet ortamında başarılı olmak için müşteri memnuniyetinden feragat etmeden müşteri riskini kontrol altına almak daha da zorlaşmıştır.

Kredi verenlerin en büyük handikaplarından bir tanesi, kredi talep edenler hakkında yeterli ve doğru veriye ulaşım imkânının her zaman mümkün olmamasıdır. Buna ek olarak, sürekli yeni ve uluslararası oyuncuların dâhil olduğu piyasalarda kredi veren kuruluşlar, artan rekabet karşısında pazar paylarını korumak için dahi yeni teknolojilere sürekli yatırım yapmaları gerekmektedir.

4.9. Risk ve Sahtekârlık

Sahtekâr abone ile Riskli abone birbirinden ayrıştırılmalıdır. Sahtekâr abone kullandığı kredinin, servis veya hizmetin bedelini ödeme niyeti olmayan abonedir. Riskli abone ise kullandığı kredi, mal ve hizmetin bedelini ödeme isteği olan ancak ödeme güçlüğü, abone memnuniyetsizliği, fatura şikâyeti, unutma, vb. nedenlerle şüpheli alacak bırakan aboneye denir. Her iki grup aboneye ait ödenmemiş borçlar, sonuç itibari ile finansal kayba neden olsa da, bu iki abone grubunun tespit edilme yöntemleri ve sonrasında yapılacak uygulamalar farklılık gösterecektir. Bu iki abonelik türünün ayrıştırılmasının bir faydası da şudur: Sahtekar abone derhal sistemden çıkarılır ve tekrar dönmesi engellenmeye çalışılırken, riskli abone uygun abonelik paketine veya kontörlü hatta yönlendirilir.

Telekomünikasyon sahtekarlığı dünya üzerinde ilk olarak 1980'lerin sonlarında görülmeye başladı ise de ülkemizdeki geçmişi 15 yıldan eski değildir. Operatörlerin kullandıkları şebekelerin analog veya dijital olması, 2G, 2.5G veya 3G olup olmaması sahtekarlık türlerinin yaygınlığını ve çeşitliliğini doğrudan etkilemektedir. Telekomünikasyon sektöründe görülen sahtekarlık tipleri her geçen gün çeşitlenmekte ve derinleşmektedir. Halihazırda kullanılan GSM (Global System for Mobile Communications) teknolojisinin bile yeterince karışık olması ve çok hızlı ilerlemesiyle meydana gelecek sahtekarlık tipleri de çeşitlenerek gelişecektir. Sahtekârlar her zaman teknolojiyi yakından takip ederek mevcut sistemlerin açığını ortaya çıkarmaya uğraşırlar. Yeni teknolojilerin gelişmesi, yemleme (phishing), truva atı, sahte başvuru gibi yeni sahtekârlık tiplerinin doğmasına yol açmıştır.

Telekomünikasyon sektöründe ilk ortaya çıkan sahtekarlık türü Abonelik Sahtekarlığıdır. Burada sahte dokümanlarla yani kimlik bilgileri değiştirilerek gerçekte var olmayan şahıs/ firma adına, haberi olmadan üçüncü şahıslar adına veya hayatta olmayan şahıslar adına yanıltıcı bilgilerle bir veya daha fazla hat aktivasyonu gerçekleştirilerek yapılır. Tanımlanan kredi dolandırıcılığının en önemli ayıracı, kurumun yanıltılarak elde edilen kredinin geri ödeme niyetinin veya elde edilen faydanın, alınan servis veya hizmetin karşılığının ödenme niyetinin bulunmamasıdır.

Bu tarz sahtekarlığın önüne geçmek için kullanılan başlıca araçlar; daha önce abonelik sahtekarlığı yaptığı tespit edilen şahıs ve firmalar için Kara/ Gri liste oluşturulması, yeni abonelik başvurularında evrakların gerçeklik kontrolünün yapılması gibi önlemlerdir. Kara/ Gri Liste uygulaması asıl olarak başvuru skorlama aşamasında kullanılabilir. Ancak, bunun yanı sıra Beyaz Liste uygulaması da mevcuttur.

Mobil telekomünikasyon operatörleri sahtekâr aboneyi tespit etmek için arama (CDR – Call Detail Records) verileri, riskli aboneyi belirlemek için ise başlıca ödeme performansını gösteren verileri kullanılır. Sahtekâr abonenin şebeke yaşı çok küçük olmasına karşın (daha çok yeni aboneler), riskli aboneler daha çok eski aboneler arasından çıkar. Sahtekâr aboneleri tespit etmek amacıyla kullanılan FMS (Fraud Management System) sistemleri anlık veri ile kural tabanlı olarak çalışır. Riskli aboneler ise geçmiş verilerden oluşturulan değişkenler kullanılarak çok değişkenli ileri analiz teknikleri ile belirli periyotlarda (örneğin günde bir defa) üretilen listelerle engellenmeye çalışılır.

Operatörlerle telekomünikasyon sahtekarları arasında eşitsiz gelişim kuralına uyan bir ilişki mevcuttur. Genel olarak, sahtekarlar oyuna hep bir adım önde başlamakta ve yaptıkları keşfedildiği anda gereken önlemlerin alınmasına kadar başka bir yerden tekrar boy göstermektedir. Dolayısı ile GSM sektöründeki sahtekarlığın tespit edilip engellenmesi için kullanılan çözümler, yeni durumlara adaptasyonu da maksimum derecede esnek uygulamalar içermelidir. Operatörlerin uygulayacağı güçlü bir anti-fraud stratejisi farklı teknolojik çözümleri, sahtekarlıklara karşı yönetim politikalarını ve sağlıklı prosedürel uygulamaları içermelidir. Ancak, dünyanın hiçbir yerindeki operatör, şebekesinde hiç sahtekarlık olayının meydana gelmemesini sağlayamaz.

Bir sahtekarlık olayının operatöre yol açacağı maliyetler sadece gelir kaybı veya gerçek para kaybı değildir. Aynı zamanda işgücü kaybına, kaynak israfına, mevcut abonenin şebekeden ayrılmasına ve firma imajının zedelenmesine yol açar.

4.10. Skorlama Modellerinin Performanslarının İzlenmesi ve Güncellenmeleri

Uygulamaya alınan skor kart için kullanılan model şampiyon (champion) olarak isimlendirilir ve sürekli olarak bu modelin yerine geçirilmesi amacıyla meydan okuyan (challenger) modeller geliştirilir. İkisinden hangisinin daha iyi sonuçlar ürettiğini bulmak için, şampiyon ve meydan okuyanın işlemesine izin verilir ve sonuçların raporları oluşur oluşmaz karşılaştırılır. Yeterince zamanın geçmesi için izin verilmelidir; böylece geçici sonuçlar iki rakibin gerçek gücünü maskelemez. Eğer şampiyon daha üstün sonuçlar üretirse, diğeri çıkarılır ve yeni mücadele eden modeller geliştirilir ve test tekrarlanır. Orijinal mücadele eden model tarafından üretilen sonuçlar büyük olasılıkla performansı iyileştirmek için hangi değişikliklerin yapılması gerektiği hakkında fikir verecektir. Eğer mücadele eden daha üstün sonuçlar çıkarmışsa, yeni şampiyon model olur. Şampiyon mücadele eden yarışması tek model ile sınırlı değildir. Birden fazla mücadelen eden model olabilir. Yönetim, aynı anda test sırasında birkaç tane mücadeleci modeli seçebilir. Bu yöntem daha hızlı sonuçlar üretir, çünkü bir şampiyon-mücadeleci deneyi sırasında birisinin bitmesini bekleme gereksinimi olmaz[95]. Daha iyi bir modelin varlığı, yarattığı geliştirme küçük bile olsa kredi verenin riskini önemli miktarlarda düşürür ve gelecekte ciddi miktarda tasarrufun oluşmasına katkıda bulunur.

[95] Hopper M.A. and E.M. Lewis, s. 36-37.

Şampiyon/ meydan okuyan modellerin tüm şartlar aynı olsa bile geliştirilme nedeni insan davranışlarının sürekli değişmesidir. Ancak, skorlama modelleri kredi sistemini etkileyecek önemli bir gelişme yaşandığında veya dönemsel olarak yenilenirler. Bu değişiklikler şunlar olabilir;

1. Ekonominin genel yapısı, genişleme veya durgunluk dönemlerine geçiş yaşanması
2. Hukuki olarak kredi kararının verilmesini etkileyen kanun ve düzenlemelerde değişiklikler yaşanması
3. Rakip firmaların stratejilerindeki değişikliklerin etkileri
4. Mevcut verilerin kaynağında değişmesi veya yeni veri kaynaklarına erişim
5. Firmanın IT altyapısında meydana gelebilecek önemli değişiklikler
6. Firmanın kredi politikalarında gerçekleşen önemli değişiklikler
7. Yeni bir ürün/ kampanyanın hayata geçmesi

Ekonomik kriz dönemlerinde kişiler ve kurumlar borç yükümlülüklerini yerine getirmekte zorlanırlar veya bazılarından kaçınma eğilimine girerler. Bu tür ekonomik durgunluk zamanlarında mevcut kredi skorlama modellerinin performanslarının aynı kalması beklenmemelidir. Çünkü kredi değerleme modelleri abonelerin geçmişteki davranışlarının gelecekte de aynı şekilde devam edeceği varsayımına dayanır. Ancak hâlihazırda ödemelerini düzenli ve zamanında yapan bir müşterinin işini kaybetmesi, firmaların faaliyetlerinin azalması, bu kişi ve kurumların ya hemen ya da ilerleyen dönemlerde borçlarını ödeme kabiliyetlerini kaybetmelerine yol açar. Böylesi durumlarda kredi yöneticileri; firmanın iş hedeflerine, piyasa şartlarına, ekonominin genel performansına ve düzenleyici kurumların (BDDK, TK gibi) yaklaşımlarını da gözeterek uygun bir risk düzeyini seçip uygulamaları bu seçime göre revize ederler.

Hızlı değişen ekonomik şartlar, skorlama modellerinin doğruluk oranlarını azaltırlar. Bunun için; işsizlik oranı, geri dönen çek/senet oranı, faiz oranı gibi zamana bağlı değişkenler modellemelere dâhil edilirler. Dışsal ekonomik değişkenlerin modellemeye dâhil edilemediği durumlarda iki çözüm üretilir: Bunlardan birinci çözüm, modellerin daha sık revize edilmesi; diğer çözüm de, eşik skor değerlerinin veya kredi limitlerinin genel ekonomik şartlara göre değiştirilmesidir.

Model doğruluğundaki iyileştirmeler, başvuruların risklilik sıralamasının daha doğru yapılmasını sağlayarak, yüksek kredi riski bulunan başvurulara kredi verilme potansiyelini azaltır. Belki de daha önemlisi, doğru model iyi kredi riskini tespit ederek satışları ve gelirleri artırabilir[96].

[96] Trinkle, Brad S. and Amelia A Baldwin., s. 124.

5. KREDİ SKORLAMA LİTERATÜR TARAMASI

5.1. Kredi Skorlamada Kullanılan Yöntemler

Kredi skorlama yöntemleri, istatistiksel teknikler ve istatistiksel olmayan teknikler şeklinde başlıca iki grup altında toplanabilir. İstatistiksel teknikler içerisinde en çok kullanılan yöntemler Doğrusal Diskriminant Analizi, Lojistik Regresyon gibi teknikler ile Karar Ağaçları ve K-enyakın komşuluk modelleridir. İstatistiksel olmayan yöntemler içerisinde ise Yapay Sinir Ağı, Genetik Algoritma ve Lineer Programlama en fazla kullanılan yöntemlerdir. Bu yöntemlerin haricinde aşağıda aktarılan çalışmalarda bu sayılanlardan daha farklı yöntemler denenmiş ve olumlu sonuçlar alınmıştır. Bunlar başlıca; Uzman Yöntemler (Expert Methods), Markow Zinciri Modelleri (Markow Chain Models), Sağkalım Analizi (Survival Analysis) ve Destek Vektör Makineleri (Support Vector Machine) dir.

Kullanılan pek çok skorlama metodunun anlamlı ve kabul edilebilir sonuçlar ürettiği görülmüştür, ancak hiçbir yöntem, abonenin (başvuran veya borçlunun) kredi riskinin sıralamasında net bir şekilde diğerlerinden daha iyi sonuç ürettiğini kanıtlayamamıştır. Ancak belirli şartlar altında bazı teknikler diğerlerinden daha avantajlı olabilmektedirler[97].

Kredi skorlama konusunda ilk kullanılan yöntem Diskriminant Analizidir. Kredi skorlama literatüründe de uzun zaman en popüler metot olmuştur. Diskriminant Analizinin kolay uygulanması, sonuçların başarısı ve modelin uygulamada anlaşılır olması, yaygın kullanımın en önemli nedenlerindendir. Ancak bu metodun varsayımlarından ve matematiksel özelliklerinden dolayı bazı eksiklikleri mevcuttur. Uygulamalı Diskriminant Analizi literatüründeki istatistiksel problemler aşağıdaki gibi sınıflandırılabilir[98,99]:

1. Değişkenlerin temel dağılım varsayımı ihlal edilir; çünkü kredi skorlamada sık kullanılan kategorik değişkenlerin hemen tamamı çoklu normal dağılım varsayımına uymaz.
2. Modelin yanlış sınıflama hatasının oranını en küçüğe indirmek için sınıflandırma fonksiyonu, müşterinin ait olduğu iyi veya kötü sınıfın deney öncesi olasılıkları ile ağırlıklandırılmalıdır.

[97] Mays E. "The Role of Credit Scores in Consumer Lending". In Elizabeth Mays (Ed.). **Credit Scoring for Risk Managers.** Thomson, South-Western, 2004. s. 8.
[98] Eisenbeis Robert A. "Problems in Applying Discriminant Analysis in Credit-Scoring Models". In Thomas Lyn C., David B. Edelman, and Jonathan N. Crook (Ed.). **Readings in Credit Scoring.** New York: Oxford University Press, 2004. s. 24.
[99] Sarantopoulos, G., s. 101.

3. Grup dağılımı eşit olmadığında bile ikinci dereceden fonksiyonların yerine doğrusal diskriminant fonksiyonlarının kullanılması
4. Analizde bireysel değişkenlerin rolünün yetersiz yorumlanması
5. Boyutluluktaki azaltmalar
6. Grupların tanımlanmasındaki problemler
7. Uygunsuz deney öncesi olasılıkların kullanımı ve/veya yanlış sınıflandırma maliyetleri
8. Modelin performansını değerlendirmek için kullanılan sınıflandırma hata oranlarının tahminindeki problemler
9. Analiz örneklemlerinin seçimi

Doğrusal regresyonda olduğu gibi Lojistik Regresyonun veriye "en iyi uyan" denklemi bulmasına rağmen, modelin türetildiği prensip oldukça farklıdır. En iyi uyum için en küçük kareler sapması kriterini kullanmak yerine, en iyi uyumu sağlayan regresyon katsayılarını veren gözlem sonuçlarından alınan olasılığı maksimum yapan, En çok olabilirlik (maksimum likelihood) yöntemini kullanır. Lojistik regresyon modeli doğrusal regresyon modelinin eleştirisini üç yöntem ile giderir. Birincisi, yanıtın (veya çıktının) iki terimli (binomial) olarak dağıldığı varsayılır. İkincisi, model kısıdı Pi, 0 ve 1 arasında yer alır. Üçüncüsü, 'logit(pi)'nin pi ye karşı eğrisi, özel bir s-şeklinde (sigmoid) eğridir ve pi = 0.5 olduğunda simetriktir[100]. Lojistik Regresyon modeli, uygun sonuçlar üretmesine rağmen bilgisayar teknolojilerinin gelişmesine kadar en çok olabilirlik metodunun hesaplama zorluklarından dolayı kullanılamıyordu.

Karar ağaçları parametrik olmayan kurallardır, özellik uzayını benzer sınıf üyeliği olasılıklarına göre bölümlere ayırır. Yöntemin kredi skorlama için uygunluğunu destekleyen birkaç neden tespit edilmiştir: Birincisi karmaşık lineer olmayan modellere uyabilir, çünkü sınırlandırılmış parametrik bir şekli yoktur. İkinci olarak, temelini teşkil eden karar süreci, sıralı bir süreç ile eş anlı bir süreçten (tüm niteliklerin modelde eş zamanlı olarak dikkate alındığı süreçler) daha iyi temsil edilebilir. Karar ağaçları yoğun olarak kredi skorlama modelleri kurmak için kullanılmışlardır ve birçok çalışma ile kıyaslandığında iyi sınıflandırma performansı göstermişlerdir[101].

[100] Sarantopoulos, G., s. 103.
[101] Sarantopoulos, G., s. 106.

Sınıflandırma ağaçları ve uzman sistemler, sorunun çözümünde kullanılan bir diğer yaklaşım grubunu oluşturur. Sınıflandırma ağaçları genellikle istatistikte, yapay zekâda ve makine öğrenmesi uygulamalarında kullanılır. Makowski (1985) kredi skorlamada sınıflandırma ağaçlarını kullanmayı öneren ilk araştırmacılardan biridir. Coffman (1986) değişkenler arasında etkileşim olduğu durumlarda sınıflandırma ağaçlarının diskriminant analizinden daha iyi çalıştığını göstermiştir[102].

Diskriminant Analizi ve Lojistik Regresyon kredi skorlama için en çok kullanılan istatistiksel yöntemlerdir, fakat sıklıkla sahip oldukları güçlü model varsayımları nedeniyle eleştirilirler. Diğer taraftan, yapay sinir ağları birleşmiş hafıza özellikleri, genelleştirme kabiliyeti ve seçkin kredi skorlama yeteneği ile çok popüler alternatif kredi skorlama yöntemi olmaktadır. Bununla birlikte, ağların karar topolojisi, potansiyel girdi değişkenlerin göreli önemlerini tespit etmenin mümkün olmaması, belli yorumsal zorluklar ve uzun model eğitme süresi uzun zamandan beri eleştirilmektedir ve bundan dolayı kredi skorlama problemlerinin çözülmesinde sınırlı olarak uygulanmıştır[103, 104].

Birçok araştırma sonuçlarının Yapay Sinir Ağlarının (YSA) hemen tüm problemleri geleneksel modelleme ve istatistiksel yöntemlerinden daha etkin olarak çözebileceğini göstermesine rağmen, bazı karşı araştırma sonuçları da istatistiksel yöntemlerin, belirli veri örneklerinde yapay sinir ağlarından çok daha iyi performans gösterdiğini ispatlamıştır. Sonuçların değişkenliği, bazen yapay sinir ağlarının sistematik bir şekilde kullanılmamasından kaynaklanmaktadır: Sadece bir veya iki yapay sinir ağı algoritmasının kullanılması veya daha iyi bir ağ yapısının mevcut olması, model eğitim zamanı ve öğrenme parametlerini yönlendirecek tüm olası optimizasyon tekniklerinin kullanılmaması bu duruma örnek verilebilir[105].

---

[102] Bugera, V., Hiroshi Konno, and Stanislav Ursayev. "Credit Cards Scoring with Quadratic Utility Functions". **Journal Of Multi-Criteria Decision Analysis**. Vol.11, 2002, s. 199.

[103] Lee, Tian-S., Chih-Chou Chiu, Chi-Jie Lu, and I-Fen Chen. "Credit scoring using the hybrid neural discriminant technique". **Expert Systems with Applications.** Vol.23, 2002, s. 245-252.

[104] Lee, Tian-S., Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu. "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines". **Computational Statistics & Data Analysis**. Vol.50, 2006, s. 1126.

[105] Bensic, M., Natasa S. and Marijana Z.S. "Modelling Small-Business Credit Scoring by Using Logistic Regression, Neural Networks And Decision Trees". **Intelligent Systems In Accounting, Finance and Management**. Vol.13, 2005, s. 138.

Sinir ağlarının standart parametrik yöntemlere göre birkaç avantajı vardır: Birincisi, potansiyel olarak gerçekçi olmayan dağılım şekillerinin varsayımlarını gerektirmez. İkincisi, gizli katman değişkenler arasındaki karmaşık doğrusal olmayan ilişkiler ve bağımlılıkların tespit edilmesine izin verir. Üçüncüsü, yöntemin paralel doğası karmaşık yüksek boyutlu veri yapısı için sıralı bir yaklaşımdan daha uygun olabilir. Son olarak, çıktı katmanı herhangi bir sayıda node içerebilir, böylece etkinleştirilmiş sonuçta kredi değerliliğinin çoklu değer alabilmesine imkân verir[106].

Yazarlar, Yapay Sinir Ağları ile Diskriminant Analizi gibi geleneksel istatistiksel yöntemleri karşılaştırıldığında, Yapay Sinir Ağlarının matematiğin baskın olduğu bir yöntemi olmaması ve karar doğruluğu açısından karşılaştırıldığında, Doğrusal Diskriminant Analizinin yapay sinir ağı modellerinden oldukça iyi olduğu sonucuna varmışlardır. Yapay Sinir Ağı modelinin özellikle finansal sıkıntı içindeki gözlemleri öngörmede doğrusal Diskriminant Analizinden çok daha doğru sonuçlar verdiği görülmüştür. Yazarlar ayrıca Yapay Sinir Ağı modellerinin şüpheli alacakları sınıflandırmada Doğrusal Diskriminant Analizi veya Lojistik Regresyondan daha doğru tahmin ettiğini raporlamışlardır[107]. Yapay Zekâ uygulamalarının erken dönemlerini kapsayan bu çalışmanın sonuçları, yakın dönemlerde yapılan araştırma sonuçları ile tam olarak örtüşmemektedir. Çünkü son dönemlerde Yapay Zekâ algoritmalarının başarı oranının genel olarak yükseldiği aşağıda aktarılacak pek çok örnekle ortaya konulmaktadır.

Lojistik Regresyon (LR) modellerinin Yapay Zekâ (YZ) ya göre başlıca 3 avantajı vardır: Bunlar; modeli yorumlamanın kolay olması, kullanılan değişkenlerin etkilerinin açıklanabilmesi ve her değişkende meydana gelecek belirli bir miktar değişikliğin sonuçlara olan etkilerinin ölçülebilmesidir. Bunların yanı sıra LR, doğrusal olmayan ilişkilerin açıklanmasında kullanılabilen sağlam bir yönteme sahiptir.

Kredi skorlama analizleri ile ilgili son dönemde yapılan pek çok çalışmada yapay zekâ modelleri, istatistiksel tekniklerine göre çok daha iyi sonuçlar ortaya koymuşlardır. Bu başarıda modellerin kapasitelerinin yanı sıra skorlama modellerinin mimari tasarımları da

[106] Sarantopoulos, G., s. 104.
[107] West, D. "Neural network credit scoring models". **Computers & Operations Research**. Vol.27, 2000, s.1133-1134.

önemli yer tutmaktadır[108]. Referans çalışmaların büyük çoğunluğunda ortaya çıkan sonuçlara göre, kredi skorlamada doğruluk açısından; değişkenler arasındaki karmaşık ve doğrusal olmayan ilişkileri yakalamada yüksek kabiliyete sahip olan Yapay Zekâ Modelleri, Diskriminant Analizi ve Lojistik Regresyon tekniklerinden daha doğru sonuçlar üretmektedir[109, 110]. Yapay Zekâ, diğer yöntemler ile kıyaslandığında daha doğru, veriye daha çok uyan (adaptive) ve daha sağlam (robust) modellerdir[111].

Literatürdeki birçok makalede seçilen bir sinir ağı öğrenme algoritması geri yayınım algoritmasıdır (back propagation algorithm - BPN). Bu algoritma ağ ağırlıklarını, modelin gerçek ve beklenen çıktıları arasındaki farklı minimize etmek için değiştirmektedir[112]. Yapay Sinir Ağı geri yayınım algoritması Kredi Skorlama modellerinde geniş bir uygulama alanı bulmuş ve daha önceki çalışmalarda en iyi sonucu vermiştir[113].

Yapay sinir ağları konu dışı nitelikleri birleştirmede veya küçük veri setlerindeki zayıf performansı nedeniyle eleştirilmektedir[114]. Kredi skorlama uygulamaları için herhangi bir sinir ağı modelinin temel bir yetersizliği, krediyi reddetme kararı için mantıklı bir açıklama getirmedeki zorluktur. Sinir ağları genellikle sonuçların açıklanması için herhangi bir mantık veya kural tabanlı açıklamalardan mahrum bırakan kara-kutu teknolojisi olarak düşünülür[115]. Çünkü skorlama modellerinin oluşturulmasında neden-sonuç ilişkisine dayalı modeller hala en popüler modellerdir, bazı durumlarda yasal zorunluluk nedeniyle Yapay Zekâ modellerinin kullanılması mümkün değildir.

Geleneksel istatistik yöntemlerine göre daha doğru sonuçlar üreten Yapay Zekâ tekniğinin; uzun eğitim prosesi, en uygun ağ mimarisinin elde edilmesinden sonra da 'kara kutu' olarak kalması ve modelde kullanılan potansiyel girdi değişkenlerinin göreli önemlerini tanımlamanın kolay olmaması gibi önemli zayıflıkları mevcuttur. Son dönemlerde kullanılan

---

[108] Hsieh, Nan-C. "Hybrid mining approach in the design of credit scoring models". **Expert Systems with Applications**. Vol.28, 2005, s. 657-658.
[109] Lee, Tian-S. ve diğerleri, 2002, s. 248.
[110] Lee, Tian-S. ve diğerleri, 2006, s. 1117.
[111] Huang, Cheng-L., Mu-Chen Chen, and Chieh-Jen Wang. **"**Credit scoring with a data mining approach based on support vector machines", **Expert Systems with Applications**. Vol.33 2007, s. 848.
[112] Sarantopoulos, G., s. 104.
[113] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan., s. 4738.
[114] Ong, Chorng-S., Jih-Jeng Huang, and Gwo-Hshiung Tzeng. "Building credit scoring models using genetic programming". **Expert Systems with Applications.** Vol.29, 2005, s. 41.
[115] West, D., s.1148.

Genetik Algoritma ve hibrid sistemleri de gelecek vaat etmektedir[116]. Veri seti ve özelliklerin varsayımlarını gerektiren geleneksel istatistiksel yöntemlerden farklı olarak, Genetik Programlama parametrik olmayan bir araçtır ve herhangi bir durum ve veri setine uygundur[117].

Geleneksel istatistik sınıflandırma tekniklerine uygun, kararın verildiği nihai olasılığı hesaplamak için temel bir olasılık modeli kabul edilmelidir. Yapay Sinir Ağları, Genetik Programlama ve Destek Vektör Makineleri gibi en yeni geliştirilen veri madenciliği yöntemleri bu sınırlandırma olmadan sınıflandırmayı yapabilmektedir. Ek olarak, bu Yapay Zekâ yöntemleri ayrıca geleneksel istatistiksel yöntemlerden daha iyi bir performans göstermişlerdir. Çünkü istatistiksel sınıflandırma modelleri sadece temel varsayımlar karşılandığında olumlu performans göstermektedirler. Geleneksel istatistik yöntemler ile karşılaştırıldığında, yapay zekâ yöntemleri (DVM, Genetik Programlama, BPN veya Karar Ağaçları gibi) girdi ve çıktı değişkenler arasında yatan ilişkilerin bilgisi gerekmemektedir[118].

Yapay Zekâ algoritmaları, yorumlamanın çok önemli olmadığı sahtekârlık modellerinde sıkça kullanılmaktadır. Sahtekârlık tespitinin yapıldığı modeller ile riskli abonelerin tespit edildiği modeller, kullanılan verinin içeriği ve metotlar açısından birbirlerinden farklılık gösterirler.

5.2. Literatür İncelemesi

Kredi skorlama kavramının çok geniş bir uygulama alanı mevcut olup burada, bizim çalışmamız ile benzerlikler taşıyan çalışmalar aktarılmıştır. Benzerlikler başlıca incelemeye konu olan ürün ile ilgilidir. Örneğin, mobil operatörler tarafından kredilendirilen faturalı aboneler ile bankalar tarafından verilen bireysel krediler için verilecek kararlar benzer kriterlere dayanırlar. Söz konusu kredilerin miktarı da ev kredisi gibi miktar olarak büyük olmayıp daha çok kredi kartı veya benzeri bireysel ürünleri kapsamaktadır. Bunun yanı sıra tez içerisinde kullanılan yöntemlerin yer aldığı karar verme yöntemlerini kullanan çalışmalar da aktarılmıştır.

[116] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan., s. 4737.
[117] Ong, Chorng-S., Jih-Jeng Huang, and Gwo-Hshiung Tzeng., s. 46.
[118] Huang, Cheng-L., Mu-Chen Chen, and Chieh-Jen Wang., s. 848, 854.

Telekomünikasyon sektöründe uygulanan davranışsal skorlama ilgili olarak literatürde çok az sayıda çalışma bulunmaktadır. Bunun en önemli nedenlerinden bir tanesi bu alandaki çalışmaların bankacılık ve diğer finans sektöründeki kadar gelişmemiş olmasıdır. Madeira ve diğerleri tarafından 2003 yılında Portekiz'de faaliyet gösteren bir telekomünikasyon firmasında risk değerlendirmesi için veri madenciliği yaklaşımı kullanılarak yapılan bir çalışma[119] bu alanda yapılan güzel bir örnektir.

Bu çalışmada mevcut aboneler risk düzeylerine göre yüksek, orta ve düşük olmak üzere üç alt gruba ayrıştırılmıştır. Değerlendirme anından önceki altı aya ait veri seti kullanılarak sonraki üç ay içerisinde hangi risk segmentinde yer alacağı tahmin edilmeye çalışılmıştır. Yeni müşterilerin risk segmentini tahmin etmek için de önceki üç aya ait veriler kullanılarak sonraki bir ay tahmin edilmiştir. Analiz veri setlerinde Çoklu Lojistik Regresyon, Yapay Zekâ ve Karar Ağaçları modelleri denenmiştir. Sonuçların doğruluk oranlarının daha yüksek çıkması ve yorumlamanın kolaylığı nedeniyle karar ağaçları tercih edilmiştir. Aboneler öncelikle dört segmente ayrıştırılmış ve her bir segment için 10 ile 60 arasında uç noda sahip Karar Ağacı modelleri geliştirilmiştir. Daha sonra buradan elde edilen model sonuçlarına, tüzel kişilere ait açıklayıcı değişkenler de eklenerek yeniden Karar Ağacı modeli uygulanarak, iki aşamalı bir skorlama sistemi geliştirilmiştir. Aboneler risk olasılıklarına göre (Probability of default) %98-%100 arası yüksek riskli, %65-%97 arası orta riskli, %0-%64 arası düşük riskli olarak kabul edilmiştir. Hâlihazırda kullanılan skor kart sonuçları ile karşılaştırıldığında, yeni kurulan iki aşamalı veri madenciliği uygulamaları ile elde edilen sonuçların doğruluk oranının tüm segmentler göz önüne alındığında %15 daha iyi olduğu görülmüştür.

Literatürde karşılaşılan kredi skorlama ile ilgili çalışmaların önemli bir kısmının aynı veri seti üzerinde yapılan çalışmalardan oluştuğu görülmüştür. Bunun önemli bir nedeni bu konuda gerekli verinin kolay elde edilememesi, hatta ülkemiz dâhil pek çok yerde verinin paylaşımı konusunda yasal engellerin bulunmasıdır. Kullanılan veri setleri; UCI Machine Learning Repository veritabanında bulunan Alman kredi veri seti (German Credit) ve Avustralya kredi veri seti (Australian Credit) dir[120]. Alman kredi veri setinde kredi verilen toplam 1.000 başvuru mevcut olup, bunların 700'ü iyi, 300'ü sorunlu krediddir. Her kayda ait kredi geçmişi, hesap bakiyesi, borç detayı ve kişisel bilgileri içeren 20 değişken mevcuttur. Avustralya kredi

[119] Madeira Sara C., Arlindo L. Oliveira, and Catarina S. Conceicao. A Data Mining Approach to Credit Risk Evaluation and Behaviour Scoring, ss. 184–188, Fernando Moura Pires, Salvador Abreu (Eds.): **EPIA 2003, LNAI 2902**, 2003. Springer-Verlag Berlin Heidelberg 2003
[120] http://archive.ics.uci.edu/ml/datasets.html (06.01.2009)

veri seti de buna benzer olup, 468'i iyi, 222'si sorunlu toplam 690 gözleme ait 14 değişkenden oluşmaktadır. Bu iki verinin aşağıda aktarılacak pek çok farklı çalışmada kullanılmasının önemli bir yararı, bu çalışmaların sonuçlarının karşılaştırılmalarını mümkün kılmasıdır.

