T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

BİREYSEL ARAÇ KREDİLERİNİN YASAL TAKİBE GİRME DURUMLARI HAKKINDA TAHMİN MODELLERİNİN OLUŞTURULMASI

Yüksek Lisans Tezi

AYBUKE DOĞAN

İSTANBUL, 2015

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

BİREYSEL ARAÇ KREDİLERİNİN YASAL TAKİBE GİRME DURUMLARI HAKKINDA TAHMİN MODELLERİNİN OLUŞTURULMASI

Yüksek Lisans Tezi

AYBUKE DOĞAN

Tez Danışmanı: Doç. Dr. M. Alper TUNGA

İSTANBUL, 2015

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

Tezin Adı: Bireysel Araç Kredilerinin Yasal Takibe Girme Durumları Hakkında Tahmin Modellerinin Oluşturulması
Öğrencinin Adı Soyadı: Aybuke Doğan
Tez Savunma Tarihi: 02.09.2015

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Fen Bilimleri Enstitüsü tarafından onaylanmıştır.

Doç. Dr. Nafiz ARICA
Enstitü Müdürü

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğunu onaylarım.

Doç. Dr. M. Alper TUNGA
Program Koordinatörü

Bu Tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans tezi olarak yeterli görülmüş ve kabul edilmiştir.

Jüri Üyeleri	İmzalar
Tez Danışmanı Doç. Dr. M. Alper TUNGA	---
Üye Yrd. Doç. Dr. Ahmet KIRIŞ	---
Üye Yrd. Doç. Dr. Yücel Batu SALMAN	---

Benden hiçbir zaman desteklerini esirgemeyen çok değerli anne ve babama…

TEŞEKKÜR

Çalışmam sırasında büyük bir sabır ve inançla yanımda olan, hiçbir koşulda benden desteklerini esirgemeyen, beni motive eden ailem ve arkadaşlarım başta olmak üzere, ihtiyacım olan her an bana yol gösteren, bilgi ve birikimini paylaşan saygıdeğer danışmanım Doç. Dr. M. Alper Tunga'ya teşekkür ederim.

İstanbul, 02.09.2015 Aybuke Doğan

ÖZET

BİREYSEL ARAÇ KREDİLERİNİN YASAL TAKİBE GİRME DURUMLARI HAKKINDA TAHMİN MODELLERİNİN OLUŞTURULMASI

Aybuke Doğan

Bilgi Teknolojileri
Tez Danışmanı: Doç. Dr. M. Alper Tunga

09 2015, 97 sayfa

Günümüzde teknoloji artık hayatın her alanında kendine vazgeçilmez bir yer edinmiştir. Her sektörde olduğu gibi finans sektöründe de teknolojik gelişmeler yakından takip edilmektedir ve uygulanmaktadır. Bu teknolojik gelişmeler ile birlikte sahip olunan bilgi miktarı hızla artmış ve veri tabanlarında gizli kalmış örüntülerin keşfedilmesi ihtiyacı yani "Veri Madenciliği" kavramı ortaya çıkmıştır. Veri Madenciliği ile bir anlam ifade etmeyen veri yığınlarının anlamlı, nitelikli ve kullanılabilir bilgiye dönüştürülmesi sağlanmaktadır. Veri madenciliği yardımı ile kurum ve kuruluşlar fark yaratarak rakiplerinden sıyrılabilmektedirler. Yani bir firma için başarılı olmanın anahtarı veri toplamanın ve bu veriyi doğru şekilde analiz etmenin önemini kavramasından geçmektedir. Geçmişe ait bir veri ne kadar iyi işlenir ve analiz edilirse geleceğe o kadar sağlam adımlar atılmış olur. Bu tez kapsamında bir tüketici finansman şirketinde kredileşen başvurulardan ödeme dönemleri içerisinde yasal takibe girmiş ve tüketici finansman şirketini kredinin zamanında ve tam olarak geri ödenmemesi riski ile dolayısı ile verilen ilgili krediden zarar etme ihtimali ile karşı karşıya bırakmış olan müşteri ve kredilerin genel yapısı ve benzer özellikleri ortaya koyulmaya çalışılmıştır.

Kurulan modeller ve oluşturulan kurallar sayesinde kredi verme sürecinde belirlenecek müşteri profilinden kaçınarak iyileştirmeler olması ve bu müşteri profilinden doğabilecek risklerin önceden tahmin edilip ona göre önlem alınmasının sağlanması amaçlanmıştır. Birinci bölümde tüketici finansman şirketinin işleyişi ve çalışmanın amacı hakkında ayrıntılı bilgiler verilmiştir.İkinci bölümde veri madenciliği hakkında genel bilgiler verilmiş olup kullanılan yöntemler ayrıntılı bir biçimde açıklanmıştır. Üçüncü bölümde kullanılan veri tanıtımı ve analizi yapılmış, modeller kurulmuştur.

Son bölümde ise kurulan modeller karşılaştırılmış ve sonuç verilmiştir. Tez kapsamında kurulan modeller SPSS Clementine 12.0 ve WEKA aracılığı ile geliştirilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Kredi Riski, Sınıflandırma Yöntemleri, WEKA, Clementine

ABSTRACT

ESTABLISHING A PREDICTIVE MODEL FOR INDIVIDUAL CAR LOANS EXPOSING TO LEGAL PROCEDURE

Aybuke Doğan

Information Technologies
Tez Danışmanı: Assoc. Prof. Dr. M. Alper Tunga

09 2015, 97 pages

In todays world, technology is essential in all fields of humanbeings lives. As all sectors, in finance sector technologic developments is also being followed and applied. Through these developments, possessed information has increased and the need for exploration of some hidden patterns in database, namely 'data mining" has appeared. By data mining, data accumulation that are not meaningful are converted into qualified and utilisable information. Owing to data mining, institutions can make a difference and pioneer to the sector.The key to succes for a company is to understand the importance of data mining and analyzing the data. Improved data storage and analyzing techniques enable companies to secure their future. Within the scope of this thesis, general characteristics and similar features of the loans and clients, which have been taken legal proceedings against within payback periods and faced with the possibility of making a loss of a loan given regarding the risk of not being paid back completely and in time among the loans in a consumer financing company. By the means of models formed and rules set, optimization in crediting processes avoiding specified client profile, foreseeing the risks that might arise due this specified client profile and taking precaution based on these risks.

In the first part, study provides an insight about operation mechanisms of a consumer financing company and detailed information about the purposes. As a second step, data mining and techniques used in the study is explained in details. In the advanced stage

of the study, data is defined, analyzed and modelled. Finally data models are compared and study has come to the conclusion. Data models have been developed by SPSS Clementine 12.0 and WEKA within the scope of the study.

Keywords: Data Mining, Credit Risk, Classification Methods, WEKA, Clementine

İÇİNDEKİLER

TABLOLAR.....xii

ŞEKİLLER.....xiv

KISALTMALAR.....xv

1.GİRİŞ.....2

2.LİTERATÜR TARAMASI.....8

3. VERİ MADENCİLİĞİNE GENEL BAKIŞ.....10

3.1 VERİ MADENCİLİĞİNİN TANIMI VE TEMEL KAVRAMLAR……10

3.2 VERİ MADENCİLİĞİNİN TARİHÇESİ.....13

3.3 VERİ MADENCİLİĞİ PROJE DÖNDÜSÜ : CRISP-DM.....15

3.3.1 İşi Anlama.....15

3.3.2 Veriyi Anlama.....16

3.3.3 Veriyi Hazırlama.....16

3.3.4 Modelleme.....16

3.3.5 Değerlendirme.....16

3.3.6 Uygulama, İzleme ve Güncelleme.....17

3.4 VERİ MADENCİLİĞİNİN KULLANIM ALANLARI.....18

3.5 VERİ MADENCİLİĞİNDE KULLANILAN MODELLER VE ÖĞRENME YÖNTEMLERİ.....19

3.5.1 Tahmin Edici Modeller.....19

3.5.2 Tanımlayıcı Modeller.....20

3.5.3 Denetimli Öğrenme.....20

3.5.4 Denetimsiz Öğrenme.....20

3.6 VERİ MADENCİLİĞİ YÖNTEM VE TEKNİKLERİ.....21

3.6.1 Sınıflama Ve Regresyon Modelleri.....21

3.6.2 Kümeleme Modelleri.....21

3.6.3 Birliktelik Kuralları Ve Ardışık Zamanlı Örüntüler.....22

4.KULLANILACAK VERİYE GENEL BAKIŞ.....23

4.1 VERİYE AİT BETİMLEYİCİ İSTATİSTİKLER.....24

4.2 KULLANILAN DEĞİŞKENLER VE AÇIKLAMALARI.....30

4.3 KULLANILAN DEĞİŞKENLERİN TÜRLERİ.....31
4.4 KULLANILAN NİTEL DEĞİŞKENLERİN KATEGORİK KARŞILIKLARI.....33
5.KULLANILAN YÖNTEMLER, UYGULAMALARI VE SONUÇLARI...37
5.1 REGRESYON TANIMI, LOJİSTİK REGRESYON, UYGULAMASI VE SONUÇLARI.....37
5.1.1 Doğrusal Regresyon.....37
5.1.2 Lojistik Regresyon.....38
5.1.2.1 ODDS.....41
5.1.2.2 ODDS ratio.....41
5.1.2.3 Lojit.....41
5.1.2.4 Parametrelerin anlamlılık testi.....42
5.1.2.4.1 Olabilirlik oran testi.....42
5.1.2.4.2 Wald testi.....43
5.1.2.4.3 Skor testi.....43
5.1.2.5 Lojistik regresyonda model seçimi.....44
5.1.2.5.1 Standart yöntem.....44
5.1.2.5.2 İleriye doğru adımsal yöntem.....44
5.1.2.5.3 Geriye doğru adımsal yöntem.....44
5.1.2.6 Modelin uyum iyiliği.....45
5.1.2.7 Lojistik regresyon uygulaması ve sonuçları.....46
5.2 KARAR AĞAÇLARI, UYGULAMASI VE SONUÇLARI.....55
5.2.1 Karar Ağaçları Tanımı Ve Temel Kavramlar.....55
5.2.1.1 Ağacın büyümesi.....56
5.2.1.2 Ayırma kriterleri.....56
5.2.1.2.1 Tek değişkenli ayırma kriterleri.....56
5.2.1.2.1.1 Bilgi kazancı.....56
5.2.1.2.1.2 Gini indeksi.....57
5.2.1.2.1.3 Kazanç oranı.....57
5.2.1.2.1.4 χ^2 testi.....57
5.2.1.2.1.5 F testi.....57
5.2.1.2.2 Çok değişkenli ayırma kriterleri.....57

5.2.1.3 Büyümenin durdurulması.....58
5.2.1.4 Budama.....58
5.2.1.5 Avantajları & dezavantajları.....58
5.2.2 CHAID Algoritması.....59
5.2.2.1 CHAID algoritması uygulaması ve sonuçları.....60
5.2.3 C&RT Algoritması.....69
5.2.3.1 C&RT algoritması uygulaması ve sonuçları.....70
5.2.4 QUEST Algoritması.....75
5.2.4.1 QUEST algoritması uygulaması ve sonuçları.....76
5.2.5 Karar Listesi.....81
5.3 BAYES TABANLI AĞLAR.....84
5.3.1 Naive Bayesian Algoritması.....84
5.4 DİĞER SINIFLANDIRMA YÖNTEMLERİ 84
5.4.1 Model Başarı Ölçütleri.....85
5.4.1.1 Verinin iki alt örnekleme ayrılması (hold out yöntemi).....85
5.4.1.2 Genel doğruluk.....86
5.4.1.3 Dengeli doğruluk.....86
5.4.1.4 Duyarlılık.....87
5.4.1.5 Seçicilik.....87
5.4.1.6 Hassasiyet.....87
5.4.1.7 Negatif tahmin değeri.....87
5.4.1.8 Yanlış pozitif oranı.....87
5.4.1.9 Yanlış negatif oranı.....87
5.4.1.10 Matthews korelasyon katsayısı.....88
5.4.1.11 F ölçütü.....88
5.4.1.12 ROC eğrisi.....88
5.4.2 WEKA'da Uygulama ve Sonuçları.....89
5.KARŞILAŞTIRMA VE SONUÇ.....94
KAYNAKÇA.....98

TABLOLAR

Tablo 1.1: 2015 yılı ikinci yarısı itibariyle BDDK verilerine göre türkiye'de faaliyet gösteren tüketici finansman şirketleri....2
Tablo 4.1: Veri kalitesi I....25
Tablo 4.2: Veri kalitesi II....26
Tablo 4.3: Betimleyici istatistikler I....27
Tablo 4.4: Betimleyici istatistikler II....28
Tablo 4.5: Betimleyici istatistikler III....29
Tablo 4.6: Değişkenler ve açıklamaları....30
Tablo 4.7: Değişken türleri....32
Tablo 4.8: "Çalışma Şekli" değişkeni kategorileri ve karşılıkları....33
Tablo 4.9: "Kredi Tür" değişkeni kategorileri ve karşılıkları....33
Tablo 4.10: "Karar Kategori" değişkeni kategorileri ve karşılıkları....34
Tablo 4.11: "Bölge" değişkeni kategorileri ve karşılıkları....34
Tablo 4.12: "LTV-Finansman Grup" değişkeni kategorileri ve karşılıkları....34
Tablo 4.13: "Ever Acc Dpd 30 Plus" değişkeni kategorileri ve karşılıkları....35
Tablo 4.14: "Ever Yasal Acc 16 M" değişkeni kategorileri ve karşılıkları....35
Tablo 4.15: "Marka" değişkeni kategorileri ve karşılıkları....35
Tablo 5.1: Değişken sayısı ve türlerine göre lojistik regresyon yöntemleri....39
Tablo 5.2: Lojistik regresyon model özeti....46
Tablo 5.3: Lojistik regresyon model katsayıları....47
Tablo 5.4: Lojistik regresyon bağımsız değişken açıklama oranı....47
Tablo 5.5: Lojistik regresyon doğru sınıflandırma oranı....48
Tablo 5.6: Lojistik regresyon değişken özeti....48
Tablo 5.7: Lojistik regresyon denetimli sınıflandırma özeti....53
Tablo 5.8: CHAID algoritması denetimli sınıflandırma özeti....60
Tablo 5.9: C&RT algoritması denetimli sınıflandırma özeti....70
Tablo 5.10: QUEST algoritması denetimli sınıflandırma özeti....76
Tablo 5.11: Karar listesi sınıflandırma özeti....81
Tablo 5.12: Hata Matrisi....86

Tablo 5.13: ROC Eğrisi Derecelendirme bilgileri.....89
Tablo 5.14: WEKA'da uygulanan algoritmalar ve performans özet bilgileri.....89
Tablo 5.15: WEKA'da uygulanan doğru sınıflandırma oranı en yüksek 15 algoritma ve performans özet bilgileri.....91
Tablo 5.16: WEKA'da uygulanan doğru sınıflandırma oranı en yüksek 15 algoritma ve bağımlı değişkenin kategorisine göre göstermiş oldukları performans özet bilgileri.....92
Tablo 6.1: Clementine 12.0'da geliştirilen algoritmalar ve sınıflandırma oranları.....94
Tablo 6.2: WEKA'da uygulanan algoritmalar ve performans özet bilgileri.....97

ŞEKİLLER

Şekil 1.1: Tüketici finansman şirketi başvuru kredileşme süreci.....6
Şekil 3.1: Veri madenciliğinde bilgiye ulaşma süreci.....13
Şekil 3.2: Veri madenciliği proje döngüsü : CRISP-DM.....18
Şekil 5.1: Lojistik regresyon değişken önem sıralaması.....53
Şekil 5.2: Lojistik regresyon kazanç grafiği.....54
Şekil 5.3: Clementine 12.0'da lojistik regresyon modellemesi.....54
Şekil 5.4: CHAID algoritması değişken önem sıralaması.....61
Şekil 5.5: CHAID algoritması karar ağacı 64
Şekil 5.6: CHAID algoritması karar ağacı düğüm 5,18,19,31,32,33,34,40,41.....66
Şekil 5.7: CHAID algoritması karar ağacı düğüm 6,7,20,21,22,23,35,36,37,42,43,44,45,46.....68
Şekil 5.8: Clementine 12.0'da CHAID algoritması modellemesi.....69
Şekil 5.9: C&RT algoritması değişken önem sıralaması.....71
Şekil 5.10: C&RT Algoritması Karar Ağacı.....73
Şekil 5.11: Clementine'da C&RT algoritması modellemesi.....75
Şekil 5.12: QUEST algoritması değişken önem sıralaması.....77
Şekil 5.13 QUEST algoritması karar ağacı.....79
Şekil 5.14: Clementine 12.0'da QUEST algoritması modellemesi.....81
Şekil 5.15: Clementine 12.0'da karar listesi algoritması modellemesi.....83

KISALTMALAR

ABD : Amerika Birleşik Devletleri
ARFF : Attribute Relationship File Format
BDDK : Bankacılık Düzenleme Denetleme Kurumu
CARMA : Continuous Association Rule Mining Algorithm
CHAID : Chi-Squared Automatic Interaction Detection
CRISP : Cross Industry Standard Process for Data Mining
CRM :Customer Relationship Management
C&RT : Classification and Regression Tree
ENIAC : Electrical Numerical Integrator And Calculator
ERP : Enterprise Resource Planning
KKB : Kredi Kayıt Bürosu
LN : Logarithm (Natural)
NCR : National Cash Register
OLAP : Online Analitik Süreç
QUEST : Quick, Unbiased, Efficient Statistical Tree
SAS : Statistical Analysis System
SEMMA : Sample,Explore,Modify,Model,Assess
SPSS : Statistical Package for the Social Sciences
SQL : Yapısal Sorgu Dili
SSE : Square Sum Error
WEKA : Waikato Environment for Knowledge Analysis

1. GİRİŞ

Bir ülkede iktisadi büyümenin en önemli girdisi, mal ve hizmet üretimindeki artışa en büyük etken olan tüketimdir. Bu noktada ise tüketimi gerçekleştiren tüketicilerin, uygun zamanda doğru ve etkili bir biçimde ihtiyaçlarını ve isteklerini karşılayabilmeleri önemli bir rol oynamaktadır.

Finansal piyasalar, arz ile talebi karşı karşıya getirerek üretim-tüketim döngüsünün devamlılığını sağlarlar. Bir ülkede finansal faaliyetler ne kadar gelişir ve sistematik bir hal alırsa ekonomik kalkınma da o oranda büyük ve hızlı olur. Bunun için finansal piyasalar, kullanılmayan fonların yatırım ve tüketime dönüşmesine aracılık ederek bu ekonomik kalkınmaya katkı sağlarlar. En büyük aracı olan bankaların yanı sıra, tüketici finansman şirketleri olarak adlandırılan, genellikle taşıt ya da dayanıklı tüketim mallarının finanse edilmesine aracılık eden, tüketicilere ödeme kolaylığı sağlayan kuruluşlar da gerekli yasal düzenlemelerin sağlanması ile birlikte 1990'lı yıllardan itibaren Türkiye ekonomisindeki yerlerini almışlardır.

2015 yılı ikinci yarısı itibariyle BDDK (Bankacılık Düzenleme Denetleme Kurumu) verilerine göre Türkiye'de faaliyet gösteren Tüketici Finansman şirketleri tablo 1.1'de gösterildiği gibidir.

Tablo 1.1: 2015 yılı ikinci yarısı itibariyle BDDK verilerine göre türkiye'de faaliyet gösteren tüketici finansman şirketleri

1	ALJ FİNANSMAN A.Ş.
2	DD FİNANSMAN A.Ş.
3	KOÇ FİAT KREDİ FİNANSMAN A.Ş.
4	KOÇ FİNANSMAN A.Ş.
5	MAN FİNANSMAN A.Ş.
6	MERCEDES BENZ FİNANSMAN TÜRK A.Ş.
7	ORFİN FİNANSMAN A.Ş.
8	PSA FİNANSMAN A.Ş.
9	ŞEKER MORTGAGE FİNANSMAN A.Ş.
10	TEB FİNANSMAN A.Ş.
11	VFS FİNANSMAN A.Ş.
12	VOLKSWAGEN DOĞUŞ FİNANSMAN A.Ş.

Kaynak:BDDK

Tüketici finansman şirketleri daha çok bireysel tüketicilere ya da küçük - orta büyüklükteki ticari oluşumlara finansman sağlayabildiği gibi üreticiler için de ihtiyaç durumunda stok finansmanını desteklerler.

Genel olarak bir tüketici finansman şirketinin işleyiş yapısı aşağıdaki şekildedir:

i. Tüketici finansman şirketi ve malı temin edecek olan satıcı (bayi) arasında önceden bir anlaşma yapılır ve sözleşme imzalanır. Böylece söz konusu şirket artık o malı satın almak isteyen tüketicileri satıcılar kanalı ile finanse etmeye hak kazanır.
ii. Tüketici kendisine ve bütçesine en uygun malı seçer ve kredili olarak satın almaya karar verir. Bu aşamada satıcı , finansman şirketinin daha önceden belirlemiş olduğu tüketiciye ait bilgi ve belgeleri eksiksiz ve doğru bir biçimde doldurarak değerlendirilmek üzere finansman şirketine göndermekle yükümlüdür.
iii. Tüketiciye ait bilgi ve belgeler finansman şirketine ulaştığında, sistem tarafından otomatik ya da gerektiğinde tahsis birimi tarafından manuel olarak incelenir ve kredinin tüketiciye verilip verilmeyeceğine , verilecek ise şartlarına karar verir, verilmeyecek ise nedenlerini değerlendirir ve mümkün olan en kısa sürede tüketiciye sonucu bildirir.