Alman ve Avustralya kredi veri seti üzerinde David West[121] tarafından yapılan çalışmanın amacı, beş sinir ağı mimarisinin (çok katmanlı algılayıcı (multilayer perceptron), uzmanların karışımı (mixture-of-experts), radyal tabanlı fonksiyon (radial basis function), öğrenen vektör nicelemesi (learning vector quantization) ve bulanık uyarlamalı rezonans (fuzzy adaptive resonance)) kredi skorlama için uygunluğunu araştırmak ve bunların performansını; Doğrusal Diskriminant Analizi, Lojistik Regresyon, k-enyakın Komşuluk, Kernel Yoğunluğu ve Karar Ağaçları yöntemleri ile karşılaştırmıştır. Bu araştırmanın sonuçları göstermiştir ki Sinir Ağı kredi skorlama modelleri, kredi skor doğruluğunda % 0.5 ten % 3'e kadar değişen ölçüde fraksiyonel olarak iyileştirme yapmıştır. Ancak kredi skorlama modellerinde Sinir Ağlarının kullanımı, ağ topolojilerini geliştirmek ve üstün eğitim yöntemleri tasarlamak için biraz modelleme yeteneğine sahip olmayı gerektirir. Çok katmanlı algılayıcı (multilayer perceptron) en yaygın kullanılan sinir ağı modeli olmasına rağmen, uzmanların karışımı (mixture-of experts) ve radyal temelli fonksiyon (radial basis function) sinir ağları da kredi skorlama uygulamaları için düşünülmelidir. Bu araştırma, ayrıca Lojistik Regresyonun Sinir Ağları modellerine iyi bir alternatif olduğunu ileri sürmektedir. Birkaç Sinir Ağı eğitim yinelemesi içeren ortalama bir durum için Lojistik Regresyon, Sinir Ağı modellerinden biraz daha doğru sonuçlar üretir. Parametrik olmayan modeller (k-en yakın komşuluk ve Kernel Yoğunluğu) ve CART ümit verici sonuçlar üretmemiştir, ancak çok büyük veri setleri için iyileşmiş sonuçlar vereceği ispat edilebilir.

Chorng-Shyong Ong ve diğerleri[122] Alman kredi veri seti ile birlikte, 307 yüksek kredibiliteye, 383 yetersiz kredibiliteye sahip gözlem için 14 değişkenden oluşan farklı bir Avusturya kredi veri setini kullanarak; Genetik Programlama (GP) ile birlikte kıyaslama amacı ile Yapay Zekâ (MLP algoritması), Karar Ağaçları (CART ve C4.5), Kaba Küme (Rough Sets) ve Lojistik Regresyon (LR) modelleri ile kredi skorlama modelleri oluşturmuştur. Daha doğru sonuçlar elde edilen ve daha esnek yapıya sahip GP modeli şu 3 nedenden dolayı seçilmiştir: Birincisi, ilgisiz değişkenler verinin yapısını bozmakta ve

[121] West, D., s.1131-1132, 1149-1150.
[122] Ong, Chorng-S., Jih-Jeng Huang, and Gwo-Hshiung Tzeng., s. 41, 44-46.

ayrıştırma fonksiyonunun doğruluğunu düşürmektedir. İkincisi, kredi skorlama modeli doğru ayrıştırma fonksiyonunu (doğrusal veya doğrusal olmayan) otomatik olarak belirlemelidir. Üçüncüsü, kredi skorlama modeli hem büyük hem de küçük veri setleri için kullanışlı olmalıdır. Ancak, Yapay Sinir Ağları ve Lojistik Regresyon modelleri de bu veri setleri üzerinde tatmin edici sonuçlar sağlamışlar ve alternatif çözümler olarak önerilmektedirler.

Chorng-Shyong Ong, ve diğerleri[123] 2005 yılında yaptıkları ve yukarıda özetlenen çalışmalarında kullandıkları veri setlerini kullanarak tekrarladıkları güncellenmiş çalışmaya, k-en Yakın Komşuluk ve iki aşamalı Genetik Programlama (GP) yöntemlerini de eklemişlerdir. İki aşamalı GP yönteminin ilk kısmında Genetik Programlama karar verici için Eğer-İse (IF-THEN) kuralları türetmesi için kullanılmaktadır. Genetik Programlamanın ikinci aşamasında, küçültülmüş veri seti, tahmin yeteneği kazandırmak amacıyla Diskriminant Fonksiyonu kurulması için kullanılır. Bu çalışma sonucunda da, ilk çalışmada GP'nın tercih nedeni olarak sıralanan üç faktöre ek olarak, 'karar vericiler için çok önemli olan veri setinin içeriğini kavramada kullanabilecekleri akıllı kuralların geliştirilmesi' nedeniyle iki aşamalı Genetik Programlama yöntemi, çalışma içerisindeki tüm diğer yöntemlere tercih edilmiştir.

İstatistik ve Sinir Ağı tabanlı yaklaşımlar en popüler paradigmalar arasındadır. Bununla birlikte, bu yöntemlerin çoğu sözde "sıkı" sınıflandırıcıları ortaya çıkarır, bunlar da beraberinde güvenilirlik ölçüsü olmayan sonuçlar üretir. Buna karşın, bulanık küme yaklaşımı kullanarak tasarlananlar gibi "yumuşak" sınıflandırıcılar; analiste daha büyük bir kavrayış sağlayan karar (alternatif karar da dahil) için bir destek ölçüsü üretir. Alman kredi veri seti üzerinde Arijit Laha[124] tarafından davranışsal skoru belirlemek için yapılan çalışmada, bulanık kurallara dayalı sınıflayıcılar kullanılarak kredi skorlama modelleri kurmaya yönelik bir yöntem önerilmektedir. İlk olarak, kural tabanı SOM (Self-Organizing Map) tabanlı bir yöntem kullanan eğitim verisinden öğrenilir. Sonra eğitim setindeki içeriksel bilgi ile entegre olan içeriksel bir sınıflandırıcı tasarlamak için bulanık k-en yakın komşuluk kuralı ile birleştirilir. Böylece daha güçlü ve niteliksel olarak daha iyi sınıflandırma yapılır. İleri sürülen yöntem, belirsizliği daha uygun bir yolla ele almak için daha bilgi verici bir sınıflandırma çerçevesi sağlamak amacıyla nitelik uzayındaki veri noktalarının içeriksel bilgisini kullanır. İçeriksel bilginin kullanılmasından dolayı sınıflandırma çıktılarının niteliksel iyileşmesi

[123] Huang, Jih-J., Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. "Two-stage genetic programming (2SGP) for the credit scoring model". **Applied Mathematics and Computation.** Vol.174, 2006, s. 1039, 1046-1047.
[124] Laha, A., s. 281, 291.

ispatlanmaktadır. Modelleme çerçevesinde gerçek yaşam kullanıcısı ile ilişkili iş sınırlarını sorunsuz bir şekilde ortaya koymanın bir yöntemi önerilmektedir. Sınıflayıcıların kalitesini iyileştirmek için içeriksel bilgiyi kullanmanın temel fikri, k-en yakın komşuluk kurallarının çeşitli farklı şekillerinin kullanımı ile takip edilebilir.

Aynı veri setleri ile yapılan bir çalışmada da Destek Vektör Makineleri (DVM) yöntemi kullanılmıştır[125]. Çünkü DVM kümeleme metodu, hâlihazırda üzerinde çok çalışılan bir araştırma alanı olup pek çok alanda kümeleme problemlerini başarı ile çözüme kavuşturmuştur. Başarılı bir şekilde kredi skorlama modelleri kurmak için bu çalışma, başvuru sahibinin girdi özelliklerinden adayın kredi skorunu değerlendirecek hibrid DVM-tabanlı kredi skorlama modelleri kurmak için üç strateji kullanmıştır: (1) Model parametrelerini optimize etmek için grid araştırma kullanılmıştır. (2) Model parametrelerini optimize etmek için grid araştırmanın kullanımı ve girdi özelliklerini seçmek için F skorları kullanılmıştır. (3) Model parametrelerini ve girdi özelliklerini eşanlı olarak optimize etmek amacıyla genetik algoritmalar kullanılmıştır. Sinir ağları, Genetik Programlama ve Karar Ağacı sınıflayıcıları ile kıyasladığında, DVM sınıflayıcısı göreli olarak daha az girdi ile eşdeğer bir sınıflandırma doğruluğunu yakalamaktadır. Ek olarak, DVM sınıflayıcı ile Genetik Algoritmanın birleştirilmesi, önerilen hibrid GA-DVM stratejisi özellik seçme görevini ve model parametreleri optimizasyonunu eşanlı olarak gerçekleştirebilir. DVM tabanlı model sınıflandırma doğruluğu açısından BPN (Back Propagation NN) ve Genetik Programlamaya göre oldukça rekabet edebilir güçtedir. GP ve BPN ile karşılaştırıldığında, DVM tabanlı kredi risk modeli eşdeğer sınıflandırma doğruluk oranını başarabilir. Bu çalışmaya göre, hibrid bir GA-DVM sistemi parametreleri ve özellik altkümelerini optimize etmek için iyi bir alternatiftir. Küçük bir özellik altkümesi ile hibrid bir GA-DVM sistemi iyi bir sınıflandırma performansı yakalayabilir. Bununla birlikte, GA-DVM stratejisi kullanılırken, GP ve BPN gibi aşırı öğrenmeden kaçınılmalıdır.

Brad S. Trinkle ve Amelia A. Baldwin[126], yukarıda detayı aktarılan Alman kredi veri seti ile yine Alman tüketicilerine ait farklı bir veri setini (toplam 3000 gözlem, iyi/kötü oranı %50, 26 bağımsız değişken) kullanarak bir çalışma yapmıştır. Yapılan bu çalışmasının sonuçlarına göre; Yapay Sinir Ağları (YSA), geleneksel doğrusal modellerden daha iyi kredi skoru sağlayan kredi skorlama modelleri geliştirme potansiyeline sahiptir. Ancak, ABD'deki kredi

[125] Huang, Cheng-L., Mu-Chen Chen, and Chieh-Jen Wang., s. 847-848, 854-855.
[126] Trinkle, Brad S. and Amelia A Baldwin., s. 123, 130, 145.

onay sürecindeki yasal sınırlandırmalar yorumlanabilir modelleri gerekli kılmaktadır. Bundan dolayı, güçlü ve çoklu gizli katman YSA'nı yorumlayabilecek teknikler geliştirildiği zaman bu problemlerin üstesinden gelinebilir.

Alman kredi veri seti üzerinde son dönemde yapılan çalışmalardan bir tanesinde de[127] genel kabul görmüş kredi skorlama yöntemlerinden Doğrusal Diskriminant Analizi (DDA), Lojistik Regresyon (LR), Çok Değişkenli Uyarlamalı Regresyon Eğrileri (ÇDURE) (Multivariate Adaptive Regression Splines - MARS), Classification and Regression Tree (CART), Case Based Reasoning (CBR) ve Yapay Sinir Ağları (YSA) kullanılmış ve bunlardan MARS, YSA ve CBR yöntemlerinin avantajları kanıtlanmıştır. Bu çalışmada kullanılan ve önerilen iki aşamalı teknikte; ilk aşamada gözlemler iyi ve kötü olarak sınıflandırılmakta ve böylece borç verenin riskini düşürüp, bu tutarı gelecekteki tasarruflarına transfer etmektedir. Yöntemin ikinci aşamasında da birinci tip hatayı azaltmak için reddedilen abonelerin bir kısmı CBR yöntemine göre geliştirilen kurallar yardımıyla koşullu olarak kabul edilmektedir. Kullanılan ikinci aşama ile kabul oranı artmaktadır.

Alman ve Avustralya kredi verisinin Hibrid yöntemler kullanılarak kredibilitesinin ölçüldüğü başka çalışmalar da yapılmıştır. Hsieh'in çalışması[128] genel olarak, kümeleme algoritmaları veya genetik algoritmalar tarafından yönlendirilen sinir ağları kullanan çok daha doğru sınıflandırıcıları yaratmak için hibrid tasarlanmış mimariler önerilmiştir. Bu çalışmada, etkin bir kredi skorlama modelinin tasarlanmasında kümeleme ve sinir ağları tekniklerine dayalı hibrid bir veri madenciliği yaklaşımı sunulmaktadır. İzole ve tutarsız kümelerdeki kümeyi temsil etmeyen örneklemleri göstermek amacıyla girdi örneklemlerini önişlemeye tabi tutmak için kümeleme teknikleri kullanılır. Kredi skorlama problemi için sinir ağlarını kurma tecrübesi kümeleme yapmanın, yüksek etkinliğe sahip ağ kurmak için çok önemli olduğunu göstermiştir. Bu yöntemi kullanarak örneklemleri bölmek, her bir grupta yer alan adayların davranışsal eğilimlerini anlamanın ve etkin kredi skor modelleri kurmanın bir yolu olacaktır ve başvurularda etkin olduğunu ispatlamaktadır. Deneysel sonuçlar böyle bir hibrid yöntemin basit ancak kredi piyasalarındaki başvuruları değerlendirmek için etkili olduğunu göstermiştir.

[127] Chuang, Chun-L. and Rong-Ho Lin. "Constructing a reassigning credit scoring model". **Expert Systems with Applications**. Vol.36, 2009, s. 1685.
[128] Hsieh, Nan-C., 2005, s. 665-662.

Hibrid yöntemler kullanılarak yapılan bir başka örnek çalışma da Tian-Shyug Lee, ve diğ, tarafından[129] Tayvan'daki bir yerel bankadan alınan 6.000 adetlik veri ve dokuz açıklayıcı değişken ile gerçekleştirilmiştir. Buradaki çalışmanın amacı, Lineer Diskriminant Analizi (LDA) yaklaşımını Sinir Ağları teknikleri ile birleştirmedeki iki aşamalı hibrid modelleme usulünü kullanarak kredi skorlama performansını keşfetmektir. Analizlerin altındaki mantıksal temel, öncelikle kredi skorlama modellerinde LDA'ni kullanmaktır. Sonra anlamlı tahmin edici değişkenler, tasarlanan Sinir Ağı modelinin girdi değişkenleri olarak görev alır. Ayrıca Diskriminant Analizinin kredi skorlama sonucu, daha iyi bir başlangıç sonucu elde etmek ve kredi skoru doğruluğunu artırmak için ekstra bilgi vermek amacıyla girdi katmanı içine dahil edilir. Şuna dikkat edilmelidir ki Diskriminant Analizini Sinir Ağları topolojisinin tasarımında destek aracı olarak kullanmak oldukça değerlidir; çünkü bu sayede içsel çalışmalardan daha fazla bilgi elde edilebilir. Ayrıca Sinir Ağı modelinin en iyi girdi değişkenlerini tespit etmede teorik bir yöntem olmadığı için, Diskriminant Analiz yöntemi, birçok potansiyel değişken düşünüldüğü zaman girdi değişkenlerinin en iyi altkümesini tespit etmek için genel kabul edilmiş bir yöntem olarak uygulanabilir. Böylece tasarlanan Sinir Ağı modelinin girdi vektörüne karar verilirken iyi bir istatistikî destek verilir. Analitik sonuçlar Sinir Ağı modellerinin; Diskriminant Analizi, Lojistik Regresyon ve Katışıksız Sinir Ağı modelleri ile kıyaslandığında doğru sınıflandırma oranında en yüksek ortalamaya sahip olduğunu ispat etmektedir ve değişenler arasındaki doğrusal olmayan ilişkileri daha iyi yakalama yeteneğine sahip olduğu tahminlerini haklı çıkarmaktadır. Ayrıca, tasarlanmış hibrid modeller yalnızca daha iyi kredi skoru doğruluğuna sahip değillerdir, aynı zamanda en düşük yüksek yanlış sınıflandırma maliyetleri ile ilişkili Tip II hatasına da sahiptir. Sonuç olarak, hibrid modeller istatistiksel destek sağlamak, girdi değişkenlerin sayısını azaltmak ve daha iyi bir başlangıç sonucu elde etmek için Diskriminant Analizi sonuçlarından yardım almıştır.

Tian-Shyug Lee ve diğerleri[130], Tayvan'da yerel bir bankaya ait 8.000 adetlik veri ile kredi kart veri setini kullanarak farklı veri madenciliği metotlarının kredi skorlama performanslarını karşılaştırmıştır. Yapılan kıyaslama sonucunda; Sınıflandırma ve Regresyon Ağacı (CART - Classification and Regression Tree) ve Çok Değişkenli Uyarlamalı Regresyon Eğrileri (Multivariate Adaptive Regression Splines - MARS) yöntemlerinin her ikisi de Diskriminant Analizi, Lojistik Regresyon, Sinir Ağları ve Destek Vektör Makineleri yöntemleri ile

[129] Lee, Tian-S. ve diğerleri, 2002. s. 245-246, 250-251.
[130] Lee, Tian-S. ve diğerleri, 2006. s. 1113-1127.

kıyaslandığında sınıflandırma doğruluğu oranında daha iyi bir ortalamaya sahip çıkmıştır. Ayrıca, CART ve MARS aynı zamanda yüksek yanlış sınıflandırma maliyeti ile ilişkili daha düşük Tip II hatasına da sahiptir, bundan dolayı da daha iyi bir bütünsel kredi skorlama yeteneğine sahiptir. Tian-Shyug Lee ve diğerleri tarafından yapılan çalışmada önemli bulunan CART ve MARS modellerinin kredi skorlama problemlerinde kullanmasının dört önemli mantıksal temeli vardır: İlk olarak, LDA ve LR dan farklı olarak, hem CART hem de MARS yaklaşımları güçlü model varsayımları olmaksızın değişkenler arasındaki karmaşık ilişkileri modelleme kabiliyeti göstermektedir. Ayrıca, Sinir Ağlarından farklı olarak, her ikisi de, birçok potansiyel değişken düşünüldüğü zaman ağaç ve temel fonksiyonlar yardımıyla "önemli" bağımsız değişkenleri belirleyebilirler. Üçüncü olarak, CART ve MARS uzun bir model eğitim sürecini gerektirmez ve böylece veri seti çok büyük olduğu zaman modelleme zamanından oldukça büyük tasarruf sağlar. Son olarak, CART ve MARS'ın diğer sınıflandırma teknikleri üzerindeki çok güçlü bir avantajı, sonuç sınıflandırma modeli kolayca yorumlanabilir. Sadece hangi değişkenlerin gözlemleri veya hedefleri sınıflandırmada önemli olduğunu göstermez, aynı zamanda belirli bir hedef veya gözlemin kurulmuş kurallar karşılandığında hangi spesifik sınıfa ait olduğunu da gösterir. Son gerçeğin önemli pazarlama uygulamaları vardır ve pazarlama profesyonellerinin daha iyi yönetsel kararlar almalarına yardımcı olur.

1994 ile 1998 yılları arasında bir Sloven bankası tarafından kullandırılan 581 kısa vadeli kredi gözlemine ait 67 değişkenli veri seti kullanılarak yapılan çalışmada da Yapay Zekâ modeli kullanılarak yapılan model sonuçları daha başarılı çıkmıştır. Öncelikle 67 değişkenden oluşan bir veri setinden optimal değişken seçimi işlemi için bir araştırma yapmanın iyi sonuçlar verdiği ispatlamıştır. Değişken seçimi için Temel Bileşenler Analizi (TBA) (Principal Component Analysis - PCA), eğitim ve test alt kümeleri oluşturmak için de Kohonen Yapay Sinir Ağları ve rastsal yönteme dayalı bir Genetik Algoritma kullanılmıştır. Kredi skorlama modellerinin geliştirilmesinde kullanılan 21 değişken, Kohonen alt-kümelerini kullanan Genetik Algoritma tarafından seçilmiştir. Modellerin geliştirilmesi için hata geri yayılımlı Yapay Sinir Ağları (back-propagation artificial neural networks) ve Lojistik Regresyon kullanılmıştır. Tahminlerin doğruluğu, k-katı çapraz doğrulama ile test edilmiş ve hata geri yayılımlı sinir ağı modeli en iyi ortalama sonuçları göstermiştir: % 79,3 doğruluk, % 17,8 tip II hatası ve % 29,9 tip I hatası[131].

[131] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan., s. 4740, 4742.

Davranışsal skorlama modelleri, aynı zamanda abone riski ile ilgili yönetim stratejilerinin geliştirilmesi için de kullanılabilir. Yoon Seong Kim, So Young Sohn[132], yine Alman kredi veri seti üzerinde oluşturdukları YSA kredi skorlama modelinin yanlış sınıflandırma desenini kullanarak aboneleri gruplara ayırmışlarıdır. Daha sonra aboneler mevcut kredi statüleri ve gruplama sonuçlarına göre dört alt gruba ayrıştırılmıştır. Her alt grubun karakteristiklerinden yapılan çıkarsamaya göre de uygun yönetim stratejileri önerilmiştir.

Zapounidis ve Doumbos[133] tarafından yapılan çalışmada, bir tanesi kredi kartı başvuru değerlemesi olmak üzere beş farklı finansal problemin çözümü için Diskriminant Analizi ile UTADIS MCDA sıralama modelleri karşılaştırılmıştır. Çok Kriterli Karar Destek (ÇKKD) (Multicriteria Decision Aid – MCDA) yöntemi, probleme iyi uyarlanmış ve kredi skorlama kapsamında yaygın olarak kullanılan birçok yöntem sunmaktadır. ÇKKD'nin iyi bilinen bir yaklaşımı UTADIS (UTilités Additives DIScriminantes) yöntemidir. Bu yöntem; parçalara ayırma yaklaşımına dayanır, sıralama problemleri içindeki önceden tanımlanmış sınıflar arasındaki yanlış sınıflandırma hatasını minimize etmek için doğrusal programlama tekniklerini kullanarak fayda fonksiyonları ve fayda profillerinin bir grubunu tahmin eder. Biri niteliksel olmak üzere yedi değişkene sahip 150 kredi kartı başvurusuna ait veri seti kullanılmıştır. Kredi kartının değerlendirilmesi iki aşamada yapılmıştır: Birinci aşamada tüm başvurular içerisinde kabul edilenler ayrıştırılmış, ikinci aşamada da reddedilen başvurular içerisinde daha fazla incelenmesi gerekenler belirlenmiştir. Bu çalışmanın sonuçlarına göre ÇKKD metodu, pek çok diğer sınıflandırma probleminde karar verme amaçlı kullanılabildiği gibi kredi skorlama için de uygulanabilecek kullanışlı ve güçlü bir araçtır ve elde edilen sonuçlar Diskriminant Analizine göre istatistiki olarak daha iyidir.

Vladimir Bugera, Hiroshi Konno and Stanislav Ursayev[134] tarafından 1995-1996 yıllarında Yunanistan Ulusal Bankasına yapılan 150 Kredi kartı başvurusuna ait 25 değişkenden yedisi kullanılarak matematiksel programlama algoritması ile başvuru skorlaması çalışması yapılmıştır. Çözüm yaklaşımı bir fayda fonksiyonunun optimize edilmesine dayanır; gösterge parametrelerinde ikinci dereceden denklemdir ve kontrol parametrelerinde doğrusaldır

[132] Kim, Yoon S. and So Young Sohn. "Managing loan customers using misclassification patterns of credit scoring model". **Expert Systems with Applications**. Vol.26, 2004, s. 567-573.
[133] Zopounidis, C. and Michael Doumpos. "Multi-group discrimination using multi-criteria analysis: Illustrations from the field of finance". **European Journal of Operational Research**. Vol.139, 2002, s. 378, 385.
[134] Bugera, V., Hiroshi Konno, and Stanislav Ursayev., s. 197-198, 206, 210-211.

(tanımlanmayı gerektirdiği için). Fayda fonksiyonunun, bazı değişkenlerdeki tek düzelik gibi niteliksel özellikleri, ek kısıtlar kullanılarak çalışmaya dahil edilmiştir. Çalışmada başvurular üçlü kümeye (kabul, ret, belirsiz) ayrıştırılmıştır. İkinci dereceden ve doğrusal fayda fonksiyonlarının birkaç sınıfının performansı çeşitli kısıtlar kullanılarak karşılaştırılmıştır. Deneyler, veri setinin büyüklüğüne göre modelin "esnekliğini" ayarlayan kısıtlar uygulamanın önemini göstermiştir. Fayda fonksiyonunun uygun bir sınıfını seçmek ve kısıtları uygulamak, önerilen yöntemin başarısında kritik rol oynamaktadır. Küçük eğitim setleri için, esnek modeller oldukça zayıf çalışmaktadır ve esnekliğin azaltılması (kısıtların uygulanması) modelin tahmin yeteneğini iyileştirebilir. Kısıtlar ile eğitim veri setinin boyutuna göre modelin esnekliğini ayarlayabiliriz. Fayda fonksiyonlarının dikkate alınan sınıfları için, en uygun fayda fonksiyonu doğrusal programlama teknikleri kullanılarak bulunabilir. Doğrusal programlama daha büyük veri setleri için kullanılabilecek hızlı algoritmalara götürür. Dikkate alınan veri seti için yapılan tespit şudur: En iyi örneklem dışı özellikler (en küçük örneklem dışı hata) kontrol değişkenleri (fayda fonksiyonunun katsayıları) üzerindeki pozitiflik sınırları ile birlikte ikinci dereceden fayda fonksiyonları tarafından verilmektedir.

Mirta Bensic ve diğerleri[135], göreli olarak küçük bir veri setini kullanarak küçük işletmelerin borçlanması ile ilgili kredi skorlama modelleri geliştirmiştir. 160 adet örneğe ait iş sahiplerinin kişisel profili, işletme aktivitesi ve finansal verisini kapsayan 31 adet değişkenden oluşan veri setinde; herhangi bir ödemesinin vadesi 46 gün ve üzerinde olanlar kötü, diğerleri iyi olarak kabul edilmiştir. Çalışmada, Lojistik Regresyon, Sinir Ağları ve CART Karar Ağacı gibi farklı yöntemler ile geliştirilen en iyi modellerin doğrulukları karşılaştırılmıştır. İleri doğrusal olmayan değişken seçim stratejisi kullanılarak geri yayınım (back propagation) algoritması, radyal tabanlı fonksiyon ağı (radial basis function network), olasılıklı (probabilistic) ve öğrenen vektör nitelendirmesi (learning vector quantization) gibi dört farklı sinir ağı algoritması test edilmiştir. Orantı kuralı ve McNemar testleri ile test edilen modeller arasında istatistiksel olarak anlamlı farklılıklar görünmemesine rağmen, Olasılıklı Sinir Ağı modeli en yüksek doğruluk oranı ve en düşük tip I hatası üretmiştir.

Kredi skorlama alanında genelde bir veri seti üzerinden sınırlı sayıdaki modellerle yapılan çalışmalarda birbiriyle çelişen sonuçlar ortaya çıkmıştır. Sonraki çalışmalar öncekilerle

[135] Bensic, M., Natasa S. and Marijana Z.S., s. 133-137.

uyuşmayabilmekte veya bir çalışmada başarılı bulunan model bir diğerinde çok kötü çıkabilmektedir. Bu eleştirilerden yola çıkılarak, gerçek hayattan alınan sekiz farklı veri seti üzerinde, birbirinden çok farklı nitelikteki sınıflandırma modelleri uygulanarak birbirleri ile kıyaslanması Baesens ve diğerleri[136] tarafından gerçekleştirilmiştir. Kullanılan veriler, Benelüx ülkeleri (Belçika, Hollanda, Lüksemburg) ile yukarıda aktarılan Alman ve Avustralya kredi veri setleri ve İngiltere'deki finansal kuruluşlara ait veri setleridir. Uygulanan teknikler; Lojistik Regresyon (LR), Lineer Diskriminant Analizi (LDA), Quadratik Disktriminant Analizi (QDA), Lineer Programlama (LP), dört farklı Destek Vektör Makineleri (DVM) uygulaması, Yapay Zekâ (YZ), Naive Bayes (NB), Tree Augmented Naive Bayes Classifiers (TANs), dört farklı Karar Ağacı algoritması, iki farklı k-en Yakın Komşuluk sınıflandırıcısıdır. Veri setleri yeterince büyük olduklarından verilerin üçte ikisi eğitim (training), üçte biri de test amaçlı olarak kullanılmıştır. Sınıflandırma performansı, doğru sınıflandırılan vakaların yüzdesi (Percentage Correctly Classified Cases - PCC) ve ROC (Receiver Operating Characteristic Curve) altında kalan alan ile değerlendirilmiştir. Kullanılan modeller çoğunluğu kabul edilebilir derecede başarılı sonuçlar üretse de, radial basis function destek vektör makineleri (RBF DVM) ve Yapay Zekâ (YZ) hem PCC hem de ROC kriterlerine göre en yüksek performansları göstermişlerdir. Bunun yanında LDA ve LR modelleri de yüksek başarı göstermiş olup, bu sonuç çoğu kredi veri setinin lineer olduğuna işaret etmektedir.

Kredi ve davranış skorlama modelleri, finansal problemlerin çözümünde kullanılan önemli araçlardır. Bu nedenle pek çok çalışma yeni kredi başvurularının kabul edilip edilmemesi ile ilgili modellerin doğruluğu üzerinde odaklanmaktadır. Ancak, davranışsal skorlama modeli pazarlama stratejilerinin geliştirilmesi amacıyla da kullanılabilir. Hsieh[137] büyük bir Tayvan bankasının kredi kartı kullanıcılarına ait 158.126 kaydını kullanarak, firmanın müşterilerine yönelik daha iyi pazarlama stratejileri geliştirilmesine neden olacak ipuçları sağlaması için, veri içerisindeki ilginç deseni ortaya çıkarmayı amaçlamıştır. Çalışma sonucunda mevcut müşteriler ortak davranış ve özelliklerine göre üç karlı müşteri grubuna bölünmüştür: tabanca (revolver) kullanıcı, aracı kullanıcı ve elverişli kullanıcı. Böylece pazarlamacılar daha sonra her bir gruptaki müşterileri yorumlayarak ve her bir grubun özelliklerine uygun yönetim stratejileri önerebilirler.

[136] Baesens B, T Van Gestel, S Viaene, M Stepanova, J Suyken and J Vanthienen "Benchmarking state-of-the-art classification algorithms for credit scoring". **Journal of the Operational Research Society**. Vol.54, 2003, s. 627, 635.
[137] Hsieh, Nan-C., 2004, s. 623–632.

G. Sarantopoulos[138] Haziran ile Kasım 2000 arasında özel bir ev eşyası alış veriş merkezine sipariş veren 77.876 başvuruyu kullanarak, başvuruları iyi ve kötü olarak ayıracak bir karar ağacı modeli oluşturmuştur. Modele alınan abonelerden altı aylık sürede hiç ödemesini geciktirmeyen aboneler iyi, üç veya daha fazla ödemesini geciktiren veya yapmayan aboneler kötü, diğerleri de ortalama abone olarak tanımlanmış; ortalama aboneler modelin iyi ve kötü aboneler arasındaki ayırım gücünü artırabilmek için analize dâhil edilmemiştir. Önerilen Karar Ağacı, kabul edilen kötü sipariş oranlarını önemli oranda düşürmüş ve toplam satışları artırmıştır. Kurulan model, karşılaştırma yapmak amacıyla oluşturulan Lojistik Regresyon modelinden önemli derecede daha iyi sonuçlar üretmiştir.

[138] Sarantopoulos, G., s. 99, 109-110, 120.

6. BİR TELEKOMUNİKASYON FİRMASINDA KREDİ SKORLAMA UYGULAMASI

Bu çalışmada yer alan uygulama, CRISP-DM metodolojisine uygun olarak yapılmıştır. Aşağıda davranışsal skorlama modelinin veri madenciliği ile yapılması süreci, önerilen metodolojinin aşamalarına uygun olarak ve detaylı açıklamaları ile yer almaktadır. Veri madenciliği uygulamaları Clementine 12.0 ve SAS DMiner 5.4 programlarında hazırlanmıştır. Proje sürecinin tasarımı EK-2 de verilmiştir.

6. 1. İş Probleminin Tanımlanması ve Amacın Belirlenmesi

Mobil operatör şebekelerindeki riskli abonelerin erken uyarı sistemleri vasıtasıyla önceden belirlenerek gerekli önlemlerin alınması ve tüm faturalı abonelerin aktive oldukları andan itibaren risk takibinin yapılabilmesi amacıyla Davranışsal Skorlama altyapısının kurulması gerekmektedir. Amaç, şüpheli alacakların önceden tespit edilmesi yoluyla, zarar büyümeden müdahale yaparak engellemektir. Ancak, bu işlem büyük veri yığınları içerisinden nitelikli bilgiye kısa sürede ulaşarak mümkün olabilir. Ayrıca riskli abonelerin davranışlarını modelleyebilmek ve değişen ihtiyaçlara göre de güncelleyebilmek gerekmektedir.

İletişim ücretlerini önceden ödeyen kontörlü aboneler, bu çalışmada bahsedilen risk değerlendirmesine dâhil değildir. Bu çalışmada geliştirilen Davranışsal Skorlama Modeli, sadece faturalı abonelerin risklerini tespit etmek amacıyla kullanılabilir.

Bankacılık alanında olduğu gibi telekomünikasyon sektöründe de abonelerin risk takipleri, bu abonelerin ait oldukları belirli tarife yapıları veya farklı alt segmentler bazında değil doğrudan abone bazında yapılmalıdır. Aksi halde kurulacak değerleme sistemlerinin performansları çok düşük olacaktır. Bundan dolayı da aboneler bazında ayrı olarak veri oluşturmak gerekmektedir. İlgili veri tabanı sistemi bu gereksinimi yerine getirebilmelidir. Bu çalışmada, her kayıt farklı bir aboneliğe aittir.

Mobil telekomünikasyon sektöründe kullanılan risk değerlendirme sistemlerinde modelleme bazının abone olması gerektiği kesindir, ancak ‘abone’ nin tanımı netleştirilmelidir. Çünkü bir gerçek kişi kendi adına birden fazla kontörlü ve/ veya faturalı hat alabildiği gibi, aynı zamanda yöneticiliğini yaptığı veya sahip olduğu işletme için de pek çok hatta sahip olabilir.