Bu noktada karar aşamasını daha ayrıntılı anlatmak gerekirse:

Bayiler, tüketici finansman şirketi ile aralarında geliştirilen bir sistem entegrasyonu sayesinde tüketicinin kredi başvurusunu finansman şirketine iletir. Tüketici finansman şirketinde işleyen sistem ise otomatik olarak karar verilen başvurular ve manuel olarak karar verilen başvurular olmak üzere 2'ye ayrılır.

i. Sisteme gelen bir başvuruda öncelikle **politika kuralları** gereği direkt red edilecek başvurular, bu kurallar çalıştırılarak ayıklanır. Politika kuralları, genel olarak başvuru sahibinin yakın tarihte herhangi bir yasal ya da yasaklı sürecinin, dolandırıcılık faaliyetinin olup olmadığının tespit edilmeye çalışıldığı kurallardır. Dönem dönem ülkenin içinde bulunduğu ekonomik ve siyasi duruma bağlı olarak ortaya çıkan önlem kuralları da politika kurallarının içerisine girer.

Politika kuralları nedeniyle red edilen bir başvurunun red kararı kesindir ve bu karar istisnai durumlar hariç manuel olarak ezilemez.

ii. Politika kuralları ile red edilen başvurular değerlendirme sürecinden çıktıktan sonra geriye kalan başvurularda ilk önce **otomatik red kuralları** çalışır.
    a. Ardından otomatik olarak red edilen başvurularda istisnai bir durum olup olmadığının kuralları çalışır.

Yani otomatik olarak red edilmesine karar verilen bir başvurunun manuel olarak tekrar gözden geçirilmesini sağlayacak olumlu bir durumunun olup olmadığı belirlenir. Var ise tahsisçinin onayına sunulmak üzere **otomatik red istisnası** adı altında manuel incelenmek üzere iletilir ve kredinin akıbeti tahsisçinin görüşüne bırakılır. Bu aşamada başvuru tahsisçinin kararı olumsuz ise **manuel red** , tahsisçinin görüşü olumlu ise **manuel onay** kodu alarak sürecini tamamlamış olur.

Otomatik red istisnası kurallarına takılmayan başvurular ise **otomatik red** kodunu alarak sistemden çıkar ve kredileşme ihtimali ortadan kalkar. Otomatik red kodunu alan başvurularda herhangi bir manuel inceleme yapılmaz.

iii. Otomatik red kurallarına takılmayan başvurularda ikinci olarak **otomatik onay kuralları** çalışır.
    b. Ardından otomatik olarak onay alan başvurularda istisnai bir durum olup olmadığının kuralları çalışır.

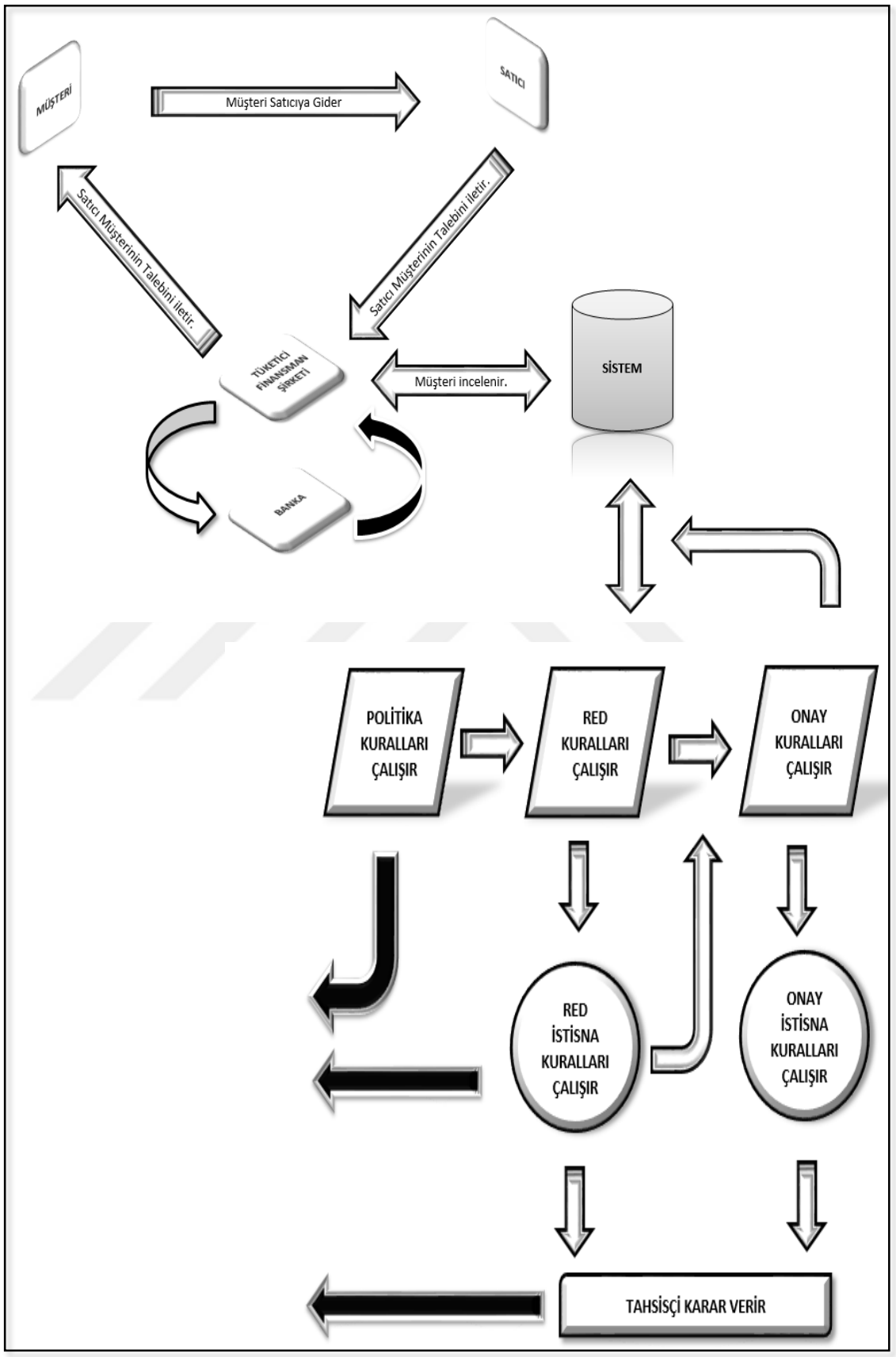
Yani otomatik olarak onay alan bir başvurunun manuel olarak tekrar gözden geçirilmesine neden olacak olumsuz bir durumunun olup olmadığı belirlenir. Var ise tahsisçinin onayına sunulmak üzere **otomatik onay istisnası** adı altında manuel incelenmek üzere iletilir ve kredinin akıbeti tahsisçinin görüşüne bırakılır. Bu aşamada başvuru tahsisçinin kararı olumsuz ise **manuel red** , tahsisçinin görüşü olumlu ise **manuel onay** kodu alarak sürecini tamamlamış olur.

iv. Sisteme gelen başvuru ana olarak politika kuralları, otomatik red kuralları ve otomatik onay kurallarının hiç birine takılmaz ve sistem kararsız kalır ise başvuru bu kez herhangi bir istisna kodu almadan direkt olarak tahsisçinin görüşüne sunulur. Ve tahsisçinin kararı olumsuz ise başvuru manuel red ,

tahsisçinin görüşü olumlu ise manuel onay kodu alarak sürecini tamamlamış olur.

v. Finansman şirketinde bir başvuru için otomatik ya da manuel olarak karar verildikten sonra, otomatik ya da manuel red kodu alan başvuru sahibinin tüketici kredisi alma ihtimali ortadan kalkar, otomatik ya da manuel onay kodu alan başvuru sahibinin tüketici kredisi alma ihtimali ise artık kendi kararına bağlıdır. Başvuru sahibi kredi şartlarını kabul eder ise bu kez onaylanan başvurunun kredileşme süreci başlatılır. Başvuru sahibi kredi şartlarını kabul etmez ise süreç sonlanmış olur ve onaylanmış başvuru kredileşmez. Şekil 1.1'de görsel olarak tüketici finansman şirketinin işleyiş yapısı ve karar aşamaları gösterilmiştir.

Şekil 1.1: Tüketici finansman şirketi başvuru kredileşme süreci



Başvuru kredileştikten sonra ise finansman şirketi açısından önemli bir süreç başlar; tüketiciye verilen bu kredinin ödeme takibinin yapılması ve kredi riskinin izlenmesi. Kredi riski, tüketiciye verilen kredinin tüketici tarafından geri ödenmemesi ihtimalidir. Tüketicinin başvurusu değerlendirilirken geçmiş bilgileri ve o anki bilgileri göz önünde bulundurularak karar verilir ve bu değerlendirme sonucunda bir risk alınır. Ancak alınan bu risk, kredinin geri ödenme sürecinde artabilir ya da azalabilir. Bu sebeple hem kredi başvurusunun değerlendirme aşamasında hem de kredi verildikten sonra takibinin yapılması aşamasında çeşitli değerlendirme modelleri geliştirilerek kredi riskini öngörüp önlem alabilmek bir finansman şirketinin en önemli amacıdır.

Bu tez kapsamında kredileşen başvurulardan ödeme dönemleri içerisinde yasal takibe girmiş ve tüketici finansman şirketini kredinin zamanında ve tam olarak geri ödenmemesi riski ile dolayısı ile verilen ilgili krediden zarar etme ihtimali ile karşı karşıya bırakmış olan müşteri ve kredilerin genel yapısı ve benzer özellikleri ortaya koyulmaya çalışılmıştır. Kurulan modeller ve oluşturulan kurallar sayesinde kredi verme sürecinde belirlenen müşteri profilinden kaçınarak iyileştirmeler olması ve bu müşteri profilinden doğabilecek risklerin önceden tahmin edilip ona göre önlem alınmasının sağlanması amaçlanmıştır.

2. LİTERATÜR TARAMASI

Veri madenciliği çeşitli tekniklerle, veri kümeleri içerisinde gizli kalmış olan anlamlı bilgilere ulaşmayı sağlayan bir analiz sürecidir. Bu bağlamda veri madenciliği teknikleri birçok alanda olduğu gibi bankacılık alanında da yaygın bir şekilde kullanılmaktadır ve bu konu üzerinde çeşitli çalışmalar yapılmaktadır.

Takipteki kredi oranının aylık bazda tahmini ve risk yönetimine yönelik Başak Tanınmış Yücememiş ve İnanç Asım Sözer tarafından yapılan ve Marmara Üniversitesi e-dergisi Cilt 3, Sayı 5 (2011)'te yayımlanan "Bankalarda Takipteki Krediler:Türk Bankacilik Sektöründe Takipteki Kredilerin Tahminine Yönelik Bir Model Uygulamasi" başlıklı çalışmada Türkiye'nin ekonomik verileri kullanılmış ve geçmiş dönem verilerinden yararlanılmıştır. Kurulan modele göre takipteki kredi oranını etkileyen en önemli faktörün bankaların geçmiş dönem performansı olduğu gözlemlenmiştir. Bir önceki dönemde takipteki kredilerini iyi yönetebilen bankaların sonraki dönemlerde ekonomideki bozulmalardan minimum seviyede etkilendiği saptanmıştır. Aynı şekilde bir önceki dönemde sanayi üretiminde ve TL'nin değerinde meydana gelen değişmelerinde bir sonraki dönemde takipteki kredi oranını ters oranda etkilediği görülmüştür.

Aslı Çalış tarafından 2013 senesinde yapılan "Veri Madenciliği Yaklaşimi İle Bireysel Müşterilerin Kredi Ödeme Performanslarinin Değerlendirilmesi" konulu Kocaeli Üniversitesi fen bilimleri enstitüsü endüstri mühendisliği anabilim dali yüksek lisans tezi kapsamında bankacılık sektöründe gerçekleştirilen bir diğer çalışma ile veri madenciliği yöntemlerinden kümeleme ve sınıflandırma algoritmaları kullanılarak bireysel kredi müşterilerinin ödeme performanslarına göre profilleri çıkartılmış ve gelecekteki kredi müşterileri için kurallar oluşturularak bireysel kredilerde kanuni takibe düşme oranının azaltılması hedeflenmiştir.SPSS Clementine kullanılarak yapılan bu çalışma kapsamında uygulamanın birinci aşamasında, mevcut kredi müşterileri için k-ortalamalar yöntemi ile kümeleme analizi yapılarak müşteriler yaş, cinsiyet, aylık gelir, medeni hal, öğrenim durumu, ödeme durumu gibi on iki farklı değişkenden yararlanılarak davranışlarına göre gruplandırılmıştır. Ödeme durumlarına göre kümeler incelendiğinde ilk kümeyi oluşturan müşterilerin tamamının kanuni takipte olduğu,

ikinci ve üçüncü kümedeki müşterilerin sırasıyla yüzde 63,64 ve yüzde 98,48'lik oranlarla ödemelerini aksatmadıkları sonucuna ulaşılmıştır. Bu kümelendirmeye göre mevcut müşterilerin yeniden kredi talep etmeleri durumunda, bulundukları kümelere bağlı olarak başvurularının değerlendirmeye alınabileceği sonucu çıkmıştır. Uygulamanın ikinci kısmında ise C&RT, C5.0, QUEST ve CHAID algoritmaları ile müşteriler sınıflandırılmış ve geleceğe yönelik tahminlerde bulunabilmek için karar ağaçları ile kural çıkarımı yapılmıştır. Algoritmalar karşılaştırıldığında CHAID algoritmasının yüzde 95,5 'lik doğru sınıflandırma oranı ile en iyi sonucu sağladığı görülmüştür. Elde edilen karar ağacında ilk dallanmanın aylık gelir ile başladığı gözlemlenmiştir. Diğer önemli dallanmayı ise kredi müşterilerinin ilgili bankanın maaş müşterisi olup olmamasının sağladığı saptanmıştır. Çünkü maaş müşterisi olan kişilerin gerektiğinde hesabına bloke konularak kredi ödenmesinin aksamasının önüne geçilebildiği tespit edilmiştir.

3. VERİ MADENCİLİĞİNE GENEL BAKIŞ

3.1 VERİ MADENCİLİĞİNİN TANIMI VE TEMEL KAVRAMLAR

Teknolojinin hızla gelişmesi ile birlikte bankacılık işlemlerinden alışveriş işlemlerine, hastane işlemlerinden sosyal paylaşım siteleri üzerinden yapılan paylaşımlara kadar varan her türlü işlem artık elektronik ortamlarda yapılmaya başlanmıştır. Bu sayede elde edilen milyarlarca verinin hem rahatlıkla depolanabilmesi hem de bu verilere gerektiği anda kolayca ulaşılabilmesi sağlanmaktadır. Ancak bu şekilde depolanmış ham veri kurum ve kuruluşlar için hiç bir anlam ifade etmemektedir, veri yığını olmaktan öteye geçmemektedir.

Rekabet ortamının gittikçe arttığı bir dünyada bir işletmenin varlığını devam ettirebilmesi, fark yaratabilmesi ve kendini bulunduğu noktadan çok daha ileri seviyelere taşıyabilmesi ancak elindeki veriyi anlamlı ve kullanılabilir bilgiye dönüştürmesi ile gerçekleşebilir. Çeşitli yazılımların sürekli gelişmesi elde edilen ham veriyi nitelikli bilgiye dönüştürebilmeyi kolaylaştırmaktadır. Ham verinin işlenerek içerisindeki gizli ilişkilerin, birlikteliklerin ve örüntülerin ortaya çıkartılması ile kullanılabilir nitelikli bilgi ortaya çıkmaktadır. Yani bir firma için başarılı olmanın anahtarı veri toplamanın ve bu veriyi doğru şekilde analiz etmenin önemini kavramasından geçmektedir. Geçmişe ait bir veri ne kadar iyi işlenir ve analiz edilirse geleceğe o kadar sağlam adımlar atılmış olur.

Veri madenciliği, diğer bir kullanım adı ile veritabanında bilgi keşfi geniş hacimli veri tabanlarında anlamı daha önceden bilinmeyen ve kestirilemeyen ancak potansiyel olarak yararlı, anlamlı ve kullanışlı bilgiye ulaşmak için kullanılan istatistiksel, matematiksel ve yapay zeka kökenli metodların tümü olarak tanımlanabilir. Veri madenciliği kısaca ham veriden nitelikli bilgiye ulaşılmasını sağlayan bir veri çözümleme sürecidir.

Frawley veri madenciliğini "Daha önceden bilinmeyen ve potansiyel olarak yararlı olma durumuna sahip verinin keşfedilmesi" olarak tanımlamıştır. Berry ve Linoff bu kavrama "Anlamlı kuralların ve örüntülerin bulunması için geniş veri yığınları üzerine yapılan keşif ve analiz işlemleri" şeklinde bir açıklama getirirken Sever ve Oğuz veri

madenciliği hakkında "Önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin büyük ölçekli veritabanlarından otomatik biçimde elde edilmesini sağlayan veri tabanlarında bilgi keşfi süreci içerisinde bir adımdır." tanımını kullanmışlardır. Gartner Group tarafından yapılan bir diğer tanıma göre ise veri madenciliği, istatistiksel ve matematiksel yöntemlerle birlikte örüntü tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni ilişki, örüntü ve eğilimlerin keşfedilmesi sürecidir.Veri Madenciliğinde kullanılan temel kavramları aşağıdaki şekilde özetlemek mümkündür:

Ham Veri: Elde edilen verinin düzenlenmeden ve işlenmeden önceki halidir.

Veri: Yorumlanmak ve sunulmak amacı ile ölçüm, sayım, gözlem ya da deney yolu ile elde edilmiş, çözümlenmiş ve özetlenmiş gerçeklerdir. Sayısal veriler *nicel (numerik),* sayısal olmayan veriler ise *nitel (kategorik)* olarak tanımlanmaktadır. Nicel veriler de kendi aralarında sayım yolu ile elde edilen ve sayma sayıları cinsinden ifade edilen *kesikli nicel veriler* ile ölçüm yolu ile elde edilen ve gerçel sayılar ile ifade edilen *sürekli nicel veriler* olmak üzere 2'ye ayrılırlar. Nitel veriler ise anlamlı bir sıralamaya tabii olmayan *nominal nitel veriler* ve anlamlı bir sıralamaya tabii olan *ordinal nitel veriler* olmak üzere 2'ye ayrılırlar.

Enformasyon: Verinin düzenlenmiş, ilişkilendirilmiş, anlamlandırılmış içinde potansiyel bilgi barındıran halidir.

Bilgi: Enformasyonun bilgiye dönüşmesi algılanması, özümsenmesi ve sonuç çıkartılmasıyla gerçekleşir.

Nitelikli Bilgi: Çeşitli modelleme ve tahmin yöntemlerini kullanarak bir karar almayı sağlayacak olan öngörüdür.

Bilgelik: Ulaşılmak istenen noktadır. Bilgilerin toplanıp sentezlenmesi ile ortaya çıkan ve yetenek tecrübe gibi kişisel niteliklerle birleştirilen olgudur.

Bu temel kavramlara veri madenciliğinde sıkça kullanılan diğer terimleri de eklemek ham veriden nitelikli bilgi elde etme sürecinin daha anlaşılır olması açısından fayda sağlayacaktır.

Veri Kaynağı: Veri tabanları, excel dosyaları, metin dosyaları, xml dosyaları, sav ve sas uzantılı çeşitli dosyalar gibi verinin tutulduğu alanlardır.

Veri Dönüştürme: Veriyi ayrıştırmak, dönüştürmek ve yüklemek işlemlerinin tümünü ifade eder. (ETL - Extract, Transform, Load)

i. **Ayrıştırma:** Veriyi, veri kaynağından yardımcı araçlar ya da kod ile alma anlamına gelmektedir.
ii. **Dönüştürme:** Bir çok farklı kaynaktan gelen çeşitli karakteristik özelliklere sahip olan verinin son hedef olan veri ambarına uygun yapıda olması için temizlenmesi, formatının hedef kayanağa uygun olarak ayarlanması ve kalitesinin arttırılması işlemidir.
iii. **Yükleme:** Veri kaynağından ayrıştırılan ve dönüştürülen verinin hedef sisteme yüklenmesidir.

Veri dönüştürme kısaca verinin kaynağından alınıp iş önceliklerine uygun bir şekilde temizlenmesi, birleştirilmesi, eşleştirilmesi, gerektiğinde ek bilgiler ile zenginleştirilerek kullanıma hazır en etkin hale getirilip işlenecek hedef kaynağa yüklenmesidir.

Veri Ambarı : Verileri düzenli bir biçimde uygun formatta saklayarak bu kayıtlar üzerinde daha kolay analiz yapılmasına olanak sağlayan özelleştirilmiş veri tabanlarıdır. "Veri ambarları, tüm operasyonel işlemlerin en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen ve tarihsel derinliği olan veri depolama sistematiği olarak tanımlanabilir." denmiştir Yaralıoğlu tarafından.

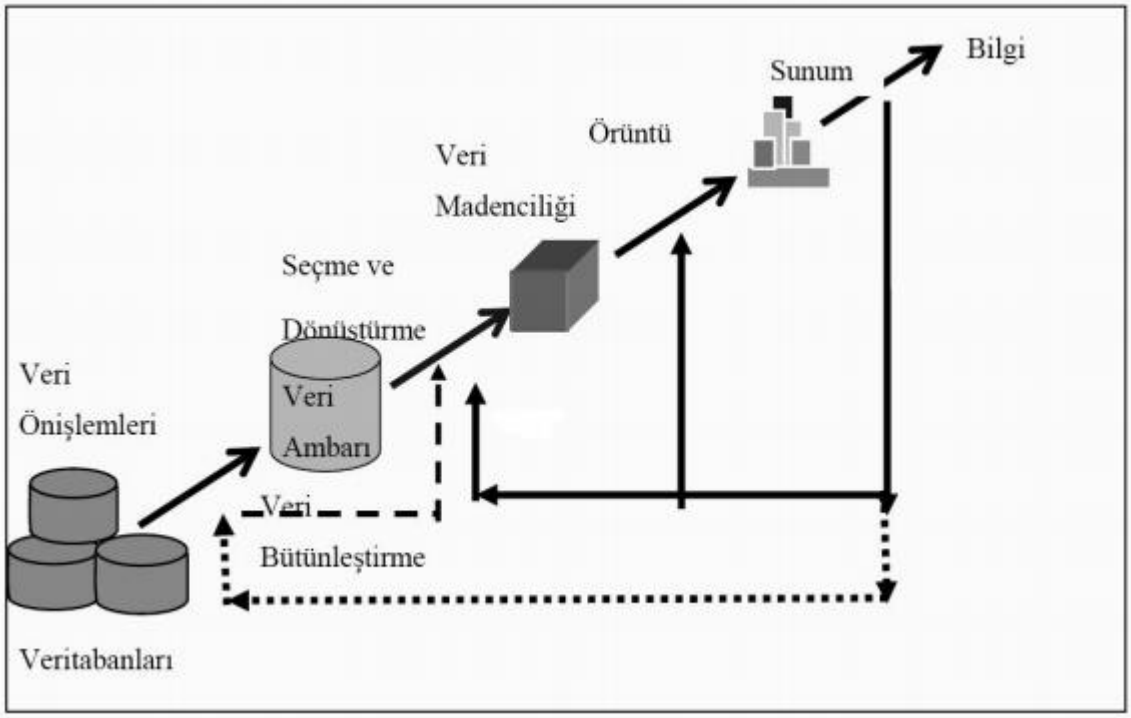
Data Mart: Veri ambarının alt kümesidir. Verinin tamamı yerine belirli bir kısmına bakış açısı sağlar. Böylece üzerinde çalışılacak konuya uygun veriye hızlı ve kolay ulaşım sağlanır. Tüm datanın karmaşıklığı ile karşı karşıya kalmaktansa ihtiyaç olan dataya erişim ve daha kolay analiz yapılmasına olanak sağlar.