Bu durumda, adına birden fazla hat aktivasyonu yapılan gerçek veya tüzel kişi 'Müşteri'; aktive edilen her bir hat da 'Abone' olarak tanımlanır. Mobil telekomünikasyon firmalarında, gerek sahtekârlık tespitini yapmak amacıyla kurulan sistemlerde gerekse de risk takibini yapmak amacıyla kurulan sistemlerde, analiz edilecek veri setinin 'Müşteri' mi yoksa 'Abone' mi olacağı durumdan duruma farklılık gösterebilmektedir. Bu çalışmada kullanılan veri seti abone bazında hazırlanmıştır, ancak abonenin ait olduğu müşterinin başka verileri de burada kullanılmaktadır.

Hemen tüm risk takip sistemlerinde abonenin Kredi Limitinin (KL) belirlenmesi gerekmektedir. Bankacılık sistemi için bu limit, müşterinin otomatik olarak daha fazla kredi alıp alamayacağını belirlerken, telekomünikasyon sektöründe de abonenin gerçekleştirebileceği iletişim hizmeti ve/ veya alabileceği cihaz tutar toplamlarını tespit etmek için kullanılabilir. Bu çalışmada kurulacak modellerin amacı, riskli olarak tanımlanan abonelerin önceden belirlenmesi olup herhangi bir KL ataması yapılmayacaktır.

Faturalı abonenin toplam riski, ödemediği faturalar ile air-time (henüz faturalaşmamış ancak bir sonraki fatura döneminde hesaplanacak iletişim hizmetleri) tutarının toplamından oluşmaktadır. Bu risk tutarının karşılaştırıldığı Kredi Limiti ise, abonenin şebekeye ilk geldiği anda atanan Statik Kredi Limiti ile daha sonraki davranışlarına göre atanan Dinamik Kredi Limitinin toplamından oluşur. Abonenin ödeme davranışlarına bağlı olarak Dinamik Kredi Limiti pozitif veya negatif olabilir. KL, minimum sıfır olabilir. Riskli olarak tanımlanan aboneler de Net Riske, yani Toplam Risk ile Kredi Limiti arasındaki farka göre belirlenir. Net risk, pozitif veya negatif değerler alabilir.

Toplam Risk = Ödenmemiş Faturalar Toplamı + Faturalaşmamış Air-time Tutarı
Kredi Limiti = Statik Kredi Limiti + Dinamik Kredi Limiti
Net Risk (Exposure) = Toplam Risk - Kredi Limiti

Abonenin net riski (exposure miktarı) anlık olarak takip edilemiyorsa günlük olarak izlenmelidir. Güncel riskin en doğru şekilde belirlenebilmesi için gerekli bileşenlerden toplam risk tutarı, sistemlerden kolaylıkla alınabilmesine karşın pek çok operatör, abonenin Kredi Limitini günlük olarak değil de aylık olarak hesaplayabilmektedir. Veri madenciliği tekniklerinin gelişmesi ile bu hesaplama öncelikle günlük yapılmalıdır. Operasyonel olarak VM modellerinin üzerinde çalıştığı datamartta yer alan değişkenler günlük olarak

güncellenmeli ve risk modelleri günlük çalışmalıdır. Gerekli altyapı oluşturulduktan sonra da gerçek zamanlı olarak hesaplama yapmak, abone risk takibinde en etkin yöntemdir.

6.2. Verilerin Seçilmesi ve Analize Hazırlanması

Kredi skorlama sisteminin en önemli parçası çoğu zaman, kullanılan istatistiksel model seçimi değildir. Analize giren verinin doğru oluşturulması, içeriğinin zenginliği, kapsamının genişliği, uygun ölçekle tanımlanmış olması gibi veri kalitesini belirleyen faktörler başarılı sonuç almanın belki de en önemli bileşenleridir.

Gelişmiş skorlama modellerinin kullanımının önündeki en büyük engel, modeller için gerekli olan veriye istenen doğrulukta, hızlı ve güvenilir bir şekilde ulaşılamamasıdır. Çoğu zaman, kullanılması gereken bazı verilerin hiç elde edilememesi de söz konusudur. Örneğin ekonomi yazınında Tüketim Fonksiyonunda girdi olarak kullanılan 'Servet' ve 'Gelir' verileri, ülkemizde ulaşılması en zor olan verilerden ikisidir. Abone kredibilitesinin değerlendirilmesinde kullanılması gereken; aylık kazanç tutarı, servet miktarı ve dağılımı (menkul ve gayrimenkul olarak), toplam harcama tutarı, diğer hane halkı geliri, kurumlara olan borç ve alacak tutarları da bu tür verilerdendir. Ancak bu verilerin alınamaması veya alınmasının kolay olmaması nedeniyle gerçekte ikincil derecede etkisi olan veriler kullanılarak modellerin oluşturulması çok yaygındır. Benzer şekilde Türkiye'de telekomünikasyon sektörünün erişebileceği kredi büro verisi mevcut olmadığından modellemelerde girdi olarak kullanılamamıştır. Çalışmada dışarıdan alınan herhangi bir veri bulunmamaktadır.

Modellemede kullanılacak veriler, gelecekte kredi başvurusu yapacak ve yine gelecekte şebekeyi kullanacak aboneleri temsil etmelidir. Böylece skorlama modelinin beklenen faydası maksimize edilmiş olur. Analiz edilen veri setinde yer alan abonelerin seçildiği coğrafi bölge, gelecekteki muhtemel aboneleri temsil etmeyi sağlayabilecek genişliğe ve çeşitliliğe sahiptir.

Örneklem seçiminde dikkat edilmesi gereken diğer bir nokta da örneklem büyüklüğünün ne kadar olacağı ve iyi-kötü kredilerin hangi unsurlara göre ayrıştırılacağıdır. Örneklemdeki iyi müşteri-kötü müşteri oranı eşit mi olmalıdır, yoksa ana kütlede olduğu şekliyle mi temsil edilmelidir? Ana kütledeki oranına göre iyi-kötü oranı belirlendiğinde kötü kredilerin açıklanmasına yetecek kadar veri örneklemde bulunmayacağından dolayı genelde bu oran

50:50 olarak kabul edilir. Örneklemde iyi-kötü değişkenlerinin dağılımları aynı değilse buna izin verecek örneklemi elde etmek için sonuçların düzeltilmesi gerekir. Uygulamada kullanılan veri setinde bu oranı, literatür kısmında üzerinde en fazla çalışma yapılan Alman kredi veri setinde olduğu gibi, %70 iyi - %30 kötü olarak alınmıştır. Veri setinde toplam 10.000 gözlem bulunmaktadır. Bu gözlemlerin 7.000'i risksiz, 3.000'i de riskli gözlemlerden oluşmaktadır.

Modele dâhil olacak veriler, bağımlı değişken değeri net olarak belirlenmiş olan iyi veya kötü aboneleri kapsar. Tanım gereği iyi veya kötü gruplarından herhangi bir tanesine girmediği için gri bölgede kalan aboneler ile hesaplarında hiç hareket olmadığı için değerlendirilemeyen aboneler modellemeye dahil edilmezler. Bunların yanı sıra, işletme politikaları gereği kredi verilmeyen aboneler ile skorlama kapsamı dışında tutulan abonelikler de analiz veri setine dahil edilmezler. Çalışmada kullanılan veri seti, bu kriterlere uygun olarak oluşturulmuştur.

Sonuçları gerçek hayata uygulanacak olan modellerde kullanılacak veri, güncel konuyla ilgili olacak kadar yeni, ancak hedef değişkeni oluşturan sonucu ortaya koyacak kadar da eski olmalıdır. Örneğin 10 yıl öncesine ait veri modellemede kullanıldığı zaman tüm kötü aboneler belirlidir, ancak verinin içeriği güncel iş ve ekonomi şartlarını yansıtmaktan uzaktır. Birkaç ay önceki verinin kullanılmak istenmesi durumunda da gerçekte iyi ve kötü abonenin hangisi olduğu net değildir. Dolayısıyla güncel zamandan altı ay ile iki yıl önceki döneme ait verinin model geliştirmede kullanılması uygundur. Ancak bu çalışmada, sektörün rekabet şartları göz önüne alındığı için 2005 yılına ait veri kullanılmıştır.

Model geliştirmede kullanılacak veri, mevsimsel etkiden arındırılarak kullanılırsa daha uygun sonuçlar ortaya çıkacaktır. Mevsimsel etkilerin varlığı durumunda farklı aylara ait veriler, analize dâhil edilmelidir. Telekomünikasyon sektöründe belirgin bir mevsimsel etkinin varlığından söz edilemeyeceği için aşağıda detayları aktarılan çalışmada kullanılan veri, statüsü Aktif (çağrı yapabilen) olan abonelerin seçilen güne ait değerlerini içermektedir.

Modelde kullanılacak verinin yetersiz olması durumunda da peşpeşe olan belirli sayıdaki aya ait veri birleştirilerek kullanılabilir. Ancak bankacılık ve hızlı tüketim gibi telekomünikasyon sektöründe de yeterli miktarda veriyi elde etmek zor değildir. Bu çalışmada kullanılan veri seti bir anlık kesit veriden oluşmaktadır.

Müşterilerden başvuru esnasında titiz olmayan süreçler ile alınan ve daha sonra düzeltilmesi çok zor olan demografik veriler tüm telekomünikasyon, sigorta, perakende ve hatta bankacılık sektöründe kirli veri olarak yer almaktadır. 2008 yılında Telekomünikasyon Kurumu (TK) tarafından yapılan girişim ile mobil operatörlerin tüm abonelerinin TC Kimlik Numaralarının MERNİS'ten sorgulanarak, operatörlerin veri tabanlarında yer alan demografik verilerin temizlenmesi mümkün olmuştur. Böylece operatörler, abonelere ait demografik bilgileri kaynağından doğrudan, temiz ve doğru olarak alma imkânına sahip olmuşlardır.

Davranışsal Skorlama Modellerinde kullanılan gözlemlere ait veriler belirli bir zaman boyutunu kapsamalıdır. Özellikle son üç ay, son altı ay ve son bir yıla ait verilerden oluşturulan değişkenlerin açıklama kabiliyetleri çok yüksektir. Çalışmada kullanılan veriler içerisinde bu tür değişkenler de yer almaktadır.

Veri seti GSM numarası - abonelik - bazında oluşturulmuştur. Sadece bir adet faturalı hattı olanlar değil, birden fazla kardeş hattı olanlar da veri setinde yer almaktadırlar. Çünkü bazı kişiler birden fazla hat kullanmaktadır, bazıları da doğrudan kendisi kullanmasa bile çıkan faturaları ödemektedir. Hatların gerçek kullanıcısı sistemde tanımlı olan kişi olmayabilir. Bundan dolayı aynı müşterinin birden fazla hattı analiz tablosunda yer almaktadır ve ayrıca değerlendirilmiştir. Bu tercih, gerçek kullanıcının hat sahipliğinden farklı olabilmesi nedeniyle ana kütlenin temsilinde yanlı davranılmaması için yapılmıştır.

6.2.1. Bağımlı Değişken Seçimi

Hedef değişkeninin doğru seçilmesi çok önemlidir, çünkü skor kartın veya skorlama modelinin gerçekte tespit ettiği aboneler bu ölçüt ile belirlenir. Hedef değerin tanımı, skor kart veya skorlama modeli tarafından gerçekte çözümlenen iş probleminin sonucunu göstermelidir.

Kredi skorlama modellerinde kullanılan bağımlı değişken, iki sonuçlu (iyi/ kötü) bir kategorik değişkendir. Kredi sektöründe, bu ayırımı yapabilmek için belirlenmesi gereken kötü abonenin, üzerinde anlaşılmış net bir tanımı bulunmamaktadır. Genel olarak, şüpheli alacak bırakan abonenin kötü abone olduğu kabul edilir, ancak herkes tarafından kabul edilen bir şüpheli alacak tanımı da mevcut değildir. Bunun da ötesinde tüm aboneler 'iyi' veya 'kötü' tanımlaması içerisine girmeyebilir veya bazı aboneleri 'kötü' olarak tanımlamak diger tüm

abonelerin iyi oldugu anlamına gelmez. Birincisi yeteri kadar gözlem bulunmadığı – örneğin faturası çıkmayanlar - için bağımlı değişken tanımlanamayabilir. İkinci olarak da gerçekte ne iyi ne de kötü olarak tanımlanabilecek durumlar bulunabilir. Örneğin ödemesini belirli bir süre geciktiren ama daha sonra ödeyen, veya arka arkaya hiç gecikmiş faturası bulunmayan ama zamanında da ödemeyen aboneler 'gri' olarak tanımlanıp modellemeye alınmayabilirler.

Farklı birçok endüstri ve kurum tarafından kullanılan ve EK – 1 de yer alan kredi skorlama modelleri ile ilgili özet tanıtımlar mevcuttur. Bu modeller ile herhangi bir kredi taksitinin 90 gün veya üzerinde ödenip ödenmeyeceği gibi kısa süreli veya abonelerin değerleme anından itibaren ilk 12 ayda iflas edip etmeyecekleri gibi daha uzun süreli sonuçlar tahmin edilmektedir. Modellerde kullanılacak gözlemlerin atanacakları hedef gruplar da bu ifadelere göre belirlenmiştir.

Borcunu veya kredi taksitini son ödeme tarihinden itibaren belirli bir süre geciktiren abone, riskli (kötü) abone olarak tanımlanır. Tanımda yer alan, 'belirli bir süre' nin ne olacağı; içinde bulunulan sektöre, sektörde yer alan farklı firmaların risk algılamasına ve ekonomik konjonktüre göre değişmekle birlikte, genelde 30 gün, 60 gün veya 90 gün gibi gecikmeler kötü abonenin belirlenmesinde ölçüt olarak kullanılmaktadır. Bunun yanı sıra belirli bir taksit adedi içerisinde, peş peşe geç ödenen veya ödemenin geciktiği durumların belirli bir sayıya ulaşması durumunda da aboneler kötü olarak tanımlanabilmektedir.

Kredi skorlama endüstrisinde kullanılan uygulamalar arasında sıkça kullanılan 90 günlük süre, bu çalışmadaki hedef değişkeninin belirlenmesinde de kullanılmıştır. Değerleme anından itibaren belirli bir sayıdaki - örneğin üç - kredi taksitinden herhangi bir tanesinin son ödeme zamanını 90 gün geçiren abone kötü, diğerleri iyi olarak tanımlanabilir. Ancak bu çalışmada kullanılan hedef değişkeni şu şekilde tanımlanmıştır: Değerleme anında çıkan ancak henüz ödenmemiş faturasını veya faturalaşmamış air-time tutarının oluşturacağı faturasını, faturaların son ödeme tarihlerinden itibaren 90 gün içerisinde ödemeyen abone kötü, diğerleri iyi olarak tanımlanmıştır. İlgili faturalar 90'ıncı günden sonra ödenmiş olsa bile abone riskli olarak değerlendirilmiştir.

6.2.2. Bağımsız Değişken Seçimi

Skorlama modelinde yer alacak tüm gözlemler gibi, herhangi bir mobil telekomünikasyon abonesini tanımlayan ilk veriler de doğal olarak demografik verilerdir. Veri setinde aboneye ait 10 adet demografik değişken yer almaktadır. Bunların bir kısmı sadece aboneyi belirlemek için veri setinde yer aldığından modellemeye katılmazlar, ancak genel olarak davranışsal skorlama modellerinde demografik değişkenlerin çok küçük bir kısmı bazen hiçbirisi anlamlı katsayılara sahip olmadıkları için yer almazlar.

Abonenin riskini belirleyen en önemli veri kaynağı; 'Faturaların kaç defa geç ödendiği?', 'Geç ödenen faturaların ortalama kaç gün geç ödendiği?', 'Ödenen maksimum fatura tutarı', 'Ödenen toplam fatura adedi' gibi ödeme performansına ait değişkenlerdir. Burada kullanılan modellerde de ödeme performansını gösteren değişkenler, modeller tarafından belirlenen en anlamlı değişkenler olmuşlardır.

Telekomünikasyon abonesine ait üçüncü önemli veri kaynağı türü; işlemsel verileri kapsayan değişkenlerdir. Burada bahsedilen işlemler, 'Tarife yapısını değiştirme sayısı', 'Hattını Roaminge (hattın yurtdışında kullanılabilmesi) açık tutma süresi' gibi abone tarafından yapılan işlemler olabileceği gibi; borcunu ödemeyen aboneye operatör tarafından otomatik olarak uygulanan 'Arama yapmaya veya arama almaya kısıtlanma' işlemlerini de kapsamaktadır. Riskli aboneleri tanımlayan önemli değişkenlerden aramaya ve aranmaya kısıtlanma adetleri uygulama veri setinde üç değişken olarak yer almaktadır.

Bir diğer önemli veri kaynağı, tüm mobil iletişim firmalarının gelirlerini kazandıkları arama verisidir. CDR (Call Detail Records) verileri, şebekede santraller tarafından sürekli üretilmekte olup günde milyarlarca kayda ulaşmaktadır. CDR kayıtları kullanılarak; 'Şebeke içi görüşme süresi', 'Roaming yapılan ülke adedi', 'Alınan SMS adedi', 'Katma değerli servis aramalarının tutarı', 'Görüşmenin yapıldığı saat dilimi', 'Kullanılan farklı hücresel şebeke birimi' gibi pek çok farklı veriye ulaşılabilir. Çalışmamızda bu tür kullanım verileri yer almamaktadır.

Sadece ön ödemeli aboneler için üretilen kontör kullanım verileri de mobil telekomünikasyon sektöründe kullanılan çok önemli bir veri kaynağıdır. Hâlihazırda Türkiye'deki her üç operatörün de portföylerinin en büyük kısmını ön ödemeli aboneler oluşturmaktadır. Faturalı

bir hattın kardeş kontörlü hattının kontör kullanma alışkanlığı veya ön ödemeli iken faturalı hatta geçen bir abonenin kontör yükleme davranışları, bu abonelerin risklerinin belirlenmesi için önemli bir kaynaktır. Bizim çalışmamızda kontör kullanım verileri de yer almamaktadır.

Yukarıda bahsedilen demografik veriler hariç üretilen tüm bu veriler, abone bazında bir araya getirilerek kullanılabilir. Her biri farklı sistemler tarafından derlenip farklı formatlarda tutulan ve birbirlerinden farklı büyüklüklere sahip olan bu verilerin çok ciddi ve titiz bir çalışmayla kodlanarak bir araya getirilmesi gerekmektedir. Bunu sağlayabilecek en önemli sistem ise gelişmiş veri ambarlarıdır. Çalışmada kullanılan veri seti, veri ambarından çekilerek kullanılmıştır.

Değişken seçimi, farklı kredi skorlama tekniklerini kullanarak en iyi performans gösteren modelin seçiminden önce yapılması gereken, çok kritik ve zorlu ama aynı zamanda da fırsatlar yaratan bir konudur. Bilindiği gibi farklı değişken seçim metotları aynı veri seti üzerinde farklı sonuçlar vermektedir[139]. Veri setinin çok büyük olduğu, binlerle ölçülen değişkenlerin yer aldığı veri setleri ile modelleme yapılması durumunda; genel olarak, öncelikle değişken azaltma olarak da isimlendirilen değişken seçme işlemi uygulanır. Değişken seçme süreci de yine karar ağaçları başta olmak üzere farklı modeller vasıtasıyla veya veri madenciliği ürünleri tarafından otomatik olarak değişken seçmek amacıyla üretilen fonksiyonlar yardımıyla yapılır. Bizim çalışmamızda kullanılan değişken adedi çalışmaya alınmadan önce uzman kişiler tarafından sınırlandırılmasına rağmen, mevcut değişkenler içerisinde hedef değişken ile olan ilişkisine göre farklı metotlar kullanılarak, modellemeye girecek değişken sayısı azaltılmıştır.

Modellemeye alınacak gözlemlere ait veriler, ödeme alışkanlığının iyi ya da kötü olduğunu yansıtabilecek kadar yeterli bilgiyi içermelidir. Uygulamada kullanılan veri seti, on adedi demografik değişken olmak üzere toplam 64 değişkenden oluşmaktadır. Bu değişkenlerden ilki olan Sıra_No, gözlemin anahtar ID alanını oluşturmaktadır. Bu değişken ile birlikte gözlemi tanımlamak üzere kullanılan toplam altı ek alan mevcuttur ve bunlar modellere dahil edilmezler. Bu değişkenlere ek olarak Doğum tarihi (Birth_Date), Doğum yeri (Birth_Place) ve Aktivasyon tarihi (Activation_Date) ile hedef değişkenini hariç tutarsak veri setinde

[139] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan., s. 4737.

toplam 54 adet açıklayıcı değişken mevcuttur. Hazırlanan ilk veri setini oluşturan tüm değişkenler ve tanımları aşağıda yer almaktadır:

Tablo 9 : Veri Setini Oluşturan Değişken İsimleri ve Anlamları

Sıra No	Değişken Adı	Değişkenin Tanımı	Değişken Tipi
1	Sıra_No	Gözlem sıra numarası	ID
2	National_ID	TC Kimlik numarası	ID
3	Customer_ID	Müşteri numarası	ID
4	GSM_ID	GSM numarası	ID
5	Name	Ad	ID
6	Surname	Soyad	ID
7	Birth_Date	Doğum tarihi	Oransal
8	Birth_Place	Doğum yeri	Kategorik
9	Activation_Date	Aktivasyon tarihi	Oransal
10	Address_City	Fatura adresinin il plaka kodu	Kategorik
11	Gender	Cinsiyet	Dikotom
12	Marital_Status	Medeni durum	Kategorik
13	Biologic_Age	Abonenin yıl olarak biyolojik yaşı (min=18)	Oransal
14	Education	Eğitim seviyesi	Kategorik
15	Tariff_Plan	Güncel tarife paketi - gruplanmış	Kategorik
16	Home_Phone_FLG	Sistemde kayıtlı ev telefonu varsa ve anlamlı ise 1, yoksa 0	Dikotom
17	Work_Phone_FLG	Sistemde kayıtlı iş telefonu varsa ve anlamlı ise 1, yoksa 0	Dikotom
18	Other_Phone_FLG	Sistemde kayıtlı diğer telefonu varsa ve anlamlı ise 1, yoksa 0	Dikotom
19	Phone_CNT	Müşterinin sistemde kayıtlı toplam irtibat telefon adedi	Oransal
20	GPRS_FLG	GPRS servisi açık ise 1, kapalı ise 0	Dikotom
21	International_FLG	Yurtiçinden yurtdışını arama servisi aktif ise 1, pasif ise 0	Dikotom
22	Automatic_Payment_FLG	Çıkan faturaları için otomatik ödeme talimatı var ise 1, yok ise 0	Dikotom
23	Roaming_FLG	Yurtdışında iken görüşme yapabilme servisi aktif ise 1, pasif ise 0	Dikotom
24	Active_Line_CNT	Müşteriye ait aktif hat adedi	Oransal
25	Inactive_Line_CNT	Müşteriye ait inaktif hat adedi	Oransal
26	Network_Age	Abonelik yaşı (ay)	Oransal
27	Black_List_FLG	Müşteri operatörün kara listesinde ise 1, değilse 0	Dikotom
28	Invoice_L3_AVR	Son 3 fatura dönemindeki çıkan ortalama fatura tutarı	Oransal
29	Invoice_L6_AVR	Son 6 fatura dönemindeki çıkan ortalama fatura tutarı	Oransal
30	Invoice_M0	Son ay çıkan fatura tutarı	Oransal
31	Invoice_M1	Sondan 1 önceki fatura dönemine ait fatura tutarı	Oransal
32	Invoice_M2	Sondan 2 önceki fatura dönemine ait fatura tutarı	Oransal
33	Invoice_M3	Sondan 3 önceki fatura dönemine ait fatura tutarı	Oransal
34	Invoice_M4	Sondan 4 önceki fatura dönemine ait fatura tutarı	Oransal
35	Invoice_M5	Sondan 5 önceki fatura dönemine ait fatura tutarı	Oransal
36	Invoice_Paid_AMT	Ödenen tüm faturaların toplam tutarı	Oransal
37	Invoice_Paid_CNT	Ödenen tüm faturaların toplam adedi	Oransal
38	Invoice_Paid_AVG	Ödenen tüm faturaların tutarlarının aylık ortalaması	Oransal
39	Invoice_Paid_Max_AMT	Tüm abonelik süresince ödenen maksimum fatura tutarı	Oransal
40	Invoice_National_L6_AVR	Son 6 fatura dönemindeki çıkan ortalama ulusal görüşme tutarı	Oransal
41	Invoice_International_L6_AVR	Son 6 fatura dönemindeki çıkan ortalama uluslararası görüşme tutarı	Oransal
42	Invoice_Roaming_L6_AVR	Son 6 fatura dönemindeki çıkan ortalama roaming görüşme tutarı	Oransal
43	Invoice_SMS_L6_AVR	Son 6 fatura dönemindeki çıkan ortalama SMS tutarı	Oransal
44	Invoice_Voice_L6_AVR	Son 6 fatura dönemindeki çıkan ortalama sesli görüşme tutarı	Oransal

45	Overdue_CNT	Aboneye çıkan tüm faturalar içerisinde, faturanın son ödeme tarihinden sonra ödenen fatura adedi.	Oransal
46	Overdue_Invoice_AVR	Ödenen tüm faturaların son ödeme tarihlerine göre ortalama geç ödenme gün sayısı	Oransal
47	Overdue_AVR	Geç ödenen faturaların son ödeme tarihlerine göre ortalama geç ödenme gün sayısı	Oransal
48	Overdue_L6_AVR	Son 6 fatura döneminde çıkan ve geç ödenen faturalarının ortalama geç ödenme gün sayısı	Oransal
49	Overdue_L6_FLG	Son 6 fatura dönemi içerisinde son ödeme tarihi geçmiş faturasının olup olmadığı. En az bir faturası son ödeme tarihinden sonra ödendi ise 1, değilse 0	Dikotom
50	Overdue_Max_L6_AMT	Son 6 fatura dönemine ait faturalardan ödemesi yapılanlar içinde, son ödeme tarihinden itibaren en geç ödenen faturanın tutarı	Oransal
51	Overdue_Sequential_L12_CNT	Son 12 ay içerisinde çıkan faturalardan arka arkaya son ödeme tarihinden sonra ödenen maksimum fatura adedi	Oransal
52	Pay_Type_Distinct_CNT	Son 6 ay içerisinde yapılan farklı ödeme yöntemlerinin adedi	Oransal
53	Pay_Type_Preferred_Method	Son 6 ay içerisinde en çok tercih ettiği ödeme tipi	Kategorik
54	PCT_Invoice	Son iki ayda gerçekleştirilen ortalama fatura tutarının, ondan önceki 4 aylık ortalamaya göre değişim yüzdesi	Oransal
55	PROPN_Invoice	Son çıkan faturanın, son altı ayda çıkan faturalarının ortalamasına oranı	Oransal
56	MO_Barred_CNT	Abonelik süresi boyunca çıkan faturaların zamanında ödenmemesinden kaynaklanan dışarıyı aramaya kısıtlanma adedi	Oransal
57	MO_Barred_L6_CNT	Son 6 ayda çıkan faturaların zamanında ödenmemesinden kaynaklanan dışarıyı aramaya kısıtlanma adedi	Oransal
58	MT_Barred_CNT	Abonelik süresi boyunca çıkan faturaların zamanında ödenmemesinden kaynaklanan dışarıdan aranmaya kısıtlanma adedi	Oransal
59	Risk_Unbilled_AMT	Air-time tutar toplamı	Oransal
60	Risk_Unpaid_AMT	Ödenmeyen fatura tutar toplamı	Oransal
61	Risk_AMT	Abonenin toplam risk tutarı	Oransal
62	Risk_RT	Toplam riskin, ödenen toplam fatura tutarına oranı	Oransal
63	Usage_Risk_RT	Riskin içerisindeki air-time oranı	Oransal
64	Target_IDD90	Riski oluşturan ödenmemiş fatura veya air-time dan kaynaklanan faturanın son ödeme tarihinden itibaren 90 günden fazla gecikmesi olup olmadığı. İlgili ay çıkan fatura 90.cı günden sonra ödenmiş olsa bile riskli olarak kabul edilir. (Hedef Değişken)	Dikotom

6.2.3. Değişkenlerin Tanımlanması

Veri madenciliği sürecinde kullanılacak her değişken, birinci aşamada, kapsadığı uygun veri tipine göre sisteme tanımlanmıştır. Açıklayıcı değişkenlerden beş tanesi kategorik, 10 tanesi dikotom, 39 tanesi de oransal ölçekle ölçülmüş verilerden oluşmaktadır. Hedef değişken de dikotom yapıdadır. Modellerde kullanılan kategorik verilerin toplu şekilde hazırlanmış frekans tabloları aşağıda yer almaktadır:

Tablo 10 : Modellerde Kullanılan Kategorik Değişkenlerin Frekans Dağılımları

Eğitim (Education)	Frekans	Yüzde (%)	Kümülâtif Frekans	Kümülâtif Yüzde (%)
Missing	4.621	46,21	4.621	46,21
İlkokul	1.161	11,61	5.782	57,82
OrtaOkul	799	7,99	6.581	65,81
Lise	2.976	29,76	9.557	95,57
Yüksekokul	167	1,67	9.724	97,24
Üniversite	239	2,39	9.963	99,63
MasterDoktora	37	0,37	10.000	100,00
Cinsiyet (Gender)	**Frekans**	**Yüzde (%)**	**Kümülâtif Frekans**	**Kümülâtif Yüzde (%)**
E (Erkek)	8.221	82,21	8.221	82,21
K (Kadın)	1.779	17,79	10.000	100
Medeni Durum (Marital_Status)	**Frekans**	**Yüzde (%)**	**Kümülâtif Frekans**	**Kümülâtif Yüzde (%)**
Bekâr	1.843	18,43	1.843	18,43
Boşanmış	391	3,91	2.234	22,34
Bilinmeyen	527	5,27	2.761	27,61
Dul	154	1,54	2.915	29,15
Evli	7.085	70,85	10.000	100
Tarife Yapısı (Tariff_Plan)	**Frekans**	**Yüzde (%)**	**Kümülâtif Frekans**	**Kümülâtif Yüzde (%)**
TP1	2.178	21,78	2.178	21,78
TP2	3.536	35,36	5.714	57,14
TP3	2.034	20,34	7.748	77,48
TP4	74	0,74	7.822	78,22
TP5	705	7,05	8.527	85,27
TP6	787	7,87	9.314	93,14
TP7	657	6,57	9.971	99,71
TP8	29	0,29	10.000	100
Tercih Edilen Ödeme Tipi (Pay_Type_Preferred_Method	**Frekans**	**Yüzde (%)**	**Kümülâtif Frekans**	**Kümülâtif Yüzde (%)**
Pay_Type_1	2.672	26,72	2.672	26,72
Pay_Type_2	4.173	41,73	6.845	68,45
Pay_Type_3	961	9,61	7.806	78,06
Pay_Type_4	2.194	21,94	10.000	100

Demografik veriler içerisinde önemli bir parametre olan eğitim (Education) değişkeninin tüm set içerisinde %46'sı eksik veridir. Eğitim alanı boş olan bu gözlemler için, boş gözlem oranının yüksek olması dolayısıyla imputasyon işlemi yapılmamış, bu şekilde missing olarak modellerde kullanılmıştır.

Cinsiyet (Gender) değişkeni açısından incelendiğinde, model veri setinin %18'i kadın, %82'si erkeklerden oluşmaktadır. Daha önce bahsedilen hat sahipliği ile hat kullanıcılığının aynı

olmayabileceğinin ipuçlarından bir tanesi de bu frekans dağılımıdır. Ancak davranış modellerinde demografik verilerin öneminin fazla olmaması, ödeme performansına dayalı yeteri kadar değişken olması ve değişken seçimi aşamasında elenmesi nedeniyle analize dahil edilmemiştir.

Abonelerin sadece %5'inin medeni durumu (Marital_Status) belirli olmamakla birlikte, modellerde anlamlı bir açıklayıcı değişken olarak yer almamıştır. Banka kredileri ile yapılan skorlama çalışmalarında önemli olduğu aktarılan bu verinin burada anlamlı çıkmamış olması, davranış skorlamada demografik bilginin öneminin az olmasının yanı sıra skorlamaya konu olan tutarın büyüklüğü ile de ilgili olabilir.

Abonelerin kullandıkları tarife yapıları (Tariff_Plan), yani ücretlendirme yöntemlerinin dağılımlarına bakıldığında dağılımın üniform olmadığını görüyoruz. Bunun başlıca sebebi, kullanıcıların farklı ücretlendirme yapıları arasındaki farklılığa karşı duyarlı olmalarıdır. Mobil hat kullanıcılarının tercih edebilecekleri uygun tarifeler, operatörler tarafından çok rekabetçi olarak çıkarılabilmekte ve tarife değişikliği SMS gibi yöntemlerle ücretsiz olarak yapılabilmektedir. Aynı şekilde mevcut hattı olan kullanıcıların tarifenin özelliklerinden faydalanmak için ek hat olarak kolayca yeni hat alıp kullanması da söz konusu olmaktadır.

Ödeme tercihlerinin (Pay_Type_Preferred_Method) abonenin risk açısından tanımlanmasında önemli bir yeri vardır. Örneğin otomatik ödeme talimatı olan abonenin faturalarını ödeme olasılığı her zaman diğer abonelik grubuna göre yüksektir. Abonenin ödemelerini, abonelik merkezlerinden kredi kartı ile mi, yoksa nakit mi ödemeyi tercih ettiği değerlendirilmiş olmaktadır. Burada sadece abonenin ödeme niyeti değil, aynı zamanda unutmadan kaynaklanan fatura ödenmemesi durumları da ortadan kalkmaktadır.

Tablo 11 : Modellerde Kullanılan Oransal Değişkenlerin Dağılım İstatistikleri

Değişken Adı	Ortalama	Std Sapma	Minimum	Maksimum	% 1.ci	% 5.ci	Alt Kartil	Üst Kartil	% 95.ci	% 99.cu
Biologic_Age	40,4	11,7	18	91	21	25	31	48	62	73
Network_Age	57,1	44,5	6	134	6	7	15	105	123	131,5
Home_Phone_FLG	0,5	0,5	0	1	0	0	0	1	1	1
Work_Phone_FLG	0,2	0,4	0	1	0	0	0	0	1	1
Other_Phone_FLG	0,7	0,5	0	1	0	0	0	1	1	1
Phone_CNT	1,3	0,7	0	3	0	0	1	2	3	3
Black_List_FLG	0,1	0,3	0	1	0	0	0	0	1	1
GPRS_FLG	0,2	0,4	0	1	0	0	0	0	1	1
International_FLG	0,2	0,4	0	1	0	0	0	0	1	1
Automatic_Payment_FLG	0,1	0,3	0	1	0	0	0	0	1	1
Roaming_FLG	0,1	0,3	0	1	0	0	0	0	1	1
Active_Line_CNT	1,6	1,1	1	14	1	1	1	2	4	6
Inactive_Line_CNT	0,4	1,0	0	11	0	0	0	0	2	5
Invoice_L3_AVR	38,4	49,8	0,0	2162,9	2,1	3,9	12,6	46,8	116,1	220,6
Invoice_L6_AVR	36,7	46,0	0,1	2004,2	2,7	5,1	13,2	44,8	106,1	199,7
Invoice_M0	40,1	66,7	0,0	2799,2	0,0	0,0	4,9	49,5	137,4	273,8
Invoice_M1	35,1	52,7	0,0	2037,7	0,0	0,0	8,7	43,5	115,3	225,8
Invoice_M2	34,6	48,6	0,0	1651,8	0,3	2,7	9,3	41,7	111,2	220,1
Invoice_M3	33,5	55,1	0,0	2769,1	0,0	0,0	9,3	40,4	107,6	207,2
Invoice_M4	34,2	63,5	0,0	3590,6	0,0	0,0	9,2	40,1	110,4	224,2
Invoice_M5	33,1	48,6	0,0	1142,6	0,9	2,9	9,5	38,6	104,3	217,3
Invoice_Paid_AMT	1521,2	2015,9	4,1	35511,8	53,6	95,8	315,8	1997,4	4981,0	9552,7
Invoice_Paid_AVG	30,8	32,0	1,1	1014,7	4,8	7,1	13,7	37,1	81,3	150,5
Invoice_Paid_CNT	53,8	43,4	1,0	125,0	5,0	6,0	13,0	101,0	122,0	123,0
Invoice_Paid_Max_AMT	101,9	127,1	2,7	3590,6	8,2	15,0	37,8	121,8	301,3	572,2
Invoice_National_L6_AVR	20,5	27,1	0,0	536,3	0,0	0,0	4,5	26,4	66,4	123,4
Invoice_International_L6_AVR	0,6	5,7	0,0	234,2	0,0	0,0	0,0	0,0	1,4	15,4
Invoice_Roaming_L6_AVR	1,0	15,7	0,0	1209,3	0,0	0,0	0,0	0,0	0,0	20,4
Invoice_SMS_L6_AVR	1,8	4,7	0,0	95,8	0,0	0,0	0,0	1,6	8,1	22,8
Invoice_Voice_L6_AVR	20,1	32,1	0,0	1648,8	0,0	0,0	4,1	25,0	65,5	126,0
Overdue_CNT	15,7	17,9	0,0	112,0	0,0	0,0	4,0	21,0	56,0	78,0
Overdue_Invoice_AVR	6,6	15,9	0,0	582,2	0,0	0,0	0,4	7,5	24,4	56,4
Overdue_AVR	12,5	29,2	0,0	1571,8	0,0	0,0	3,5	14,3	36,5	79,7
Overdue_L6_AVR	8,7	13,5	0,0	152,0	0,0	0,0	0,0	11,0	34,0	64,5
Overdue_L6_FLG	0,66	0,47	0	1	0,0	0,0	0,0	1,0	1,0	1,0
Overdue_Max_L6_AMT	37,7	72,8	0,0	2769,1	0,0	0,0	0,0	48,4	143,7	284,1
Overdue_Sequential_L12_CNT	3,1	3,1	0,0	12,0	0,0	0,0	1,0	5,0	10,0	11,0
Pay_Type_Distinct_CNT	1,44	0,59	1	4	1	1	1	2	2	3
PCT_Invoice	57,3	2779,1	-100,0	275100	-100	-100	-45	43	172	626
PROPN_Invoice	0,9	0,7	0,0	5,1	0,0	0,0	0,2	1,3	2,1	3,0
MO_Barred_CNT	4,3	6,6	0	52,0	0,0	0,0	0,0	5,0	19,0	31,0
MO_Barred_L6_CNT	1,1	1,6	0	6,0	0,0	0,0	0,0	2,0	5,0	6,0
MT_Barred_CNT	0,2	0,6	0	9,0	0,0	0,0	0,0	0,0	1,0	3,0
Risk_Unbilled_AMT	29,1	57,1	0	3896,7	1,0	2,1	7,3	33,6	98,3	199,6
Risk_Unpaid_AMT	35,9	66,1	0	2799,2	0,0	0,0	0,0	45,3	132,5	267,6
Risk_AMT	65,0	108,5	0	4837,8	1,8	2,4	10,6	78,9	220,4	438,9
Risk_RT	0,07	0,09	0	1	0,00	0,00	0,02	0,09	0,26	0,45
Usage_Risk_RT	0,60	0,32	0	1	0,06	0,16	0,36	1,00	1,00	1,00
Target_IDD90	0,30	0,46	0	1	0	0	0	1	1	1

Çalışmada kullanılan gözlemlerin sahip oldukları verilerin doluluk oranı da çok yüksektir. Oransal verilerin tamamı dolu olup, hiçbir tanesinin içerisinde boş (missing) alan bulunmamaktadır. Bu nedenle herhangi bir değişkene imputasyon yapılmasına gerek duyulmamıştır.

Davranışsal skorlama modelleri olması nedeniyle abonelik yaşı minimum altı ay olan aboneler analize dahil edilmiştir. Aktif ve aktif olmayan hat adedi toplamı 20'nin üzerinde olan aboneler de, iş amaçlı olarak kullanılan hatların bireysel türden aktive edildiği varsayımıyla analize katılmamıştır.

Son altı ayda herhangi bir nedenle hiç fatura üretmemiş aboneler (aktivasyonu yaptıktan itibaren hiç CDR gerçekleşmediği için fatura çıkmayan veya son altı ay içerisinde kısıtlanmış oldukları için fatura çıkmayan ve çok kısa bir süre önce hatlarını tekrar kullanıma sokan aboneler gibi) analiz veri setine alınmamıştır. Aynı şekilde son altı ay içerisinde herhangi bir faturası çıkmış bile olsa hiçbir faturası ödenmemiş olan aboneler de analize dahil edilmemiştir.

6.2.4. Değişken Eleme Aşaması

Daha sonra modellerde kullanılacak değişkenlerin seçimi farklı Karar Ağaçları ve Lojistik Regresyon modelleri kullanılarak yapılmıştır. Değişken azaltılması olarak da isimlendirilen bu aşama, veri madenciliği sürecini daha etkinleştirmek için çok sık kullanılır. Özellikle açıklayıcı değişken sayısının yüzlerce hatta binlerce yer aldığı günümüz datamartları kullanılarak oluşturulacak modeller için neredeyse bir zorunluluktur. Bunun için modelde kullanılacak tüm değişkenler, kendi içlerinde alt gruplara ayrıştırıldıktan sonra veya doğrudan hedef değişkeni ile açıklama oranlarını ölçebilecek metotlar kullanılarak en önemlileri seçilir. Bu aşamada, anlamlılık oranı düşük veya ilişkisiz olan açıklayıcı değişkenler tespit edilerek bir sonraki aşamaya geçmeden elenirler.

Veri incelemesi aşamasında hedef değişken ile olan ilişkilerine göre tüm değişkenlerden, C5.0 veya Lojistik Regresyon modeline göre anlamlı çıkmayan değişkenler elenmiştir. Bu iki modelden herhangi birinde anlamlı çıkan değişken bırakılmıştır. Değişken eleme ile ilgili olarak aşağıdaki düzeltmeler, alan bilgisine istinaden yapılmıştır.

- Abonelik yaşı (Network_age) davranış skorlama analizlerinde önemli bulunmasına rağmen bu değişkenle çok yüksek korelâsyona sahip olan toplam ödenen fatura adedi (Invoice_Paid_CNT) değişkeni tercih edilmiş, Abonelik yaşı verisi analiz dışında tutulmuştur.
- Son altı ayda çıkan fatura tutarları (Invoice_MX) için, bu değişkenlerin tamamı anlamlı çıkmadığından (örneğin birinci, üçüncü ve beşinci aylardaki fatura tutarları anlamlı değişken olarak çıkmış, ancak ikinci ve dördüncü aylara ait olanlar önemsiz bulunmuştur) hiçbir tanesi analizlere alınmamıştır.
- Geç ödeme davranışlarını ölçen birçok benzer parametre değişken setinde yer aldığından, geç ödeme alışkanlıklarını ölçen parametrelerden biri olan Overdue_AVR değişkeni analizden çıkarılmıştır.
- Değişken seçimi aşamasında elenmesine rağmen, son üç veya altı ayda çıkan veya ödenen fatura ortalamaları sektörde sık kullanılan bir parametre olması nedeniyle önemli olduğuna karar verilip Invoice_L6_AVR değişkeni analize dahil edilmiştir.
- Ödeme tiplerinin abone davranışlarını belirlemede anlamlı bir veri olduğu bilindiğinden bu alandaki iki değişkenden en çok tercih edilen ödeme davranışı (Pay_Type_Preferred) değişkeni, bu aşamada modeller tarafından seçilmemesine rağmen analizlerde kullanılmıştır.

Tüm bu istatistiksel ve alan bilgisine dayalı düzenlemelerden sonra 54 değişkenin çok önemli bir kısmı elenmiş ve geriye 30 adet açıklayıcı değişken kalmıştır.

6.2.5. Değişken Dönüşümü

Tarife bilgisini içeren ve kategorik Tariff_Plan değişkeni, sekizden üçe indirgenmiştir. Burada TP2 ve TP6 tarifeleri Grup1; TP1, TP4, TP5 ve TP7 tarifeleri Grup2; TP3 ve TP8 tarifeleri de Grup3 olarak tekrar gruplanmıştır. Modellerde de sekiz kategorili değişken yerine üç kategorili yeni değişken kullanılmıştır.

Eğitim seviyesini gösteren Education değişkeni de kategorik veri tipinden, sıralı veri tipine dönüştürülmüştür. Kullanılan modellerin hiçbir tanesinde açıklayıcı değişkenlerin normal dağılması zorunluluğu bulunmamasına rağmen kategorik veri tipinin yanısıra deneme

yapılabilmesi amacıyla oluşturulmuştur. Bu değişkende bulunan boş alanlar için herhangi bir doldurma işlemi yapılmamış, ancak boş olan kayıtlara kategorik olarak "Boş" ifadesi atanmıştır.

Lojistik Regresyon modelinde çoklu normal dağılım varsayımı geçerli olmamasına karşın, değişkenlere logaritmik dönüşüm yapıldıktan sonra model performansında ciddi bir artış görülmüştür. Bundan dolayı da Lojistik Regresyon modeli uygulanmadan önce 18 oransal değişkenden 11'inin logaritmik dönüşümü yapılmıştır. Böylece toplam açıklayıcı değişken adedi 41'e yükselmiştir. Ancak, dönüşüm yapılan değişkenler Karar Ağaçları ve Yapar Sinir Ağ modellerinde kullanılmamıştır. Tüm modellere açıklayıcı değişken olarak girmek üzere seçilen değişkenlerin listesi yer almaktadır:

Tablo 12 : Modellerde Kullanılan Açıklayıcı Değişkenler

1	Gender		
2	Marital_Status		
3	Education		
4	Tariff_Plan		
5	Phone_CNT		
6	Automatic_Payment_FLG		
7	Roaming_FLG		
8	Active_Line_CNT		
9	Inactive_Line_CNT		
10	Black_List_FLG		
11	Invoice_L6_AVR		
12	Invoice_Paid_CNT		
13	Invoice_Paid_AVG	31	Invoice_Paid_AVG_LogN
14	Invoice_Paid_Max_AMT	32	Invoice_Paid_Max_AMT_LogN
15	Invoice_National_L6_AVR	33	Invoice_National_L6_AVR_LogN
16	Invoice_Voice_L6_AVR	34	Invoice_Voice_L6_AVR_LogN
17	Overdue_Invoice_AVR	35	Overdue_Invoice_AVR_LogN
18	Overdue_L6_AVR	36	Overdue_L6_AVR_LogN
19	Overdue_L6_FLG		
20	Overdue_Max_L6_AMT	37	Overdue_Max_L6_AMT_LogN
21	Overdue_Sequential_L12_CNT		
22	Pay_Type_Preferred_Method		
23	PROPN_Invoice	38	PROPN_Invoice_LogN
24	MO_Barred_CNT		
25	MO_Barred_L6_CNT		
26	MT_Barred_CNT		
27	Risk_Unpaid_AMT	39	Risk_Unpaid_AMT_LogN
28	Risk_AMT	40	Risk_AMT_LogN
29	Risk_RT	41	Risk_RT_LogN
30	Usage_Risk_RT		

6.2.6. Veri Setinin Bölümlenmesi

Geliştirilen bir skor kartın veya skorlama modelinin gerçek veri ile doğrulanması da çok önemlidir. Böylece modelin başka bir veri üzerinde de aynı seviyede doğru tahmin yapıp yapamadığı ortaya çıkar. Kurulan bir modelin doğrulanması ile aşırı öğrenmiş bir model (over-modelling) olup olmadığı veya modellemede ilgisiz bir analiz veri setinin kullanılıp kullanılmadığı test edilmiş olur.

Doğrulama yapmak için kullanılacak veri seti iki şekilde seçilebilir: Birincisi, model geliştirmek için toplanan verilerin bir kısmı modellemede kullanılmayıp doğrulama amacıyla ayrılır ki en çok kullanılan yöntemdir. İkinci yöntem ise, benzer abonelere ait başka bir döneme ait verinin kullanılmasıdır. Bizim çalışmamızda yeteri kadar veri mevcut olduğundan doğrulama ve test veri setleri ayrılmıştır. Tüm veri setinin %60'ı eğitim, %30'u doğrulama ve %10'u da test amacıyla üç alt parçaya ayrıştırılmıştır. 10.000 gözlemin parçalandıktan sonraki dağılımı aşağıdaki gibidir:

Tablo 13 : Analiz Veri Setinin Bölümlenmiş Durumu

		Target_IDD90		Toplam	Riskli Abone Oranı
		Risksiz	Riskli		
Partition	Eğitim	4.176	1.774	5.950	29,82%
	Test	2.121	930	3.051	30,48%
	Doğrulama	703	296	999	29,63%
Toplam		7.000	3.000	10.000	30,00%

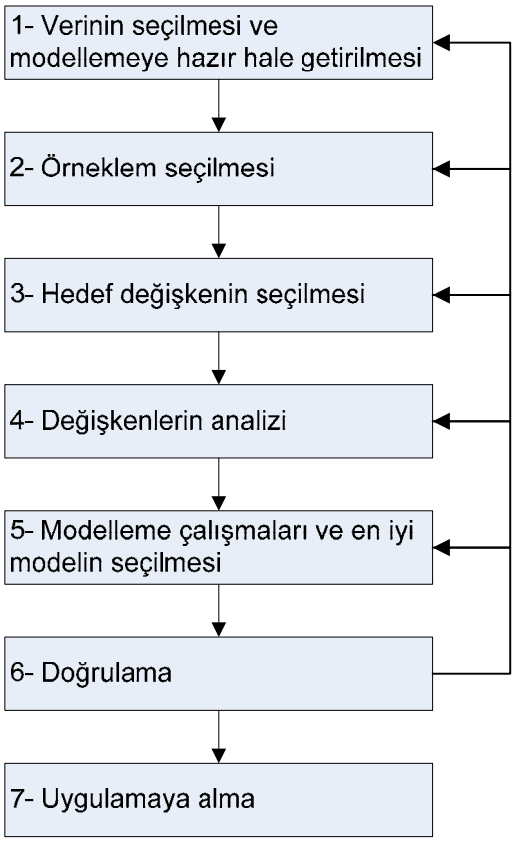
Verinin veri madenciliğinde kullanılan modellerin geçerliliğinin test edilmesinde çok önemli bir yöntem olan alt gruplara ayırma, bu çalışmada da kullanılan Yapay Sinir Ağları ve Karar Ağaçları gibi veriden öğrenmeye dayanan algoritmalar için özellikle tercih edilmiştir.

6.3. Modelleme

6.3.1. Model Geliştirme Süreci

Etkili bir ayırım fonksiyonu oluşturabilmek için iki konunun dikkate alınması gerekmektedir: Bunlardan birincisi, değişkenler arasındaki ilişkiler ve hedef değişken ile olan ilişkilerinin doğrusal olup olmadığıdır. İkincisi de sınıflandırma modelinin doğruluğunu artırmak için

ilgisiz değişkenlerin analiz verisinden çıkarılmasıdır[140]. İyi bir istatistiksel süreç; veriden en iyi tahmini yapabilmeli, eldeki tüm veriyi modellemeye dâhil edebilmeli, verinin yapısından etkilenmemeli, yorumlaması ve anlaşılması kolay olmalıdır. Aşağıdaki şekil modelleme aşamalarını kısaca özetlemektedir. Burada, ikinci ile altıncı adımlar arasında her zaman geri dönüşlerle beslenen bir süreç mevcuttur.



Şekil 16 : Model Geliştirme Süreci

Bu çalışmada, Lojistik Regresyon, Karar Ağaçları ve Yapay Sinir Ağ modelleri kullanılmıştır. Her model için en iyi sonucu veren algoritma tercih edilmiştir; Lojistik Regresyon için Logit fonksiyonu, Karar Ağacı için CHAID algoritması, Yapay Sinir Ağı algoritmaları içinde de en performanslı sonucu veren Çok Katmanlı Algılayıcı (Multilayer Perception – MLP) algoritması tercih edilmiştir.

[140] Ong, Chorng-S., Jih-Jeng Huang, and Gwo-Hshiung Tzeng., s. 41.

6.3.2. Segmentasyon

Skorlama modelleri kurulmadan önce aboneler mutlaka segmentlere ayrıştırılmalıdır. Segmentasyon; kurumsal ve bireysel şeklinde olabileceği gibi, daha önce kredi geçmişi olan ve olmayan aboneler veya başka bölümler tarafından yapılan segmentasyon modeller ile ayrıştırılan aboneler de olabilir. Skorlama sonuçlarının etkinliğini artırmak için abonelerin daha homojen alt gruplara ayrıştırılması başarıyı çok önemli ölçüde artırır. Çünkü her segment için önemli olan değişkenler farklıdır. Örneğin, firma aboneliklerinin değerlendirilmesinde finansal tablolardan elde edilen rasyolar veya faaliyet alanı, sermaye miktarı gibi faktörler kullanılırken; bireysel abonelerin kredibilitelerinin değerlendirilmesinde aylık kazançları, ödedikleri sigorta primleri, son çalışılan işyerindeki çalışma süreleri, sahip olunan aracın özellikleri, ikamet edilen evin adresi veya son oturulan evde ne kadar süreden beri ikamet edildiği gibi özellikler ön plana çıkmaktadır. Bunun da ötesinde her segment için aynı özellik mevcut olmayabilir. Örneğin, bireysel abonelerin finansal tabloları yoktur.

Banka gibi kuruluşlar skor sonuçlarını genel olarak tek başlarına kullanmazlar. Mutlaka finansal tablolarında olmayan değişkenlere de ulaşarak bunları da kalitatif olarak değerlendirmeye katarlar. Dolayısıyla her değerlendirme süreci tamamen otomatize olmayıp manuel müdahalenin yer aldığı kısımlar bulunabilir. Bizim çalışmamızda kullanılan skorlama süreci bu tür müdahalelerden bağımsızdır.

Modelde kullanılan veri, analiz yapılan operatörün şebekesindeki genel ana kütleyi temsil etmemektedir. Bunun birinci nedeni, gerçek hayatta da uygulaması yapılan, doğruluğu artırmak için modelleme öncesinde segmentasyon uygulanması olup, çalışmaya sadece belirli bir segmente ait aboneler dahil edilmiştir. Diğer neden ise, rakipler tarafından yapılabilecek muhtemel kıyaslamaları engellemektir. Ancak aşağıdaki şartları sağlayan gözlemler mevcut anakütleden rastsal olarak SAS Enterprise Guide programı vasıtasıyla seçilmiştir.

- Analiz tarihinden 3 yıl önceki dönem itibari ile güncel olan veriler
- Belirli bir coğrafi bölgede aktivasyonu yapılan aboneler
- Abonelik yaşı altı aydan büyük aboneler (son altı ay içerisinde hiç aktivitesi olmayan – faturası çıkmayan aboneler vb. veri setinden çıkarılmıştır)
- En az bir faturası ödenmiş aboneler

- Sadece bireysel hat kullanan aboneler seçilmiştir. Bu nedenle, aktif veya inaktif hat adedi toplamı 20 nin üzerinde olan kayıtlar veri seti içerisinden temizlenmişlerdir

6.3.3. Kullanılan Modeller

6.3.3.1. Lojistik Regresyon Modeli

6.3.3.1.1. Giriş

Regresyon analizinde temel amaç, bağımsız değişkenler yardımıyla bağımlı veya açıklanan değişkenin hangi değeri alacağını bulmaktır. Bağımlı ve bağımsız değişkenlerin hangi ölçekle ölçülmüş oldukları, kurulan modellerin ve katsayıların nasıl bulunacağını belirler. Birçok uygulamada, bağımlı değişken sadece iki sonuçlu olaylarla ilgilenebilir ve 0 veya 1 değerini alabilir[141]. Lojistik Regresyon (LR), dikotom bağımlı değişkenin birden fazla oransal veya nominal açıklayıcı değişken kullanılarak tahmin edilmesini sağlayan matematiksel bir modeldir[142]. Bizim örneğimizdeki gibi abonenin borcunu ödeyip öde(ye)meyeceği de iki sonuçlu olaylara örnek olarak verilebilir. Bu gibi durumlarda bağımlı değişken dikotomdur ve 1 değeri π olasılığını ve 0 değeri 1-π olasılığını gösterir.

Doğrusal Regresyon analizinde amaç, bağımlı değişkenin tahmin edici bir grup bağımsız değişken karşısında alacağı ortalamayı bulmaktır. Lojistik Regresyon modelinde ise temel amaç, tesadüfi değişken Y'nin π olasılıkla 1 ve 1-π olasılıkla 0 değerini alabileceğini göstermektir.

Tesadüfi Bernoulli değişkeni aşağıdaki farklı olasılık dağılımına sahiptir:

$P_r(Y = y) = \pi^y (1-\pi)^{1-y}$;

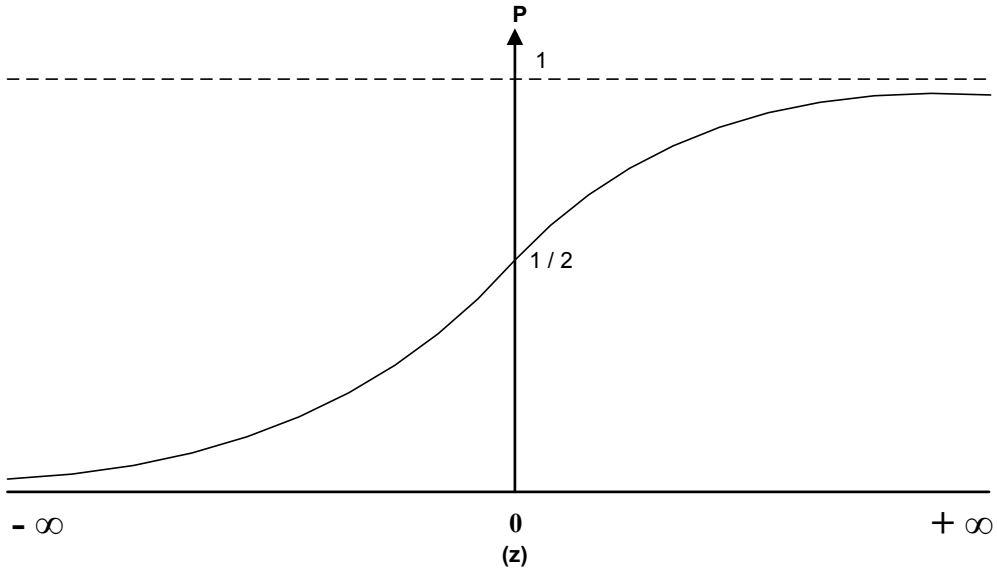
y=0,1.

Dikotom hedef değişkeni tahmin etmek için birçok başka yöntem mevcuttur. Ancak iki önemli sebepten dolayı LR modeli çok kullanılan popüler bir analiz aracıdır[143, 144]:

[141] Le, Chap T., **Applied Categorical Data Analysis.** New York: John Wiley & Sons Ltd., 1998. s. 113.
[142] Kleinbaum, David G., **Logistic Regression A self Learning Text.** New York : Springer, 1994. s. 5.
[143] Kleinbaum, David G., s. 6-7.
[144] Le, Chap T., s. 116.

1 – Lojistik fonksiyonunun (f(z)) her vaka için 0 ile 1 arasında değerler alması, bu yöntemin en önemli özelliğidir. Olasılık tahmin edilecek olaylarda LR, çoğunlukla ilk seçenek olarak kullanılır. Bu uygulamada olduğu gibi her yeni abonenin risk derecesini veya borcun geri ödeme olasılığını belirleyerek bireysel riski bulabiliriz.

2 – Lojistik eğrisi, Şekil 17'de de görüldüğü gibi artan S şeklindedir ve üst sınırı vardır. Bu şekil birçok risk faktörünün kombine etkisini yansıtmaktadır. Bu durum, sağlık alanındaki ilgili verileri modellerken de çok kullanışlıdır.



Şekil 17 : Lojistik Fonksiyon

Olasılıklar sıfırdan bire doğru büyürken Logit fonksiyonu -∞ ile +∞ arasında değerler alır. Modelin bağımlı değişkenleri ile bağımsız değişkenleri arasında doğrusal bir ilişki olduğu halde, olasılıklarla bağımlı değişkenler arasındaki ilişki doğrusal değildir.

LR fonksiyonu da aşağıdaki gibi ifade edilir:

$$f_z = \frac{1}{1+e^{-z}},$$

LR fonksiyonundan yola çıkarak oluşturacağımız Lojistik Model aşağıda yer almaktadır:

$$z_i = \beta_0 + \sum_{j=1}^{k} \beta_j \chi_{ji}$$

X'ler risk faktörlerini veya bağımsız değişkenleri göstermektedirler.

Lojistik Regresyon modeli ifadesindeki regresyon kısmı, açıklayıcı değişkenlerin değerlerinin lineer kombinasyonunun var olması ve regresyon katsayılarının varlığından, lojistik ifadesi de bağımlı değişkenin olasılığının lojistik transformasyon denklemine sahip olmasından kaynaklanmaktadır:

$$Y = \ln(\frac{\chi}{1-\chi})$$

LR modelindeki gözlemler için lojistik fonksiyon, yani bağımsız değişkenlerin verilen özelliklerine göre π_i olasılığı, aşağıdaki gibi lineer model olarak tanımlanabilir:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \chi_i$$

veya

$$\pi = \frac{1}{1+e^{-(\beta_0+\beta_1\chi_i)}}$$

π risksiz olma olasılığını gösterirse

$$e^{Z_i} = \frac{\pi_i}{1-\pi_i}$$

olarak ifade edilebilir.

Bağımlı değişken nominal ölçekli olduğunda En Küçük Kareler (EKK) tekniğiyle elde edilen tahminler yetersiz kalmaktadır. Çünkü bağımlı değişken nominal ölçekli olduğundan normal dağılmamaktadır. Dolayısıyla, tahmin edilen varyanslar minimum değildir[145]. EKK tekniğiyle oluşturulan doğrusal regresyon modelinde hataların normal dağılmaması, tahmin edilen olasılıkların 0 – 1 aralığında kalmaması, düşük R^2 ve farklı varyans problemlerine yol açmaktadır. Doğrusal regresyon modeli içerisinde EKK yöntemi uygulayarak bu problemlerin üstesinden gelmek için bazı yollar mevcuttur. Örneğin tahmin edilen olasılıkların 0 – 1 aralığında kalmasını sağlamak için kısıt konulabilir, ağırlıklı regresyon analizi kullanılarak eşvaryans, örnek hacmi büyütülerek normallik varsayımı sağlanabilir. Ancak doğrusal EKK modelinin en büyük sakıncası bağımlı değişken ile bağımsız değişkenler arasında doğrusal bir ilişki olduğunu varsaymasıdır. Çoğu durumda ise değişkenler arasındaki doğal yapı sonucu doğrusal olmayan modeller daha açıklayıcı olmaktadır.

[145] Albayrak, A. Sait, **Uygulamalı Çok Değişkenli İstatistik Teknikleri**. Ankara: Asil Yayın Dağıtım, 2006. s. 439.

Lojistik Regresyon modelinin parametreleri, analitik olarak elde edilemediğinden iteratif bir yöntem olan Maksimum Likelihood (ML) tekniğiyle tahmin edilmektedir. Çünkü analiz yöntemi açısından bakıldığında EKK'nin aksine LR modelinde çözüm, önce bir çözümün kabulünden başlanarak bunun yeniden gözden geçirilerek iyileştirilmesi çabasına dayanmaktadır. Olabilirlik fonksiyonundaki değişim ihmal edilebilir seviyenin altına indiğinde tekrar etme (iterasyon) işlemine son verilir.

LR modeli ortalama π ve varyansı $\pi(1-\pi)$ olan Bernoulli dağılımına uyar. Basit LR modelinde Likelihood fonksiyonu aşağıdaki gibidir[146] :

$$L = \prod_{i=1}^{n} \Pr((Y_i = y_i)$$

$$L = \prod_{i=1}^{n} \frac{\left[e^{(\beta_0+\beta_1\chi_i)}\right]^{y_i}}{1+e^{(\beta_0+\beta_1\chi_i)}}$$

$y_i = 0,1;$

6.3.3.1.2. Odds Oranı (Üstünlük oranı)

Lojistik regresyonda kullanılan Odds oranı nisbi risklerin ölçülebilmesine imkan verir. Odds Oranı (Üstünlük Oranı), olma ihtimalinin olmama ihtimaline oranı olarak tanımlanabilir. Odds oranı aşağıdaki şekilde formülize edilir:

p : olayın olma (gerçekleşme) ihtimali

1 - p : olayın olmama (gerçekleşmeme) ihtimali

$$OddsOranı = \frac{p}{1-p}$$

Bağımsız değişkenlerin oransal olup olmamasından bağımsız olarak, üstünlük oranının doğal logaritması bağımsız değişkenlerin doğrusal bir fonksiyonu olarak ifade edilebilir[147] :

$$\ln(Odd; X = x) = \beta_0 + \beta_1(x)$$

$$\ln(Odd; X = x+1) = \beta_0 + \beta_1(x+1)$$

[146] Le, Chap T., s. 117.
[147] Le, Chap T., s. 118.

Yukarıdaki iki denklemden elde edilen değer aşağıdaki denklemde yerine konursa, X bağımsız değişkenindeki 1 birimlik artıştan kaynaklanan Odds oranı bulunmuş olur:

$$e^{\beta_i} = \frac{(Odd; X = x+1)}{(Odd; X = x)}$$

X değişkenindeki değerin m birim artmasından kaynaklanan (X=x+m değerinin X=x değerine göre)

Odds oranı $= e^{m\beta_1}$

olur. Odds oranı; $e^{\hat{\beta}_1}$ olarak tüm paket program çıktılarında mevcuttur. Odds oranının güven aralığı da (%95 olasılıkla) şu şekilde ifade edilir:

$$e^{\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)}$$

Bununla birlikte Odds oranını yorumlamayı güçleştiren en önemli etken bağımsız değişkenin ölçüldüğü ölçek olup, Odds oranı aynı değişken için ölçeğe göre değişmektedir. Odds oranının doğal logaritması Logit olarak bilinir. Odds oranı ve olasılıklar aynı olayı farklı şekilde gösterme yöntemleridir.