Yapısal Sorgu Dili (SQL) : Veri tabanındaki verilere ulaşmak için kullanılan sorgulama dilidir, komutlar bütünüdür.

Online Analitik Süreç (OLAP) : Çok boyutlu veri analizini destekleyen sorgulama metodudur. Karmaşık problemlere cevap vermek, daha detaylı analizler yapabilmek adına ilişkisel veri tabanlarında depolanan veriden çok boyutlu veri küpü kurmaya

olanak sağlayan bir teknolojidir.Veri madenciliğinde bilgiye ulaşma süreci şekil 3.1'de görsel olarak gösterilmiştir.

Şekil 3.1: Veri madenciliğinde bilgiye ulaşma süreci



*Kaynak:*www.sertacogut.com

3.2 VERİ MADENCİLİĞİNİN TARİHÇESİ

Veri madenciliğinin temelini ilk sayısal bilgisayar olan ve 1946 yılında 2. Dünya Savaşı sırasında ABD ordusu için Amerikalı bilim adamları John Mauchly ve J. Presper Eckert tarafından geliştirilen ENIAC (Electrical Numerical Integrator And Calculator) 'a kadar dayandırmak mümkündür. Bilgisayar ve yazılım uzmanlarının geliştirdikleri ürünler zamanla kullanıcıların istek ve ihtiyaçları ile şekillenerek bilgisayarların bugünkü halini almasına katkı sağlamışlardır. Bilgisayarların etkin kullanılması verilerin depolanması ile başlamaktadır. İlk başlarda yalnızca karmaşık hesaplamalar yapmak amacıyla geliştirilen bilgisayarlar zaman içerisinde kullanıcıların ihtiyaçları doğrultusunda veri depolama işlemleri için de kullanılmaya başlanmıştır. Ve böylece veri tabanları ortaya

çıkmıştır. Zaman içerisinde veri tabanlarının genişleme eğiliminde olması donanımsal olarakta bu verilerin tutulacağı ortamların genişlemesini ve veri ambarı kavramının ortaya çıkmasını gerektirmiştir. Saklanmak istenen verilerin fiziksel sürücülerde tutulması ile gittikçe büyüyen veri tabanı organizasyonunun yönetilmesi güç bir hal almaya başlamıştır. Bu da veri modelleme kavramını ortaya çıkartmıştır. İlk olarak basit veri modelleri olan *Hiyerarşik* ve *Şebeke* veri modelleri geliştirilmiştir. Ancak çoklu ilişki kurmanın mümkün olmadığı bu veri modelleri kullanıcıların ihtiyaçlarını tam olarak karşılayamamıştır ve günümüzde en sık kullanılan ve gelişimi devam eden *Varlık-İlişki*, *İlişkisel* ve *Nesne-Yönelimli* veri modelleri geliştirilmiştir.

İhtiyaçlar ile şekillenen ve gelişen teknoloji ile birlikte verilerin saklanması, düzenlenmesi, organize edilmesi sağlansa da hedeflenen sonuca ulaşmak bir sorun haline gelmeye başlamıştır.

1960'lı yıllarda bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlanması ile birlikte veri madenciliği kavramı ortaya çıkmıştır. O yıllarda yeterince vakit ayrılarak bir tarama yapıldığında istenilen veriye ulaşılabileceği gerçeği kabul edilmiştir. Bu işleme veri madenciliği yerine önceleri *Veri Taraması* ve *Veri Yakalanması* gibi isimler verilmiştir.

Veri madenciliği ismi ilk olarak 1990'lı yıllarda ortaya atılmıştır. Ve böylece kökeninde istatistik, makine öğrenimi, veritabanları, otomasyon, pazarlama, araştırma gibi kavramlar yatan yaklaşımlar getirilmeye başlanmıştır. O zamana kadar verilerin değerlendirilmesi ve analizi konusunda hizmet veren bir yöntemler topluluğu olan istatistik, bilgisayarların veri analizi için kullanılmasıyla birlikte hız kazanmaya başlamıştır. Ve 1990'lardan sonra istatistik veri madenciliği ile ortak bir platforma taşınmış olup sıkı bir çalışma birlikteliği ortaya konmuştur. İstatistiğin yanısıra veri madenciliği, veri tabanları ve makine öğrenimi disipliniyle birlikte yol almıştır. Günümüzdeki yapay zeka çalışmalarının temelini oluşturan makine öğrenimi, bilgisayarlara karmaşık örüntüleri algılama ve veriye dayalı akılcı kararlar verebilme becerisi kazandırmaktır. İlk zamanlarda bilgisayarlar, insan öğrenmine benzer bir yapıda inşa edilmeye çalışılmıştır ancak 1980'lerden sonra bu yaklaşım değiştirilerek bilgisayarlar daha spesifik konularda kestirim algoritmaları üretmeye yönelik inşa

edilmiştir. Böylece uygulamalı istatistik ve makine öğrenim kavramları veri madenciliği altında bir araya getirilmiştir.

3.3 VERİ MADENCİLİĞİ PROJE DÖNGÜSÜ : CRISP-DM

Veri madenciliği proje yönetim döngüsü veya veri madenciliği hayat döngüsü olarakta ifade edilen çeşitli yöntemler arasında en yaygın olanı üyeleri Daimler-Benz, SPSS ve NCR olan bir konsorsiyum tarafından geliştirilen ve pek çok veri madenciliği aracı tarafından şu an yaygın olarak baz alınmakta olan CRISP (Cross Industry Standard Process for Data Mining) yöntemidir. IBM SPSS, STATSOFT, STATISTICA gibi veri madenciliği araçları bu yöntemi baz almaktadırlar. Bir diğer yaygın kullanım ağına sahip olan SAS ise SEMMA(Sample,Explore,Modify,Model,Assess) sürecini baz almaktadır.

CRISP döngüsü 6 adımdan oluşmaktadır. Bu adımlar:

3.3.1 İşi Anlama (Business Understanding)

Veri madenciliğinin en önemli ve en zor kısmı olan ilk adım hedeflerin belirlenmei ve buna uygun bir planın çıkartılmasıdır. Bu aşamada sorunun tanımlanması, ne tür bir veri madenciliği tekniğinin uygulanacağına karar verilmesi, ne tür bir analiz yapılacağının belirlenmesi gerekmektedir. Çünkü amacı açıkça belirlenmemiş, süreci bilinmeyen ya da eksik/yanlış ifade edilen, sorun ile birebir örtüşmeyen bir veri madenciliği çalışması sorunu çözmeye yetmeyeceği gibi zaman ve maliyet kaybı başta olmak üzere başka problemlere de yol açabilir.

3.3.2 Veriyi Anlama (Data Understanding)

Veriyi toplarken veri kümesinde ne tür değişkenlerin bulunduğunun tespit edilmesi, bu değişkenlerin ve değerlerinin ne ifade ettiğinin bilinmesi, hangi amaçla kullanıldıklarının anlaşılması gerekmektedir. Çünkü veriyi tanımadan yapılacak bir analiz yanlış ya da anlamsız bir model kurmaya sebep olabilir. Veriye hakim olmadan yaratılacak bir model görünürde anlamlı dahi olsa gerçekleri yansıtmaz ve uygulanabilir değildir ve hatalı kararlar alınarak daha büyük problemler çıkmasına sebep olabilir.

3.3.3 Veriyi Hazırlama (Data Preparation)

Sahip olunan ham veriden veri madenciliğinde kulanılabilecek veriyi elde etmek için yapılan tüm işlemler veri hazırlığıdır ve büyük bir zaman ve özveri gerektirmektedir. Verilerin temizlenmesi (tutarsız ve hatalı verilerin ayıklanması), bütünleştirilmesi (farklı türde ama aynı anlama gelen verilerin tek türe dönüştürülmesi), indirgenmesi (veri sayısı ya da değişken sayısının azaltılması), dönüştürülmesi (içeriğinin korunarak şeklinin değiştirlmesi,normalizasyon) ihtiyaç varsa yeni değişkenlerin oluşturulması işlemleri dikkatlice yapılmalıdır.

3.3.4 Modelleme (Modeling)

Tanımlanan sorun için en uygun modelin kurulabilmesi, bir çok model tekniğinin denenmesi ile mümkün olmaktadır. Her bir modelin tekniği farklı olabileceğinden modelin kurulması aşamasında veri hazırlama adımına tekrar dönülmesi gerekebilir. En iyi olduğu düşünülen model bulunduğunda öğrenme süreci tamamlanmış olur ve belirlenen kurallar yeni örneklere uygulanarak model geliştirilir.

3.3.5 Değerlendirme (Evaluation)

Kalitesi yüksek ve anlamlı bir model kurulduktan sonra uygulama aşamasına geçmeden önce modelin değerlendirilerek belirlenen hedefe uyup uymadığı, problemin çözümüne katkı sağlayıp sağlamadığı değerlendirilir, modeli kurana kadar uygulanan adımlar

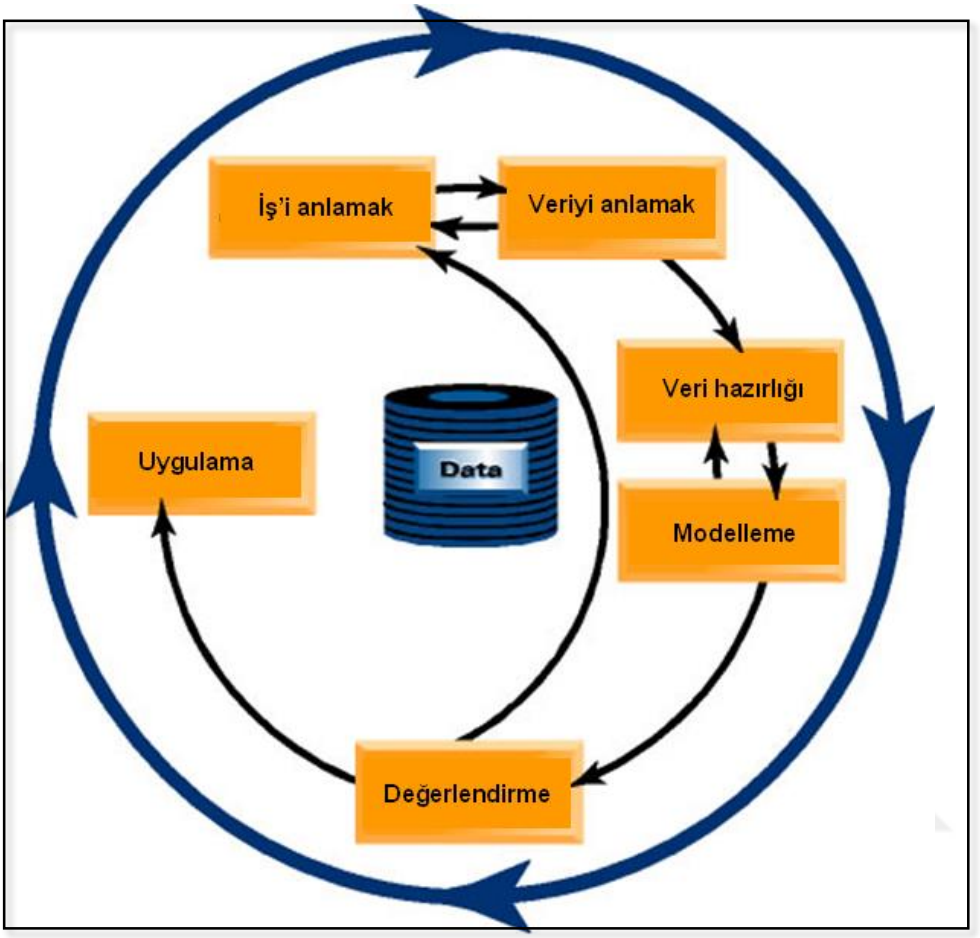
tekrar gözden geçirilerek eksik ya da gözden kaçan bir şey olup olmadığının tespiti yapılır. Modeli değerlendirirken örneğin danışmanlı öğrenmede seçilen algoritmaya uygun veriler hazırlandıktan sonra verilerin büyük bir kısmı eğitim kümesi olarak seçilir ve geri kalan kısmı modelin geçerliliğinin testi için ayrılır. Modelin kurulması eğitim kümesi kullanılarak gerçekleştirildikten sonra test kümesi ile de modelin doğruluk düzeyi belirlenmektedir.

3.3.6 Uygulama, İzleme ve Güncelleme (Deployment)

Kurulan en iyi modelin değerlendirmesi yapılıp nihai karar verildikten sonra elde edilen sonucun düzenlenmesi ve bunun kullanıcın anlayabileceği bir bilgiye dönüştürülmesi gerekmektedir. Hedefin ne olduğuna göre değişmekle birlikte bu bilgiye dönüşüm bir raporlamadan ibaret olabileceği gibi modelin entegre edileceği başka bir sistem düzenlenmesi de olabilir. Burada önemli olan modelin ve uygulamasının başlangıçta belirlenen hedeflerle örtüşerek probleme çözüm sağlamasıdır. Bir model uygulandıktan sonra her şey bitmiş demek değildir. Gelişen ve değişen dünyaya dolayısı ile de verilere bağlı olarak kurulan model bir zaman sonra anlamsızlaşabilir.Bu sebeple yeni veriler ile model belirli periyotlarda test edilmeli gerekiyor ise yeniden düzenlenmelidir.

Veri madenciliği proje döngüsü şekil 3.2’de görsel olarak gösterilmiştir.

Şekil 3.2: Veri madenciliği proje döngüsü : CRISP-DM



Kaynak:Wikipedia

3.4 VERİ MADENCİLİĞİNİN KULLANIM ALANLARI

Veri madenciliği karar verme mekanizmasına ihtiyaç duyulan bir çok alanda kullanılmaktadır. Operasyonel kararlardan ziyade, stratejik ve politik karar verme süreçlerinde önemli bir yere sahiptir. Birçok kurum ve kuruluş CRM (Müşteri İlişklileri Yönetimi) ve ERP (Kurumsal Kaynak Planlaması) gibi birtakım uygulamalar ve çeşitli teknikler vasıtası ile veri madenciliği yapmaktadırlar. Özellikle pazarlama, bankacılık ve sigortacılık alanlarında sıkça kullanılan veri madenciliği istatistik ile iç içe olması sebebiyle tıp alanında da oldukça sık kullanılmaktadır. Veri madenciliği ile şirketler,

kurum ve kuruluşlar müşterilerinin özelliklerini ve davranışlarını irdeleyerek daha etkin fiyatlandırma politikaları uygulayarak karlılıklarını arttırabilirler, mevcut müşterilerin elde tutulması yeni müşterilerin kazanılması için etkin politikalar yaratabilirler, kredi riskini tahmin edebilirler, sahtekar müşteri tespiti yapabilirler, riskli müşteri profili belirlenerek ona göre bir fiyatlandırma ya da uzak durma politikası izleyebilirler, hastalık riskini ilk aşamada tespit ederek kontrolünü yapabilirler. Bunlar veri madenciliği sayesinde yapılabilcek sadece birkaç örnektir. Bilginin bu kadar değerli olduğu çağımızda bilgiye ulaşıp onu en verimli şekilde kullanabilmek veri madenciliği ile daha kolay bir hal almıştır.

3.5 VERİ MADENCİLİĞİNDE KULLANILAN MODELLER VE ÖĞRENME YÖNTEMLERİ

Veri madenciliği modelleme yöntemleri ;

i. Tahmin Edici (Predictive)
ii. Tanımlayıcı (Descriptive)

olmak üzere 2'ye ayrılırlar.

3.5.1 Tahmin Edici Modeller

Tahmin edici modellerde, sonuçları bilinen verilerden yola çıkılarak bir model geliştirilir ve geliştirilen bu modelden yararlanılarak sonuçları bilinmeyen veri kümesi için sonuç tahmin edilir. Regresyon analizi, karar ağaçları/yapay sinir ağları/zaman serileri analizi/K-en yakın komşu algoritması/karar destek makineleri/bayes algoritması gibi sınıflandırma yöntemleri tahmin edici modeller arasında yer almaktadır.

Örnek: Bir bankanın mevcut müşterilerinin özellikleri ve batma durumları göz önüne alınarak bir model kurulması ve bu modele göre potansiyel müşterilerinin özellikleri değerlendirilerek müşterilerin batma olasılıklarının tahmin edilmesi.

3.5.2 Tanımlayıcı Modeller

Tanımlayıcı modellerde ise tahmin edici modelin aksine karar vericilere yol göstermek için veri kümesinde bulunan gizli örüntüler tanımlanır. Birliktelik kuralları, kümeleme analizi, faktör analizi, istisna analizi, ardışık zamanlı örüntü analizi gibi yöntemler tanımlayıcı modeller arasında yer almaktadır.

Örnek: Bir firmanın kampanyasından sıkça yararlanan yaşlı müşterileri ile bu kampanyadan sıkça yararlanmayan genç müşterilerin satın alma örüntülerinin birbirine benzerlik gösterdiğinin belirlenmesi.

Öğrenme Yöntemleri ise ;

i. Denetimli
ii. Denetimsiz

olmak üzere 2'ye ayrılır.

3.5.3 Denetimli Öğrenme

"Örnekten öğrenme olarak da isimlendirilen denetimli öğrenmede, analizi yapan kişiler tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı, verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin belirli kural cümleleri ile ifade edilmesidir." Yani veri kümesi içerisinden bir eğitim kümesi seçilerek model bu eğitim kümesi üzerinde kurulur ve ayrılan test kümesi ile de modelin doğruluğu incelenir. Kurulan modelin doğruluğu yeterli görülür ise yeni gelen örnekler üzerinde model uygulanır ve bu örneklerin hangi sınıflara ait olduğu belirlenir. Regresyon analizi ve karar ağaçları yöntemleri bu gruba örnek verilebilir.

3.5.4 Denetimsiz Öğrenme

Denetimsiz öğrenmede ise sınıflar önceden belirlenmemiştir, verilerin özelliklerine göre sınıflar oluşturulmaktadır. Yani bağımlı değişken yoktur. Faktör ve kümeleme analizi yöntemleri bu gruba örnek verilebilir.

3.6 VERİ MADENCİLİĞİ YÖNTEM VE TEKNİKLERİ

Veri madenciliği modellerini işlevlerine göre aşağıdaki şekilde gruplamak mümkündür.

i. Sınıflama (Classification) ve Regresyon (Regression) Modelleri
ii. Kümeleme (Clustering) Modelleri
iii. Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns)

3.6.1 Sınıflama ve Regresyon Modelleri

En yaygın kullanılan veri madenciliği yöntemlerinden olan sınıflama tekniği ile sınıfı belirlenmiş olan mevcut verilerden kurulan model yardımı ile sınıf belli olmayan verilerin sınıfı tahmin edilir. En yaygın sınıflama teknikleri;

i. Yapay Sinir Ağları (Artificial Neural Networks),
ii. Genetik Algoritmalar (Genetic Algoriythms),
iii. K-En Yakın Komşu (K-Nearest Neighbor),
iv. Karar Ağaçları (Decision Trees),
v. Bellek Temelli Nedenleme (Memory Based Reasoning),
vi. Naive-Bayes,Lojistik Regresyon (Logistic Regression)'dur.

Sınıflama ve regresyon modellerini birbirinden ayıran en temel özellik sınıflama yönteminde kategorik değerler tahmin edilirken regresyon yönteminde genellikle sürekli değerlerin tahmin edilmesidir.

3.6.2 Kümeleme Modelleri

Denetimsiz öğrenme metodu olan kümeleme, heterojen olan veriyi homojen alt gruplara ayırarak özetleme işlemidir. Kümelemede amaç kendi içerisinde oldukça benzerlik gösteren ancak grupların birbirinden çok farklı özellikler sergilediği alt kümeler oluşturmaktır. Sınıflama ile kümelemeyi birbirinden ayıran en temel özellik kümlemede daha önceden belirlenmiş sınıfların olmamasıdır. Verilerin kümelenmesi işlemi birbirlerine olan benzerliklerine istinaden yapılmaktadır. Belirlenen sınıfların ne anlama geldiği analistin yorum gücüne bağlıdır. Kümeleme genellikle bir başka veri

madenciliği tekniğinin ilk basamağı olarak kullanılmaktadır. En yaygın kümeleme teknikleri;

i. Bölme Yöntemleri (Partitioning Methods),
ii. Hiyerarşik Yöntemler (Hierarchical Methods) BIRCH, ROCK vb. algoritmalar
iii. Izgara Tabanlı Yöntemler (Grid Based Methods) STING vb. algoritmalar
iv. Model Tabanlı Yöntemler (Model Based Methods) EM, COBWEB vb. algoritmalar
v. Yoğunluk Tabanlı Yöntemler (Density Based Methods)'dir.

3.6.3 Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Çok büyük veri setlerinde gözle görünen birliktelikler dışında kalan ve hemen farkedilmeyen bazı ilişkilerin ortaya konulması birliktelik, bir olayın ardından gerçekleşen bir diğer olayın tespit edilmesi ise ardışık zamanlı örüntüleri ifade etmektedir. Birliktelik kuralların en tipik kullanıldığı örnek, müşterilerin yaptıkları alışverişteki ürünler arasındaki birlikteliklerin bulunarak müşterilerin satın alma analizinin yapıldığı *market sepeti* uygulamasıdır. Veriler arasındaki ilişkiler aşağıdaki şekilde ifade edilmektedir:

Eğer <bazı şartlar sağlanırsa> sonra <bazı niteliklerin değerlerini tahmin et>

Örnek: Bir markette müşteri peynir satın alıyor ise ekmek satın alma olasılığı nedir?

En sık kullanılan birliktelik kuralları;

i. GRI (The Generalized Rule Induction),
ii. Apriori,
iii. CARMA (Continuous Association Rule Mining Algorithm) 'dir.

4. KULLANILACAK VERİYE GENEL BAKIŞ

Bu tez kapsamında kullanılan veriler yalnızca taşıt kredisi veren özel bir tüketici finansman şirketinin müşterilerine ait bilgilerdir. 01.01.2010 – 31.12.2013 tarihleri arasında bu şirkete yapılan 396.326 adet başvuru ve başvuruda bulunan her bir müşteriye ait demografik, finansal ve başvuru özelinde bilgileri içeren 66 adet değişken çalışma kapsamında değerlendirilmek üzere sözkonusu şirketin veri tabanından alınmıştır.

Tez konusu "Bireysel Araç Kredilerinin Yasal Takibe Girme Durumları Hakkında Tahmin Modellerinin Oluşturulması" olduğundan ve temel olarak yasal takibe girmiş müşteriler inceleneceğinden, başvuru datasından;

i. Tüzel kişilere ait olan 40.294 adet başvuru,
ii. Yalnızca başvuru aşamasında kalmış ve kredileşmemiş 207.327 adet bireysel başvuru,
iii. Kredileşmiş ancak finansal datası eksik olan 36.236 adet bireysel başvuru,
iv. Tüm datası var olan ancak datanın çekilme tarihi itibari ile riske[1] konu olmamış 19.393 adet bireysel başvuru çıkartılmıştır.[1]

Verinin alındığı şirkette daha önce yapılan çalışmalara, gözlemlere ve geçmiş verilere dayanılarak bir taşıt kredisinin yasal takibe girme süresinin ortalama olarak kredinin verildiği tarihten itibaren ilk 12 ay içerisinde gerçekleştiği ortaya konulmuştur. Bu da müşterinin 12 aylık bir taksit sürecinden geçmesi demektir. Resmi olarak yasal takip süreci ise 4 taksidin üst üste ödenmediği zaman başlamaktadır. Bu sebeple sistemde 12 aylık bir ödeme geçmişi olan müşterinin kesin olarak yasal takibe girip girmediği ilk 16 taksidinde belli olmaktadır. Çalışma yapılırken risk kriteri olarak 16 taksit seçilmesinin sebebi hem müşterilerin ilk 12 ayda yasal takibe girme oranları daha fazla olduğu için

[1] Risk değerlendirmesi yapabilmek için bir başvurunun datanın veri tabanından çekilme tarihi itibari ile en az 16 taksitlik bir ödeme performansı geçmişine sahip olması gerekmektedir.

çok sayıda gözlem yapabilmek hem de yasal takibe girme sürecinin resmi olarak başladığından emin olmak içindir.

Data temizlendikten sonra geriye kalan 93.076 adet veriden ilk 16 taksidinde yasal takibe girmiş olan 1416 verinin hepsi, bu 1416 veri alındıktan sonra da geriye kalan veriden 1416 adet ilk 16 taksidinde yasal takibe girmemiş olan müşteri rastgele seçilmiştir. Böylece 2832 adet müşteri ve bu müşterilere ait verilerden oluşan örneklem kümesi belirlenmiştir.

Çalışma kapsamında örneklem üzerinde öncelikle lojistik regresyon modeli uygulanmış ve 66 adet değişkenden anlamlı olmayanları çıkartılarak veri 26 değişkene kadar indirgenmiştir. Bu 66 adet değişken içerisinde bulunan ve başvuru sırasında bilgisi müşteriden alınan gelir, borç, ev sahibi olup olmama vb. müşterinin manipule edebileceği ve çalışmanın sonucunu yanıltabilecek bilgiler istatistiki olarakta anlamsız bulunarak çıkartılmıştır. Çoğunluğu müşterinin KKB[2] verilerine dayalı olmak üzere bir kısmı da başvuru bilgilerini içeren ve herhangi bir kullanıcı müdahelesine açık olmayan tamamı gerçek değerlerini yansıtan 26 adet değişken ile modellemeler yapılmıştır.

4.1 VERİYE AİT BETİMLEYİCİ İSTATİSTİKLER

Tablo 4.1 ve 4.2'de gösterildiği gibi verinin kalitesine bakıldığında bütün değişkenlere ait 2832 adet verinin de geçerli olduğu görülmektedir. Veride herhangi bir kayıp ya da boş gözlem bulunmamaktadır.

[2] Kredi Kayıt Bürosu (KKB), Türkiye'de bir kişinin bankalar ile olan hertürlü ilişki verilerinin tutulduğu ve bu verilere istinaden kişinin bireysel kredi notunun hesaplanarak riskinin ortaya konulduğu sistemdir. Bireysel kredi notu bir kişinin kaç bankaya borcunun olduğu, borçlarının zamanında ödenip ödenmediği, kalan borç miktarı, kredi kartı ya da kredi limitleri, hakkında yasal ya da idari takip kararının bulunup bulunmaması vs. gibi her türlü bankalar ile olan ilişki göz önünde bulundurularak hesaplanmaktadır. Ve bu sisteme üye olan finansal kuruluşlar belirli bir ücret karşılığında müşterilerinin bu bilgilerine ulaşarak onları sorgulayabilmektedirler ve böylece bir müşterinin kredibilitesi ortaya çıkmaktadır.

Tablo 4.1: Veri kalitesi I

Alan Adı	Tip	Aykırı Değerler	Uç Değerler	Aksiyon	Kayıp Gözlem	Metod
CalismaSekli	Set	-	-	-	Never	Fixed
BasvuruSayisi	Range	41	14	None	Never	Fixed
KrediTur	Set	-	-	-	Never	Fixed
Marka	Set	-	-	-	Never	Fixed
KararKategori	Set	-	-	-	Never	Fixed
Vade	Range	0	0	None	Never	Fixed
AracYasi2	Range	70	26	None	Never	Fixed
LTV_FinansmanGrup	Set	-	-	-	Never	Fixed
Bolge	Set	-	-	-	Never	Fixed
EverAccDpd30Plus	Set	-	-	-	Never	Fixed
EverYasalAccount16M	Flag	-	-	-	Never	Fixed
Borcluluk	Range	198	0	None	Never	Fixed
CardOrODmaxOpenLimit	Range	52	6	None	Never	Fixed
Kural3_nbOfRejAppL3M	Range	53	9	None	Never	Fixed
Kural4_nbOf2UnpaideDel24M	Range	45	23	None	Never	Fixed
Kural5_nbOf3UnpaidDel24M	Range	0	31	None	Never	Fixed
Kural6_Loan	Range	8	29	None	Never	Fixed
Kural7_Loan	Range	7	2	None	Never	Fixed
Kural11_CardTotalOpenLimit	Range	44	21	None	Never	Fixed
Kural11_CardTotalOpenBalance	Range	56	24	None	Never	Fixed
Kural12_LoanTotalOpenLimit	Range	35	21	None	Never	Fixed
Kural12_LoanTotalOpenBalance	Range	44	18	None	Never	Fixed
Kural12_LoanMaxClosedLimit	Range	18	13	None	Never	Fixed
Kural13_CardTSFirstOpened	Range	18	0	None	Never	Fixed
Kural14_CardTSLastOpened	Range	47	10	None	Never	Fixed
Kural14_LoanTSLastOpened	Range	50	13	None	Never	Fixed

*Kaynak:*Clementine 12.0

Tablo 4.2: Veri kalitesi II

Alan Adı	Tamamlanma Oranı	Geçerli Gözlem Sayısı	Boş Değer	Boş Dizi	Alfabe Dışı karakter	Boş Gözlem
CalismaSekli	100	2832	0	0	0	0
BasvuruSayisi	100	2832	0	0	0	0
KrediTur	100	2832	0	0	0	0
Marka	100	2832	0	0	0	0
KararKategori	100	2832	0	0	0	0
Vade	100	2832	0	0	0	0
AracYasi2	100	2832	0	0	0	0
LTV_FinansmanGrup	100	2832	0	0	0	0
Bolge	100	2832	0	0	0	0
EverAccDpd30Plus	100	2832	0	0	0	0
EverYasalAccount16M	100	2832	0	0	0	0
Borcluluk	100	2832	0	0	0	0
CardOrODmaxOpenLimit	100	2832	0	0	0	0
Kural3_nbOfRejAppL3M	100	2832	0	0	0	0
Kural4_nbOf2UnpaideDel24M	100	2832	0	0	0	0
Kural5_nbOf3UnpaidDel24M	100	2832	0	0	0	0
Kural6_Loan	100	2832	0	0	0	0
Kural7_Loan	100	2832	0	0	0	0
Kural11_CardTotalOpenLimit	100	2832	0	0	0	0
Kural11_CardTotalOpenBalance	100	2832	0	0	0	0
Kural12_LoanTotalOpenLimit	100	2832	0	0	0	0
Kural12_LoanTotalOpenBalance	100	2832	0	0	0	0
Kural12_LoanMaxClosedLimit	100	2832	0	0	0	0
Kural13_CardTSFirstOpened	100	2832	0	0	0	0
Kural14_CardTSLastOpened	100	2832	0	0	0	0
Kural14_LoanTSLastOpened	100	2832	0	0	0	0

*Kaynak:*Clementine 12.0

Nicel değişkenlere ait minimum, maksimum, toplam, değer aralığı, ortalama, standart hata, standart sapma, varyans, çarpıklık, basıklık, medyan ve mod değerlerini içeren betimleyici istatistikler ise tablo 4.3,4.4 ve 4.5'te ayrıntılı bir biçimde gösterilmiştir.

Tablo 4.3: Betimleyici istatistikler I

Alan Adı	Tip	Min	Max	Toplam
CalismaSekli	Set	0	7	-
BasvuruSayisi	Range	1	6	3051
KrediTur	Set	0	8	-
Marka	Set	0	34	-
KararKategori	Set	0	6	-
Vade	Range	6	60	115062
AracYasi2	Range	0	7	1111
LTV_FinansmanGrup	Set	0	10	-
Bolge	Set	0	6	-
EverAccDpd30Plus	Set	0	1	-
EverYasalAccount16M	Flag	0	1	-
Borcluluk	Range	0	101000	2E+07
CardOrODmaxOpenLimit	Range	0	100000	2,2E+07
Kural3_nbOfRejAppL3M	Range	0	4	75
Kural4_nbOf2UnpaideDel24M	Range	0	20	1336
Kural5_nbOf3UnpaidDel24M	Range	0	6	180
Kural6_Loan	Range	0	9	1876
Kural7_Loan	Range	0	9	925
Kural11_CardTotalOpenLimit	Range	0	399000	5,8E+07
Kural11_CardTotalOpenBalance	Range	0	215185	3,3E+07
Kural12_LoanTotalOpenLimit	Range	0	636981	6,5E+07
Kural12_LoanTotalOpenBalance	Range	0	415416	4,9E+07
Kural12_LoanMaxClosedLimit	Range	0	300000	2,7E+07
Kural13_CardTSFirstOpened	Range	0	8728	6801974
Kural14_CardTSLastOpened	Range	0	5615	1200228
Kural14_LoanTSLastOpened	Range	-8	2370	536228

*Kaynak:*Clementine 12.0

Tablo 4.4: Betimleyici istatistikler II

Alan Adı	Değer Aralığı	Ortalama	Standart Hata	Tip	Standart Sapma
CalismaSekli	-	-	-	Set	-
BasvuruSayisi	5	1,077	0,007	Range	0,389
KrediTur	-	-	-	Set	-
Marka	-	-	-	Set	-
KararKategori	-	-	-	Set	-
Vade	54	40,629	0,217	Range	11,545
AracYasi2	7	0,392	0,02	Range	1,081
LTV_FinansmanGrup	-	-	-	Set	-
Bolge	-	-	-	Set	-
EverAccDpd30Plus	-	-	-	Set	-
EverYasalAccount16M	-	-	-	Flag	-
Borcluluk	101000	7112,47	483,8	Range	25746
CardOrODmaxOpenLimit	100000	7903	176,3	Range	9383,24
Kural3_nbOfRejAppL3M	4	0,026	0,004	Range	0,196
Kural4_nbOf2UnpaideDel24M	20	0,472	0,026	Range	1,406
Kural5_nbOf3UnpaidDel24M	6	0,064	0,007	Range	0,378
Kural6_Loan	9	0,662	0,021	Range	1,107
Kural7_Loan	9	0,327	0,011	Range	0,584
Kural11_CardTotalOpenLimit	399000	20309	624,1	Range	33210,3
Kural11_CardTotalOpenBalance	215185	11744,7	408	Range	21714
Kural12_LoanTotalOpenLimit	636981	22952	702,5	Range	37386,8
Kural12_LoanTotalOpenBalance	415416	17326,6	547	Range	29107
Kural12_LoanMaxClosedLimit	300000	9590,15	270,8	Range	14413,3
Kural13_CardTSFirstOpened	8728	2401,83	28,89	Range	1537,31
Kural14_CardTSLastOpened	5615	423,809	10,3	Range	547,98
Kural14_LoanTSLastOpened	2378	189,346	5,485	Range	291,892

*Kaynak:*Clementine 12.0

Tablo 4.5: Betimleyici istatistikler III

Alan Adı	Varyans	Çarpıklık	Basıklık	Medyan	Mod
CalismaSekli	-	-	-	-	6
BasvuruSayisi	0,151	6,591	53,948	1	1
KrediTur	-	-	-	-	8
Marka	-	-	-	-	11
KararKategori	-	-	-	-	5
Vade	133,293	-0,82	-0,075	48	48
AracYasi2	1,169	3,255	10,925	0	0
LTV_FinansmanGrup	-	-	-	-	6
Bolge	-	-	-	-	0
EverAccDpd30Plus	-	-	-	-	0
EverYasalAccount16M	-	-	-	-	1
Borcluluk	6,6E+08	3,375	9,397	62	101000
CardOrODmaxOpenLimit	8,8E+07	2,792	14,059	4569	1000
Kural3_nbOfRejAppL3M	0,039	9,837	126,93	0	0
Kural4_nbOf2UnpaideDel24M	1,977	5,741	48,853	0	0
Kural5_nbOf3UnpaidDel24M	0,143	9,149	107,7	0	0
Kural6_Loan	1,226	4,359	27,842	0	0
Kural7_Loan	0,341	3,476	34,968	0	0
Kural11_CardTotalOpenLimit	1,1E+09	3,984	23,747	8100	0
Kural11_CardTotalOpenBalance	4,7E+08	3,681	17,034	3643	0
Kural12_LoanTotalOpenLimit	1,4E+09	4,692	41,682	11000	0
Kural12_LoanTotalOpenBalance	8,5E+08	4,104	27,504	7809,5	0
Kural12_LoanMaxClosedLimit	2,1E+08	5,813	77,932	5000	0
Kural13_CardTSFirstOpened	2363317	0,65	0,119	2160	0
Kural14_CardTSLastOpened	300282	2,537	10,118	226	0
Kural14_LoanTSLastOpened	85201	2,562	8,344	73	0

*Kaynak:*Clementine 12.0

4.2 KULLANILAN DEĞİŞKENLER VE AÇIKLAMALARI

Çalışmada kullanılan değişkenler ve açıklamaları ayrıntılı bir biçimde tablo 4.6'da gösterilmiştir.

Tablo 4.6: Değişkenler ve açıklamaları

CalismaSekli	Basvuru Sahibinin Çalışma Şekli Bilgisi
BasvuruSayisi	Başvuru Sahibinin Kaç Adet Taşıt İçin Başvuruda Bulunduğunun Bilgisi
KrediTur	Başvuru Sahibinin Ne Tür Bir Krediye Başvurduğunun Bilgisi
Marka	Başvuru Sahibinin Hangi Marka Taşıt İçin Krediye Başvurduğunun Bilgisi
KararKategori	Başvurunun Sistemde Aldığı Karar Bilgisi
Vade	Başvuru Sahibinin Krediyi Geri Ödemek İçin Talep Ettiği Ay Sayısı
AracYasi2	Başvuru Sahibinin Almak İstediği Taşıtın Yaş Bilgisi
LTV_FinansmanGrup	Başvuru Sahibinin Alacağı Taşıt Fiyatının Yüzde Kaçı İçin Kredi Talep Ettiğinin Bilgisi
Bolge	Başvuru Sahibinin Yaşadığı Bölge Bilgisi
EverAccDpd30Plus	Kredinin Geri Ödenmeye Başlandığı İlk 3 Taksidinde Hiç 30 Gün Ve Üstü Gecikmeye Düşüp Düşmediğinin Bilgisi
EverYasalAccount16M	Kredinin Geri Ödenmeye Başlandığı İlk 16 Taksidinde 120 Gün Ve Üstü Gecikmeye (Yasal Takip) Düşüp Düşmediğinin Bilgisi
Borcluluk	Başvuru Sahibin KKB Verilerine Göre Güncel Toplam Kredi Kartı Borcunun Toplam Kredi Kartı Limitine Oranının Bilgisi
CardOrODmaxOpenLimit	Başvuru Sahibinin KKB Verilerine Göre Sahip Olduğu Açık Maksimum Kredi Kartı Ya Da Overdraft Limiti Bilgisi
Kural3_NbOfRejAppL3M	Başvuru Sahibinin KKB Verilerine Göre Son 3 Ayda Red Edilen Kredi Başvurusunun Olup Olmadığının Bilgisi
Kural4_nbOf2UnpaidDel24M	Başvuru Sahibinin KKB Verilerine Göre Son 2 Yılda Herhangi Bir Ödemesinde Hiç 2 Taksit Gecikmeye Düşüp Düşmediğinin Bilgisi
Kural5_nbOf3UnpaidDel24M	Başvuru Sahibinin KKB Verilerine Göre Son 2 Yılda Herhangi Bir Ödemesinde Hiç 3 Taksit Gecikmeye Düşüp Düşmediğinin Bilgisi
Kural6_Loan	Başvuru Sahibinin Tüm KKB Verilerine Göre Herhangi Bir Ödemesindeki En Kötü Performansının Bilgisi
Kural7_Loan	Başvuru Sahibinin Tüm KKB Verilerine Göre Son 1 Yılda Herhangi Bir Ödemesindeki En Kötü Performansının Bilgisi
Kural11_CardTotalOpenLimit	Başvuru Sahibinin KKB Verilerine Göre Sahip Olduğu Tüm Kredi Kartlarının Toplam Güncel Açık Limit Bilgisi
Kural11_CardTotalOpenBalance	Başvuru Sahibinin KKB Verilerine Göre Sahip Olduğu Tüm Kredi Kartlarının Toplam Güncel Açık Borç Bilgisi
Kural12_LoanTotalOpenLimit	Başvuru Sahibinin KKB Verilerine Göre Sahip Olduğu Tüm Kredilerin Toplam Güncel Açık Limit Bilgisi
Kural12_LoanTotalOpenBalance	Başvuru Sahibinin KKB Verilerine Göre Sahip Olduğu Tüm Kredilerin Toplam Güncel Açık Borç Bilgisi

Kural12_LoanMaxClosedLimit	Başvuru Sahibinin KKB Verilerine Göre Kapatmış Olduğu En Yüksek Kredi Limiti Bilgisi
Kural13_CardTSFirstOpened	Başvuru Sahibinin KKB Verilerine Göre İlk Kredi Kartı Açılışından İtibaren Geçen Süre Bilgisi
Kural14_CardTSLastOpened	Başvuru Sahibinin KKB Verilerine Göre Son Kredi Kartı Açılışından İtibaren Geçen Süre Bilgisi
Kural14_LoanTSLastOpened	Başvuru Sahibinin KKB Verilerine Göre İlk Kredi Açılışından İtibaren Geçen Süre Bilgisi

4.3 KULLANILAN DEĞİŞKENLERİN TÜRLERİ

Kullanılan değişkenlerden,

i. CalismaSekli,
ii. KrediTur,
iii. Marka,
iv. KararKategori,
v. LTV_FinansmanGrup,
vi. Bolge,
vii. EverAccDpd30Plus değişkenleri kategorik geri kalan diğer değişkenler ise niceldir.

Nicel değişkenlerden aldığı değerler çok geniş aralıklarda değişen ve zaman zaman çok büyük değerler alan,

i. Borçluluk,
ii. CardOrODmaxOpenLimit,
iii. Kural11_CardTotalOpenLimit,
iv. Kural11_CardTotalOpenBalance,
v. Kural12_LoanTotalOpenLimit,
vi. Kural12_LoanTotalOpenBalance,
vii. Kural12_LoanMaxClosedLimit,
viii. Kural13_CardTSFirsOpened,
ix. Kural14_CardTSLastOpened ve
x. Kural14_LoanTSLastOpened

değişkenlerine lojistik regresyon modelinde ln (doğal logaritmik) dönüşümü uygulanmıştır. Bunun sebebi değişkenlerin orjinal değerlerinde kaldıkları zaman modelin bunu algılayamaması ve bu değişkenler anlamlı çıktığı halde katsayılarının 0

olup sabit terime gereğinden fazla yüklenilmesine sebep olmalarıdır. Diğer modellemelerde bu değişkenler orjinal değerleri ile kullanılmıştır, herhangi bir dönüşüm uygulanmamıştır.

Lojistik regresyon analizinde anlamlı bulunan ve veri setinde kullanılan 26 değişken ve türleri tablo 4.7'de gösterilmiştir.