6.3.3.1.3. Modelin Parametrelerinin Test Edilmesi ve Uygunluğunun Değerlendirilmesi

Veriye uyan bir LR modeli kurulduktan ve değişkenlerin katsayıları tahmin edildikten sonra LR modelinin uygunluğunu değerlendirirken 3 farklı yol kullanılır[148]:

a) Tüm modelin bir bütün olarak anlamlılığının test edilmesi
b) Modele dahil edilmesi düşünülen yeni bir değişkenin gerçekte anlamlı bir katkı sağlayıp sağlayamayacağını belirlemek için tekil olarak yapılan test
c) Birden fazla bağımsız değişkenin modele olan katkısını beraberce değerlendirmek.

LR modeli bütün olarak test edilirken "log-likelihood" istatistiği kullanılır. Log-likelihood değeri negatiftir ve -2 ile çarpıldığında k modeldeki parametre sayısını göstermek üzere, n-k serbestlik derecesiyle X^2 dağılımına uyar[149]. Modelin genel geçerliliğini test eden sıfır ve karşıt hipotezler aşağıdaki gibi yazılabilir:

[148] Le, Chap T., s. 130.
[149] Albayrak, A. Sait, s. 450.

H_0 : Tüm bağımsız değişkenler birlikte değerlendirildiğinde bağımlı değişkendeki değişimi açıklayamamaktadırlar.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

H_1 : Teorik model verileri iyi temsil etmektedir.

Bu hipotez -2LogL ile test edilir. Likelihood oran testi;

$$X^2_{LR} = 2\left[\ln L(\hat{\beta}) - \ln L(0) \right]$$

-2LogL istatistiği, modele ilave edilen bağımsız değişkenlerin modele olan katkılarının araştırılmasında da kullanılırlar. Diğer bir anlatımla, -2LogL istatistiği LR katsayılarının anlamlılıklarının sınanmasında da kullanılmaktadırlar. Bu nedenle -2LogL istatistiği bağımlı değişkendeki açıklanmayan varyansın anlamlılığını gösterir. Bu istatistik sapan ki-kare istatistiği olarak da bilinir. Bu istatistiğin anlamlı olmaması LR analizinde istenen durumu göstermektedir.

Regresyon analizindeki F testinin tüm denklemin anlamlılığını test ettiği gibi Hosmer ve Lemeshow G istatistiği olarak bilinen model ki-kare istatistiği, LR modelini genel olarak test eder. Bağımsız değişkenlerden hiçbirisinin bağımlı üstünlük oranıyla (odds ratio) anlamlı doğrusal bir ilişki göstermediğini ileri süren sıfır hipotezini test eder. Diğer bir anlatımla bu istatistik, sabit terimin dışındaki tüm logit katsayılarının sıfıra eşit olup olmadığını sınamaktadır. LR analizinde Hosmer ve Lemeshow G istatistiğinin anlamlı olması arzu edilir.

Hosmer Lemeshow testi için hipotezler aşağıdaki gibi yazılır:

H_0 : Tahmin denklemi anlamlıdır.

H_1 : Tahmin denklemi anlamlı değildir.

Bağımsız değişkenlerin ayrı ayrı test edilmesinde Ho hipotezi şu şekilde ifade edilir:"Diğer bağımsız değişkenler modelde iken yeni eklenen Xi değişkeni bağımlı değişkenin tahmininde ek bir katkı sağlamamıştır". Diğer bir deyişle;

$$H_0 : \beta_i = 0$$

Test, X^2 dağılımına uyan 1 serbestlik dereceli -2LogL istatistiği ile yapılır. Bununla birlikte burada alternatif olarak aşağıdaki istatistik de kullanılabilir:

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)},$$

z_i standart normal dağılıma uyar.

Regresyon analizindeki R^2 istatistiğine benzeyen ve geniş kabul gören bir istatistik LR analizinde bulunmamaktadır. R^2, bağımlı değişkenin açıklanan varyansının yüzdesini göstermektedir, ancak LR analizinde bağımlı değişkenin varyansı bu değişkenin olasılık dağılımına bağlıdır. Diğer bir anlatımla iki gruplu bir bağımlı değişkenin varyansı, grup frekansları eşit olduğu zaman (%50 * %50 = 0.25) maksimum olacaktır. Bu nedenle Regresyon Analizindeki R^2 değeri ile Lojistik Regresyon analizindeki R^2 değerini birbirleri ile karşılaştırmak uygun değildir. Bununla birlikte literatürde LR analizi için birkaç R^2 istatistiğine yer verilmektedir[150].

Cox ve Snell R^2 : Olabilirlik esasına göre çoklu R^2 istatistiğine benzemektedir. İstatistiğin maksimum değerinin genelde 1'den küçük olması bu istatistiğin yorumunu güçleştirmektedir.

Nagelkerke R^2 : Cox ve Snell R^2 istatistiğinin 0-1 aralığında değerler almasını sağlamak amacıyla geliştirilmiştir.

6.3.3.1.4. Çoklu Lojistik Regresyon Analizi

Birden fazla bağımsız değişkenin kullanıldığı LR modellerinin açıklayıcılığı daha fazla değişken kullanmanın getirdiği avantaja sahiptir. Değişkenler, dikotom, kategorik veya oransal değişken olabilir. Çoklu LR modelinin genel ifadesi şu şekildedir:

$$\pi_i = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{k} \beta_j \chi_{ji}\right)}}, \quad i=1,2,\ldots,n$$

veya

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \sum_{j=i}^{k} \beta_j \chi_{ji}$$

şeklinde ifade edilebilir.

Bu modelin dayandığı Likelihood fonksiyonunun ifadesi de şudur:

[150] Albayrak, A. Sait, s. 460.

$$L = \prod_{i=1}^{n} \frac{\left(\exp\left[\beta_0 + \sum_{i=1}^{k} \beta_j \chi_{ji} \right] \right)^{y_i}}{1 + \exp\left[\beta_0 + \sum_{i=1}^{k} \beta_j \chi_{ji} \right]}; \; y_i = 0,1.$$

Aynı şekilde odds oranı $\exp(\beta_i)$ ile ifade edilir ve şu iki durumu göstermek üzere kullanılır:

a) X_i dikotom olmak üzere, X_i'nin odds oranını gösterir (X_i=1 gerçekleşme ihtimali ve X_i=0 gerçekleşmeme ihtimali)
b) X_i sürekli değişken olmak üzere, X_i değişkenindeki 1 birimlik artışın odds oranını gösterir.

Çoklu LR modelinde eğer bağımsız değişkenler arasında yüksek korelasyon varsa bunlar arasından en iyi tahmin ediciler seçilir. Ayrıca benzer çalışmalarda kullanılan değişkenler bu konuda örnek alınabilir. Ayrıca bağımsız değişkenler arasında faktoriyel çarpım ($\chi_i \chi_2$) veya değişkenin yüksek dereceden kuvveti (χ^2) modelin daha güçlü olmasını sağlayabilir. Ancak modelin lineer yapısının ciddi zarar gördüğü düşünülüyorsa değişkenin kuvvetinin modele sokulmasından vazgeçilebilir.

Kurulan Lojistik Regresyon modelinin doğruluğunun anlaşılması ile ilgili pek çok yol mevcuttur:

- **Pseudo R^2 İstatistiği:** R^2 istatistiği doğrusal regresyon modelinde bağımlı değişkendeki değişimin açıklanan oranını göstermektedir, ancak LR için aynı şey geçerli değildir. Bununla birlikte R^2 istatistiği yerine benzer özellikleri gösteren Pseudo R^2 istatistiği geliştirilmiştir
- **Kümeleme ve Doğrulama:** Çapraz tablolar ile gözlenen ve model sonucu üretilen gözlemlerin birbirlerine ne derecede uydukları ölçülerek, modelin var olan datayı hangi oranda doğru tanımlayabildiği ölçülür.

LR modelinin kurulmasında en iyi tahmin edicilerin modele dahil edilmesi için değişkenlerin adımsal olarak seçilmesi ile ilgili pek çok prosedür mevcuttur. Bu süreçler, hedef değişken ile anlamlı katsayılara sahip değişkenlerin otomatik olarak seçerler. Aşağıda bunlardan çok kullanılanlardan ikisi aktarılmıştır:

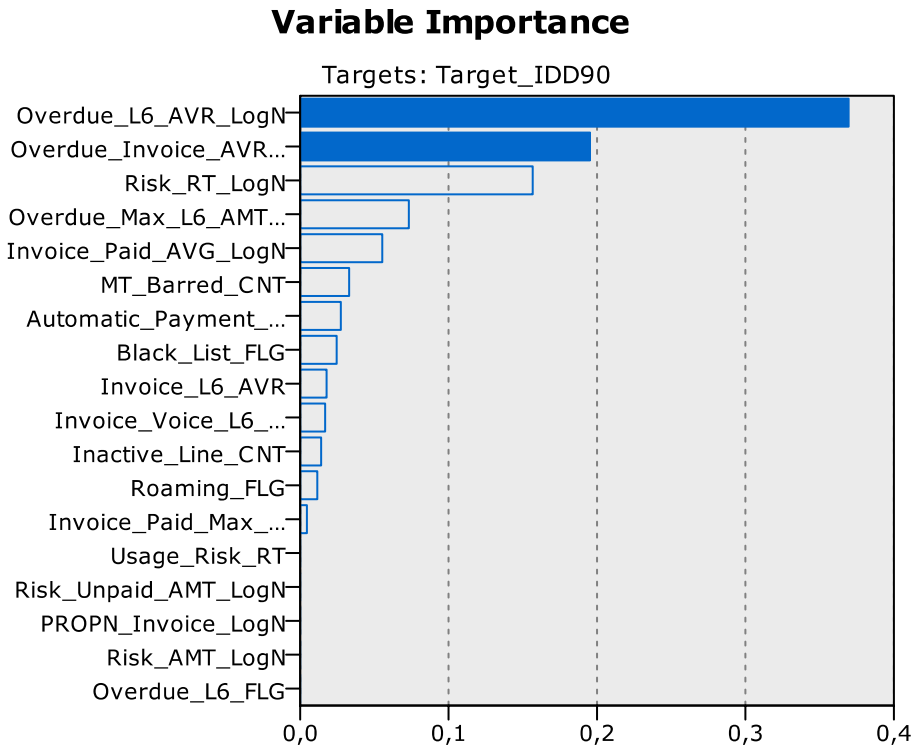
Forward Stepwise metodu, hiçbir tahmin edicinin olmadığı model ile başlar. Her adımda, modele en fazla uyan değişkeni en yüksek anlamlılık oranına uygun olarak modele dahil eder. Sırasıyla diğer değişkenleri modele dahil eder, anlamlılık değeri belirli bir değerin altına(0,05) düşünceye kadar yeni değişkenleri alır. Yeni alınan değişkenlerin anlamlılığı 0,05 oranının (bu değer başka bir oran olarak da belirlenebilir) altına düştüğü anda, modele yeni değişken almaya son verir.

Backward Stepwise metodunda ise ilk başta tüm tahmin edici değişkenleri modele dahil eder. Her adımda en az açıklayıcılığı olan değişkeni modelden çıkarır. Bu süreç geriye kalan tüm değişkenlerin açıklayıcılığının anlamlılığı belirlenmiş bir katsayıyı aşıncaya kadar devam eder. Eğer her iki metod ile de aynı değişkenler modelde yer alıyorsa modelin iyi kurulmuş olduğu kendiliğinden doğrulanmış olur.

6.3.3.1.5. Lojistik Regresyon Analizinin Uygulanması

LR modelinin varsayımları arasında kullanılan değişkenlerin normal dağılması zorunluluğu bulunmamaktadır. Ancak dağılımı çarpık değişkenlerden bazılarının doğal logaritması alındıktan sonra tekrar kurulan logit modelinin bir öncekinden daha iyi tahmin yaptığı görüldüğünden, Lojistik Regresyon modeli, veri dönüşümü yapıldıktan sonra oluşturulmuştur. Değişkenlerin modele alınması ile ilgili olarak backward stepwise metodu tercih edilmiştir.

LR sonuçlarına ilişkin ilk grafik olan Şekil 18, Duyarlılık Analizine aittir. Bu listede yer alan değişkenlerin modele katkılarının önem sırasını göstermektedir. Duyarlılık analizinde belirlenen değişkenlerin önemliliği konusunda değerlendirme yapılmalıdır. Duyarlılık analizi değişkenlerin güvenilirliğini ya da yararlılığını ölçmez.



Şekil 18 : Lojistik Regresyon Modelinde Kullanılan Değişkenlerin Modeli Açıklama Gücü Grafiği

Son dönemde çıkan faturaların zamanında ödenip ödenmemesi ile ilgili olarak analize sokulan Overdue_L6_AVR değişkeninin logaritmik dönüşüm yapıldıktan sonra, abone ödeme davranışlarını belirleyen en önemli değişken olduğu görülmektedir. İkinci sırada önemli çıkan değişken de benzer şekilde Overdue_Invoice_AVR (Ödenen tüm faturaların son ödeme tarihlerine göre ortalama geç ödenme gün sayısı) yine faturaların zamanında ödenip ödenmemesi ile ilgilidir. Buna göre fatura tutarından bağımsız olarak, zamanında ödeme yapma verisinin riskli aboneyi belirlemede en önemli parametreler olduğu görülmektedir.

Üçüncü önemli değişken olan Risk_RT_LogN, toplam riskin, ödenen toplam fatura tutarına oranının logaritmasını göstermektedir. Buna göre, riski değerlendirilecek abonelerden şebekede yaptığı toplam ödeme tutarı daha yüksek olan abone daha az risklidir.

LR modelinde kullanılan değişkenlerin modeli açıklama güçleri Tablo 14'de verilmiştir. İlk değişken olan Overdue_L6_AVR_LogN'ın %36.96, ikinci Overdue_Invoice_AVR_LogN'ın %19.53 ve üçüncü değişken olan Risk_RT_LogN'ın da %15.67 açıklama sağladığı görülmektedir.

Tablo 14 : Lojistik Regresyon Modelinde Kullanılan Değişkenlerin Modeli Açıklama Gücü

Değişkenler	Önem Derecesi	Sıralama
Overdue_L6_AVR_LogN	36,96%	1
Overdue_Invoice_AVR_LogN	19,53%	2
Risk_RT_LogN	15,67%	3
Overdue_Max_L6_AMT_LogN	7,32%	4
Invoice_Paid_AVG_LogN	5,54%	5
MT_Barred_CNT	3,30%	6
Automatic_Payment_FLG	2,74%	7
Black_List_FLG	2,46%	8
Invoice_L6_AVR	1,78%	9
Invoice_Voice_L6_AVR6_LogN	1,68%	10
Inactive_Line_CNT	1,42%	11
Roaming_FLG	1,16%	12
Invoice_Paid_Max_AMT_LogN	0,45%	13
Overdue_L6_FLG	0,00%	14
Risk_AMT_LogN	0,00%	15
PROPN_Invoice_LogN	0,00%	16
Risk_Unpaid_AMT_LogN	0,00%	17
Usage_Risk_RT	0,00%	18

Not: Son 5 değişkenin açıklama güçleri %1'in altında kalmıştır.

Aşağıdaki Tablo 15'de LR modeline girdi olarak kullanılmak üzere kullanılan kategorik değişkenlerin frekans dağılımları mevcuttur. Tarife değişkeni üç alt grupta, cinsiyet de kayıp değer olmadığı için iki alt grupta toplanmıştır. Az kategorili olan bu iki değişkenin frekansları arasında da büyük bir farklılık olmamasına rağmen son aşamada modelde anlamlı değişkenler olarak yer almamışlardır.

Tablo 15 : Kategorik Değişkenlerin Kodlanması

		Frekans	Parametre Kodlaması	
			(1)	(2)
Tariff_gr	Grup1	2.544	1	0
	Grup2	2.197	0	1
	Grup3	1.209	0	0
Gender	E	4.879	1	
	K	1.071	0	

a. This coding results in indicator coefficients.

Model X^2 istatistiği (Hosmer lemeshow G) Lojistik Regresyon modelini genel olarak test etmektedir. Bu test regresyon analizindeki F istatistiğine benzer ve anlamlı olması gerekir. Hosmer-Lemeshow X^2 istatistiği, modelin veriyi yeterli şekilde belirleyip belirleyemediğini ölçer.

H_0 : Sabit terim hariç tüm bağımsız değişkenler birlikte değerlendirildiğinde bağımlı değişkendeki değişimi açıklayamamaktadırlar

$$H_0 = \beta_1 = \beta_2 = ... = \beta_k = 0$$

H_1 : Teorik model verileri iyi temsil etmektedir.

Tablo 16'da yer alan, Hosmer-Lemeshow istatistiğinin olasılık değeri (0,041) 0,05'ten küçük olduğu için modelin datayı yeterli şekilde açıklayamadığı Ho hipotezi reddedilir.

Tablo 16 : Hosmer ve Lemeshow Test

Adım	Chi-square	df	Sig.
11	16,131	8	0,041

Modelin verileri tam temsil etmesi durumunda olabilirlik katsayısı (L) 1'e ve - 2LogL istatistiği de 0'a eşittir. -2LogL istatistiği bağımlı değişkendeki açıklanamayan varyansın anlamlılığını gösterir. Bu istatistik sapan X^2 değeri olarak bilinir ve anlamlı olması istenir.

Belirlilik katsayısı (R^2), analizde kullanılan datasetten elde edilen regresyon doğrusunun verileri ne kadar iyi temsil ettiğini gösteren bir ölçüdür. Aşağıdaki Tablo 17'de yer alan Pseudo R^2 istatistiği mevcut model ile herhangi bir tahmin edici değişkenin dışarıda tutularak oluşturulan "null" model arasındaki olasılığa dayanır. Son adımda (son modelde) %37,3'tür. Bu oran, bağımlı değişken ile bağımsız değişkenler arasında yaklaşık %37,3'lük bir ilişkinin olduğunu göstermektedir. Cox ve Snell R^2 istatistiğinin maksimum değerinin birden küçük olması yorumu güçleştirmektedir. Nagelkerke R^2 istatistiği, Cox ve Snell R^2 nin maksimum değerinin bir olmasını sağlamak amacıyla geliştirilmiştir. Son modelde yer alan orana göre bağımlı değişken ile bağımsız değişkenler arasında yaklaşık %52,9'luk bir ilişkinin olduğu görülmektedir.

Tablo 17 : Lojistik Regresyon Model Özeti

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
11	4476,354(a)	,373	,529

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,000.

Tablo 18'de yer alan sınıflandırma tablosu, oluşturulan LR modelinin pratik sonuçlarını göstermektedir. Herbir abone için regresyondan elde edilen olasılık değeri 0.5'ten büyükse riskli abone, 0.5'ten küçükse de risksiz abone grubuna girmektedir. Buna göre, 4.176 risksiz abonenin %90.5'una karşılık gelen 3.779'u, 1.774 riskli aboneninin de %65.8'ine karşılık gelen 1.167'si doğru olarak tahmin edilmiştir. Genel olarak tüm abone grubunun %83,1'i doğru gruba girmiştir.

Tablo 18 : Lojistik Regresyon Modeli Sınıflandırma Tablosu

		Tahmin			
		Risksiz	**Riskli**	**Toplam**	**Doğruluk Yüzdesi (%)**
Gözlem	**Risksiz**	3.779	397	4.176	90,5
	Riskli	607	1.167	1.774	65,8
Genel Doğuluk Yüzdesi (%)					83,1

a. Eşik değeri 0.5'tir.

Tablo 19'da herbir tahmin edici parametrenin etkisi yer almaktadır. Tablonun altında yer alan değişkenler ilk adımda modele alınan değişkenleri sırasıyla göstermektedir. Son modelde sabit terim hariç 18 açıklayıcı değişken yer almaktadır.

Büyük örnekler için LR katsayılarının testi X^2 dağılımına uyan Wald istatistiği ile yapılmaktadır. Serbestlik derecesi bire eşit olan bir bağımsız değişkenin Wald istatistiği, ilgili değişkenin LR katsayısının standart hatasına oranının karesine eşittir. Kategorik değişkenlerin Wald istatistiği grup sayısının bir eksiği (df=G-1) ile X^2 dağılımına uymaktadır. Wald istatistiklerinin anlamlılık düzeyleri, Odds oranları sütunundan önceki Sig. alanında verilmiştir. Wald istatistiğinin anlamlılık seviyesi 0.05'ten daha küçükse o parametrenin model içerisinde yer almasının anlamlı olduğu söylenir. Son adımdaki LR katsayılarının tamamı %5 anlamlılık düzeyinde Wald istatistiğine göre anlamlıdır.

Wald değerleri regresyon modelinde yer alan her değişkenin modele olan katkısını vermektedir. Wald değeri ne kadar büyük olursa değişkenin önemi de artmaktadır. Buna göre, Risk_AMT_LogN, Risk_RT_LogN, Overdue_L6_AVR_LogN, Risk_Unpaid_AMT_LogN, Usage_Risk_RT ve Automatic_Payment_FLG değerleri abonenin riskinin ölçülmesinde modelde diğer değişkenlere göre daha etkin olarak yer almışlardır.

Tablo 19 : Lojistik Regresyon Modelinde Kullanılan Değişkenler

	B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
							Lower	Upper
Automatic_Payment_FLG	,874	,118	54,764	1	,000	2,397	1,901	3,021
Roaming_FLG	-,466	,151	9,579	1	,002	,627	,467	,843
Inactive_Line_CNT	,098	,041	5,588	1	,018	1,103	1,017	1,196
Black_List_FLG	,512	,124	17,010	1	,000	1,668	1,308	2,128
Invoice_L6_AVR	,004	,001	5,929	1	,015	1,004	1,001	1,006
Overdue_L6_FLG	-,617	,243	6,426	1	,011	,540	,335	,869
MT_Barred_CNT	,285	,068	17,532	1	,000	1,330	1,164	1,519
Usage_Risk_RT	2,643	,306	74,510	1	,000	14,060	7,715	25,624
Invoice_Paid_AVG_LogN	,770	,111	47,654	1	,000	2,159	1,735	2,686
Invoice_Paid_Max_AMT_LogN	,338	,098	11,978	1	,001	1,403	1,158	1,699
Invoice_Voice_L6_AVR6_LogN	-,219	,040	29,510	1	,000	,803	,742	,869
Overdue_Invoice_AVR_LogN	,316	,043	53,878	1	,000	1,371	1,260	1,492
Overdue_L6_AVR_LogN	,687	,060	130,508	1	,000	1,989	1,767	2,238
Overdue_Max_L6_AMT_LogN	,152	,066	5,265	1	,022	1,164	1,022	1,326
PROPN_Invoice_LogN	,421	,098	18,373	1	,000	1,524	1,257	1,847
Risk_Unpaid_AMT_LogN	,856	,083	106,343	1	,000	2,354	2,000	2,770
Risk_RT_LogN	,837	,060	197,449	1	,000	2,309	2,055	2,595
Risk_AMT_LogN	-1,545	,096	257,091	1	,000	,213	,177	,258
Constant	-1,789	,354	25,540	1	,000	,167		

a. Variable(s) entered on step 1: Gender, Phone_CNT, Automatic_Payment_FLG, Roaming_FLG, Active_Line_CNT, Inactive_Line_CNT, Black_List_FLG, Invoice_L6_AVR, Invoice_Paid_CNT, Overdue_L6_FLG, Overdue_Sequential_L12_CNT, MO_Barred_CNT, MO_Barred_L6_CNT, MT_Barred_CNT, Usage_Risk_RT, Tariff_gr, Invoice_Paid_AVG_LogN, Invoice_Paid_Max_AMT_LogN, Invoice_National_L6_AVR_LogN, Invoice_Voice_L6_AVR6_LogN, Overdue_Invoice_AVR_LogN, Overdue_L6_AVR_LogN, Overdue_Max_L6_AMT_LogN, PROPN_Invoice_LogN, Risk_Unpaid_AMT_LogN, Risk_RT_LogN, Risk_AMT_LogN.

LR katsayılarının anlamları doğrusal regresyon katsayılarının anlamları ile örtüşmez. B sütunundaki değerler tahmin edicilerin modele katkılarının test edilmesi için uygun olmakla birlikte Exp(B) sütunundaki değerlerin yorumlanması daha kolaydır. Exp(B) sütunu Odds oranını göstermektedir. Exp(B), modeldeki diğer değişkenlerin sabit tutulması durumunda, tahmin edicideki bir birim değişimin neden olduğu Odds oranındaki değişim oranını göstermektedir. Örneğin, Inactive_Line_CNT değişkeninin Odds oranı olan 1,103, diğer tüm şartlar aynı iken, halihazırda sahip olduğu aktif olmayan hat adedi iki olan müşterinin, sahip olduğu aktif olmayan hat adedi bir olan müşteriye göre 1,103 kat üstün olduğunu söyler. Odds oranı, X bağımsız değişkeni değiştiğinde olayın olma ihtimalinin artıp azaldığını göstermektedir. Eğer Odds oranı birden küçük ise, ilgili değişken arttığında olayın olma olasılığı azalmaktadır. Aynı şekilde, Eğer Odds oranı birden büyük ise, ilgili değişken arttığında olayın olma olasılığı da artmaktadır. Buna göre, aktif olmayan hat adedinin artması – beklendiği üzere - abonenin riskli olma olasılığını artırmaktadır.

Diğer değişkenler sabit iken toplam riskinin tamamı ödenmemiş faturalardan kaynaklanan abonenin risksiz olma olasılığı, toplam riskinin tamamı air-time dan oluşan aboneye göre 14 kat fazladır. Otomatik ödeme talimatı olan bir abonenin risksiz olma olasılığı, otomatik ödeme talimatı olmayan bir aboneye göre 2.4 kat daha iyidir. Son 6 ayda hiç kısıtlanmayan abonenin, en az bir defa kısıtlanan aboneye göre risksiz olma ihtimali de yaklaşık olarak iki kat (1 / 0,540) daha yüksektir.

6.3.3.2. Karar Ağacı Tekniği

6.3.3.2.1. Giriş

Karar ağaçları, sınıflama ve tahmin amacıyla kullanılan güçlü ve popüler yöntemlerdir. Anlaşılmasının kolay olması ve görsel olarak ifade edilebilmesi, bu tekniğin en önemli tercih nedenleridir. Karar ağacı bir ağaç yapısı şeklinde olan sınıflayıcıdır. Veriyi herhangi bir kayba yol açmadan alt gruplara ayırmaktadır. Her bir düğüm (node) ya hedef özelliğinin ve gözlemlerin sınıfını gösteren bir yaprak düğümüdür, ya da alt dalların bağlı olduğu bir gövde düğümüdür. Bir karar ağacı, ağacın kökünden başlayarak ve bir yaprak düğümüne ulaşıncaya kadar ilerlenerek verilen bir gözlemi sınıflamak için kullanılabilir. Her yaprak amaca bağlı olarak segmentasyon veya tahmin için kullanılabilir.

Karar Ağacı, bir kurallar kümesinden oluşur. İç içe geçmiş Eğer / İse (IF / THEN) kurallarının dizisidir. Bu kurallar entropi, bilgi kazanımı, Gini indeksi ki-kare testi, ölçüm hatası, sınıflandırma oranı gibi istatistiksel kriterleri kullanarak bir veri setinin sınıflandırılması için faydalanılan kriterler tarafından oluşturulur. Bu yöntemler bağımlı değişkenin veri tipine göre değişmektedir.

Karar Ağacı modeli kurmak için yapılan iki işlem, ağaç oluşturma ve sonradan budama işlemidir. Bu işlemlerde yapılacak seçimler tecrübeye ve veri setine bağlı olmak üzere şunlardır; düğümün parçalanması için kullanılacak algoritma, düğümün maksimum bölünme sayısı, derinlik, minimum son yaprak büyüklüğüdür. Genel olarak kullanılan beş karar ağacı algoritması da şunlardır; CART, CHAID, ID3, C4.5 ve C5.0. Karar ağaçlarını oluşturmak için geliştirilmiş algoritmaların çoğu olası karar ağaçlarının yukarıdan aşağıya çalıştırılan bir çekirdek algoritmanın varyasyonlarıdır.

Bir karar ağacı her bir düğüm saf hale gelene kadar büyütülebilir, başka bir ifade ile yaprak düğümleri daha ileri bölünemez ve her bir yaprak düğümü içindeki üyeler sadece bir sınıfa ait olur. En büyük sınıflandırma ağacı eğitim veri setinde % 100 doğruluk verir, fakat bu aşırı uyumun bir sonucudur ve test veri setinde zayıf bir tahmin verir. Ağaç karmaşıklığı yaprakların sayısının, bölümlerin sayısının ve ağacın derinliğinin bir fonksiyonudur. İyi uyumlu bir ağaç, eğitim veri seti ile doğrulama ve test veri setleri arasında düşük bir yanlılık ve düşük bir varyansa sahiptir. Bir ağacın doğru boyutlandırılması, yani aşırı öğrenmesinden kaçınmak, ileriye yönelik büyümesini durdurarak veya ağacı tam boyutuna kadar büyütüp sonradan geriye budayarak yapılabilir. Bütün ağacı budayarak küçültmek; eğitim veri seti için tahmin hatasını artırır, ancak genel olarak test veri seti için daha iyi bir tahmin gücü sağlar.

6.3.3.2.2. CHAID Yöntemi

Uygulamada kullanılan veri setine en fazla uyan Karar Ağacı algoritması CHAID metodudur. CHAID algoritması homojenlik testinin önem derecesi testine dayalı olarak kesin bir durdurma kriteri üretmek için ki-kare testini kullanır; hipotez X^2'nin büyük değerleri için reddedilir. Eğer homojenlik belirli bir nod için ret edilirse, bölünme devam eder, aksi takdirde nod son aşama (terminal nod) olur. CART algoritmasından farklı olarak, CHAID fazla dalları atma mekanizmasından ziyade ki-kare testinin önem derecesine dayalı durdurma kriteri kullanarak ağacın büyümesini durdurmayı tercih eder[151].

CHAID karar ağaçları algoritmasının içinde X^2 uzaklığı kullanılır. Bir *g* dağılımı ve bir *f* hedefi arasındaki Ki-kare uzaklığı aşağıdaki gibi ifade edilir ve Pearson istatistiğinin genelleştirilmiş şekline benzer. Bu uzaklık kategorik verinin olduğu durumlarda betimsel ve öngörüsel problemler için entropi uzaklığına alternatif olarak kullanılır. Zorunlu olarak temelini oluşturan bir olasılık modeli gerektirmez[152].

$$\chi^2 \mathrm{d} = \sum_i \frac{(f_i - g_i)^2}{g_i}$$

CHAID için özgün gerekçe, değişkenler arasındaki istatistiksel ilişkiyi keşfetmektir. Bir karar ağacı kurarak bunu gerçekleştirir, böylece yöntem aynı zamanda sınıflandırma aracı olarak da kullanılabilir. CHAID yöntemi, ki-kare testinden birkaç açıdan faydalanır; birincisi, hedef

[151] Giudici, P. s. 107.
[152] Giudici, P. s. 189.

değişken üzerinde önemli derecede farklı etkisi bulunmayan sınıfları birleştirir; sonra en iyi ayrım seçilir ve son olarak da nod üzerinde ek bir ayrım yaratmaya değip değmeyeceğine karar verilir. Ki-kare testi kategorik verilere uygulanır. Bu nedenle klasik CHAID algoritmasında girdi değişkenler kategorik olmalıdır. Sürekli değişkenler bölmelere ayrılmalı veya yüksek, orta, düşük gibi ordinal sınıflar ile yer değiştirilmelidir[153]. Sürekli değişkenler de otomatik olarak sınıflandırıldıktan sonra ağacın dalları oluşturulmaktadır.

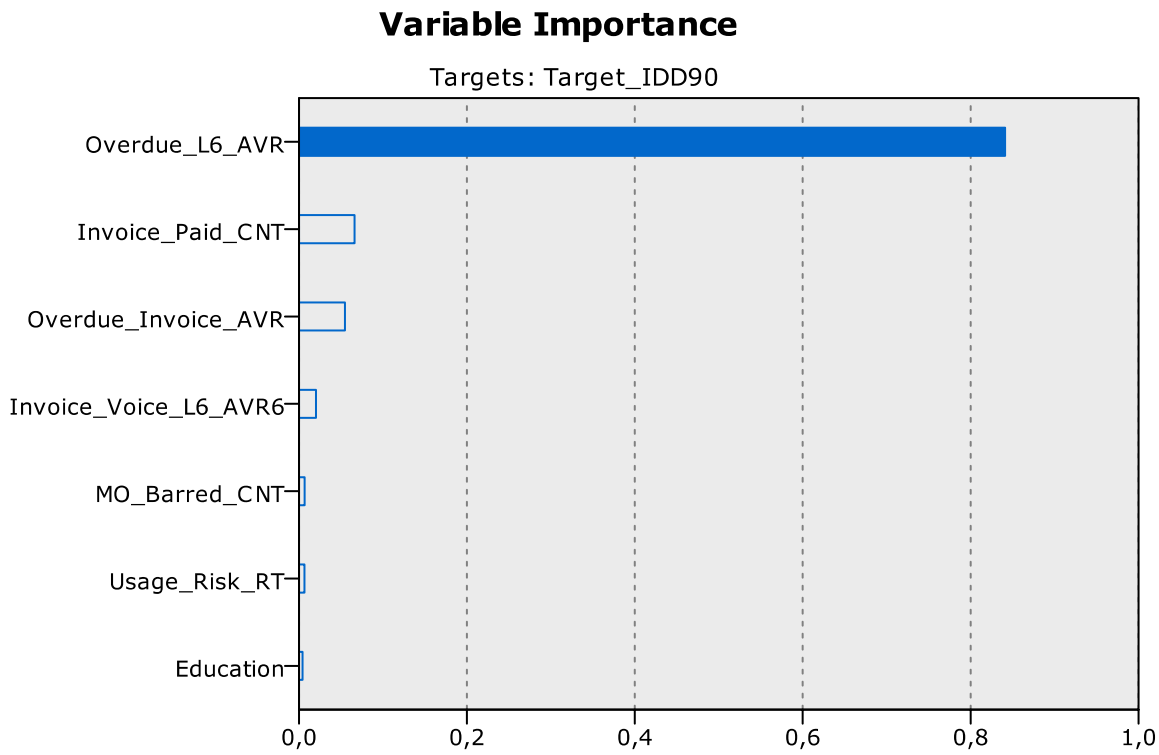
6.3.3.2.3. Karar Ağacı Modellerinin Uygulanması

Uygulama veri seti üzerinde C&RT, Quest ve CHAID algoritmaları ile farklı karar ağaçları modelleri denenmiştir. Tüm modellerde Minimum ara node= 400, Minimum terminal node=200 olarak belirlenmiştir. Oluşturulan bu modellerden doğruluk oranı en yüksek olan CHAID algoritması ile oluşturulan model tercih edilmiştir. CHAID yöntemine göre yapılan modellerin özet tabloları aşağıdadır:

Bu şekilden de görüleceği gibi, abonenin riskli olup olmayacağını birinci sırada belirleyen ödeme performansını gösteren bir değişkendir. Overdue_L6_AVR, son altı fatura döneminde çıkan ve geç ödenen faturalarının ortalama geç ödenme gün sayısını göstermektedir.

İkinci derecede önemli değişken Invoice_Paid_CNT, abone tarafından ödenen tüm faturaların toplam adedini göstermesine rağmen aynı zamanda ne kadar uzun süreden beri aynı şebekede abone olarak bulunduğunu da bildirmektedir. Karar Ağacı incelendiğinde de görüleceği gibi ödenen fatura adedi ne kadar yüksek ise abonenin riskli olarak tanımlanma olasılığı da o kadar düşüktür.

[153] Berry, Michael J.A. and Gordon S. Linoff. s. 183.



Şekil 19 : Karar Ağacı Modelinde Kullanılan Değişkenlerin Modeli Açıklama Gücü Grafiği

Model, yedi adet değişkeni anlamlı bulmuş ve bu değişkenlerin aldığı değerlere göre karar ağacını oluşturmuştur. Aşağıda bu değişkenlere CHAID modeli tarafından atanan önem dereceleri bulunmaktadır. Buna göre model büyük oranda bir değişkene dayalı olarak çalışmaktadır. Bu durumun bir dezavantaj olduğu söylenebilir. Ancak diğer modeller tarafından da bu değişkenin anlamlı çıkması bu dezavantajın etkisini çok azaltmaktadır.

Tablo 20 : Karar Ağacı Modelinde Kullanılan Değişkenlerin Modeli Açıklama Gücü

Değişkenler	Önem Derecesi	Sıralama
Overdue_L6_AVR	84,10%	1
Invoice_Paid_CNT	6,63%	2
Overdue_Invoice_AVR	5,48%	3
Invoice_Voice_L6_AVR6	2,03%	4
Usage_Risk_RT	0,66%	5
MO_Barred_CNT	0,68%	6
Education	0,42%	7

Tablo 29'de görüldüğü gini Karar Ağacı modelinde en anlamlı çıkan değişken, fatura ödeme alışkanlığını gösteren; son altı fatura döneminde çıkan ve geç ödenen faturalarının ortalama geç ödenme gün sayısı (Overdue_L6_AVR) parametresi olmuştur. Sonucu belirleyen üçüncü en önemli değişken de fatura ödeme zamanı ile ilgilidir. Ancak, özellikle abonelerin son dönemlerde çıkan faturalarının ödemelerinin zamanında yapılıp yapılmamasının tüm abonelik

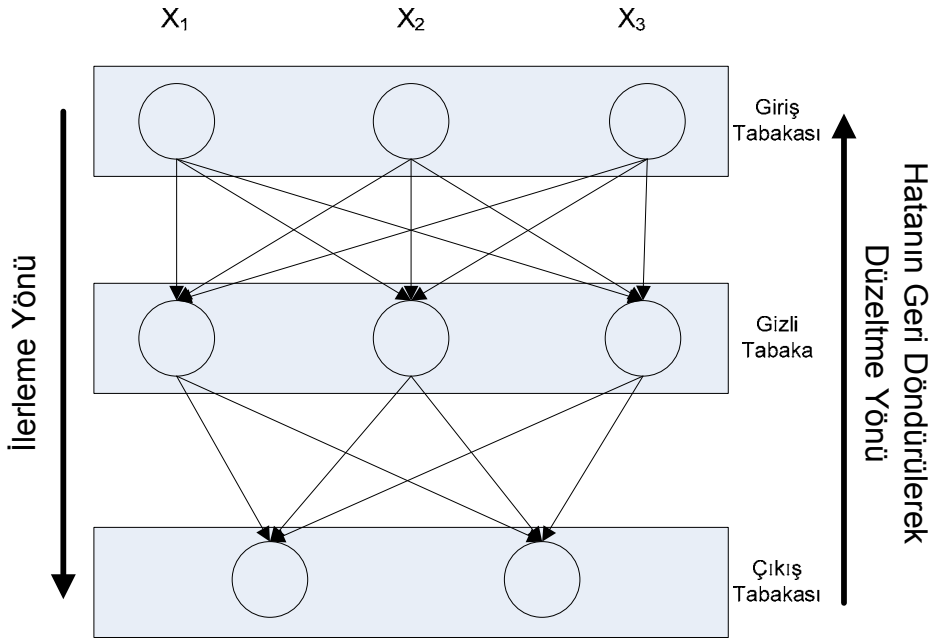
boyunca yapılan ödeme alışkanlığına göre davranış skorlama sonucu üzerinde çok daha önemli olduğu ortaya çıkmıştır. İkinci sırada önemli çıkan değişken olan ödenen fatura adedi, abonenin şebekede ne kadar uzun süreden bu yana bulunduğunu göstermektedir.

CHAID algoritmasına göre oluşturulan Karar Ağacına ait kurallar seti EK-3'de, Karar Ağacının görüntüsü de EK-4'de yer almaktadır. KA modelinin basit olduğu kadar etkili bir yöntem olduğu buradaki doğru sınıflandırma oranlarına bakılarak ileri sürülebilir.

6.3.3.3. Yapay Sinir Ağı Tekniği

6.3.3.3.1. Giriş

Veri madenciliğinin dayandığı Yapay Zekâ (YZ) (Artificial Intelligence - AI), istatistiğin tersine höristik (heuristics) temele dayanır ve insan zekâsının işleyiş temellerini taklit ederek sonuca gider. Bu alanın ilerlemesi için de yüksek hızdaki bilgisayar altyapısının gelişmesi gerekmiştir. Yapay Zekâ, iş problemlerine insan zekâsına benzer yaklaşımlarla çözüm üretmeye çalışır. Makine öğrenmesi, istatistik ve yapay zekâyı birleştirerek, veri kalitesine bağlı olarak birbirinden farklı çözümler üretmeye çalışır.



Şekil 20 : Yapay Sinir Ağ Modeli Şekli

Yapay sinir ağları geleneksel istatistiksel skor kart tekniklerini sınırlayan birçok kısıttan kaynaklanan sıkıntı çekmezler ve aşağıda sayılan avantajları mevcuttur[154]:

- Kurum içi öngörüsel model kurmaya izin verir
- Modelin çoklu çıktı problemleri ikili çıktı modelleri kadar kolaydır
- Skor kart oluşturma sürecini otomatik hale getirir
- İş alanı uzmanlarının skor kart hazırlama sürecine katılmalarına imkân sağlar
- Tahmin doğruluğunu artırır
- Hızlı ve doğru skor kart uygulamasına imkân verir
- Tahmin edici değişkenler arasındaki doğrusal olmama ve karşılıklı bağımlılık durumlarını hesaba katar

6.3.3.3.2. MLP Yöntemi

Çok katmanlı algılayıcı (Multi-Layer Perceptron-MLP) içinde, her bir gizli katman nöronu, bir önceki katmandaki nöronun çıktılarının ağırlıklandırılmış kombinasyonuna dayalı bir girdi alır. Sonuç katmanında yer alan nöronlar, sırasıyla çıktıyı üretmek için birlikte hareket ederler. Bu tahmini değer, daha sonra doğru çıktı ile karşılaştırılır ve iki değer arasındaki fark (hata) sırasıyla güncellenen ağ içine geri beslenir. Bu hatanın ağa geri beslenmesi geri yayılım (back-propagation) olarak adlandırılır[155].

Pek çok yazar, YSA modellerinde bir adet gizli katman kullanmanın karmaşık sistemlerde bile yeterli doğruluğu sağlayacağını savunmakta, bir diğer kısmı da ağırlık adedinin eğitim veri setinde yer alan değişkenlerden daha fazla olmamasını tavsiye etmektedirler[156]. Bir yapay sinir ağının gizli katmanı içindeki nöronların optimum sayısını belirlemek için kullanılabilecek formal bir kılavuz yoktur. Yapay sinir ağı modelleri, hata yüzeyinin eğimi yerel bir minimuma ulaştığı zaman biten öğrenme nedeniyle doğru olmayabilir. Bir yapay sinir ağının hata yüzeyi doğrusal bir modelden daha karmaşıktır ve küresel minimumdan daha yukarıda bir yerel minimum ile sonuçlanabilir[157].

---

[154] Pearce W., "Neural Scoring". In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. s. 140.
[155] Clementine 12.0 Ders Notları, 2.2
[156] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan., s. 4738.
[157] Trinkle, Brad S. and Amelia A Baldwin., 145.

Eğitim ve test alt kümelerinin uygun şekilde seçilmesi, optimum YSA mimarisinin tasarımı için önemlidir. Buradaki kilit nokta yükümlülükleri yerine getirilen ve getirilmeyen borçların dağılımları arasındaki ilişkidir. Ödenmeyen borçların yüzdesinin ödenen borçlar ile karşılaştırıldığında genel olarak küçük olması gerçeğinden dolayı rassal örneklem seçildiğinde kredi başvuruları hakkında önemli bir bilgi kaybı oluşabilir[158].

6.3.3.3.3. YSA Modellerinin Uygulanması

Uygulamada yer alan veri setine farklı Yapay Sinir Ağ algoritmaları uygulanmıştır. İki çeşit denetlemeli sinir ağı modeli kullanılmıştır: Multi-Layer Perception (MLP) ve Radial Basis Function Network (RBFN). MLP için kullanılan İleri beslemeli geri yayınım modeli (feed-forward back-propagation network) ile bulunan sonuçlar daha tatmin edici olmuştur.

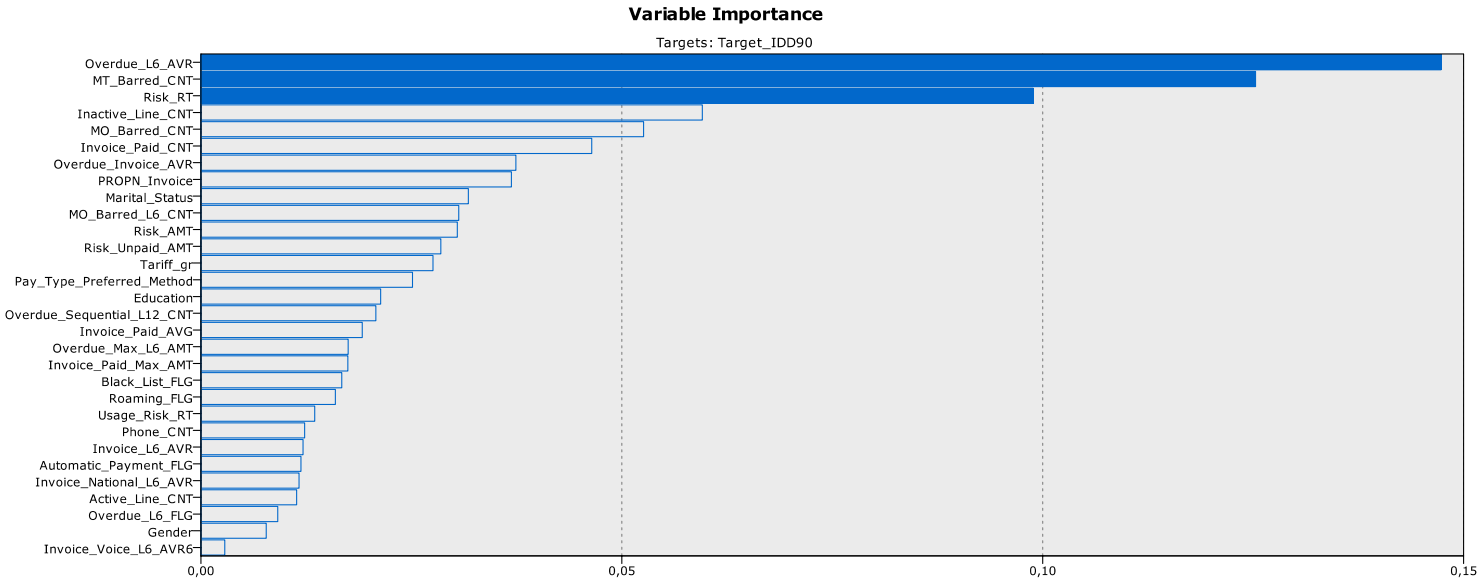
Uygun yapay sinir ağı mimarisinin belirlenmesinde, daha önce literatürde kredi skorlamada yaygın olarak kullanılan algoritmalar kullanılmış, en uygun sonucu almak için de model performansları arasındaki farklılığın kabul edilebilir seviyelere ininceye kadar deneme yapılmıştır. Yapılan denemelerde gizli katman sayısı, gizli katmanlardaki düğüm adedi, öğrenme oranı, momentum terimi, aktivasyon fonksiyonu, devir sayısı gibi çeşitli parametrelerin kombinasyonları denenerek, test seti üzerinde de yüksek performans gösteren modeller oluşturulmuş ve bunlar içerisinden en iyisi alınmıştır. Modellerin aşırı öğrenmesinin önüne geçmek için doğrulama ve test veri setleri de kullanılmış ve modellerin her üç veri setinde birbirine yakın sonuçlar verip vermediğine dikkat edilmiştir.

Katman sayısının artırılması sinir ağının daha kompleks ilişkileri öğrenebilmesini sağlar, ancak modelin eğitim zamanının uzaması yanında aşırı öğrenmeye de yol açabilir. Aşırı öğrenmenin önüne geçmek için iki katman kullanılmıştır. Birinci katmanda 20, ikinci katmanda 15 nöron bulunmaktadır. Girdi katmanında da 45 nöron bulunmaktadır.

En iyi sonuç veren Yapay Sinir Ağ modeli ile ilgili özet sonuçlar aşağıda yer almaktadır. Modelde kullanılan en anlamlı parametre diğer ikisinde de olduğu gibi Overdue_L6_AVR değişkeni olmuştur. Bu çıktı, veri setinin iyi hazırlanmış, modellerin de birbirleri ile tutarlı sonuçlar ürettiğinin bir göstergesi sayılabilir.

[158] Šušteršicˇ, M., Dušan Mramor, and Jure Zupan., s. 4741.

Şekil 21'de verildiği gibi, değişkenlerin önem derecelerine baktığımızda ikinci sırada aranmaya kısıtlanma adedi (MT_Barred_CNT) yer almaktadır. Aboneler çıkan faturalarını ödemedikleri durumda uygulanan önemli aksiyonlardan bir tanesi aranmaya kısıtlanmadır. Bir abonenin daha önceki faturalarda aranmaya kısıtlanma zamanına kadar ödeme yapmaması hareketini tekrarlama eğiliminin, ödemesini geciktirmediği için kısıtlanmayan aboneye göre daha yüksek olduğu söylenebilir.



Şekil 21 : Yapay Sinir Ağ Modelinde Kullanılan Değişkenlerin Modeli Açıklama Gücü Grafiği

Risk_RT değişkeni LR da olduğu gibi burada üçüncü önemli değişken olarak yer almıştır. Dolayısıyla aynı miktarda riski olan aboneleri birbirleri ile karşılaştırırken hangisinin yaptığı ödeme miktarı daha fazla ise onun daha az riskli olduğu ileri sürülebilir.

YSA modelinde yer alan değişkenlerin açıklama güçleri Tablo 21'de verilmiştir. Yukarıda aktarılan benzerliklerin de ötesinde dönüştürülüp dönüştürülmediğinden bağımsız olarak LR modelinde anlamlı olarak çıkan 18 değişkenin tamamı YSA modelinde de yer almaktadır.

Tablo 21 : Yapay Sinir Ağı Modelinde Kullanılan Değişkenlerin Modeli Açıklama Gücü

Değişkenler	Önem Derecesi	Sıralama
Overdue_L6_AVR	14,73%	1
MT_Barred_CNT	12,53%	2
Risk_RT	9,89%	3
Inactive_Line_CNT	5,96%	4
MO_Barred_CNT	5,26%	5
Invoice_Paid_CNT	4,64%	6
Overdue_Invoice_AVR	3,74%	7
PROPN_Invoice	3,69%	8
Marital_Status	3,18%	9
MO_Barred_L6_CNT	3,06%	10
Risk_AMT	3,05%	11
Risk_Unpaid_AMT	2,85%	12
Tariff_gr	2,76%	13
Pay_Type_Preferred_Method	2,51%	14
Education	2,14%	15
Overdue_Sequential_L12_CNT	2,08%	16
Invoice_Paid_AVG	1,92%	17
Invoice_Paid_Max_AMT	1,75%	18
Overdue_Max_L6_AMT	1,75%	19
Black_List_FLG	1,67%	20
Roaming_FLG	1,60%	21
Usage_Risk_RT	1,35%	22
Phone_CNT	1,23%	23
Invoice_L6_AVR	1,21%	24
Automatic_Payment_FLG	1,19%	25
Invoice_National_L6_AVR	1,17%	26
Active_Line_CNT	1,14%	27
Overdue_L6_FLG	0,91%	28
Gender	0,78%	29
Invoice_Voice_L6_AVR6	0,28%	30

6.4. Değerlendirme

Üretilen modellerin gerçek iş problemine ne ölçüde çözüm getirdiği değerlendirildikten sonra uygulamaya alınabilir. Bu değerlendirme, modelin ayırım gücünü ve doğruluğunu ölçmeli, ayrıca model ile doğrulama yapılmalıdır.

6.4.1. Modellerin Performanslarının Karşılaştırmalı Olarak Değerlendirilmesi

Modelin ayırım gücü kadar, hata tipleri de model seçiminde kullanılırlar. Kredi skorlama modelinde iki çeşit hata yapılabilir. Bunlardan birincisi gerçekte riskli olan abonenin risksiz olarak değerlendirilmesidir. Tip-I hata (α hatası) da denen bu durumda, kredi veren tarafın maliyeti kredi miktarının ve/ veya faiz tutarının kaybından oluşur. İkinci olarak da model

gerçekte riski düşük olan abonenin yüksek riskli olarak değerlendirilmesidir ki buna da Tip-II veya (β hatası) denir. Bu ikinci tip hatadan kaynaklanan potansiyel kayıplar; başlıca gelir kayıplarını içerir ve pazar payı düşüşlerine de yol açabilir.

Tablo 22 : Model Sınıflandırma Tablosu

	Tahmini Risksiz	Tahmini Riskli	Toplam Gerçek Olasılık
Risksiz	**A (doğru negatif)**	**B (Yanlış pozitif)**	**A + B**
Riskli	**C (yanlış negatif)**	**D (Doğru pozitif)**	**C + D**
Toplam Tahmini	**A + C**	**B + D**	**A + B + C + D**

Aşağıdaki sınıflandırma ölçütleri tablodaki ilişkilerden türetilmiştir:

Doğru Sınıflandırma Oranı : 100 * (A + D) / (A + B + C + D)
Yanlış Sınıflandırma Oranı : 1 - (100 * (A + D) / (A + B + C + D))
Duyarlılık (Doğru Pozitif Oran) : 100 * D / (C + D)
Belirlilik (Doğru Negatif Oran) : 100 * A / (A + B)
1 - Belirlilik (Yanlış Pozitif Oran) : 100 * B / (A + B)

Modelin toplam yaptığı doğru tahmin oranına Doğru Sınıflandırma (Accuracy) Oranı denir. Buradaki doğruluk; risksiz abonenin risksiz, riskli abonenin de riskli olarak tahmin edilmelerini kastetmektedir. Bu oranın tam tersi olan Yanlış Sınıflandırma (Misclassification) Oranı, doğru olarak sınıflandırılmayan aboneleri kapsamaktadır, dolayısıyla bu iki oranın toplamı bire eşittir.

Doğru Sınıflandırılmış Gözlemlerin Yüzdesi (DSGY) (PCC-Percentage Correctly Classified) veri setinin bir örneklemi içindeki doğru sınıflandırılmış örneklerin oranını ölçer. Bir dizi örnekte, DSGY en uygun performans kriteri olmayabilir. Açıkça söylenmeden, yanlış-pozitif ve yanlış-negatif öngörüler için eşit yanlış sınıflandırma maliyetini kabul eder. Bu varsayım problemlidir, çünkü çoğu gerçek yaşamdaki problemler için, bir tip sınıflandırma hatası diğerinden çok daha fazla maliyetli olabilir. Bir değerlendirme kriteri olarak DSGY'nin kullanımı ile ilgili diğer bir üstü kapalı varsayım da örneklemler arasındaki sınıf dağılımı (sınıf öncelikleri) zamanla sabit varsayılır ve göreli olarak dengelenir. Böylece, sadece DSGY kullanmak sıklıkla yetersiz çıkmaktadır, çünkü sınıf dağılımları ve yanlış sınıflandırma

maliyetleri nadiren üniformdur. Bununla birlikte, sınıf dağılımlarını ve yanlış sınıflandırma maliyetlerini hesaba katmak oldukça zordur, çünkü pratikte nadiren kesin olarak belirtilebilmektedir ve çoğunlukla değişime maruz kalmaktadır[159].

Duyarlılık (Sensitivity) veya doğru pozitif oranı, gerçekte riskli olan abonelerin model tarafından ne kadarlık bir başarıyla tespit edildiğini ölçer. Buna benzer şekilde doğru negatif oranı olarak da isimlendirilen Belirlilik (Specificity) katsayısı da, modelin gerçekte risksiz olan aboneler arasından risksiz olanları ayırma kabiliyetini ölçer.

Bu çalışmada; Lojistik Regresyon Modeli, Karar Ağacı ve istatistik modelleri dışında tanımlanan Yapay Sinir Ağı modeli kurulmuştur. Her üç modelin sonuçlarının karşılaştırıldığı tablo ve grafikler aşağıda yer almaktadır.

Birinci olarak her bir modelin tüm veri seti içerisindeki doğruluk oranları ayrı şekilde değerlendirilerek Tablo 23, 24 ve 25'te sunulmuştur. Tablolarda yer alan gözlemler; eğitim, doğrulama ve test veri setlerini içermektedir. Her üç modelde de gözlemlere atanan olasılıklar eşik değeri 0,5'e göre gruplara atanmıştır. Modellerin doğru sınıflandırma oranlarının yanı sıra duyarlılık ve belirlilik oranları da hesaplanarak birbirleri ile karşılaştırılmışlardır.

Tablo 23 : Lojistik Regresyon Modeli Sınıflandırma Tablosu

		Gerçek Durum		
		Risksiz	**Riskli**	**Toplam**
LR Model Sonucu	**Risksiz**	6.346	1.035	7.381
	Riskli	654	1.965	2.619
	Toplam	7.000	3.000	10.000

Doğru Sınıflandırma Oranı : % 83.11
Yanlış Sınıflandırma Oranı : % 16.89
Duyarlılık (Doğru Pozitif Oran) : % 65.50
Belirlilik (Doğru Negatif Oran) : % 90.66
1 - Belirlilik (Yanlış Pozitif Oran) : % 9.34

[159] Baesens B. ve diğerleri, s. 631.

Tablo 24 : Yapısal Sinir Ağı Modeli Sınıflandırma Tablosu

		Gerçek Durum		
		Risksiz	**Riskli**	**Toplam**
YSA Model Sonucu	**Risksiz**	6.360	1.067	7.427
	Riskli	640	1.933	2.573
	Toplam	7.000	3.000	10.000

Doğru Sınıflandırma Oranı : % 82.93
Yanlış Sınıflandırma Oranı : % 17.07
Duyarlılık (Doğru Pozitif Oran) : % 64.43
Belirlilik (Doğru Negatif Oran) : % 90.86
1 - Belirlilik (Yanlış Pozitif Oran) : % 9.14

Tablo 25 : Karar Ağacı Modeli Sınıflandırma Tablosu

		Gerçek Durum		
		Risksiz	**Riskli**	**Toplam**
KA Model Sonucu	**Risksiz**	6.287	1.164	7.451
	Riskli	713	1.836	2.549
	Toplam	7.000	3.000	10.000

Doğru Sınıflandırma Oranı : % 81.23
Yanlış Sınıflandırma Oranı : % 18.77
Duyarlılık (Doğru Pozitif Oran) : % 61.20
Belirlilik (Doğru Negatif Oran) : % 89.81
1 - Belirlilik (Yanlış Pozitif Oran) : % 10.19

Modeller tüm verilere uygulandığında en yüksek doğruluk oranı LR (%83.11) modeline aittir. Bunu sırasıyla YSA (%82.93) ve KA (%81.23) modelleri izlemektedir. Bu modellerin aralarındaki farklılık beklendiği gibi düşüktür.

Aynı şekilde, gerçekte riskli olan abonelerin en fazla oranda doğru tahminini gösteren Duyarlılık oranı açısından da LR modeli en yüksek orana (%65.50) sahiptir. Duyarlılık oranı açısından da LR modelini ikinci olarak YSA (%64.43), üçüncü olarak da KA (%61.20) takip etmektedir. Riskli abonelerin tespitinin çok önem kazandığı durumlarda veya dönemlerde de LR modeli, diğerlerine göre daha avantajlıdır. Skorlama modellerinin müşteri veya abone bazında kullanılmasının önem kazandığı durumlarda da her gözlem için farklı bir skor atayan LR ve YSA modelleri, KA modeline göre daha avantajlıdır.

Risksiz olan abonelerin doğru olarak sınıflandırılmasının ölçütü olan Belirlilik oranına baktığımızda ise ilk sırada YSA modeli (% 90.86), ikinci sırada LR modeli (% 90.66) ve üçüncü sırada da KA modeli (% 89.81) bulunmaktadır. Kredi Risk Departmanlarının amacı

mümkün olduğunca çok sayıda riskli aboneyi tespit etmek iken aynı zamanda da risksiz olan aboneye de müdahale etmemektir. Risksiz aboneye müdahale oranı arttıkça kredi verilebilecek veya potansiyel müşteri kitlesinin firmadan uzaklaştırılması anlamına gelir. Bunun birinci sakıncası doğrudan ilgili işlemlerden kaynaklanacak gelir kaybı iken ikinci ve daha önemli sonucu ise firmanın imajının olumsuz etkilenmesi dolayısıyla gelecekteki kar fırsatlarının kaçırılmasıdır. Bizim örneğimizde ilk iki model arasındaki farklılık sadece %0,2'dir. Bu fark nedeniyle YSA'nın tercih edilmesi beklenmemelidir.

Modeller ikinci olarak, veri bölümlemesine göre tüm modellerin her bir alt veri seti için doğruluk oranları belirlenerek değerlendirilmişlerdir. Modellerin doğruluk oranlarının farklı veri setlerine göre doğru ve yanlış tahmin oranları Tablo 26, 27 ve 28'de yer almaktadır:

Tablo 26 : LR Modelinin Alt Veri Setleri İçindeki Doğruluk Oranlarının Karşılaştırılması

Model Sonucu	Eğitim Verisi		Test Verisi		Doğrulama Verisi	
	Adet	Oran(%)	Adet	Oran(%)	Adet	Oran(%)
Doğru	4.946	83,13%	2.529	82,89%	836	83,68%
Yanlış	1.004	16,87%	522	17,11%	163	16,32%
Toplam	5.950	100,00%	3.051	100,00%	999	100,00%

Tablo 27 : YSA Modelinin Alt Veri Setleri İçindeki Doğruluk Oranlarının Karşılaştırılması

Model Sonucu	Eğitim Verisi		Test Verisi		Doğrulama Verisi	
	Adet	Oran(%)	Adet	Oran(%)	Adet	Oran(%)
Doğru	4.968	83,50%	2.493	81,71%	832	83,28%
Yanlış	982	16,50%	558	18,29%	167	16,72%
Toplam	5.950	100,00%	3.051	100,00%	999	100,00%

Tablo 28 : KA Modelinin Alt Veri Setleri İçindeki Doğruluk Oranlarının Karşılaştırılması

Model Sonucu	Eğitim Verisi		Test Verisi		Doğrulama Verisi	
	Adet	Oran(%)	Adet	Oran(%)	Adet	Oran(%)
Doğru	4.806	80,77%	2.488	81,55%	829	82,98%
Yanlış	1.144	19,23%	563	18,45%	170	17,02%
Toplam	5.950	100,00%	3.051	100,00%	999	100,00%

Eğitim veri setinde gerçekleşen doğruluk oranının, doğrulama ve özellikle test veri setinde de gerçekleşmesi modelin tutarlılığı açısından çok önemlidir. Eğitim veri setinde en yüksek doğruluk oranı %83.50 ile YSA modeline ait iken, doğrulama ve özellikle de test veri setindeki en yüksek doğruluk oranı %82.89 ile LR modeline aittir.

Modellerin her üç alt veri setindeki doğruluk oranları arasındaki farklılıklar alınıp, bu farklılıkların mutlak değerlerinin ortalamaları karşılaştırıldığında da minimum fark yine LR modeline aittir. Bu açıdan da bakıldığında LR modelinin seçilen veri seti için en tutarlı model olduğu görülmektedir.

6.4.2. Modellerin Performanslarının Toplu Şekilde Karşılaştırması

Tablo 29, her 3 modelin sonuçlarını alt veri setleri ile birlikte değerlendirmektedir. Burada yer alan önemli istatistiklerin anlamları ve bu kriterlere göre farklı modellerin karşılaştırılmaları aşağıda yer almaktadır. En çok kullanılan ROC indeksi, Gini katsayısı ve Lift oranı daha alt bölümlerde ayrıntılı olarak yer almaktadır.

Kolmogorov – Smirnov istatistiği, iki değerli hedef değişken için riskli ve risksiz skorların kümülatif dağılımları arasındaki maksimum dikey ayrımı ölçer. Aşağıdaki tabloya göre, her üç veri seti için maksimum ayrım LR modelinde gerçekleşmiştir. En düşük fark da KA modeline aittir.

Tablo 29 : Tüm Modellerin Alt Veri Setleri İçin Karşılaştırma İstatistikleri

İstatistikler	Lojistik Regresyon			Yapay Sinir Ağı			Karar Ağacı		
	Eğitim	Test	Doğrulama	Eğitim	Test	Doğrulama	Eğitim	Test	Doğrulama
Kolmogorov-Smirnov Statistic	0,63	0,64	0,63	0,63	0,61	0,62	0,56	0,59	0,59
Average Squared Error	0,12	0,12	0,12	0,12	0,13	0,13	0,13	0,13	0,13
ROC Index	0,89	0,89	0,89	0,89	0,88	0,88	0,86	0,87	0,87
Average Error Function	0,38	0,38	0,37	0,39	0,42	0,41	0,41	0,4	0,4
Percent Capture Response	13,5	13,7	15,19	13,33	13,61	13,53	13,68	12,91	13,15
Frequency of Classified Cases	5.950	3.051	999	5.950	3.051	999	5.950	3.051	999
Divisor for ASE	11.900	6.102	1.998	11.900	6.102	1.998	11.900	6.102	1.998
Error Function	4.476	2.349	731	4.684	2.588	821	4.851	2.470	792
Gain	190,87	178,49	207,09	190,87	178,6	190,54	186,66	175,8	188,6
Gini Coefficient	0,78	0,78	0,79	0,78	0,75	0,76	0,73	0,74	0,75
Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0,62	0,64	0,62	0,63	0,61	0,6	0,56	0,58	0,58
Lift	2,7	2,74	3,04	2,67	2,72	2,71	2,74	2,58	2,63
Maximum Absolute Error	0,99	1	0,99	1	1	1	0,99	0,99	0,95
Misclassification Rate	0,17	0,17	0,16	0,17	0,18	0,17	0,19	0,18	0,17
Sum of Frequencies	5.950	3.051	999	5.950	3.051	999	5.950	3.051	999
Root Average Squared Error	0,35	0,35	0,34	0,35	0,36	0,35	0,36	0,36	0,36
Percent Response	80,5	83,55	89,99	79,5	82,96	80,18	81,6	78,73	77,92
Sum of Squared Errors	1.429,5	731,5	229,8	1.424,6	790,2	250,1	1.576,4	797,0	255,4
Number of Wrong Classifications	1004	522	163	982	558	167	1144	563	170

Yanlış sınıflandırma adetlerinin toplamlarına baktığımızda en düşük adet LR modeli tahminlerinde 1689 (1004+522+163), ikinci en düşük adet YSA model tahminlerinde 1707 (982+558+167) ve en yüksek yanlış tahmin adedi de 1877 (1144+563+563) ile KA modeline aittir. Bu adetlerle bağlantılı olarak yanlış sınıflandırma oranları açısından da en düşük oranlar; %17 (eğitim), %17 (test) ve %16(doğrulama) ile LR modeline aittir. Bu istatistiklerle tutarlı olarak toplam hata fonksiyonu, ortalama hata fonksiyonu ve ortalama hata kareleri istatistikleri için de LR en düşük ortalamalara sahiptir.