Tablo 4.7: Değişken türleri

Alan Adı	**Tip**
CalismaSekli	Nitel
BasvuruSayisi	Nicel
KrediTur	Nitel
Marka	Nitel
KararKategori	Nitel
Vade	Nicel
AracYasi2	Nicel
LTV_FinansmanGrup	Nitel
Bolge	Nitel
EverAccDpd30Plus	Nitel
EverYasalAccount16M	İki Kategorili
Borcluluk	Nicel
CardOrODmaxOpenLimit	Nicel
Kural3_nbOfRejAppL3M	Nicel
Kural4_nbOf2UnpaideDel24M	Nicel
Kural5_nbOf3UnpaidDel24M	Nicel
Kural6_Loan	Nicel
Kural7_Loan	Nicel
Kural11_CardTotalOpenLimit	Nicel
Kural11_CardTotalOpenBalance	Nicel
Kural12_LoanTotalOpenLimit	Nicel
Kural12_LoanTotalOpenBalance	Nicel
Kural12_LoanMaxClosedLimit	Nicel
Kural13_CardTSFirstOpened	Nicel
Kural14_CardTSLastOpened	Nicel
Kural14_LoanTSLastOpened	Nicel

*Kaynak:*Clementine 12.0

4.4 KULLANILAN NİTEL DEĞİŞKENLERİN KATEGORİK KARŞILIKLARI

Çalışmada kullanılan kategorik değişkenler, açıklamaları ve maskelendikleri değerler tablo 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 ve 4.15'te gösterilmiştir.

Tablo 4.8: "Çalışma Şekli" değişkeni kategorileri ve karşılıkları

Çalışma Şekli	Açıklama	Kategorik Değeri
Çalışmıyor	Çalışmıyor	0
Ek Gelir Sahibi	Ek Gelir Sahibi	1
Emekli	Emekli	2
Ev Hanımı	Ev Hanımı	3
Kamu	Kamu Sektörü Çalışanı	4
Öğrenci	Öğrenci	5
Özel Sektör	Özel Sektör Çalışanı	6
Serbest Meslek	Serbest Meslek Sahibi	7

*Kaynak:*Clementine 12.0

Tablo 4.9: "Kredi Tür" değişkeni kategorileri ve karşılıkları

Kredi Tür	Açıklama	Kategorik Değeri
ARAÇ+5 YILDIZ	Araç & 1 Yıllık İhtiyaç Paketi	0
ARAÇ+ADKS	Araç & 1 Yıllık İhtiyaç Paketi &Uzun Dönem Bakım Kredisi	1
ARAÇ+CPI+5 YILDIZ	Araç & 1 Yıllık İhtiyaç Paketi & Kredi Koruma Paketi	2
ARAÇ+CPI+5 YILDIZ+ADKS	Araç & Kredi Bakım Paketi & 1 Yıllık İhtiyaç Paketi &Uzun Dönem Bakım Kredisi	3
ARAÇ+CPI+ADKS	Araç & Kredi Bakım Paketi &Uzun Dönem Bakım Kredisi	4
ARAÇ+FKS	Araç & Ferdi Kaza Sigortası Paketi	5
ARAÇ+FKS+KREDİ KORUMA	Araç & Ferdi Kaza Sigortası Paketi& Kredi Koruma Paketi	6
ARAÇ+KREDİ KORUMA	Araç & Kredi Koruma Paketi	7
KLASİK KREDİ	Yalnızca Taşıt Kredisi	8

*Kaynak:*Clementine 12.0

Tablo 4.10: "Karar Kategori" değişkeni kategorileri ve karşılıkları

Karar Kategori	Açıklama	Kategorik Değeri
Accept	Oto Onay	4
Accept Refer	Ono Onay İstisnası	5
Decline	Oto Red	2
Decline Refer	Oto Red İstisnası	3
No Policy	Log Kayıtları-Etkisiz Kural	0
Policy Decline	Politika Kuralından Red	1
Refer	Gri Alan (Otomatik Olarak Red Ya Da Kabul Kararı Verilemiyor ,Çekimser)	6

*Kaynak:*Clementine 12.0

Tablo 4.11: "Bölge" değişkeni kategorileri ve karşılıkları

Bölge	Açıklama	Kategorik Değeri
Marmara	Marmara Bölgesi	0
Akdeniz	Akdeniz Bölgesi	1
Anadolu	İç Anadolu Bölgesi	2
Ege	Ege Bölgesi	3
Karadeniz	Karadeniz Bölgesi	4
Güneydoğu	Güneydoğu Anadolu Bölgesi	5
Doğu	Doğu Anadolu Bölgesi	6

*Kaynak:*Clementine 12.0

Tablo 4.12: "LTV-Finansman Grup" değişkeni kategorileri ve karşılıkları

LTV-Finansman Grup	Açıklama	Kategorik Değeri
1-LTV<=30	Kredi Oranı Taşıt Fiyatının %30'u Ya Da Daha Azı	0
2-31<=LTV<=40	Kredi Oranı Taşıt Fiyatının %31'i İle %40'ı Arasında	1
3-41<=LTV<=50	Kredi Oranı Taşıt Fiyatının %41'i İle %50'si Arasında	2
4-51<=LTV<=60	Kredi Oranı Taşıt Fiyatının %51'i İle %60'ı Arasında	3
5-61<=LTV<=70	Kredi Oranı Taşıt Fiyatının %61'i İle %70'i Arasında	4

6-71<=LTV<=75	Kredi Oranı Taşıt Fiyatının %71 İle %75'i Arasında	5
7-76<=LTV<=80	Kredi Oranı Taşıt Fiyatının %76'sı İle %80'i Arasında	6
8-81<=LTV<=85	Kredi Oranı Taşıt Fiyatının %81'i İle %85'i Arasında	7
9-86<=LTV<=90	Kredi Oranı Taşıt Fiyatının %86'sı İle %90'ı Arasında	8
99-91<=LTV<=95	Kredi Oranı Taşıt Fiyatının %91'i İle %95'i Arasında	9
999-LTV=>96	Kredi Oranı Taşıt Fiyatının %96'sı Ya Da Daha Fazlası	10

Tablo 4.13: "Ever Acc Dpd 30 Plus" değişkeni kategorileri ve karşılıkları

Ever Acc Dpd 30 Plus	**Açıklama**
0	İlk 3 taksidinde 30 gün üstü gecikmeye düşmemiş kredi
1	İlk 3 taksidinde 30 gün üstü gecikmeye düşmüş kredi

*Kaynak:*Clementine 12.0

Tablo 4.14: "Ever Yasal Acc 16 M" değişkeni kategorileri ve karşılıkları

Ever Yasal Acc 16 M	**Açıklama**
0	İlk 16 taksidinde yasal takibe girmemiş kredi
1	İlk 16 taksidinde yasal takibe girmiş kredi

*Kaynak:*Clementine 12.0

Tablo 4.15: "Marka" değişkeni kategorileri ve karşılıkları

Marka	**Kategorik Değeri**
ALFA ROMEO	0
AUDI	1
BMW	2
CHEVROLET	3
CHRYSLER	4
CITROEN	5
DACIA	6
DAIHATSU	7
FIAT	8

FORD	9
HONDA	10
HYUNDAI	11
INFINITY	12
ISUZU	13
JAGUAR	14
JEEP	15
KIA	16
LAND ROVER	17
MAZDA	18
MERCEDES	19
MINI	20
MITSUBISHI	21
NISSAN	22
OPEL	23
PEUGEOT	24
RENAULT	25
SAAB	26
SEAT	27
SKODA	28
SMART	29
SUBARU	30
SUZUKI	31
TOYOTA	32
VOLKSWAGEN	33
VOLVO	34

*Kaynak:*Clementine 12.0

5. KULLANILAN YÖNTEMLER, UYGULAMALARI VE SONUÇLARI

5.1 REGRESYON TANIMI, LOJİSTİK REGRESYON UYGULAMASI VE SONUÇLARI

5.1.1 Doğrusal Regresyon

İki veya daha çok değişken arasındaki ilişkiyi ölçmek varsa bu ilişkinin yönünü ve gücünü ortaya koymak için kullanılan analiz yöntemine regresyon çözümlemesi denilmektedir. Regresyon çözümlemesinde açıklanan değişkene bağımlı değişken, açıklayan değişken/lere ise bağımsız değişken/ler adı verilmektedir.

Bağımlı ve bağımsız değişken arasındaki ilişki doğrusal ise bu ilişki doğrusal regresyon yöntemi kullanılarak analiz edilir.

Bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkinin analiz edilmesine *basit doğrusal regresyon* adı verilmektedir.

$$Y_i = \beta_0 + \beta_i X_i + \varepsilon_i \qquad i = (1 \dots \dots \dots . . n) \qquad (5.1)$$

Bir bağımlı değişken ve 1'den fazla bağımsız değişken arasındaki ilişkinin analiz edilmesine ise *çoklu doğrusal regresyon* adı verilmektedir.

$$Y_i = \beta_0 + \beta_{1i} X_{1i} + \cdots + \beta_{ki} X_{ki} + \varepsilon_i \qquad i = (1 \dots \dots \dots . . n) \qquad (5.2)$$

$$Y_i = \sum \beta_{ki} X_{ki} + \varepsilon_i \qquad i = (1 \dots \dots \dots . . n) \qquad (5.3)$$

Burada Y_i bağımlı değişkeni,

β_0 sabit değeri,

β_{ki} ise X_i değişkeninin Y_i değişkeninin açıklama derecesini,

ε_i ise hata terimini göstermektedir.

İki değişken arasında doğrusal bir bağıntı var ise regresyon denklemi bağımlı değişkenin gerçek değeri ile tahmini değeri arasındaki farkı en küçük yapabilen kestiricilerin elde edilmesi ile bulunur. Varyans kavramı göz önünde bulundurulduğunda bunun ölçüsü *artık kareler toplamı'*dır. Bu toplamı en küçükleyen β_0 ve β_1 kestiricilerinin elde edilmesine ilişkin kestirim yöntemi *en küçük kareler yöntemi* adını alır. Ve modelin aşağıdaki varsayımları sağlaması gerekir.

ε_i hata terimi;

i. Normal dağılıma sahiptir.
ii. Ortalaması 0 (sıfır)'a eşittir.
iii. Hata terimleri arasında otokorelasyon(içsel bağıntı) yoktur.
iv. Hata terimlerinin varyansı her X bağımsız değişkeni için sabittir.

X bağımsız değişkeni;

i. Hata terimi ile arasında kovaryans 0(sıfır)'a eşittir, yani ilişkili değildir.
ii. Stokastik (değişken) değildir, tekrar eden örneklerde sabittir.
iii. Varyansı sonlu bir sayıdır.
iv. Birden fazla bağımsız değişken olması durumunda aralarında çoklu bağlantı (ilişki) yoktur.

5.1.2 Lojistik Regresyon

Bağımlı değişkenin nicel olduğu durumlarda doğrusal regresyon analizi kullanılır. Ancak bağımlı değişgenin kategorik olduğu durumlarda ise bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi ortaya koymak için lojistik regresyon kullanılır. Lojistik regresyonda bağımsız değişkenler nicel ya da kategorik olabilir. Lojistik regresyon normallik varsayımı gerektirmez. Lojistik regresyonda bağımlı ve bağımsız değişkenler arasında doğrusal bir ilişki olması gerekmez, üstel ya da polinom ilişkisi de olabilir. Lojistik regresyon doğrusal olmayan ilişkiyi koruyarak ilişkinin formunu koruyan logaritmik dönüşümler yapar.

Lojistik regresyonun bağımlı ve bağımsız değişkenlerinin ve bunların kategorilerinin sayısına göre uygulanacak yöntem tablo 5.1'de gösterildiği şekilde farklılık gösterir.

Tablo 5.1: Değişken sayısı ve türlerine göre lojistik regresyon yöntemleri

Bağımlı Değişken Kategori Sayısı	**Bağımsız Değişken Sayısı**	**Bağımsız Değişkenin Kategori Sayısı**	**Uygulanacak Yöntem**
2	1	2	Binominal Lojistik Regresyon
2	1	2+	Binominal Lojistik Regresyon
2	2+	Çeşitli	Çok Değişkenli Lojistik Regresyon
2+ (Nominal)	Tek/Çok	Çeşitli	Multinominal Lojistik Regresyon
2+ (Ordinal)	Tek/Çok	Çeşitli	Ordinal Lojistik Regresyon

İkili lojistik regresyon matematiksel olarak aşağıdaki şekilde ifade edilmektedir.

$$P(Y_i = 1) \tag{5.4}$$

$$E(Y_i) = 1 \times P(Y_i = 1) + 0 \times P(Y_i = 0) = P(Y_i = 1) \tag{5.5}$$

i. Gözlemin 1 değerini alma olasılığının beklenen değeri denklem 5.6'da gösterildiği gibi ifade edilir.

$$E(Y_i) = P(Y_i = 1) = \sum_{k=0}^{p} \beta_k X_{ik} \tag{5.6}$$

Regresyon denklemi denklem 5.7'de gösterildiği gibi ifade edilir.

$$E(Y_i) = P(Y_i = 1) = \sum_{k=0}^{p} \beta_k X_{ik} \tag{5.7}$$

Bağımlı değişkeni [0,1] aralığında, bağımsız değişkenleri ise sınırsız değer alan bu fonksiyonda eşitlik sağlanamaz ve olasılık değeri ±∞ arasında olacak şekilde logaritmik dönüşüme uğratılır.

$$E(Y_i) = \log\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{p} \beta_k X_{ik} \tag{5.8}$$

$$E(Y_i) = exp\,\frac{\sum_{k=0}^{p} \beta_k X_{ik}}{(1+\exp(\sum_{k=0}^{p} \beta_k X_{ik}))} \tag{4.9}$$

$$Log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_{\text{i}} \tag{5.10}$$

Burada;

P : İstenilen durumun gerçekleşme olasılığını

β_0: Sabit değeri

$\beta_i \; i = (1 \dots \dots \dots n)$: Her bir bağımsız değişkenin katsayısını

$X_i \quad i = (1 \dots \dots \dots n)$: Her bir bağımsız değişkeni ifade etmektedir.

İkili lojistik regresyonun varsayımları aşağıdaki gibidir:

i. $Y_i \in (0,1)$
ii. $P(Y_i = 1 | \, X_i) = \; P_i$
iii. $Y_1 \dots \dots \dots \dots Y_n$ değerleri bağımsızdır.
iv. Bağımsız değişkenler arasında çoklu bağlantı yoktur.

Logaritmik dönüşümün özellikleri ise aşağıdaki gibidir:

i. P arttıkça $\log(P)$ de artar.
ii. $0 \; \leq P \; \leq 1$ iken $-\infty \leq \log(P) \leq +\infty$ aralığında değer alabilir.
iii. $P > 0{,}5$ iken $\log(P) < 0 \;\; , \;\; P < 0{,}5$ iken $\log(P) > 0$ olur.

Çoklu lojistik regresyon ise aşağıdaki şekilde ifade edilmektedir.

Bağımlı değişken 0,1,2 gibi 3 kategoriye sahip olsun. Bu durumda 2 farklı grup lojistik regresyon söz konusu olur. 0 kategorisi baz alındığında 2 no'lu kategoriyi 1 no'lu kategori ile karşılaştıran fonksiyonlar aşağıdaki gibidir.

$$g_1(X) = \log\left(\frac{P\left(Y=\frac{1}{X}\right)}{P\left(Y=\frac{0}{X}\right)}\right) = \beta_{10} + \beta_{11}X_1 + \cdots + \beta_{1p}X_p \tag{5.11}$$

$$g_2(X) = \log\left(\frac{P\left(Y=\frac{2}{X}\right)}{P\left(Y=\frac{0}{X}\right)}\right) = \beta_{20} + \beta_{21}X_1 + \cdots + \beta_{2p}X_p \tag{5.12}$$

$$P_k(X) = \frac{\exp(g_k(X))}{\sum_{i=0}^{2} \exp(g_i(X))} \tag{5.13}$$

Lojistik regresyon yönteminin amacı bağımlı değişkenin sonucu tahmin edebilecek en sade modeli bulmaktır. Bulunan modelin uygun olup olmadığı model χ^2 testi ile, her bir bağımsız değişkenin anlamlı olup olmadığı ise Wald istatistiği ile yorumlanır.

Hem ikili lojistik regresyon için hem de çoklu lojistik regresyon için bağımsız değişkenin katsayı değişimi bağımlı değişkenin gerçekleşme olasılığı üzerindeki etkisidir.

Lojistik regresyonda sık kullanılan bazı terimleri tanımlamak gerekirse:

5.1.2.1 ODDS (bahis)

Bir olayın gerçekleşme olasılığının, gerçekleşmeme olasılığına oranıdır.

$$Odds\ \frac{p_i(x)}{1-p_i(x)} \tag{5.14}$$

5.1.2.2 ODDS ratio (bahis oranı)

İki odds'un birbirine oranıdır. İki değişken arasındaki ilişkinin özetidir.

5.1.2.3 Lojit

Odds ratio'nun doğal logaritmasıdır. Odds ratio asimetriktir ve logaritması alınarak simetrik hale dönüştürülür.

Lojit katsayıları doğrusal regresyon analizindeki "β" katsayısına karşılık gelmektedir.

$$logit(p_i) = \log(\frac{p_i}{1-p_i}) \tag{5.15}$$

Bu denklem bir bağıntı fonksiyonu olarak ele alındığında ve x 'ler bağımsız değişkenleri göstermek üzere lojit model denklem 5.16'da gösterildiği gibidir:

$$\log\left(\frac{p_i}{1-p_i}\right) = Z_i = \sum_{k=0}^{p} \beta_k X_{ik} \tag{5.16}$$

p_i 0' a yaklaştığında logit(p_i) -∞'a yakınsar.

p_i 1' e yaklaştığında ise logit(p_i) +∞'a yakınsar. Bu durumda;

$$\frac{p_i}{1-p_i} = \frac{1+e^{z_i}}{1+e^{-z_i}} \tag{5.17}$$

p_i bağımlı değişkenin 1 değerini alma olasılığını $1-p_i$ ise bağımlı değişkenin 0 değerini alma olasılığını göstermek üzere bahis oranı yukarıdaki eşitlik ile elde edilir. Bahis oranı 1' e yakın çıkan değişkenler bağımlı değişkenin(Y_i) değişimine önemli bir katkısı olmayan değişkenlerdir.Bahis oranı 1'e yakın olan değişkenlerin katsayıları da anlamlı bulunmaz ise , söz konusu değişkenlerin modelde önemli bir etken olmadığı söylenebilir. Ancak katsayılar anlamlı bulunur ise bahis oranının 1'den büyük olması ilgili değişkenlerin modelde önemli bir etken olduğunu gösterir.

0' a yakın çıkan bahis oranı değerleri ise, katsayının anlamlı olması şartıyla, değişkeninin önemli bir etmen olduğunu ama bağımlı değişkenin (Y_i) düşük değerler almasına sebep olduğu yani negatif bir etki sağladığı söylenebilir. (Özdamar, 1999:487) Bağımlı değişkeninin lojit değişkenine dönüşümünün ardından lojistik regresyon en çok olabilirlik tahminini (maximum likelihood estimation) kullanır. Lojistik regresyon analizinde katsayıların tahmin edilmesinin ardından, önemli bulunan değişkenlerin anlamlılığı değerlendirilmektedir. Bu anlamlılık olabilirlik oran testi (likelihood ratio test), Wald testi (Wald test) ve skor testi (score test) ile değerlendirilir.

5.1.2.4 Parametrelerin anlamlılık testi

5.1.2.4.1 Olabilirlik oran testi

Kestirilen ve doymuş (gözlenen) modelin kıyaslanmasında sapma(deviance) olarak adlandırılan eşitlikten yararlanılmaktadır.

$$D = -2log\left[\frac{Kestirilen\ Modelin\ Olaabilirliği}{Doymuş\ Modelin\ Olabilirliği}\right] \qquad (5.18)$$

Olabilirlik fonksiyonunun yazılması ile ;

$$D = \sum_{i=1}^{n} d_i^2 = -2\sum_{i=1}^{n}(y_i \log\left(\frac{P_i}{y_i}\right) + (1-y_i)\log(1-P_i)/(1-y_i)) \qquad (5.19)$$

Biçiminde ifade edilen sapma ölçütü, k modelde yer alan parametre sayısını göstermek üzere (n-k) serbestlik dereceli χ^2 tablo değeri ile kıyaslanmaktadır.

Lojistik model için sapma, doğrusal regresyon analizinde kullanılan SSE (hataların kareleri toplamı) ile aynı anlama gelmektedir. Aynı zamanda bu eşitlik hipotez testi amaçlı olarak da kullanılabilmektedir. Bu hipotez testi olabilirlik oran testi olarak adlandırılmaktadır. Modelde yer alan parametrelerin anlamlı olup olmadıklarının belirlenmesinde, bağımsız değişkeni içeren modelin sapması bağımsız değişken içermeyen modelin sapması ile karşılaştırılmaktadır.

D değerindeki değişim G istatistiği olarak adlandırılmaktadır. G istatistiği aşağıdaki biçimde gösterilmektedir.

G= D (bağımsız değişken içermeyen model)-D(bağımsız değişken içeren model)

G= -2ln [bağımsız değişkensiz olabilirlik / bağımsız değişkenli olabilirlik]

5.1.2.4.2 Wald testi

Wald testi eğim parametresi β_1 'in en çok olabilirlik tahmini ile bu tahminin standart hatasının karşılaştırılması temeline dayanmaktadır. β_1 'in standart hatası kovaryans matrisindeki köşegen elemanların kareköklerinin alınması ile elde edilir.

$$W = \frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)} \sim N(0,1) \tag{5.20}$$

Wald istatistiği standart normal dağılım göstermektedir. Bu istatistiğin karesi alınır ise 1 serbestlik dereceli χ^2 dağılımı elde edilir.

$$W^2 = \left(\frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)}\right)^2 \sim \chi_1^2 \tag{5.21}$$

5.1.2.4.3 Skor testi

Skor testi ise en çok olabilirlik denklemlerinin türevlerinin koşullu dağılımlarına dayanmaktadır ve standart normal dağılım göstermektedir.

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i-\bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim N(0,1) \tag{5.22}$$

5.1.2.5 Lojistik regresyonda model seçimi

Lojistik regresyonda model seçimi;

i. Standart (direkt,tam,enter)
ii. Adımsal (aşamalı,stepwise)
    a) İleriye Doğru Adımsal
    b) Geriye Doğru Adımsal

olmak üzere 2 temel yöntemle yapılır.

5.1.2.5.1 Standart yöntem

Bu yöntemde tüm ortak değişkenler aynı anda bir blok olarak regresyon modeline alınmaktadır ve ardından her bir blok için kestirimler hesaplanmaktadır.

5.1.2.5.2 İleriye doğru adımsal yöntem

Bu yöntemde ilk önce sabit terim modele alınır ve ardından en anlamlı bulunan değişkenden başlamak üzere anlamlı değişken kalmayınca kadar modele değişkenler alınmaya devam edilir.Değişkenleri elemek için Wald istatistiği, durum indeksi ya da olabilirlik oran istatistiği kullanılır.

5.1.2.5.3 Geriye doğru adımsal yöntem

İleriye doğru adımsal yöntemin aksine bu yöntemde ilk önce tüm değişkenler modele alınır daha sonra modele en az katkıyı sağlayan değişkenden başlamak kaydıyla değişkenler modelden çıkartılır.

Lojistik regresyon analizi yapılırken uygun tüm değişkenlerin modele dahil edildiğinden ve uygun olmayan tüm değişkenlerin modelden çıkartıldığından emin olmak gerekir. Çünkü uygun değişkenlerin modele alınmaması hata terimi arttırıp modelin yetersiz olmasına sebep olacağı gibi uygun olmayan değişkenlerin modelde tutulması da modelin yorumlanmasını zorlaştırabilir. Veri içerisinde tekrarlayan kayıtların

olmadığından, kayıp veri olmadığından,uç değerler olmadığından ve ölçüm hatalarının olabildiğince az olduğundan emin olmak gerekir. Çünkü bu tip veriler katsayıların tahmininde yanlılığa ve buna bağlı olarak modelin yetersiliğine sebep olabilir. Örneklem büyüklüğünün yeterli olduğundan emin olmak gerekir. Çünkü yetersiz sayıda veri içeren bir örneklemle oluşturulan modelin güvenilirliği azalır.

Modelin anlamlı olup olmadığının belirlenmesinde G istatistiği kullanılmaktadır.

G= -2ln [bağımsız değişkensiz olabilirlik / bağımsız değişkenli olabilirlik]

Kuralacak hipotezler ise aşağıdaki şekildedir:

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

$H_s = En\ az\ bir\ tanesi\ 0'dan\ farklıdır.$

Sıfır hipotezi red edilir ise kurulan modelin anlamlı olduğu, kabul edilir ise kurulan modelin anlamlı olmadığı söylenir.

5.1.2.6 Modelin uyum iyiliği

Uyum iyiliği kurulan modelin etkinliğinin bir ölçütüdür ve bu ölçütün belirlenmesinde Hosmer-Lemeshow testi ve sınıflandırma tabloları (classification tables) kullanılmaktadır.Hosmer-Lemeshow testinde tahmin edilen olasılık değerleri yüzdeliklerine göre ya da almış oldukları kesin değerlere göre gruplandırılmaktadır. Hosmer-Lemeshow uyum iyiliği istatistiği Cˆ, gözlenen ve teorik frekanslardan oluşan 2×g tablosu için χ^2 değeri olarak bulunur ve aşağıdaki biçimde hesaplanır:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n_k\bar{\pi}_k)^2}{n_k\bar{\pi}_k(1-\bar{\pi}_k)} \quad (5.23)$$

$$O_k: Gözlenen\ Frekans\ \Rightarrow\ O_k = \sum_{j=1}^{n_k} y_j \quad (5.24)$$

$$\bar{\pi}_k: Tahmin\ edilen\ olasılık\ değerinin\ ortalamasıdır\ \Rightarrow\ \bar{\pi}_k = \sum_{j=1}^{n_k} \frac{m_j\,\hat{\pi}_j}{n_k} \quad (5.25)$$

Sınıflandırma Tabloları, bağımlı değişkenin çapraz sınıflandırılması ile elde edilir. Sınıflandırma tablosunda, bağımlı değişkeninin gözlenen ve kestirilen lojistik

olasılıklarından türetilen (0 veya 1) değerleri yer almaktadır. Türetilen bağımlı değişken değerlerinin elde edilmesinde bir kesim değerinin (c) tanımlanması zorunludur. Kestirilen olasılık değeri c'den büyük olduğunda türetilen bağımlı değişken 1, tersi durumda 0 değerini alır. Kesim değeri için en çok kullanılan değer 0,5' dir. Ele alınan kestirim değeri 0,5 değerini geçtiğinde 1, aksi durumda 0 grubuna atama yapılır. Ancak kestirim değerleri birbirine yakın olduğunda değerlerin farklı gruplara atanması bu yöntemin bir dezavantajdır.Örneğin kestirim değeri 0,51 olan değer 1 grubuna atanırken, 0,49 olan değer 0 grubuna atanır.

5.1.2.7 Lojistik regresyon uygulaması ve sonuçları

Bu tez kapsamında uygulanan lojistik regresyon model seçiminde standart yöntem kullanılmıştır. Toplamda 2832 adet kayıt içeren örneklemden rastgele yüzde 66'sı eğitim grubu olarak seçilmiştir. Tablo 5.2'de de gösterildiği gibi eğitim grubu olarak seçilen 1821 kayıt ile model kurulmuştur.

Tablo 5.2: Lojistik regresyon model özeti

Unweighted Cases(a)		**N**	**Percent**
Selected Cases	Included in Analysis	1821	100,0
	Missing Cases	0	0,0
	Total	1821	100,0
Unselected Cases		0	0,0
Total		1821	100,0
a.If weight is in effect,see classification table for the total number of cases.			

*Kaynak:*Clementine 12.0

Model için kurulan hipotez:

$H_0 = Model\ katsayıları\ anlamsızdır.$

$H_s = Model\ katsayıları\ anlamsız\ değildir.$

Tablo 5.3'de görüldüğü gibi anlamlılık düzeyi (Sig) <0,05 olduğundan H_0 hipotezi red edilir ve kurulan modelin anlamlı olduğu anlaşılır.

Tablo 5.3: Lojistik regresyon model katsayıları

		Chi-square	df	Sig.
Step 1	**Step**	1004,388	76	0,000
	Block	1004,388	76	0,000
	Model	1004,388	76	0,000

*Kaynak:*Clementine 12.0

Tablo 5.4'te modeldeki bağımsız değişkenler bağımlı değişkendeki değişimin yaklaşık olarak yüzde 57'sini açıkladığı görülmektedir. Bu oran gerçek veriler ile yapılan bir çalışma için normaldir.

Tablo 5.4: Lojistik regresyon bağımsız değişken açıklama oranı

Step	- 2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1519,592 (a)	0,424	0,565
a.Estimation terminated at iteration number 20 because maximum iterations has been reached.Final Solution cannot be found.			

*Kaynak:*Clementine 12.0

Tablo 5.5'te gösterildiği gibi kurulan modelin genel doğru sınıflama oranı yüzde 80,1'dir. İlk 16 ayında yasala girmeyen müşterileri doğru sınıflama oranı yüzde 78,3 iken, ilk 16 ayında yasala giren müşterileri doğru sınıflama oranı ise yüzde 81,7'dir.

Tablo 5.5: Lojistik regresyon doğru sınıflandırma oranı

	Observed		Predicted		
			EverYasalAccount16M		Percentage Correct
			0.0	1.0	
Step 1	EverYasalAccount16M	0.0	702	194	78,3
		1.0	169	756	81,7
	Overall Percentage				80,1
a.The cut value is 0,500					

*Kaynak:*Clementine 12.0

Modeli oluşturacak değişkenlerin belirlenmesi ise anlamlılık düzeylerine bakılarak anlaşılmaktadır.

$H_0 = Değişken\ katsayıları\ anlamsızdır.$

$H_s = Değişken\ katsayıları\ anlamsız\ değildir.$

Anlamlılık düzeyi (Sig) ≤ 0,05 olduğunda H_0 hipotezi red edilir ve o değişkenin model için anlamlı olduğuna karar verilir. Bu şekilde değerlendirme yapıldığında tablo 5.6'da gösterilen işaretli değişkenlerin model için anlamlı olduğu söylenebilir.

Tablo 5.6: Lojistik regresyon değişken özeti

	β	S.E.	Wald	df	Sig.	Exp(β)
CalismaSekli			15,618	6	0,016	
CalismaSekli(1)	-22,324	16191,43	0	1	0,999	0
CalismaSekli(2)	-0,433	0,884	0,24	1	0,624	0,649
CalismaSekli(3)	-0,316	0,228	1,925	1	0,165	0,729
CalismaSekli(4)	-1,488	0,865	2,958	1	0,085	0,226
CalismaSekli(5)	-0,899	0,249	13,052	1	0	0,407
CalismaSekli(6)	-0,232	0,158	2,149	1	0,143	0,793
BasvuruSayisi	0,322	0,186	2,987	1	0,084	1,38
KrediTur			11,439	5	0,043	
KrediTur(1)	0,615	0,556	1,225	1	0,268	1,85
KrediTur(2)	1,055	0,573	3,387	1	0,066	2,873
KrediTur(3)	0,115	0,399	0,083	1	0,774	1,122
KrediTur(4)	-0,311	0,424	0,537	1	0,464	0,733

KrediTur(5)	0,463	0,17	7,407	1	0,006	1,589
Marka			39,696	24	0,023	
Marka(1)	-2,591	56845,14	0	1	1	0,075
Marka(2)	-0,388	48195,01	0	1	1	0,679
Marka(3)	-18,805	40198,2	0	1	1	0
Marka(4)	-18,345	40198,2	0	1	1	0
Marka(5)	-18,567	40198,2	0	1	1	0
Marka(6)	-18,134	40198,2	0	1	1	0
Marka(7)	-17,864	40198,2	0	1	1	0
Marka(8)	-15,909	40198,2	0	1	1	0
Marka(9)	-18,772	40198,2	0	1	1	0
Marka(10)	-37,724	56485,14	0	1	0,999	0
Marka(11)	-17,741	40198,2	0	1	1	0
Marka(12)	2,515	56485,14	0	1	1	12,365
Marka(13)	-18,138	40198,2	0	1	1	0
Marka(14)	1,913	45363,21	0	1	1	6,777
Marka(15)	1,906	48541,89	0	1	1	6,723
Marka(16)	-16,227	40198,2	0	1	1	0
Marka(17)	-18,861	40198,2	0	1	1	0
Marka(18)	-18,626	40198,2	0	1	1	0
Marka(19)	-18,536	40198,2	0	1	1	0
Marka(20)	3,993	45026,18	0	1	1	54,243
Marka(21)	-21,907	40198,2	0	1	1	0
Marka(22)	1,277	44395,09	0	1	1	3,586
Marka(23)	-19,549	40198,2	0	1	1	0
Marka(24)	-17,029	40198,2	0	1	1	0
KararKategori			59,427	6	0	
KararKategori(1)	-1,013	0,461	4,825	1	0,028	0,363
KararKategori(2)	14,796	40192,97	0	1	1	2664733,448
KararKategori(3)	-0,459	0,928	0,245	1	0,621	0,632
KararKategori(4)	0,299	0,197	2,291	1	0,13	1,348
KararKategori(5)	-1,37	0,209	42,848	1	0	0,254
KararKategori(6)	-0,486	0,17	8,161	1	0,004	0,615
Vade	0,041	0,007	31,153	1	0	1,042
AracYasi2	0,115	0,067	2,925	1	0,087	1,122
LTV_FinansmanGrup			77,251	10	0	
LTV_FinansmanGrup(1)	-0,826	1,029	0,644	1	0,422	0,438
LTV_FinansmanGrup(2)	-2,167	1,118	3,76	1	0,052	0,114
LTV_FinansmanGrup(3)	-2,38	0,925	6,615	1	0,01	0,093
LTV_FinansmanGrup(4)	-1,538	0,864	3,166	1	0,075	0,215
LTV_FinansmanGrup(5)	-1,162	0,845	1,89	1	0,169	0,313

LTV_FinansmanGrup(6)	-0,59	0,852	0,479	1	0,489	0,554
LTV_FinansmanGrup(7)	-0,048	0,827	0,003	1	0,954	0,953
LTV_FinansmanGrup(8)	-0,031	0,827	0,001	1	0,97	0,969
LTV_FinansmanGrup(9)	0,168	0,886	0,036	1	0,85	1,183
LTV_FinansmanGrup(10)	-0,332	0,984	0,114	1	0,736	0,717
Bolge			10,776	6	0,096	
Bolge(1)	0,632	0,412	2,353	1	0,125	1,881
Bolge(2)	1,051	0,427	6,065	1	0,014	2,86
Bolge(3)	0,883	0,45	3,855	1	0,05	2,419
Bolge(4)	0,841	0,441	3,642	1	0,056	2,319
Bolge(5)	0,884	0,474	3,478	1	0,062	2,421
Bolge(6)	1,018	0,451	5,084	1	0,024	2,766
EverAccDpd30Plus(1)	-3,231	0,335	92,811	1	0	0,04
Borcluluk	0,012	0,078	0,025	1	0,873	1,012
CardOrODmaxOpenLimit	0,152	0,076	3,992	1	0,046	1,164
Kural3_NbOfRejAppL3M	0,682	0,417	2,674	1	0,102	1,977
Kural4_nbOf2UnpaidDel24M	0,024	0,055	0,187	1	0,666	1,024
Kural5_nbOf3UnpaidDel24M	0,13	0,185	0,495	1	0,482	1,139
Kural6_Loan	0,021	0,066	0,103	1	0,748	1,021
Kural7_Loan	0,071	0,127	0,313	1	0,576	1,074
Kural11_CardTotalOpenLimit	-0,029	0,134	0,048	1	0,827	0,971
Kural11_CardTotalOpenBalance	0,06	0,076	0,617	1	0,432	1,062
Kural12_LoanTotalOpenLimit	-0,947	0,199	22,756	1	0	0,388
Kural12_LoanTotalOpenBalance	1,015	0,203	24,905	1	0	2,761
Kural12_LoanMaxClosedLimit	0,003	0,021	0,02	1	0,888	1,003
Kural13_CardTSFirstOpened	-0,134	0,064	4,316	1	0,038	0,875
Kural14_CardTSLastOpened	-0,111	0,033	11,134	1	0,001	0,895
Kural14_LoanTSLastOpened	0	0,038	0	1	0,994	1
Constant	19,354	40198,2	0	1	1	254181267,1

Tablo 5.6'da gösterilen ve anlamlı çıkan değişkenleri sırasıyla yorumlamak gerekirse;

i. Çalışma şekli 'öğrenci' olan müşterilerin ilk 16 ayda yasal takip'e girmeme olasılıkları yasal takip'e girme olasılıklarına göre 2,46 kat daha fazladır.[3]

ii. Kredi türü 'Araç+FKS' olan müşterilerin ilk 16 ayda yasal takip'e girme olasılıkları yasal takip'e girmeme olasılıklarına göre 1,589 kat daha fazladır.

[3] β katsayısı negatif değerli olduğu durumlarda Exp(β) değeri 1/ Exp(β) olarak yorumlanır ve yorum tersine çevrilir.

iii. Karar Kategorisi ‘Politika Kuralından red’ olan müşterilerin ilk 16 ayda yasal takip’e girmeme olasılıkları yasal takip’e girme olasılıklarına göre 2,75 kat daha fazladır.

iv. Karar Kategorisi ‘Oto Onay İstisnası’ olan müşterilerin ilk 16 ayda yasal takip’e girmeme olasılıkları yasal takip’e girme olasılıklarına göre 3,93 kat daha fazladır.

v. Karar Kategorisi ‘gri alan’ olan müşterilerin ilk 16 ayda yasal takip’e girmeme olasılıkları yasal takip’e girme olasılıklarına göre 1,62 kat daha fazladır.

vi. Vade değişkeninde meydana gelen 1 birimlik artış ilk 16 ayda yasal takip’e girme olasılığını 1,042 kat arttırmaktadır.

vii. LTV’si [yüzde 41, yüzde 50] aralığında olan müşterilerin ilk 16 ayda yasal takip’e girmeme olasılıkları yasal takip’e girme olasılıklarına göre 10,75 kat daha fazladır.

viii. Bölgesi ‘İç Anadolu Bölgesi’ olan müşterilerin ilk 16 ayda yasal takip’e girme olasılıkları yasal takip’e girmeme olasılıklarına göre 2,86 kat daha fazladır.

ix. Bölgesi ‘Ege Bölgesi’ olan müşterilerin ilk 16 ayda yasal takip’e girme olasılıkları yasal takip’e girmeme olasılıklarına göre 2,41 kat daha fazladır.

x. Bölgesi ‘Doğu Anadolu’ olan müşterilerin ilk 16 ayda yasal takip’e girme olasılıkları yasal takip’e girmeme olasılıklarına göre 2,76 kat daha fazladır.

xi. EverAccDpd30Plus değişkeni ‘1’ olan müşterilerin ilk 16 ayda yasal takip’e girmeme olasılıkları yasal takip’e girme olasılıklarına göre 25 kat daha fazladır.

xii. CardOrMaxOpenLimit değişkeninde meydana gelen 1 birimlik artış ilk 16 ayda yasal takip’e girme olasılığını 1,164 kat arttırmaktadır.

xiii. LoanTotalOpenLimit değişkeninde meydana gelen 1 birimlik artış ilk 16 ayda yasal takip’e girmeme olasılığını 2,57 kat arttırmaktadır.

xiv. LoanTotalOpenBalance değişkeninde meydana gelen 1 birimlik artış ilk 16 ayda yasal takip’e girme olasılığını 2,76 kat arttırmaktadır.

xv. CardTSFirstOpened değişkeninde meydana gelen 1 birimlik artış ilk 16 ayda yasal takip’e girmeme olasılığını 1,14 kat arttırmaktadır.

xvi. CardTSLastOpened değişkeninde meydana gelen 1 birimlik artış ilk 16 ayda yasal takip’e girmeme olasılığını 1,11 kat arttırmaktadır.

xvii. Modelde sabit terim anlamsız bulunmuştur.

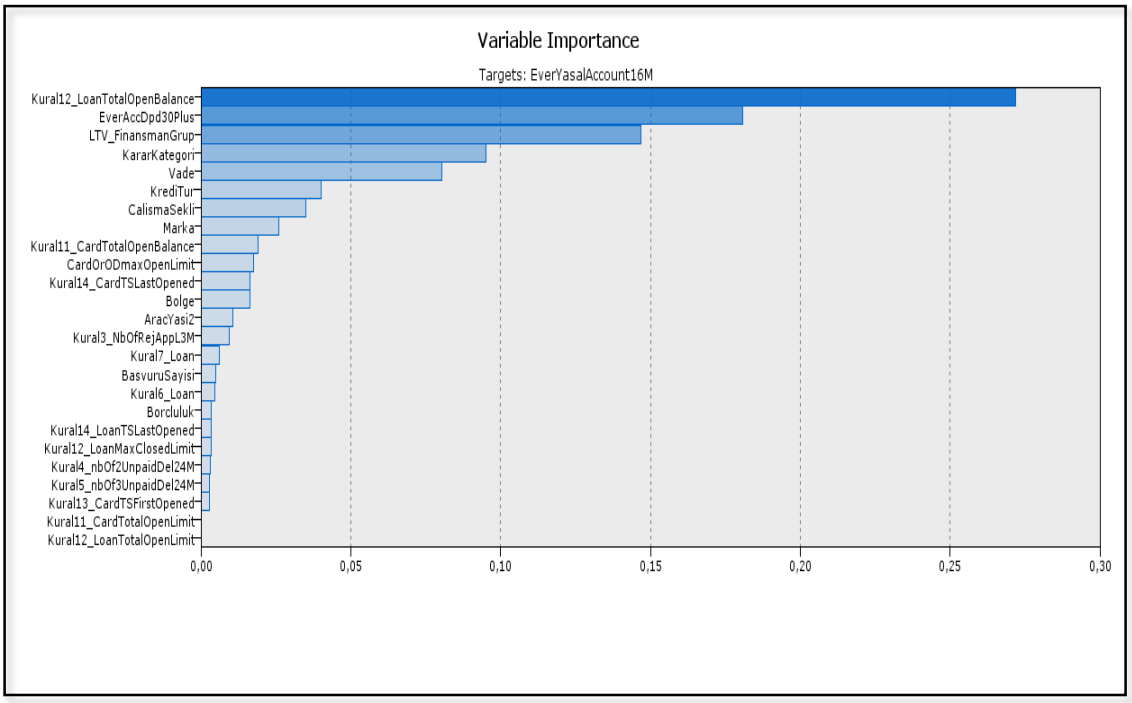
Buna göre lojistik regresyon denklemi aşağıdaki şekilde ifade edilir:

$$Y_1 = -0{,}899X_1 + 0{,}463X_2 - 1{,}013X_3 - 1{,}37X_4 - 0{,}486X_5 + 0{,}041X_6 - 2{,}38X_7 + 1{,}051 + 0{,}883X_9 + 1{,}018X_{10} - 3{,}231X_{11} + 0{,}152X_{12} - 0{,}947X_{13} + 1{,}015X_{14} - 0{,}134X_{15} - 0{,}111X_{16}$$

$$Y_1 = -0{,}899(CalismaSekli_5) + 0{,}463(KrediTur_5) - 1{,}013(KararKategori_1) - 1{,}37(KararKategori_5) - 0{,}486(KararKategori_6) + 0{,}041(Vade) - 2{,}38(LTV_{FinansmanGrup3}) + 1{,}051(Bolge_2) + 0{,}883(Bolge_3) + 1{,}018(Bolge_6) - 3{,}231(EverAccDpd30Plus) + 0{,}152(CardOrODmaxOpenLimit) - 0{,}947(Kural12LoanTotalOpenLimit) + 1{,}015(Kural12LoanTotalOpenBalance) - 0{,}134(Kural13CardTSFirstOpened) - 0{,}111(Kural14CardTSLastOpened)$$

Modele giren değişkenlerin modeli etkileme önem dereceleri ise şekil 5.1'de gösterildiği gibidir.