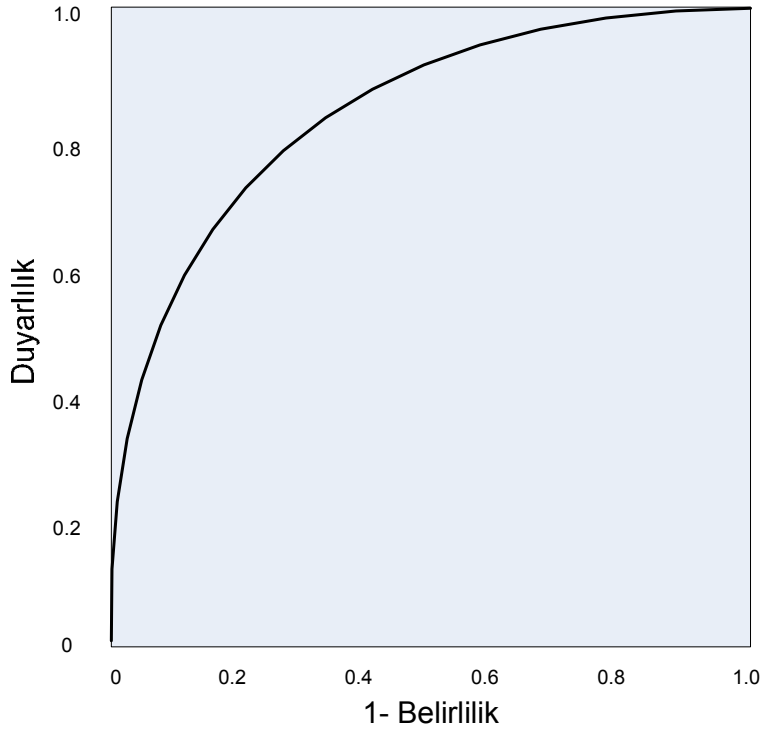
6.4.2.1. ROC İstatistiği

Kredi skorlama modellerinin istatistiksel performansını ölçmek için kullanılan pek çok metod mevcuttur. Bunlardan en çok kullanılanı ROC (Receiver Operating Characteristics) metodudur. ROC grafiği iki şıklı hedef değişkeni olan modellerin genel performanslarını karşılaştırmak için kullanılır. ROC eğrisi modelin tahmini doğruluğunu ölçen grafiksel bir gösterimdir.

Alıcı Operasyon Karakteristiği eğrisi (ROC-Receiver operating characteristic) Y ekseninde duyarlılığa karşın X ekseninde sınıflandırma eşiklerinin çeşitli değerlerinin bulunduğu iki boyutlu grafiksel bir gösterimdir. Temel olarak sınıflandırma dağılımı ve yanlış sınıflandırma maliyetine dikkat etmeksizin bir sınıflandırıcının davranışını gösterir, böylece sınıflandırma performansının bu faktörlerden bağlantısını etkin olarak keser[160].

Eğri üzerindeki her bir nokta bir kesim noktası olasılığını gösterir. Kesim noktası seçimi duyarlılık ve belirlilik katsayıları arasındaki bir değişimi gösterir. Dik eksendeki değer, doğru tespit edilen riskli abonelerin kendi içerisindeki yüzdesini verirken, yatay eksen de yanlış tespit edilen risksiz aboneleri gösterir. İdeal olarak hem duyarlılık hem de belirlilik için yüksek değerlerin olması istenir; böylece model hem risklileri hem de risksizleri daha doğru tahmin edebilir. Düşük bir kesim noktası genel olarak daha fazla yanlış pozitif, yüksek bir kesim noktası da daha fazla yanlış negatif anlamına gelir. İş bilgisi burada devreye girerek amaca en uygun kesim noktası seçilir.

[160] Baesens B. ve diğerleri, s. 631.



Şekil 22 : ROC (Receiver Operating Characteristic) Eğrisi

ROC eğrisi gruplama performansını iki boyutlu olarak ölçer ve konkav olduğu durumda da Kolmogorov-Smirnov istatistiğinden elde edilen bilgileri görsel olarak sunar. ROC içbükey ise ROC eğrisi ve diyagonal çizgi arasındaki en büyük fark, bir sabit çarpımı kadar Kolmogorov-Smirnov istatistiğine eşittir. Eğer ROC eğrisi içbükey değilse, böyle genel bir uygunluk yoktur. Her bir olası t düzeyi kesim noktası için a ve b hataları hesaplanarak oluşturulur. Hataların iki grubu da ROC eğrisinin koordinatlarına denk düşer[161]. Blöchlinger ve Leippold (2006), kar maksimizasyonunun gerçekleştirilebileceği nokta ve fiyatlama eğrisinin belirlenmesi gibi optimal kredi kararlarının verilmesinde ROC eğrisini NBD (Net Bugünkü Değer) analizi ile birleştirilerek türetilebileceğini göstermiştir.

Bir modelin performansını değerlendirmenin alternatif iki yolu olan ROC eğrisi ve Lift eğrileridir. Bu eğriler farklı modellerin birçok uygulama durumu için nasıl performans gösterdiği ile ilgili kavrayış sağlayabilirler. ROC eğrisi, sinyal tespit teorisinden meydana gelmiştir. Doğru pozitif oranına (y-ekseni) karşın yanlış pozitif oranını (x-ekseni) çizer. Eğer bu boşlukta iki model var ise, birisi her ikisinin bağlantı çizgisi üzerinde herhangi bir performansı elde edebilir; çizgi üzerindeki istenen pozisyona oransal olarak bazı olasılıklar ile birlikte rassal olarak modellerin kullanılması ile bu işlem yapılabilir. Bu eğri, değerleme

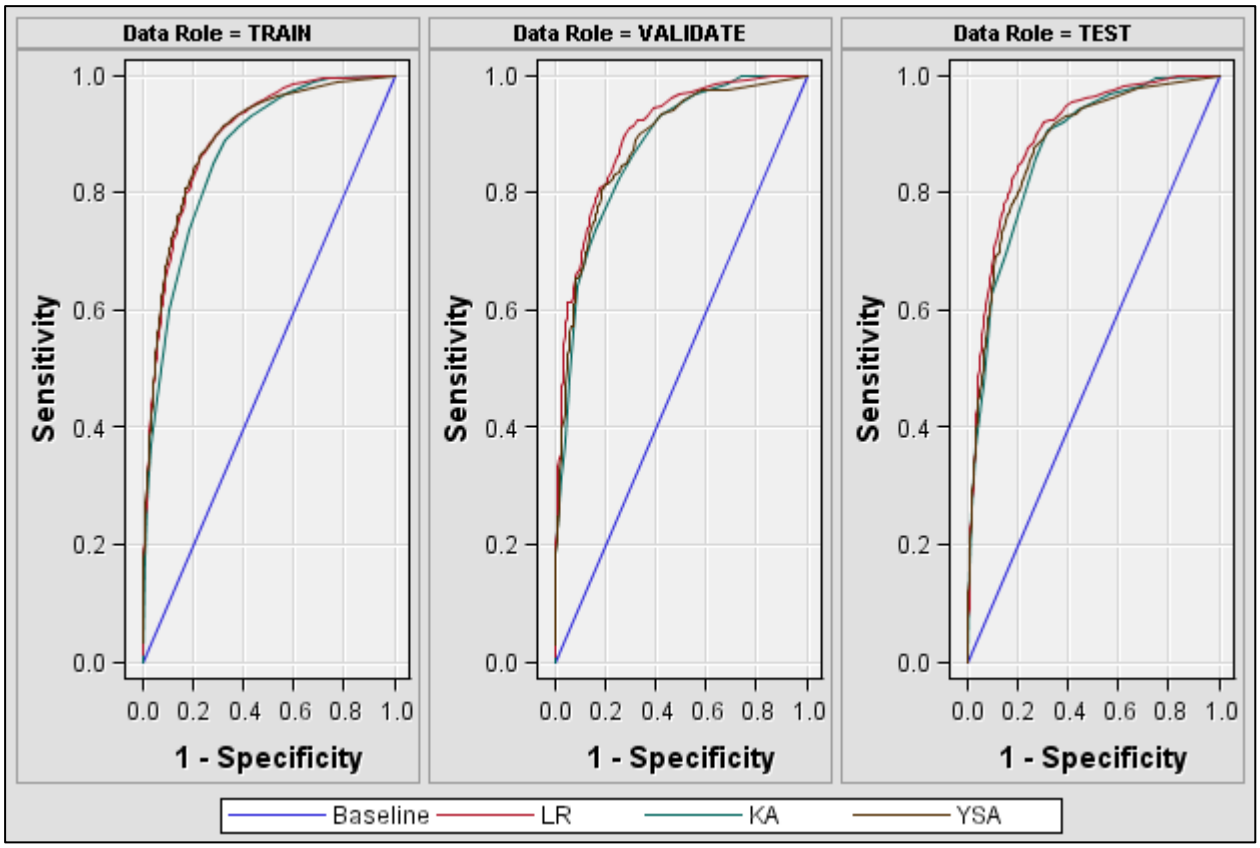
[161] Blöchlinger, A. and Markus Leippold., s. 853.

yapan analiste tahmin zamanında varsayılan sınıf dağılımına bağlı olarak en uygun modeli seçmesine imkân verir[162].

ROC eğrisi üzerinde sağ üst köşeden sol alt köşeye doğru eşik skoru düşmektedir. Böylece daha fazla abone olumlu olarak değerlendirilmektedir. Bu durum, kredi portföyünü büyütürken maruz kalınan zararı da artırır.

Şekil 23'te her alt veri seti için farklı modellerin ROC eğrileri yer almaktadır. Bu şekilde her üç model de birbirine çok yakın sonuç vermiş olsa da LR ve YSA modellerine ait eğrilerin daha üst tarafta olduğu seçilebilmektedir. Üstteki tabloda yer alan ROC indekslerine baktığımızda da LR modelinin her üç veri seti için de aynı alana (%89) sahip olduğu görülmektedir. ROC indeksi, ROC eğrisi altında kalan alanı göstermektedir. Her üç alt veri setinin ortalama olarak da ayrı olarak da ROC indeksleri arasındaki maksimum değer LR modeline aittir. Bu nedenle modelin diğerlerine göre daha tutarlı olduğu ileri sürülebilir. Diğerlerinde ise en yüksek oran sadece YSA'nın eğitim veri seti için %89 olarak gerçekleşmiştir. Aşağıdaki şekil incelendiğinde özellikle yanlış pozitif oranının arttığı sağ üst kısımda LR modelinin diğer iki modelden daha iyi tahminler yaptığı söylenebilir.

[162] Weiss, Sholom M. and Zhang T. "Performance Analysis and Evaluation", In Ye N. (Ed.). **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. s. 432.



Şekil 23 : Modellerin Alt Veri Setlerine Göre ROC Eğrisi Karşılaştırmaları

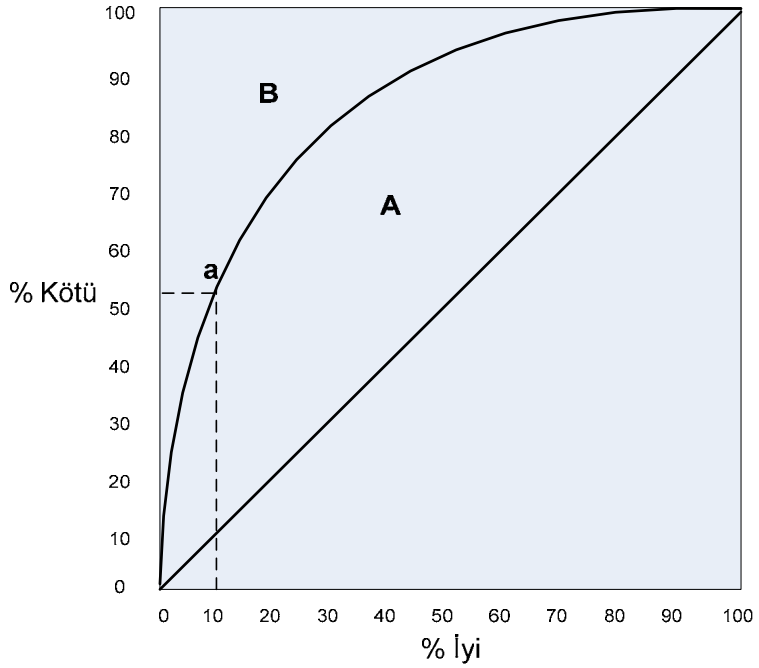
6.4.2.2. Gini İstatistiği

Modelin ayırım gücü, iyi ve kötü olarak tanımlanan gözlemlerin model tarafından ne kadar uygun bir şekilde ait oldukları gruplara atandıklarını ölçerek belirlenir. Bu konuda kullanılan yöntemlerden bir tanesi de Gini istatistiğidir. Gini istatistiği skorun tüm aralığı boyunca, iyiler ve kötülerin kümülatif yüzdesini karşılaştırarak ayrımı ölçer. Şekil 24'e göre Gini katsayısının hesaplanması şu şekilde olur:

Gini = 100 * A / (A + B)

Bu katsayı farklı modellerin güçlerini karşılaştırmalı olarak değerlendirmek için uygun bir ölçü olarak kullanılır. Rasgele değerlendirilen bir modelde, yani riskli ve risksiz abone ayırımının başarısız olduğu modelin Gini katsayısı 0 değerini alırken, teorik olarak maksimum performanslı bir modelin Gini katsayısı 100'e eşit olur. Bu katsayılar modelin riskli ve risksiz aboneleri yeteri kadar iyi ayırdığını göstermektedir.Yani Gini katsayısının sıfır olması, iyi ve kötü abone ayırımının rastsal yapıldığını, bir olması durumunda da iyi ve kötü abone ayırımının mükemmel yapıldığı söylenebilir. Eğitim, doğrulama ve test veri setleri

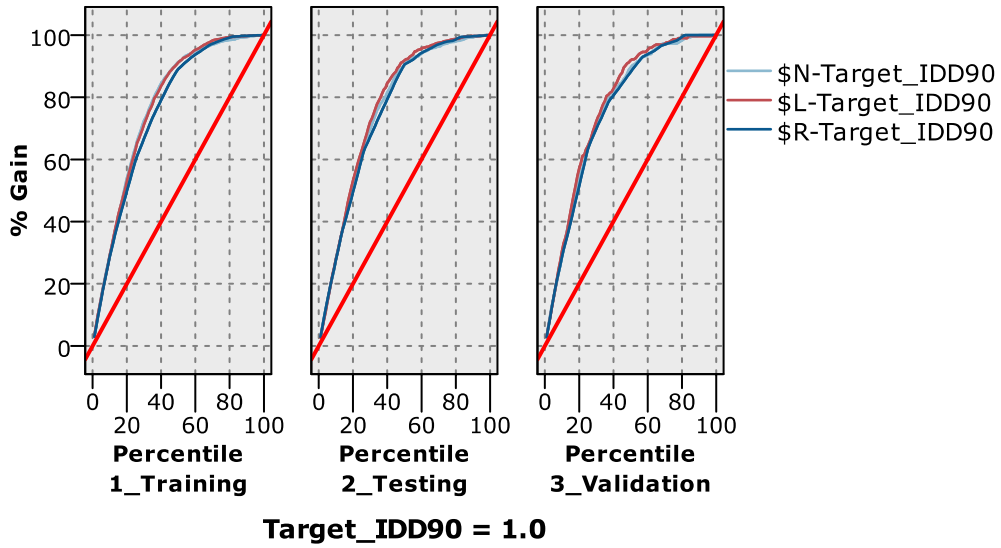
içerisinde en yüksek katsayılar yine LR eğrisine aittir.Tablo 29'a göre en yüksek Gini katsayısı eğitim veri seti için %78, test veri seti için %78 ve doğrulama veri seti için %79 ile LR modeline aittir.



Şekil 24 : Gini Katsayısı Şekli

6.4.2.3. Gain İstatistiği

Modellerin birbirleri ile karşılaştırılmasında kullanılan üçüncü yöntem Gain katsayısıdır. Şekil 25'te üç modelin eğitim, doğrulama ve test veri setlerine göre Gain eğrileri bir arada yer almaktadır. Şekile dikkatli bakıldığında, LR ve YSA modellerine ait Gain eğrilerinin neredeyse üst üste çakıştığı; KA modeline ait eğrinin ise diğer ikisinden daha içeride olduğu görülebilir. Tablo 29'da da Gain katsayısı eğitim ve test veri setleri için LR ve YSA modelleri neredeyse aynı değerleri alırken, doğrulama veri setinde LR modeli daha yüksek katsayıya sahiptir.



Şekil 25 : Modellerin Eğitim, Doğrulama ve Test Veri Setleri ile Hazırlanmış GAIN Eğrileri

Bu eğrinin oluşturulduğu veri seti yüzde beşlik dilimler için aşağıda yer almaktadır. Bu tablo incelendiğinde de, modellerin sonuçları arasında çok büyük bir farklılığın olmadığı söylenebilir. Her üç veri seti ortalamaları karşılaştırıldığında genel olarak YSA modeline ait lift oranı daha yüksektir, ancak ilk %5'lik dilime bakıldığında LR ve KA nın Lift oranı %15, YSA modelinin Lift oranı da %14.9'dur.

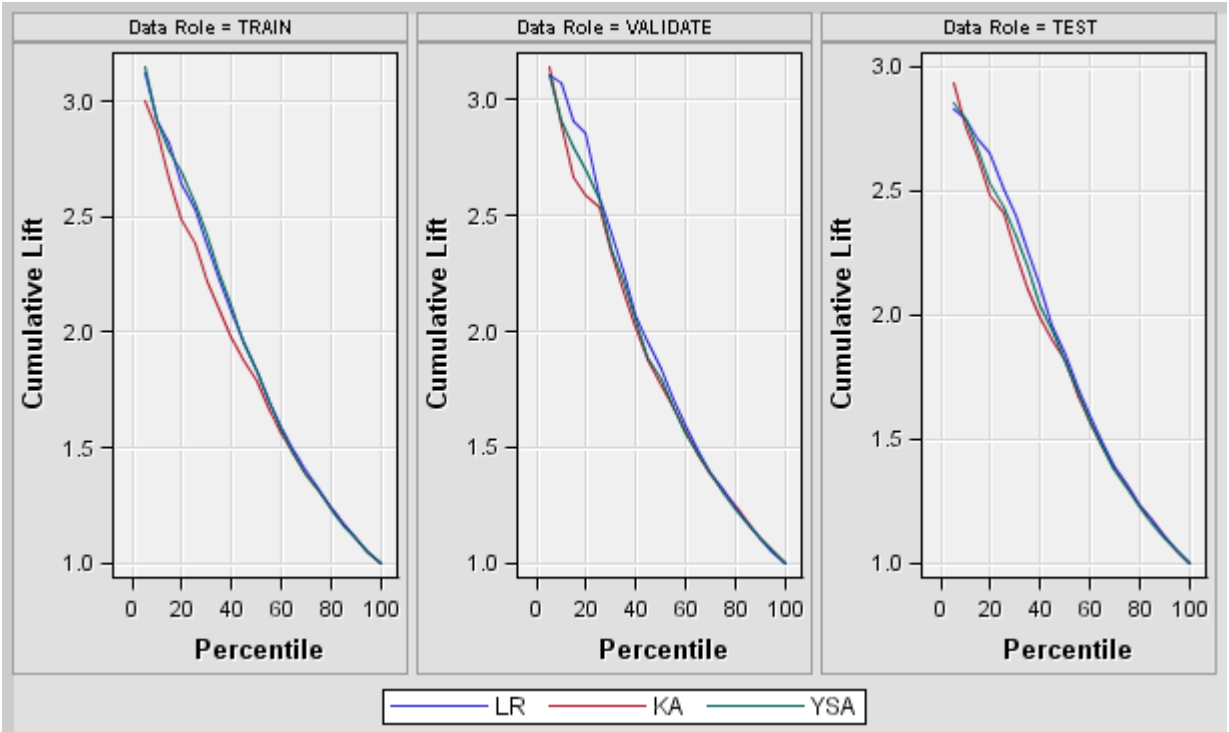
Tablo 30 : GAIN Eğrisinin Altında Kalan Alanın Yüzdelik Dilimlere Göre Dağılımı

Yüzdelik Dilim	LR Modeli			YSA Modeli			KA Modeli		
	Eğitim	Test	Doğrulama	Eğitim	Test	Doğrulama	Eğitim	Test	Doğrulama
5	15,6	14,1	15,2	15,7	14,2	15,2	15,0	14,6	15,4
10	29,1	27,8	30,4	29,1	27,8	29,1	28,7	27,6	28,6
15	42,2	40,5	43,2	41,7	40,0	41,9	39,8	39,5	39,7
20	52,9	53,0	57,1	53,8	50,5	53,7	49,8	49,5	51,6
25	63,2	62,7	64,2	63,8	60,6	63,9	59,6	60,1	63,2
30	71,3	72,0	72,3	72,4	69,6	70,3	66,7	67,3	69,9
35	78,0	78,8	78,7	78,9	76,7	77,4	73,4	73,5	75,9
40	83,7	84,5	82,8	84,4	81,5	82,1	79,1	79,4	80,6
45	88,2	88,2	87,8	88,2	87,4	84,8	84,6	85,4	84,3
50	91,2	91,8	91,9	91,3	90,3	89,5	89,1	90,6	88,1
55	93,2	93,8	93,9	93,5	92,9	91,6	91,6	92,1	91,7
60	95,2	95,7	95,3	94,8	94,1	93,6	93,7	94,3	93,8
65	96,8	96,6	97,0	95,9	95,6	95,6	95,6	95,7	95,4
70	98,3	97,5	97,3	97,0	96,3	97,3	97,2	97,0	97,0
75	98,9	98,2	98,6	97,9	97,3	97,3	98,3	97,9	97,9
80	99,4	98,7	99,0	98,4	98,0	98,3	99,2	98,6	99,4
85	99,5	99,5	99,7	99,0	98,7	99,3	99,6	99,4	100,0
90	99,7	99,7	99,7	99,4	99,2	100,0	99,8	99,6	100,0
95	99,8	99,8	99,7	99,8	99,8	100,0	99,9	99,8	100,0
100	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

6.4.2.4. Lift İstatistiği

Birçok uygulama için öngörünün amacı bazı aksiyonların alınması (posta ile reklam gönderimi gibi) istenen sınıf üyelerinin tespit edilmesidir. Eğer öngörü tahmini sınıf olasılıklarına dayalı sıralama şeklinde yapılacaksa, sınıflandırmadan daha esnek olur. Lift eğrisi sonra kümülatif Doğru Pozitif kapsamını (y-ekseni) dereceye göre sıralı örneklere karşı (x-ekseni) çizer. Rassal bir sıralama sonuçları bu çizimdeki doğru diyagonal çizgi üzerindedir. Bir modelin lift eğrisi genellikle bu çizginin üzerindedir ve ne kadar yüksekse o kadar iyidir[163]. Tablo 29'a göre Lift istatistiği her üç veri setinin ortalamasına göre LR modeli diğer iki modelden daha iyidir. Bu kritere göre de, riskli abonelerin tespit edilmesinde LR modeli daha başarılıdır.

Şekil 26'da LR modeline ait lift eğrisinin eğitim veri seti için YSA'na ait lift eğrisi ile hemen hemen üst üste çakışmışken, diğer veri setlerinde özellikle ilk kısımlarda diğer modellere ait lift eğrilerinden biraz daha yukarılarda olduğu görülmektedir. Bu durum özellikle riskli abonelerin yer aldığı yüksek olasılıklı ilk bölgede LR modelinin gücünü göstermektedir.



Şekil 26 : Modellerin Veri Setlerine Göre Kümülatif Lift Katsayılarının Karşılaştırılmaları

[163] Weiss, Sholom M. and Zhang T. s. 432

7. SONUÇ

Bu çalışmada, mobil operatörlerin faturalı aboneleri için örnek bir veri seti üzerinde Davranışsal Skorlama Modeli geliştirilmiştir. Davranışsal Skorlama kavramı asıl olarak bankacılık ve finans sektörüne özgü olsa da son yıllarda müşterilerinin kredibilitesini ölçmeye çalışan tüm firmalar tarafından kullanılmaktadır. Davranışsal Skorlamanın veri madenciliği yöntemleri kullanılarak yapılmasının amacı; çok fazla miktarda, birbirinden farklı yapılarda ve dağınık olarak depolanan veriyi, uygulanabilir stratejik kararlar almak için yararlı bilgiye dönüştürmektir.

Kredi skorlama, kredi talep edenlerin finansal yükümlülüklerini yerine getirme kabiliyetlerinin değerlendirilerek adayları iyi veya kötü olarak gruplara ayırıp derecelendirmek için kullanılan matematiksel bir araçtır. Kredi skorlama uygulamaları Başvuru Skorlama ve Davranışsal Skorlama olarak iki alt başlığa ayrılmaktadır. Başvuru Skorlama Modellerinde amaç kredi almak için başvuru yapan adaya kredinin verilip verilmeyeceğini belirlemektir. Davranışsal Skorlama Modeli ise mevcut müşteriler için kullanılır. Davranışsal Skorlama Modeli, finans sektöründe başlıca aboneye yeni bir finansman kaynağının yaratılıp yaratılmamasının veya mevcut borcun yeniden finansman kararının verilmesinde kullanılırken; telekomünikasyon sektöründe yer alan müşteriler için abone riskinin anlık olarak takip edilmesi ve gerekli aksiyonların alınması amacıyla kullanılır.

Skorlama sistemi, kredi veren bir işletmeye hem sağlam hem de tutarlı stratejiler kurmasına imkân veren bir derecelendirme mekanizması sağlar. Ancak, skorlama sistemi tek başına tam bir cevap değildir ve ancak doğru kullanıldığında fayda sağlayan bir yoldur. Skorlama modellerinin doğru çalışmasının yanı sıra sağlam ve tutarlı kredi stratejilerinin ve uygulamalarının kurulması gerekmektedir.

Davranışsal skorlama sonuçları kullanılarak firmaya uygun risk stratejileri hayata geçirilebilir. Skorlamanın günlük frekansta yapılabildiği bir altyapı, erken uyarı sistemi için uygundur ve elde edilen skor sonuçlarına dayanılarak belirlenen riskli müşteriler için uygun önlemler alınabilir. Etkin bir uygulama ile de riskli abonelerden kaynaklanacak şüpheli alacak miktarı minimuma indirilebilir.

Veri madenciliği, istatistiksel analiz, makine öğrenmesi, yapay zeka ve veri görselleştirme tekniklerini kullanarak büyük veri yığınları arasındaki gizli kalıpları ve ilişkileri keşfetmenin en gelişmiş, hızlı ve esnek yöntemlerine sahiptir. Daha iyi ve nitelikli kararlar alabilmek için geleceği öngörmeyi ve hızlı aksiyon almayı sağlayacak araçlar sunar. Çok büyük veri yığınlarına sahip mobil telekomünikasyon sektöründe riskli abonelerin tespit edilmesi de ancak güçlü bir proje metodolojisine sahip, doğru veri altyapısının üzerine kurulmuş güçlü ve çok çeşitli analitik yöntemler sağlayan veri madenciliği yöntemleri ile başarılabilir.

Uygulamada veri madenciliği modellerinde kullanılabilecek kadar yeterli gözlem ve değişken mevcuttur. Çalışmada, Lojistik Regresyon (LR), Karar Ağacı (KA) ve Yapay Sinir Ağı (YSA) modelleri üretilmiştir. Modeller tüm verilere uygulandığında en yüksek doğruluk oranı LR (%83.11) modeline aittir. Bunu sırasıyla YSA (%82.93) ve KA (%81.23) modelleri izlemektedir. Bu modellerin genel performansları arasındaki farklılık beklendiği gibi düşük gerçekleşmiştir.

Duyarlılık oranı açısından da en yüksek değerler LR modeline ait olduğu için; riskli abonelerin tespitinin çok önem kazandığı durumlarda veya dönemlerde de LR modeli, diğerlerine göre daha avantajlıdır. Kredi Risk Departmanlarının amacı mümkün olduğunca çok sayıda riskli aboneyi tespit etmek iken aynı zamanda da risksiz olan aboneye de müdahale etmemektir. Belirlilik oranına baktığımızda ise ilk sırada YSA modeli yer almasına rağmen LR modelinin sonucu ile arasındaki fark çok azdır.

Eğitim veri setinde gerçekleşen doğruluk oranının, doğrulama ve özellikle test veri setinde de gerçekleşmesi modelin tutarlılığı açısından çok önemlidir. Eğitim veri setinde en yüksek doğruluk oranı %83.50 ile YSA modeline ait iken, doğrulama ve özellikle de test veri setindeki en yüksek doğruluk oranı %82.89 ile LR modeline aittir.

ROC indekslerine baktığımızda, her üç alt veri setinin ortalama olarak da ayrı olarak da ROC indeksleri arasındaki maksimum değer LR modeline aittir. Eğitim, doğrulama ve test veri setleri içerisinde en yüksek Gini katsayıları yine LR eğrisine aittir. Lift istatistiği açısından da her üç veri setinin ortalamasına göre LR modeli diğer iki modelden daha iyidir. Kolmogorov – Smirnov istatistiğine göre de, her üç veri seti için maksimum ayrım LR modelinde gerçekleşmiştir. Bu istatistiğe göre en düşük fark da KA modeline aittir.

Yanlış sınıflandırma adetlerinin toplamlarına baktığımızda en düşük adet LR modeli tahminlerinde (1.689), ikinci en düşük adet YSA model tahminlerinde (1.707) ve en yüksek yanlış tahmin adedi de (1.877) ile KA modelinde gerçekleşmiştir. Bu adetlerle bağlantılı olarak alt veri setleri için yanlış sınıflandırma oranları açısından da en düşük oranlar; %17 (eğitim), %17 (test) ve %16 (doğrulama) ile LR modeline aittir. Bu istatistiklerle tutarlı olarak toplam hata fonksiyonu, ortalama hata fonksiyonu ve ortalama hata kareleri istatistikleri için de LR modeli en düşük ortalamalara sahiptir.

LR modeli, abonelerin ayırım gücünün yeterliliği ve tahmin başarısının yüksekliği kadar, tüm veri setlerinde birbirine benzer sonuçlar ile tutarlı bir yöntem olduğunu da kanıtlamıştır. YSA modeli, özellikle eğitim veri setinde çok başarılı iken doğrulama ve test veri setlerinde aynı başarıyı gösterememiştir. KA modelinin, kredi skorlama süreçlerinde kullanılabilecek basit olduğu kadar etkili bir yöntem olduğu da buradaki doğru sınıflandırma oranlarına bakılarak ileri sürülebilir.

Modeller oluşturulduktan sonra sonucu belirleyen önemli değişkenler incelenmiş ve çok önemli bulgulara ulaşılmıştır. Overdue_L6_AVR (Son 6 fatura döneminde çıkan ve geç ödenen faturalarının ortalama geç ödenme gün sayısı) değişkeni her üç model tarafından da birinci sırada anlamlı değişken olarak seçilmiştir. Buna göre fatura tutarından bağımsız olarak, zamanında ödeme yapma verisinin, mobil telekomünikasyon sektöründeki riskli aboneyi belirlemede en önemli parametre olduğu ileri sürülebilir. Özellikle abonelerin değerlemeden önceki son dönemlerde çıkan faturalarının ödemelerinin zamanında yapılıp yapılmamasının tüm abonelik boyunca yapılan ödeme alışkanlığına göre Davranış Skorlama sonucu üzerinde çok daha önemli olduğu da ortaya çıkmıştır.

Modellerde açıklama gücü en üst sıralarda yer alan değişken olan Invoice_Paid_CNT, abone tarafından ödenen tüm faturaların toplam adedini vermesine rağmen, aynı zamanda ne kadar uzun süreden beri aynı şebekede abone olarak bulunduğunu da göstermektedir. Modellerdeki katsayılar ve ayırım incelendiğinde, ödenen fatura adedi ne kadar yüksek ise abonenin riskli olarak tanımlanma olasılığının da o kadar düşük olduğu söylenebilir.

Risk_RT değişkeni bir diğer önemli değişken olarak modeller tarafından belirlenmiştir. Buna göre, aynı miktarda riski olan aboneleri birbirleri ile karşılaştırırken hangisinin yaptığı ödeme miktarı daha fazla ise onun daha az riskli olduğu ileri sürülebilir.