Şekil 5.1: Lojistik regresyon değişken önem sıralaması



Tablo 5.7'de gösterildiği gibi eğitim grubu ile oluşturulan model verilerin yüzde 80,07'sini doğru gruplamaktadır. Oluşturulan bu model test grubuna uygulandığında ise verilerin yüzde 78,73'ünün doğru sınıflandığı görülmektedir. Bu iki oranın birbirine çok yakın olması modelin başarılı olduğunun bir göstergesidir.

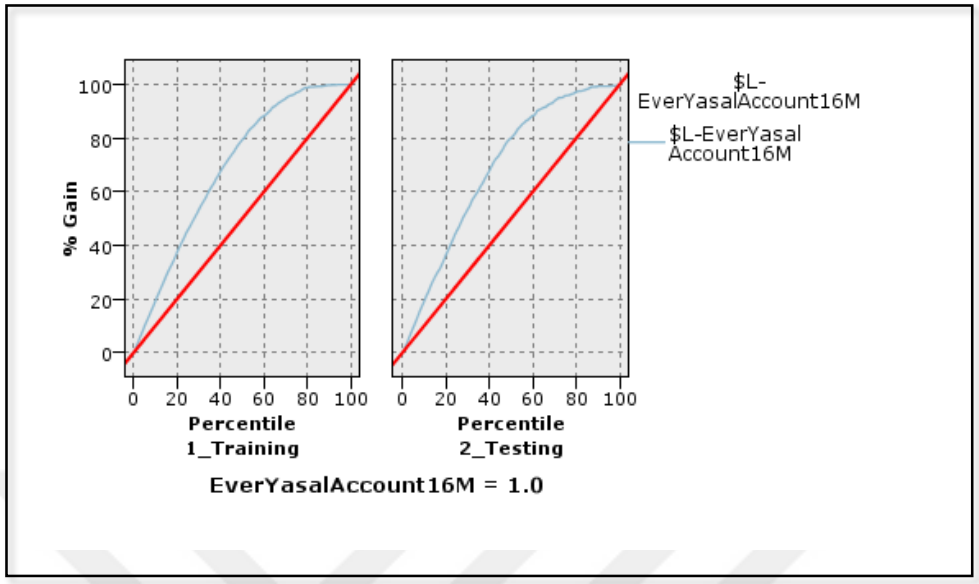
Tablo 5.7: Lojistik regresyon denetimli sınıflandırma özeti

Sınıflandırma	**1_Eğitim**		**2_Test**	
Doğru	1458	80,07%	796	78,73%
Yanlış	363	19,93%	215	21,27%
Toplam	1821		1011	

*Kaynak:*Clementine 12.0

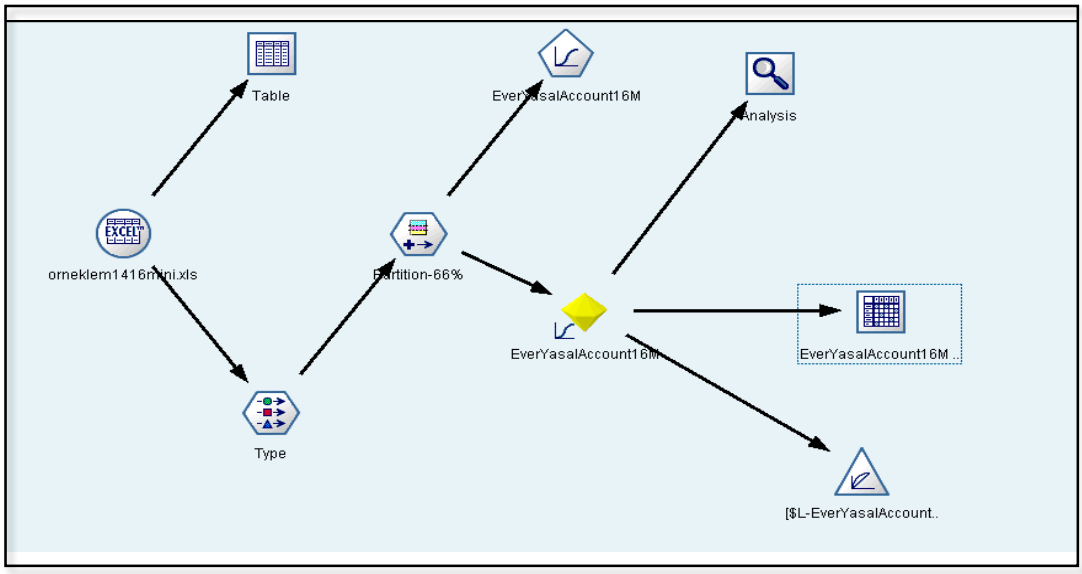
Şekil 5.2'de hem eğitim grubunda hem test grubunda verilerin yüzde 80'i kullanılarak modelleme yapılabildiği ve yüzde 20 kazanç sağlandığını görülmektedir.

Şekil 5.2: Lojistik regresyon kazanç grafiği



Lojistik regresyon modeli Clementine 12.0 programında şekil 5.3'te gösterildiği şekilde modellenmiştir.

Şekil 5.3: Clementine 12.0'da lojistik regresyon modellemesi



5.2 KARAR AĞAÇLARI, UYGULAMASI VE SONUÇLARI

5.2.1 Karar Ağaçları Tanımı Ve Temel Kavramlar

Karar ağaçları akış şemalarına benzeyen her bir niteliğin bir düğüm olarak temsil edildiği bir sınıflandırma yöntemidir.“Karar ağaçları, tek bağımlı değişken ve çok sayıda bağımsız değişkene sahip olmaları açısından regresyon modellerine benzerler. Bununla birlikte, ek olarak, veriden regresyon modellerine alternatif olabilecek farklı ve kullanışlı örüntüler keşfederler.” (Louis Anthony Cox, “Data Mining and Causal Modelling of Customer Behaviours”, Telecommunication Systems, Volume 21, 2002, s. 356)(Oded Maimon, Lior Rokach, Data Mining and Knowledge Discovery Handbook, Ramat-Aviv, Springer 2005, s. 183-184) Karar ağaçları düşük maliyetli ve veri tabanı ile kolay entegre edilebilir olmaları,anlaşılmasının ve yorumlanmasının kolay olması, eksik veya hatalı veriler ile tahminlme yapabiliyor olmaları ve güvenilirliklerinin iyi olması özellikleri ile en yaygın kullanılan sınıflama tekniklerinden biridir. Karar ağacı yönteminde iki basamak vardır. Bunlardan ilki öğrenme basmağıdır. Öğrenme basamağında eğitim verisi model oluşturmak amacı ile algoritma tarafından çözümlenir, ardından oluşturulan model test verisi tarafından modelin doğruluğunu belirlemek amacıyla kullanılır. Bu da ikinci basamak olan sınıflama basamağıdır.

Karar ağacında başlangıç düğümüne kök düğüm, ara safhalardaki düğümlere ara düğüm, ağacın bittiği düğüme ise son düğüm denmektedir. Her bir düğümdeki gözlem sayısı o düğümün büyüklüğünü, ağaçtaki dallanma sayısı ise ağacın derinliğini ifade eder.

Karar ağacı yönteminde ilk olarak bilgi kuramı, olasılık teorisi, istatistiksel yöntemler, optimizasyon ya da yapay zeka algoritmaları ile her bir sınıf için en ayırt edici kırılmayı sağlayan bağımsız değişken seçilir. O değişken üzerinden koşullandırmayı yani dalları oluşturacak değişkenler yaprak düğümlere taşınır. Böylece kök(ana) düğümden yaprak düğümler elde edilir. Düğüm oluşturma işleme algoritmanın çeşidinden çeşidine bağlı olarak değişebilen belirli durdurma kuralları devreye girene kadar ya da ağaçta daha fazla kırılmaya yapılmayacak duruma gelene kadar devam eder.

5.2.1.1 Ağacın büyümesi

Ağaç oluşturulurken öncelikli amaç veri setindeki tüm gözlemleri n sayıda sınıfa ayırmaktır. N sayıdaki sınıfın gözlemlerini en iyi şekilde kıracak olan bağımsız değişkenin bulunup gözlemlerin bu değişken üzerinden k tane düğüme ayrılması ana hedeftir.uygun kırılma sayısı değişkenin tipine göre belirlendiği için değişken tiplerinin ne şekilde belirlendiği kritik bir noktadır. N kategorili değişkenler için (n-1) sayıda koşul oluşturarak n ya da daha az kırılma gerçekleştirirken, sürekli değişkenler küçükten büyüğe doğru sıralanır ve kırılma ölçütü $x_1 \leq x_2 \leq \cdots \ldots \ldots \ldots \leq x_n \, olmak$ üzere $x < ort(x_1, x_{n+1})$'dir. En uygun kırılma bulunduğunda kırılma gerçekleşir ve bu kırılmalar her düğüm için durdurma kriteri devreye girene kadar devam eder.

5.2.1.2 Ayırma kriterleri

Dallanmayı devam ettirecek olan değişkenlerin seçiminde temel olarak aşağıdaki kriterler kullanılmaktadır.

5.2.1.2.1 Tek değişkenli ayırma kriterleri

Tek değişkenlilikten kasıt ara düğümün (iç düğümün) tek bir özniteliğin değerine göre bölünmesidir.

5.2.1.2.1.1 Bilgi kazancı(information gain)

Sistemin düzensizliğini ifade eden entropi ilkesine dayanır.Bağımlı değişkenin kategorik olduğu durumlarda kullanılır. Düzensizliğin azalması ile bilgi kazanımı artar.algoritma düzensizliğin azaltılmasını amaçlar.herhangi bir düğüm için en fazla bilgiyi kazandırabilecek değişken ayırıcı değişken olarak seçilir.

5.2.1.2.1.2 Gini indeksi

Bağımlı değişkenin kategorik olması durumlarında kullanılır.bilgi kazanımı yaklaşımına benzer.Gini indeksi bir olayın gerçekleşme ve gerçekleşmeme olasılıklarının karesinin hesaplanıp 1'den çıkarılmasıyla elde edilir. Yani hedef niteliklerin değerlerinin olasılık dağılımları arasındakiuzaklığı ölçer.En düşük gini indeksine sahip değişken seçilerek ağaç dallandırılır.

5.2.1.2.1.3 Kazanç oranı (gain ratio)

Bilgi kazancının normalleştirilmesidir.Öncelikle bilgi kazancı tüm nitelikler için hesaplanır.ancak yalnızca ortalama bilgi kazancı kadar ya da daha fazla performans gösteren nitelikler dikkate alınır ve en iyi kazanç oranını elde eden nitelik seçilir.kazanç oranı doğruluk açısından ve daha basit sınıflandırma yapması açısından bilgi kazancına göre daha iyi performans gösteririr.

5.2.1.2.1.4 χ^2 testi

Bağımlı değişkenin kategorik olduğu durumlarda kullanılır.

5.2.1.2.1.5 F testi

Bağımlı değişkenin sürekli olduğu durumlarda kullanılır. Ki-kare ve F testleri homojen olarak değerlendirilebilecek tüm değerleri birleştirir ve geri kalan tüm değerleri heterojen olarak değerlendirir.

5.2.1.2.2 Çok değişkenli ayırma kriterleri

Bağımsız değişkenlerin kombinasyonlarına dayalı olarak gerçekleştirilir.burada en optimal çözüme ulaşmak daha çok deneme yanılma yöntemine dayalı olan doğrusal programlama ya da diskriminant analizi ile mümkün olur.

5.2.1.3 Büyümenin durdurulması

Teorik olarak düğümlerdeki gözlemler ayrılamayacak durumda özdeş olduğunda büyümenin durması gerekmektedir.ancak ağacın derinliğinin daha fazla artmasını engellemek adına ağacın derinliği belirlenen sınıra ulaştığında, her düğümde gözlenen gözlem sayısı belirlenenden daha az sayıya düştüğünde ve istenen başarıya ulaşıldığında durdurma gerçekleştirilebilir. Çünkü karar ağacının büyüklüğü modelin kalitesini belirler, çok küçük ve çük büyük fazla dallanmış ağaçların veriyi temsil yeteneği düşük olabilir.

5.2.1.4 Budama

Karar ağacı oluşurken çok fazla dallanmış, bazı dallarında aşırı sapmalar meydana gelmiş, ya da aşırı öğrenme gerçekleşmiş olabilir.bu durumda fazla güvenilir olmayan dalların ayıklanması,budanması ağacın daha kolay yorumlanmasını sağlar. Budama ön budama ve son budama olarak ikiye ayrılır. Ön budama ağacın dallandırma aşamasında ayırma kriterlerine belirli bir eşik değerinin verilerek daha fazla dallanmasını engelleyerek gerçekleştirilir. Son budamada ise bütün karar ağacı oluşturulduktan sonra bütün üzerinden küçültme işlemi gerçekleştirilir.

5.2.1.5 Avantajları & dezavantajları

i. Karar ağaçları kendini açıklayıcı özelliğe sahiptir ve yoğun olduğu zaman bile takip etmesi kolaydır. Yani, karar ağacı makul sayıda yapraklara sahip olduğu zaman, profesyonel olmayan kullanıcılar tarafından bile kolayca anlaşılabilirdir.
ii. Karar ağaçları kurallar kümesine dönüştürülebilir.
iii. Karar ağaçları hem nominal (kategorik) hem de sayısal (sürekli) girdi öznitelikleri ile işlem yapabilir.
iv. Karar ağacının gösterimi herhangi bir ayrık değerli sınıflandırıcıyı ifade etmek için çok zengindir.
v. Karar ağaçları hatalar ihtiva eden veri kümelerini işleme yeteneğine sahiptir.
vi. Karar ağaçları eksik değerlere (missing values) sahip veri kümelerini işleme konusunda başarılıdır.

vii. Karar ağaçları parametrik olmayan bir metot olarak kabul edilir. Bundan dolayı, karar ağaçları uzay dağılımı ve sınıflandırıcının yapısı hakkında varsayımlara sahip değildir.

Bununla beraber, karar ağaçları bazı dezavantajlara da sahiptir.

i. Birçok karar ağacı algoritmaları, ID3 ve C4.5 gibi, hedef özniteliğin sadece kesikli ayrık değerler (discrete values) gerektirmektedirler.
ii. Karar ağaçlarının aç gözlü özelliği bir diğer dezavantaj sağlamaktadır. Bu dezavantaj ise karar ağaçlarının eğitme kümesine, alakasız özniteliklere ve ses'e karşı aşırı duyarlı olmasıdır.
viii. Karar ağacı "böl ve yönet" metodu kullandığı gibi, eğer birkaç çok tane çok alakalı öznitelik mevut ise çok iyi bir performans göstermektedirler, fakat eğer birçok karmaşık etkileşimler mevcut ise daha az performans gösterirler. Bunun sebeplerinden bir tanesi diğer sınıflandırıcılar bir karar ağacını kullanarak ifade etmesi çok zor olan bir sınıflandırıcıyı kompakt bir şekilde tarif edebilirler.

5.2.2 CHAID (chi-squared automatic ınteraction detection)

1980 yılında G.V.Kass tarafından geliştirilen chaid algoritması hem sınıflama hem de regresyon amacı ile kullanılmaktadır. Chaid yani ki-kare otomatik ilişki tespiti algoritması hem sürekli hem de kategorik bağımlı ve bağımsız değişkenler ile çalışabilir. Dallanma yaparken değişken seçiminde kategorik değişkenlerle çalıştığı zaman χ^2 testini ,sürekli değişkenlerle çalıştığı zaman ise F testini kullanır.kayıp değerler ise ayrı bir kategoride değerlendirirlir. Test sonucunda olasılık değeri en küçük olan yani önem değeri en büyük olan değişken dallanma için seçilir.bu algoritma bağımlı değişken ile dallanan değişken arasındaki bağımlılığı test eder. Eğer bağımlılık var ise ağaç büyür.eğer bağımlılık yok ise büyüme durur.çünkü amaç bağımlı değişkenin dallanarak açıklanmasıdır ve bağımlı değişken ile bağımsız değişkenlerin ayrık olmaları bağımsız değişkenin bağımlı değişkeni açıklamadığını gösterir.chaid algoritması kategorik değişkenler ile çalışmayı tercih etmektedir. Bu sebeple bağımsız değişkenler sürekli ise algoritma bu değişkenleri bölerek kategorik hale getirir.

5.2.2.1 CHAID Algoritması Uygulaması ve Sonuçları

CHAID karar ağacı modelinin performans ölçümleri tablo 5.8'de gösterildiği gibidir. 2832 adet kayıttan oluşan örneklemin rastgele belirlenmiş olan yüzde 66'sı eğitim kümesi olarak kullanılmış ve kurallar oluşturulmuştur. Verinin geri kalanı ise bu kuralların test edildiği test kümesi olarak kullanılmıştır. Doğru sınıflama oranı eğitim kümesinde yüzde 77,21 iken test kümesinde yüzde 72,4'e gerilemiştir.

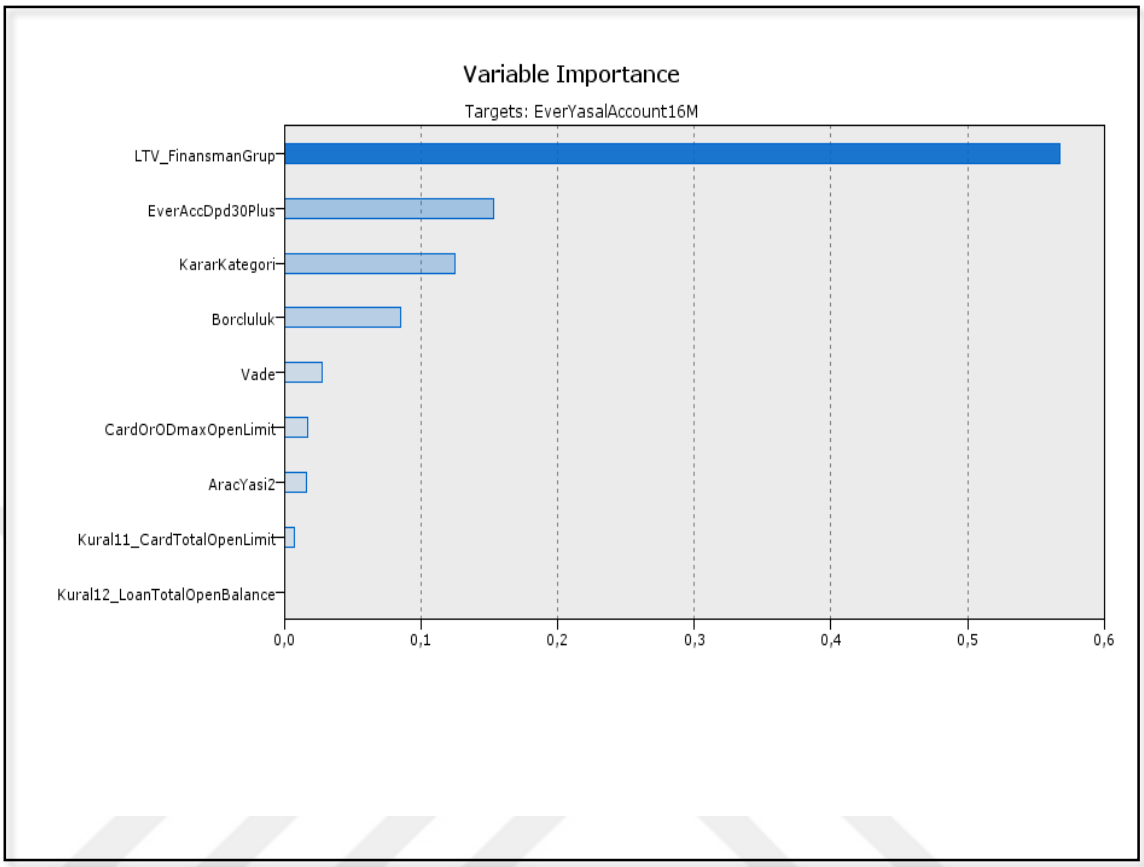
Tablo 5.8: CHAID algoritması denetimli sınıflandırma özeti

Sınıflandırma	**1_Eğitim**		**2_Test**	
Doğru	1406	77,21%	732	72,40%
Yanlış	415	22,79%	279	27,60%
Toplam	1821		1011	

*Kaynak:*Clementine 12.0

Müşterilerin ilk 16 ayda yasal takip'e girme eğilimlerine etki eden faktörler ve önem sıraları şekil 5.4'te gösterildiği gibidir.