Kredi skorlama alanında yapılabilecek pek çok geliştirme mevcuttur. Ancak ülkemizde kredi skorlama sürecinin daha bilimsel temellere oturabilmesi ve elde edilen faydanın artırılabilmesi için aşağıda sayılan geliştirmelerin hayata geçirilmesi büyük önem taşımaktadır:

- Aynı sektör içerisinde kredi kuruluşları arasındaki işbirliğinin artırılması gerekmektedir
- Abone verilerinin paylaşılması alanındaki yasal engeller ortadan kaldırılmalıdır
- Yasal düzenlemeler ile Kredi Değerleme Büroları Türkiye'de de kurulmalıdır
- MERNIS gibi kamuya ait veritabanlaından daha fazla faydalanılmalıdır
- Abone hakkında önemli bilgileri barındıran Adres Doğrulama Sistemi bir an önce tamamlanıp özel sektörün de kullanımına açılmalıdır

KAYNAKÇA

Kitaplar ve Süreli Yayınlar

1. Albayrak, A. Sait, **Uygulamalı Çok Değişkenli İstatistik Teknikleri**. Ankara: Asil Yayın Dağıtım, 2006.
2. Baesens B, T Van Gestel, S Viaene, M Stepanova, J Suyken and J Vanthienen "Benchmarking state-of-the-art classification algorithms for credit scoring". **Journal of the Operational Research Society**. Vol.54, 2003, ss. 627–635
3. Bailey M., "An Introduction to the Principles", In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. ss. 1-6.
4. Bensic, M., Natasa S. and Marijana Z.S. "Modelling Small-Business Credit Scoring by Using Logistic Regression, Neural Networks And Decision Trees". **Intelligent Systems In Accounting, Finance and Management**. Vol.13, 2005, ss. 133–150
5. Berry, Michael J.A. and Gordon S. Linoff. **Data Mining Techniques for Marketing, Sales, and Customer Relationship Management.** Indianapolis: Wiley Publishing, 2004.
6. Betts J., "Credit Scoring Systems". In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. ss. 27-34.
7. Blöchlinger, A. and Markus Leippold. "Economic benefit of powerful credit scoring". **Journal of Banking & Finance.** Vol.30, 2006, ss. 851–873
8. Bugera, V., Hiroshi Konno, and Stanislav Ursayev. "Credit Cards Scoring with Quadratic Utility Functions". **Journal Of Multi-Criteria Decision Analysis**. Vol.11, 2002, ss. 197–211
9. Chandler G. "Generic and customized scoring models: A comparison". In Elizabeth Mays (Ed.). **Credit Scoring for Risk Managers.** Thomson, South-Western, 2004. ss. 13-48.
10. Chuang, Chun-L. and Rong-Ho Lin. "Constructing a reassigning credit scoring model". **Expert Systems with Applications**. Vol.36, 2009, ss. 1685-1694
11. Cios Krzysztof J., Witold Pedrycz, Roman W. Swiniarski and Lukasz A. Kurgan**. Data Mining: A Knowledge Discovery Approach**. New York: Springer, 2007
12. Clementine 12.0 Ders Notları, 2.2
13. Close, Diane B., Arnold D. Robbins, Paul H. Rubin, Richard Stallman, Piet van Oostrum, **The AWK Manual,** Edition 10, 1995.
14. Crook, J. and John Banasik. "Does reject inference really improves the performance of application scoring models?" **Journal of Banking & Finance**. Vol.28, 2004, ss. 857–874
15. Das G. and Dimitrios Gunopulos, "Time Series Similarity and Indexing", In Ye N. (Ed.) **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. ss. 279-304.
16. Dekker B. "Lending Strategies". In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. ss. 77-85.

17. Dekker B. “In-house Scorecard Development”, In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. ss. 155-162.

18. Didimo W. and Giuseppe Liotta, “Graph Visualization and Data Mining”, In Cook, D.J. and Holder L.B. (Ed.). **Mining Graph Data**. New Jersey: John Wiley & Sons, 2007. s. 35-64.

19. Eisenbeis Robert A. “Problems in Applying Discriminant Analysis in Credit-Scoring Models”. In Thomas Lyn C., David B. Edelman, and Jonathan N. Crook (Ed.). **Readings in Credit Scoring.** New York: Oxford University Press, 2004. ss. 17-32.

20. Feelders, A.J. “Credit Scoring and Reject Inference With Mixture Models”. **International Journal of Intelligent Systems in Accounting, Finance & Management**. Vol.9, 2000, ss. 1–8

21. Giudici, P. **Applied Data Mining: Statistical Methods for Business and Industry**. West Sussex: John Wiley & Sons Ltd., 2003.

22. Gündüz Ş. ve Eşref Adalı. “Web kullanıcılarının davranışları için örüntü bulma ve modelleme”, **İTÜ Dergisi/d Mühendislik**, Cilt:3, Sayı:6, Aralık 2004,15-24,

23. Han, J. and Micheline Kamber. **Data Mining: Concepts and Techniques.** Second edition, San Francisco: Morgan Kaufmann Publications, 2006.

24. Hand D., Heikki Mannila, and Padhraic Smyth. **Principles of Data Mining.** Massachusetts: A Bradford Book The MIT Press, 2001.

25. Hopper M.A. and E.M. Lewis, “Behaviour Scoring and Adaptive Control Systems”. In Thomas Lyn C., David B. Edelman, and Jonathan N. Crook (Ed.). **Readings in Credit Scoring.** New York: Oxford University Press, 2004. ss. 33-46.

26. Hsieh, Nan-C. “An integrated data mining and behavioral scoring model for analyzing bank customers”. **Expert Systems with Applications.** Vol.27, 2004, ss. 623–633

27. Hsieh, Nan-C. “Hybrid mining approach in the design of credit scoring models”. **Expert Systems with Applications**. Vol.28, 2005, ss. 655–665

28. Huang, Jih-J., Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. “Two-stage genetic programming (2SGP) for the credit scoring model”. **Applied Mathematics and Computation.** Vol.174, 2006, ss.1039–1053

29. Huang, Cheng-L., Mu-Chen Chen, and Chieh-Jen Wang. **“**Credit scoring with a data mining approach based on support vector machines”, **Expert Systems with Applications**. Vol.33 2007, ss. 847–856

30. Hyde R. **The Art of Assembly Language Programming.** San Francisco: No Starch Press, 2003

31. IBM Global Technology Services, **The toxic terabyte: How data-dumping threatens business efficiency**, London: 2006

32. Johnson R.W. “Legal, Social, and Economic Issues in Implementing Scoring in the United States”. In Thomas Lyn C., David B. Edelman, and Jonathan N. Crook (Ed.). **Readings in Credit Scoring.** New York: Oxford University Press, 2004. ss. 5-15.

33. Kim, Yoon S. and So Young Sohn. “Managing loan customers using misclassification patterns of credit scoring model”. **Expert Systems with Applications**. Vol.26, 2004, ss. 567–573

34. Kindred D., “What is Scorecard?”. In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. ss. 7-13.

35. Kleinbaum, David G. **Logistic Regression A self Learning Text.** New York : Springer, 1994.

36. Kredi Kayıt Bürosu, **Annual Report 2007**, İstanbul: 2008.

37. Laha, A. “Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring”. **Advanced Engineering Informatics** Vol.21, 2007, ss. 281–291

38. Larose, Daniel T. **Discovering Knowledge in Data an Introduction to Data Mining.** New Jersey: John Wiley & Sons, 2005

39. Le, Chap T., **Applied Categorical Data Analysis.** New York: John Wiley & Sons Ltd., 1998.

40. Lee, Tian-S., Chih-Chou Chiu, Chi-Jie Lu, and I-Fen Chen. “Credit scoring using the hybrid neural discriminant technique”. **Expert Systems with Applications.** Vol.23, 2002, ss. 245–254

41. Lee, Tian-S., Chih-Chou Chiu, Yu-Chao Chou, and Chi-Jie Lu. “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines”. **Computational Statistics & Data Analysis**. Vol.50, 2006, ss. 1113–1130

42. Madeira Sara C., Arlindo L. Oliveira, and Catarina S. Conceicao. A Data Mining Approach to Credit Risk Evaluation and Behaviour Scoring, ss. 184–188, Fernando Moura Pires, Salvador Abreu (Eds.): **EPIA 2003, LNAI 2902**, 2003. Springer-Verlag Berlin Heidelberg 2003

43. Mays E. “The Role of Credit Scores in Consumer Lending”. In Elizabeth Mays (Ed.). **Credit Scoring for Risk Managers.** Thomson, South-Western, 2004. ss. 3-12.

44. Mehmet Sayal, “Time Series Analysis and Mining Techniques”, In John Wang (Ed.). **Encyclopedia of Data Warehousing and Mining.** Hershey: Idea Group Inc, 2006. ss. 1120-1124.

45. Mladenić D. “Text Mining-Machine Learning on Documents”. In John Wang (Ed.). **Encyclopedia of Data Warehousing and Mining.** Hershey: Idea Group Inc, 2006. ss. 1109-1112.

46. Olson, David L. and Dursun Delen. **Advanced Data Mining Techniques.** Berlin: Springer, 2008.

47. Ong, Chorng-S., Jih-Jeng Huang, and Gwo-Hshiung Tzeng. “Building credit scoring models using genetic programming”. **Expert Systems with Applications.** Vol.29, 2005, ss. 41–47

48. Özkan, Y. **Veri Madenciliği Yöntemleri.** İstanbul: Papatya Yayıncılık Eğitim. 2008.

49. Pearce W., “Neural Scoring”. In Murray Bailey (Ed.). **Credit Scoring, The Principles and Practicalities.** Second Edition. Bristol: White Box Publishing, 2004. ss. 138-143.

50. Ramakrishnan, R. and Johannes Gehrke. **Database Management Systems.** Third Edition, New York: The McGraw-Hill Companies, 2003.

51. Ridgeway G., “Strategies and Methods for Prediction”. In Ye N. (Ed.). **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. ss. 160-192.

52. Sarantopoulos, G. “Data Mining In Retail Credit”. **Operational Research. An International Journal.** Vol.3, No.2, 2003, ss. 99-122

53. SAS Institute Inc. **Appliying Data Mining Techniques Using SAS Enterprise Miner Course Notes.** Cary, NC : 2005.

54. Sumathi S. and S.N. Sivanandam. **Introduction to Data Mining and its Applications.** Berlin: Springer, 2006.

55. Šušteršicˇ, M., Dušan Mramor, and Jure Zupan. “Consumer credit scoring models with limited data”. **Expert Systems with Applications**. Vol.36, 2009, ss. 4736-4744

56. Trinkle, Brad S. and Amelia A Baldwin. “Interpretable Credit Model Development via Artificial Neural Networks”. **Intelligent Systems in Accounting, Finance and Management.** Vol.15, 2007, ss. 123–147

57. Webb, Geoffrey I. “Association Rules”, In Ye N. (Ed.). **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. ss. 25-40.

58. Weiss, Sholom M. and Zhang T. “Performance Analysis and Evaluation”, In Ye N. (Ed.). **The Handbook of Data Mining.** New Jersey: Lawrence Erlbaum Associates, 2003. ss. 425-440.

59. West, D. “Neural network credit scoring models”. **Computers & Operations Research**. Vol.27, 2000, ss. 1131-1152

60. Wilkinson, L. **The Grammar of Graphics (Statistics and Computing).** Second Edition, New York: Springer, 2005.

61. Yao, J. T., Y. Y. Yao and Y. Zhao, “Foundations of Classification”. In Lin Tsau Y., Setsuo Ohsuga, Churn-Jung Liau, Xiaohua Hu (Eds.), **Foundations and Novel Approaches in Data Mining**. Berlin Heidelberg: Springer, 2006. ss. 75-98.

62. Zopounidis, C. and Michael Doumpos. “Multi-group discrimination using multi-criteria analysis: Illustrations from the field of finance”. **European Journal of Operational Research**. Vol.139, 2002, ss. 371–389

İnternet Kaynakları

1. Kredi Kayıt Bürosu http://www.kkb.com.tr (14.03.2009)
2. http://en.wikipedia.org/wiki/Data (23.02.2009)
3. http://www.asciitable.com(14.02.2009)
4. Pendse, N. Business Application Research Center, 2007. http://www.olapreport.com/consolidations.htm (10.05.2009)
5. CRoss Industry Standard Process for Data Mining, The CRISP-DM consortium August 2000, http://www.crisp-dm.org/Overview/index.htm (18.04.2009)
6. http://www.crisp-dm.org/SIG/index.htm (18.04.2009)
7. http://www.kdnuggets.com/polls/2008/data-mining-applications.htm (26.06.2009)
8. http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm (30.06.2009)
9. http://archive.ics.uci.edu/ml/datasets.html (06.01.2009)
10. Newman D. G. Engineering Economic Analysis, 2nd Edition, San Jose : Engineering Press, Inc. http://www.msstate.edu/dept/AgEdExp/4163/decision/decision.htm (24.02.2009)
11. Ergenekon.tc - Burak Arıkan, http://ergenekon.tc/ (25.05.2009)

EKLER

EK - 1 : GENEL AMAÇLI OLARAK GELİŞTİRİLEN SKORLAMA MODEL ÖRNEKLERİ

Skorlama Sistem Adı	Dağıtıcı Firma	Model Geliştiricisi	Tip	Tanım / Tahmin	Model Adedi	Tahmin Periyodu
Başvuru Skorlama Modeli	Fair, Isaac	Fair, Isaac	Başvuru skorlaması	Başvuru Risk Modelleri borç verisinin ulusal havuzuna dayandırılır ve borç verenlere maliyet etkin bir kredi riski değerlendirme aracı sağlamak için tasarlanır. Çok çeşitli porftöylerin kredi riskini içerir, belirli sınıra kadar tekrar yenilenebilir, dolaylı, doğrudan veya mevcut öz sermaye karşılığında verilen krediler örnek verilebilir. Deneysel olarak özellikle kredi çıkarma kararlarında kulllanmak için geliştirildiler.	14 model	12 ay
ASSIST® 2.0	TransUnion	Fair, Isaac	Sigorta riski	Muhtemel göreli zarar oranı açısından riskin derecesine göre başvuruları ve poliçe sahiplerini sınıflandırır	11 model (5 otomatik model, 6 özel model)	12 ay
Authentication Solutions Level One Score	Experian	Experian	Sahtekarlık / Doğrulama	Müşteriye ait isim, sosyal güvenlik numarası ve telefon numarası gibi bilgiler doğrulanmaktadır	1 Model	12 ay
Otomatik Risk Modeli	Experian	Experian	Endüstriye özel risk	Sonraki 24 ay boyunca otomatik kredi veya kira bedeli üzerindeki önemli derecede geri ödememe veya zararlı kredi davranışının olabilirliğini öngörür	7 skor kart	24 ay
CollectScore℠	Experian	Fair, Isaac	Tahsilat	Ödenmemiş hesapları muhtemel geri ödeme miktarlarının düzeylerine göre sınıflandırır	2 genel, 1 özelleştirilmiş model	6 ay
Cross View	Experian	Experian	Risk	Borç ve kredi verisi kullanılarak, 90 günün üzerinde gecikme yaşanıp yaşanmayacağını tahmin eder	2 skor kart	12 ay
DELPHI	TransUnion	Experian	İflas	12 ay içerisinde iflas etme olasılığını tahmin eder	9 model	12 ay
EDAS (Enhanced Delinquency Alert System)	Equifax	Experian	İflas / Risk	12 ay içerisinde iflas etme olasılığını tahmin eder. Ayrıca ciddi ödeme güçlüğü durumlarını tahminde kullanılır.	6 model	12 ay
Equifax Gelir Tahmincisi	Equifax	Equifax	Gelir tahmini	Müşteriye ait brüt yıllık geliri tahmin eder	Birçok farklı model	
Equifax Risk Scoru '98	Equifax	Equifax	Genel kredi riski	Müşterinin 90 günden daha fazla ödemeyi geciktirip geciktirmeyeceğini tahmin eder	Birçok farklı model	24 ay

EK - 1 : GENEL AMAÇLI OLARAK GELİŞTİRİLEN SKORLAMA MODEL ÖRNEKLERİ (devam)

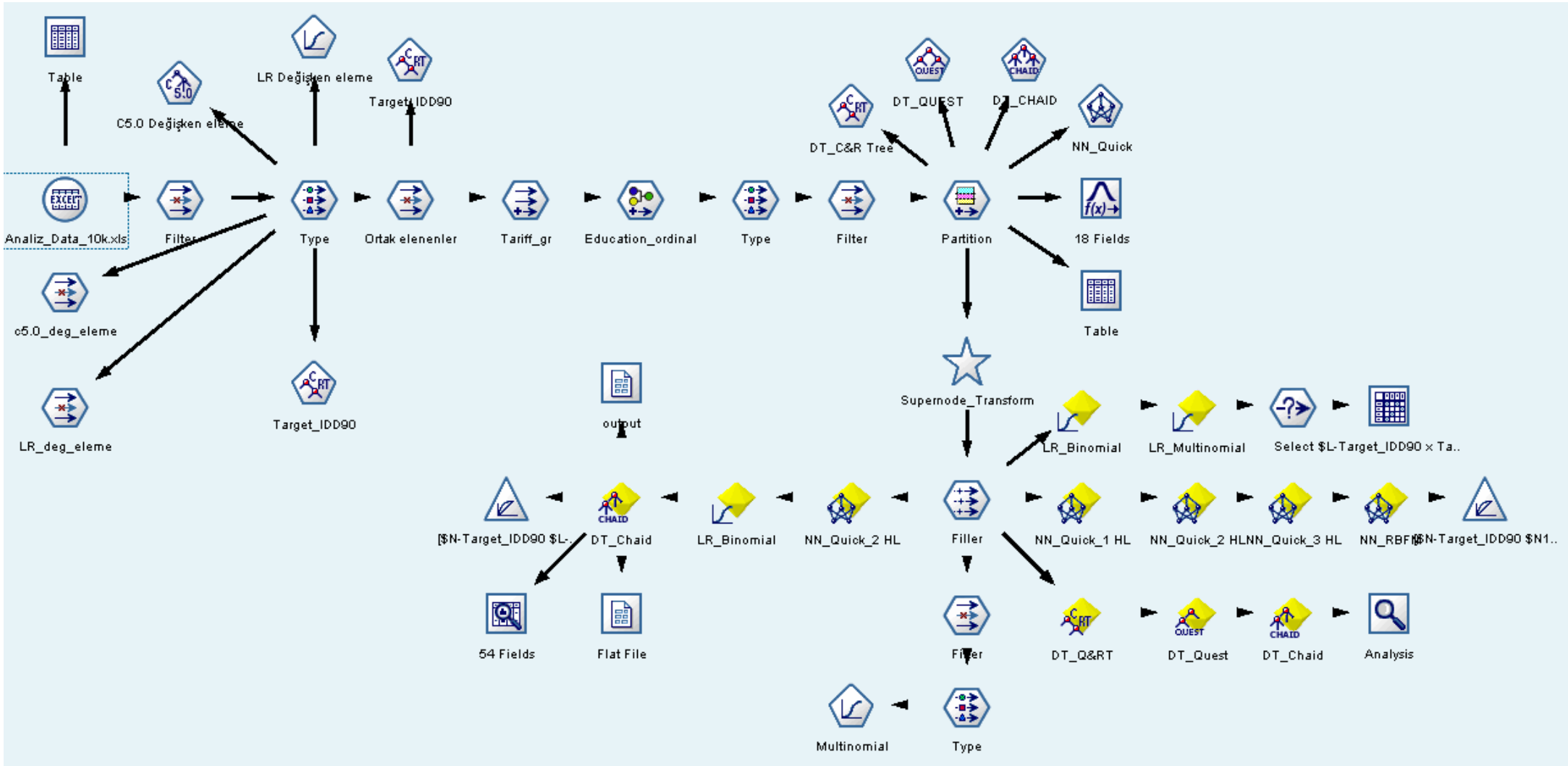
Skorlama Sistem Adı	Dağıtıcı Firma	Model Geliştiricisi	Tip	Tanım / Tahmin	Model Adedi	Tahmin Periyodu
Equifax Telco Skoru '98	Equifax	Equifax	Telekomunikasyon kredi riski	Telekomünikasyon firması aboneleri için ödemelerde ciddi gecikme olup olmayacağını tahmin eder.	Birçok farklı model	12 ay
Equifax Kablosuz Skoru '98	Equifax	Equifax	Kablosuz erişim kredi riski	Mobil operatör aboneleri için ödemelerde ciddi gecikme olup olmayacağını tahmin eder.	Birçok farklı model	6 ay
PRECISION[SM]	Experian, Fair, Isaac (PreScore, ScoreNet)	Fair, Isaac	Risk	Müşterileri, gelecekteki kredi yükümlülüklerini yerine getirememe olasılığına göre sınıflandırır. Bir sonraki nesil FICO® skoru, bütün kredi risk spektrumu üzerinde daha belirginleştirilmiş bir değerlendirme sağlar	18 model	24 ay
Uzman Modeller	Magnum Communications, Ltd ve Cypress Software Sytems	Scoring Solutions, Inc	Risk	Uzman tabanlı jenerik risk modelleri	Ürüne özel modeller	
Sahtekarlık Kalkan Skoru	Experian	Experian	Sahtekarlık	Hilekarlık ve kredibilite değişkenlerini tek skorda birleştirir, hilekarlık riski ve potansiyel birinci ödemeyi yapmama riski üzerinde tam bir bakış açısı sağlar.	1 model	12 ay
GEM[SM]	TransUnion	Scoring Solutions, Inc	Risk	Gaz ve elektrik endüstrisinde kullanılan model, müşterinin ödeme güçlüğüne düşmesi ile ciddi gecikme veya kayıp olasılığını tahmin eder	3 model (kredi dosyasının derinliği ve ödeme güçlüğünün ciddiyetine göre)	12 ay
HORIZON[SM]	TransUnion, Fair, Isaac (PreScore, ScoreNet)	Fair, Isaac	İflas	İflas zarar oranının (iflastan kaybedilen zarar miktarının "iyi" müşterilerden gelen net gelire oranı) derecesine göre müşterileri sınıflandırır	11 model	18 ay
In the Market Model	Experian	Experian	Yanıt	Posta gönderiminden önce otomotiv kiralama veya kredi alma pazarında olabilecek bireyleri tespit eder	2 skor kart	5 ay
Recovery Score - Retail	Experian	Experian	Recovery	6 ay içerisinde yapılabilecek tahsilat seviyesini yüksekten düşüğe doğru tahmin eder	1 skor kart	6 ay

EK - 1 : GENEL AMAÇLI OLARAK GELİŞTİRİLEN SKORLAMA MODEL ÖRNEKLERİ (devam)

Skorlama Sistem Adı	Dağıtıcı Firma	Model Geliştiricisi	Tip	Tanım / Tahmin	Model Adedi	Tahmin Periyodu
ROI[SM] (Revenue Opportunity Indicator)	Experian, Fair, Isaac (PreScore, ScoreNet)	Fair, Isaac	Gelir	Gelecek 12 ay boyunca yeniden kullanılabilir kredi üzerinden üretilebilecek göreli gelir miktarının derecesine göre müşterileri sınıflar	4 model	12 ay
SENTRY[SM]	TransUnion, Fair, Isaac (ScoreNet)	Fair, Isaac	Attrition	Skorlamadan sonra üç veya beş ay içerisinde müşterinin banka kartını kapatma veya hesabı sıfırlama olasılığına göre müşterileri sınıflar	10 model	3 - 5 ay
Küçük işletme yeni hesap risk skoru	Fair, Isaac / D&B	Fair, Isaac	Ticaret kredi riski	D&B'den gelen iş hakkındaki ve müşteri raporlama ajanslarından gelen işin esasları hakkındaki veriye dayalı olarak, işletmeden başka bir işletmeye verilen borçları içeren işlemlerin riskini derecelendirir	6 model	24 ay
SPECTRUM®	TransUnion	Scoring Solutions, Inc	Risk	Kablosuz iletişim sektörü için bir risk modelidir. Bir müşterinin borcunu önemli derecede geciktirme veya zarar ile sonuçlanma olasılığını öngörür.	3 model	6 ay
Telekomunikasyon, Eenrji Kablo Risk Modeli	Experian	Experian	Risk	Kablosuz aboneliklerde son ödeme tarihini 90 gün veya daha fazla geçip geçmeyeceğini tahmin etmektedir.	4 skor kart	12 ay
Tele-Risk	Experian	Experian	Endüstriye özel risk	Telekomünikasyon sektöründe bulunan bir müşterinin sonraki 12 ay boyunca borcunu önemli derecede geciktirme veya zararlı kredi davranışında bulunma olasılığını öngörür.	2 skor kart	12 ay
TELESCOPE[SM]	TransUnion	Scoring Solutions, Inc	Risk	Telekomunikasyon endüstrisi için geliştirilmiş risk modelidir. Telekomünikasyon firması aboneleri için ödemelerde ciddi gecikme olup olmayacağını veya kayba yol açıp açmayacağını tahmin eder.	5 model	12 ay
TransRecovery	TransUnion	TransUnion Modelleme Servisi	Collections / Recovery	Tahsilat sürecindeki bir müşteriden 12 ay içinde ($50 +) tahsilat yapabilme olasılığını öngörür	3 model	12 ay
Vista® Hesap Yönetimi Risk Scor Servisi	Experian	Fair, Isaac	Küçük işletme kredi riski	Gelecekteki iflas, elden çıkarma, yükümlülüğünü yerine getirememe ve borcununu önemli derecede geciktirme olasılığına dayalı olarak küçük işletme kredisi müşterilerini sınıflandırır	3 model	24 ay

Kaynak : Chandler G. "Generic and customized scoring models: A comparison". In Elizabeth Mays (Ed.). Credit Scoring for Risk Managers. Thomson, South-Western, 2004. s. 19-31

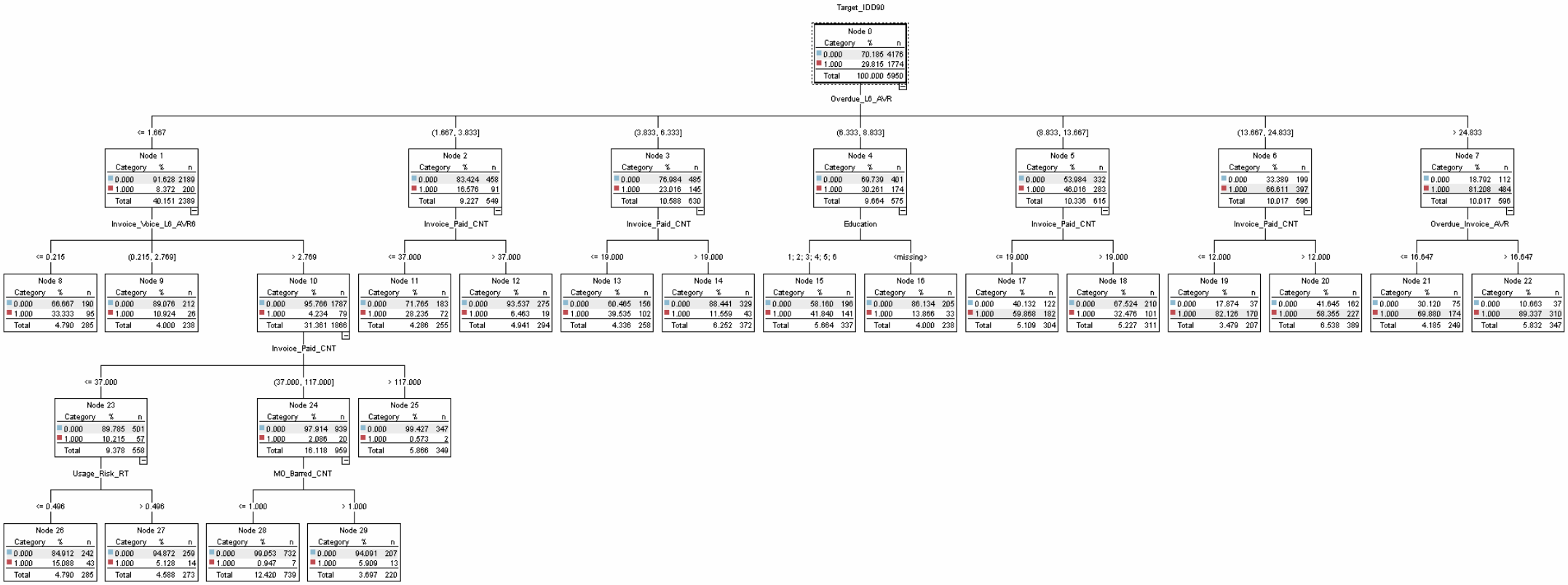
EK - 2 : VERİ MADENCİLİĞİ UYGULAMA SÜRECİ



EK - 3 : CHAID MODELİ KURAL SETİ

```
Overdue_L6_AVR <= 1,667 [ Mode: 0 ]
        Invoice_Voice_L6_AVR6 <= 0,215 [ Mode: 0 ] => 0,0
        Invoice_Voice_L6_AVR6 > 0,215 and Invoice_Voice_L6_AVR6 <= 2,769 [ Mode:
        0 ] => 0,0
        Invoice_Voice_L6_AVR6 > 2,769 [ Mode: 0 ]
                Invoice_Paid_CNT <= 37 [ Mode: 0 ]
                        Usage_Risk_RT <= 0,496 [ Mode: 0 ] => 0,0
                        Usage_Risk_RT > 0,496 [ Mode: 0 ] => 0,0
                Invoice_Paid_CNT > 37 and Invoice_Paid_CNT <= 117 [ Mode: 0 ]
                        MO_Barred_CNT <= 1 [ Mode: 0 ] => 0,0
                        MO_Barred_CNT > 1 [ Mode: 0 ] => 0,0
                Invoice_Paid_CNT > 117 [ Mode: 0 ] => 0,0
Overdue_L6_AVR > 1,667 and Overdue_L6_AVR <= 3,833 [ Mode: 0 ]
        Invoice_Paid_CNT <= 37 [ Mode: 0 ] => 0,0
        Invoice_Paid_CNT > 37 [ Mode: 0 ] => 0,0
Overdue_L6_AVR > 3,833 and Overdue_L6_AVR <= 6,333 [ Mode: 0 ]
        Invoice_Paid_CNT <= 19 [ Mode: 0 ] => 0,0
        Invoice_Paid_CNT > 19 [ Mode: 0 ] => 0,0
Overdue_L6_AVR > 6,333 and Overdue_L6_AVR <= 8,833 [ Mode: 0 ]
        Education in [ "1" "2" "3" "4" "5" "6" ] [ Mode: 0 ] => 0,0
        Education IS MISSING [ Mode: 0 ] => 0,0
Overdue_L6_AVR > 8,833 and Overdue_L6_AVR <= 13,667 [ Mode: 0 ]
        Invoice_Paid_CNT <= 19 [ Mode: 1 ] => 1,0
        Invoice_Paid_CNT > 19 [ Mode: 0 ] => 0,0
Overdue_L6_AVR > 13,667 and Overdue_L6_AVR <= 24,833 [ Mode: 1 ]
        Invoice_Paid_CNT <= 12 [ Mode: 1 ] => 1,0
        Invoice_Paid_CNT > 12 [ Mode: 1 ] => 1,0
Overdue_L6_AVR > 24,833 [ Mode: 1 ]
        Overdue_Invoice_AVR <= 16,647 [ Mode: 1 ] => 1,0
        Overdue_Invoice_AVR > 16,647 [ Mode: 1 ] => 1,0
```

EK – 4: CHAID ALGORİTMASI KARAR AĞACI DİYAGRAMI



ÖZGEÇMİŞ

1974 yılında Malatya'da doğdu, ilköğrenimini Kahramanmaraş'ta tamamladı. 1997 yılında İTÜ İşletme Mühendisliği Bölümünden mezun oldu. Aynı yıl İTÜ SBE'de İşletme Yüksek Lisansına başladı. Aralık 1998'den itibaren İşletme Mühendisliği Bölümünde Muhasebe ve Finansman Anabilim Dalında Araştırma Görevlisi olarak üç yıl görev yaptı. Bu süreç içerisinde "Principles of Accounting" ve "Introduction to Computer Science" derslerinin uygulamalarına katıldı, İTÜ SEM'de bilgisayar eğitmenliği yaptı. "Bank Rating Systems and An Empirical Evaluation Model for Commercial Banks in Turkey: A CAMEL Approach" isimli çalışmasını 9-13 Mart 2001 tarihleri arasında Berlin'de düzenlenen İTÜ ve TUB Üçüncü Geleneksel Çalışma Seminerinde sundu. Bu çalışma İşletme Mühendisliği Tartışma Yazıları içerisinde yayımlanmıştır. Daha sonra çalışma hayatına Pazar Araştırması şirketlerinde başladı, buralarda uluslararası firmaların pazar araştırma projeleri başta olmak üzere, pazarlama ve sosyal araştırma projelerinde kısmen veya tüm aşamalarda yer aldı. Halihazırda Türkiye'de faaliyet gösteren bir mobil telekomünikasyon operatöründe Kredi Değerleme Müdürü olarak çalışmaktadır. İkinci yüksek lisans çalışmasının, ilgi ve deneyiminin olduğu İstatistik alanında olmasını tercih etmiştir.