Şekil 5.4: CHAID algoritması değişken önem sıralaması



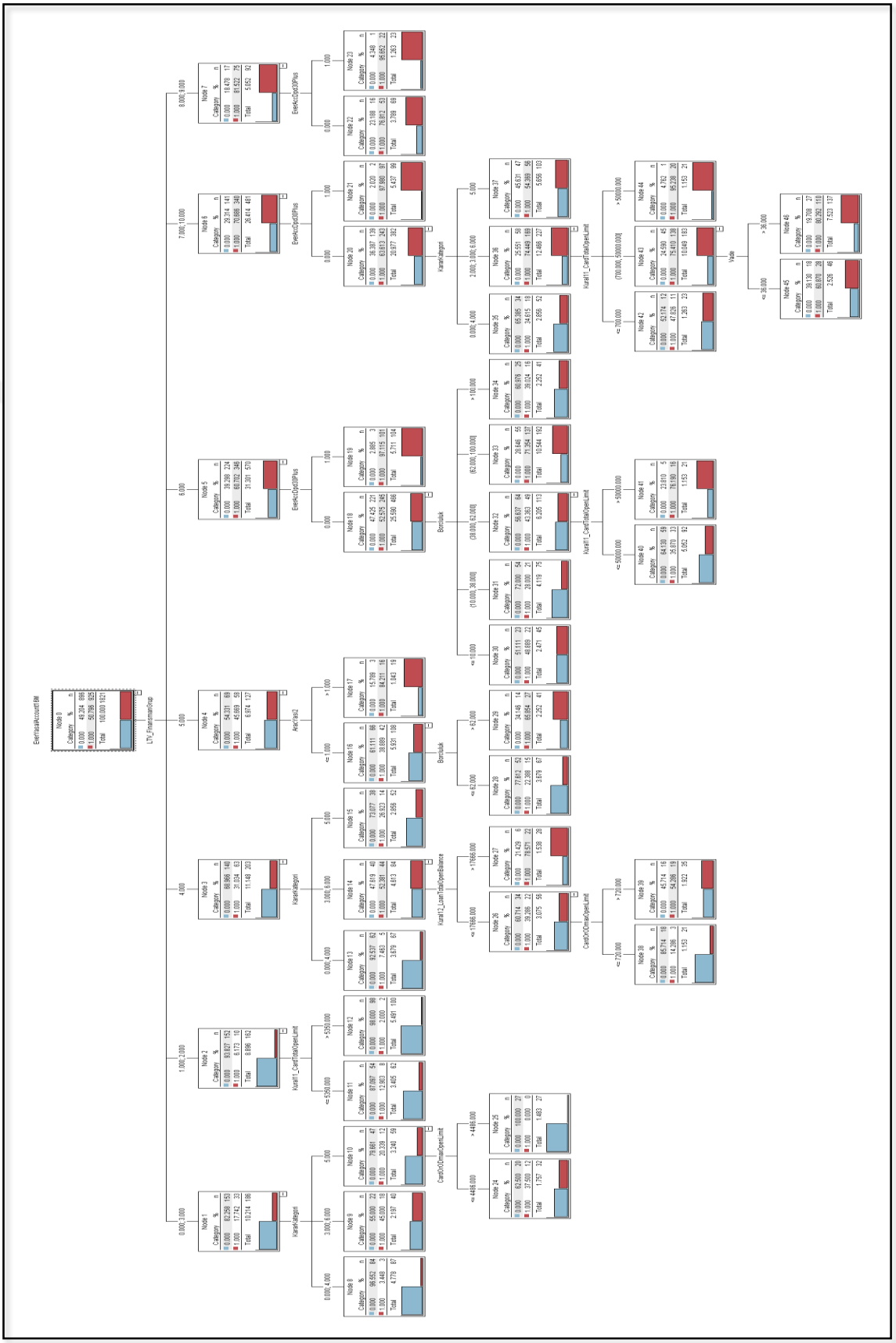
CHAID algoritmasına göre müşterinin yasal takibe girme eğilimini belirleyen en önemli faktör müşterinin alacağı taşıt fiyatının yüzde kaçı için kredi talep ettiğinin bilgisinin tutulduğu "LTV" değişkenidir ve önem oranı yüzde 57'dir.En önemli 2. Faktör yüzde 16'lık önem oranına sahip olan müşterinin krediyi ödemeye başladıktan sonra ilk 3 taksidinde 30 gün ve üstü gecikmeye düşüp düşmediğinin bilgisinin tutulduğu "EverAccDpd30Plus" değişkenidir.En önemli 3. Faktör ise başvurunun sistemde aldığı karar bilgisinin tutulduğu "Karar Kategori" değişkenidir ve önem oranı yüzde 13'tür.Müşterinin kredi kartı borçluluğunu gösteren "borçluluk" değişkeni yüzde 9'luk önem oranı ile müşterinin yasal takip'e girme eğilimini etkileyen 4. faktördür.Yüzde 3'lük önem oranı ile "vade" değişkeni 5. Önemli faktör olurken müşterinin açık maksimum kredi kartı ya da overdraft limit bilgisinin tutulduğu "CardOrODMaxOpenLimit" değişkeni yüzde 2'lik önem derecesi ile 6. Faktör olarak grafikte yerini almaktadır."AraçYaşı" değişkeni ise en önemli 7. Faktördür ve önem derecesi yüzde 1,6'dır. 8. sırada yüzde 1 önem oranı ile müşterinin güncel toplam kart

limiiti bilgisinin tutulduğu “CardTotalOpenLimit” değişkeni vardır. Müşterinin güncel toplam kredi borcunu gösteren “LoanTotalOpenBalance” değişkeni ise yüzde 0,4’lük önem oranı ile en son sırada yer almaktadır.

CHAID algoritması karar ağacı kural setini yazmak gerekirse aşağıdaki şekilde yazılabilir.

LTV_FinansmanGrup = 0 or LTV_FinansmanGrup = 3 [Mode: 0] 186)
KararKategori = 0 or KararKategori = 4 [Mode: 0] => 0,0 (87; 0,966)
KararKategori = 3 or KararKategori = 6 [Mode: 0] => 0,0 (40;0,55)
KararKategori = 5 [Mode: 0] (59)
CardOrODmaxOpenLimit <= 4.486 [Mode: 0] => 0,0 (32; 0,625)
CardOrODmaxOpenLimit > 4.486 [Mode: 0] => 0,0 (27; 1,0)
LTV_FinansmanGrup = 1 or LTV_FinansmanGrup = 2 [Mode: 0] (162)
Kural11_CardTotalOpenLimit <= 5.350 [Mode: 0] => 0,0 (62; 0,871)
Kural11_CardTotalOpenLimit > 5.350 [Mode: 0] => 0,0 (100; 0,98)
LTV_FinansmanGrup = 4 [Mode: 0] (203)
KararKategori = 0 or KararKategori = 4 [Mode: 0] => 0,0 (67; 0,925)
KararKategori = 3 or KararKategori = 6 [Mode: 1] (84)
Kural12_LoanTotalOpenBalance <= 17.666 [Mode: 0] (56)
CardOrODmaxOpenLimit <= 720 [Mode: 0] => 0,0 (21; 0,857)
CardOrODmaxOpenLimit > 720 [Mode: 1] => 1,0 (35; 0,543)
Kural12_LoanTotalOpenBalance > 17.666 [Mode: 1] => 1,0 (28; 0,786)
KararKategori = 5 [Mode: 0] => 0,0 (52; 0,731)
LTV_FinansmanGrup = 5 [Mode: 0] (127)
AracYasi2 <= 1 [Mode: 0] (108)
Borcluluk <= 62 [Mode: 0] => 0,0 (67; 0,776)
Borcluluk > 62 [Mode: 1] => 1,0 (41; 0,659)
AracYasi2 > 1 [Mode: 1] => 1,0 (19; 0,842)
LTV_FinansmanGrup = 6 [Mode: 1] (570)
EverAccDpd30Plus = 0 [Mode: 1] (466)
Borcluluk <= 10 [Mode: 0] => 0,0 (45; 0,511)
Borcluluk > 10 and Borcluluk <= 38 [Mode: 0] => 0,0 (75; 0,72)
Borcluluk > 38 and Borcluluk <= 62 [Mode: 0] (113)
Kural11_CardTotalOpenLimit <= 50.000 [Mode: 0] => 0,0 (92; 0,641)
Kural11_CardTotalOpenLimit > 50.000 [Mode: 1] => 1,0 (21; 0,762)
Borcluluk > 62 and Borcluluk <= 100 [Mode: 1] => 1,0 (192; 0,714)
Borcluluk > 100 [Mode: 0] => 0,0 (41; 0,61)
EverAccDpd30Plus = 1 [Mode: 1] => 1,0 (104; 0,971)
LTV_FinansmanGrup = 7 or LTV_FinansmanGrup = 10 [Mode: 1] (481)
EverAccDpd30Plus = 0 [Mode: 1] (382)
KararKategori = 0 or KararKategori = 4 [Mode: 0] => 0,0 (52; 0,654)
KararKategori = 2 or KararKategori = 3 or KararKategori = 6 [Mode: 1] (227)
Kural11_CardTotalOpenLimit <= 700 [Mode: 0] => 0,0 (23; 0,522)
Kural11_CardTotalOpenLimit > 700 and
Kural11_CardTotalOpenLimit <= 50.000 [Mode: 1]
Vade <= 36 [Mode: 1] => 1,0 (46; ,609)
Vade > 36 [Mode: 1] => 1,0 (137; ,803)
Kural11_CardTotalOpenLimit > 50.000 [Mode: 1] => 1,0 (21; 0,952)
KararKategori = 5 [Mode: 1] => 1,0 (103; 0,544)
EverAccDpd30Plus = 1 [Mode: 1] => 1,0 (99; 0,98)
LTV_FinansmanGrup = 8 or LTV_FinansmanGrup = 9 [Mode: 1] (92)
EverAccDpd30Plus = 0 [Mode: 1] => 1,0 (69; 0,768)
EverAccDpd30Plus = 1 [Mode: 1] => 1,0 (23; 0,957)

Şekil 5.5: CHAID algoritması karar ağacı



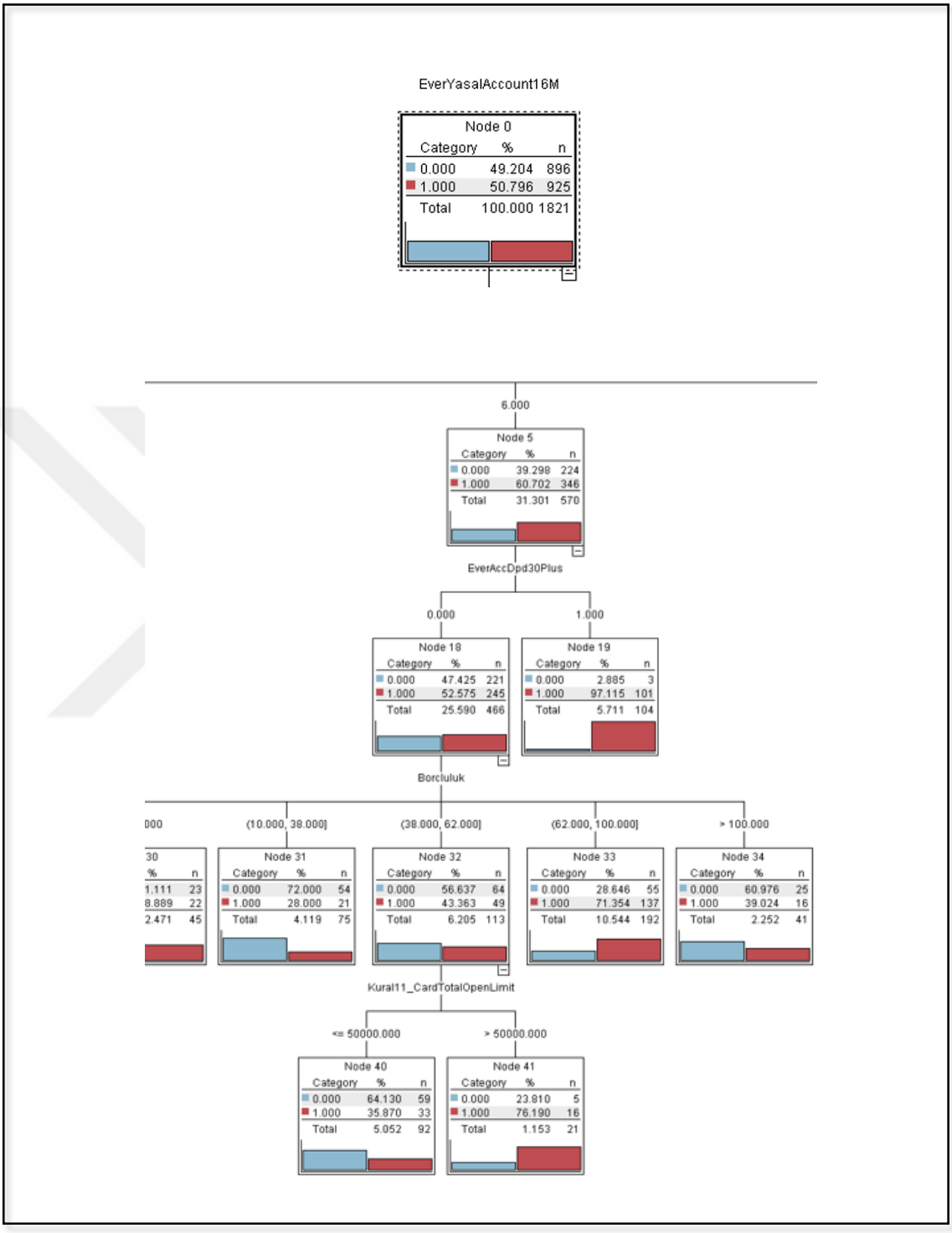
Şekil 5.5'te görüldüğü gibi elde edilen karar ağacı 5 seviye ve 46 alt düğümden oluşmaktadır.

Karar ağacı incelendiğinde müşterilerin yasal takip'e girip girmediğini gösteren bağımlı değişken ilk önce "LTV" faktörü üzerinden 7 dala ayrılmıştır. Bu 7 dal daha sonra "Karar Kategori", "CardTotalOpenLimit","AraçYaşı" ve "EverAccDpd30Plus" faktörleri yardımıyla alt dallara bölünmüştür. Bu alt dallar ise kendi içlerinde "CardOrODMaxOpenLimit", "LoanTotalOpenBalance", "Borçluluk", "KararKategori" ve "CardTotalOpenLimit" faktörleri ile son kırılımlarını gerçekleştirmişlerdir.

Bu tez kapsamında yasal takibe girme eğilimi olan müşterilerin özellikleri ortaya konulmaya çalışıldığından yasal takibe girme oranı üzerinde etkisi olan düğümlerin yorumu üzerinde yoğunlaşılmıştır.

Şekil 5.6'da görüldüğü gibi en başta tüm müşteriler içerisinde yaklaşık yüzde 51 yasal takibe gitme oranı var iken bu oran LTV'Sİ yüzde 71- yüzde 75 aralığında olan müşterilerin seçildiği düğüm 5'te yüzde 60,702'ye çıkmaktadır. Bir alt kırılım olan düğüm 19'da, ilk 3 taksidinde 30 gün ve üzeri gecikmeye düşen müşterilerde ise bu oran yüzde 97,115'e kadar yükselmektedir. İlk 3 taksidinde 30 gün ve üzeri gecikmeye düşmeyen müşterilerde ise borçluluğu yüzde 62-100 aralığında olan müşterilerde en başta yüzde 51 olan oran yüzde 71,354 seviyelerine yükselmektedir.(Düğüm 33)

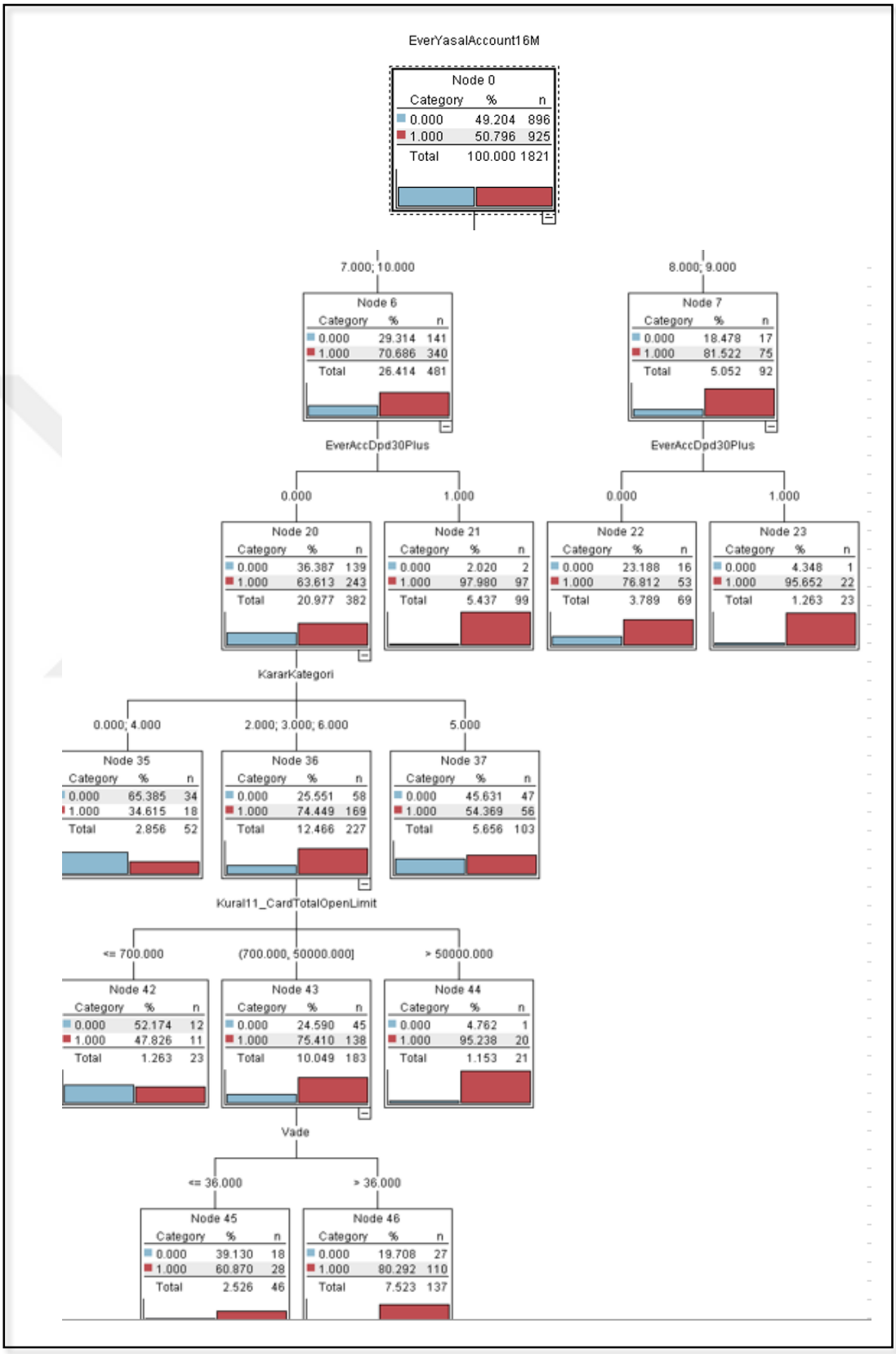
Şekil 5.6: CHAID algoritması karar ağacı düğüm 5,18,19,31,32,33,34,40,41



Şekil 5.7'de görüldüğü gibi en başta tüm müşteriler içerisinde yaklaşık yüzde 51 yasal takibe gitme oranı var iken bu oran LTV'Sİ yüzde 76- yüzde 80 ile yüzde 91-95 aralığında olan müşterilerin seçildiği düğüm 6'da yüzde 70,686'ya çıkmaktadır. Bir alt kırılım olan ilk 3 taksidinde 30 gün ve üzeri gecikmeye düşen müşterilerde ise bu oran yüzde 97,980'e kadar yükselmektedir. İlk 3 taksidinde 30 gün ve üzeri gecikmeye düşmeyen müşterilerde ise karar kategorisi oto onay,oto red ya da gri alan olan müşterilerde yüzde 70,686 olan oran yüzde 74,449'a yükselmektedir. Bu müşterilerden açık kredi kartı toplam limiti 50.000 TL'nin üzerinde olan müşterilerin yüzde 95,238'i yasal takibe girmektedir. Açık kredi kartı toplam limiti (700 TL, 50.000 TL] aralığında olan müşterilerde ise yüzde 74,449 olan oran yüzde 75,410'a yükselmektedir. Bu müşterilerden ise vadesi 36 aydan fazla olan müşterilerin yasal takip'e girme oranı yüzde 80,292'dir.

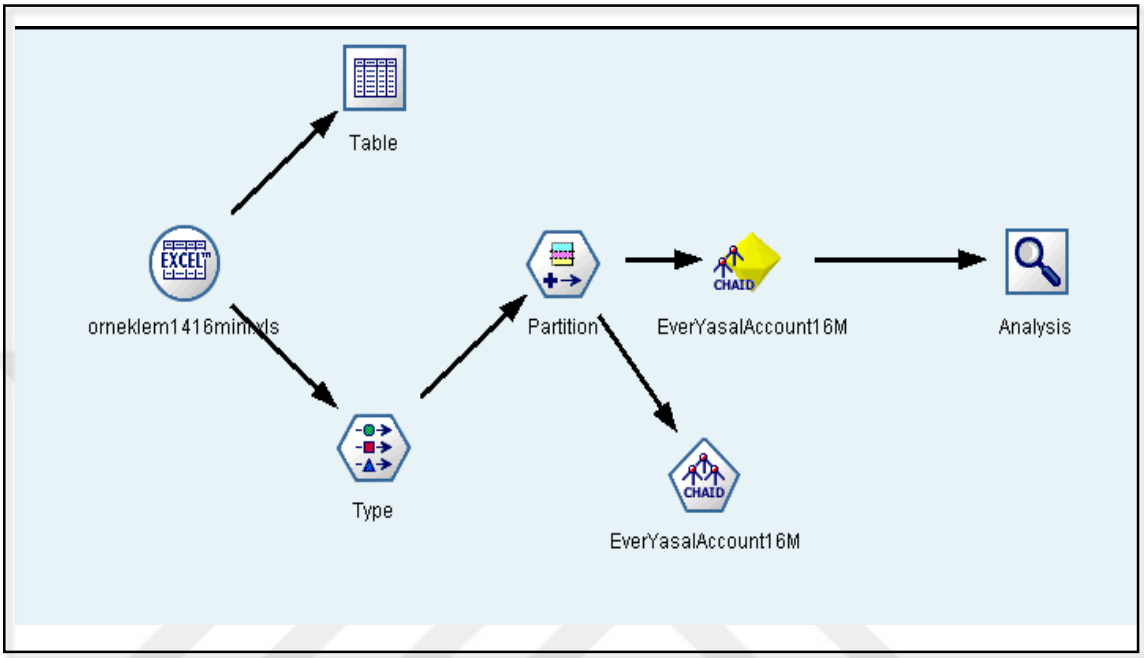
En başta tüm müşteriler içerisinde yaklaşık yüzde 51 yasal takip'e gitme oranı var iken bu oran LTV'Sİ yüzde 81-90 aralığında olan müşterilerin seçildiği düğüm 7'de yüzde 81,522'ye çıkmaktadır. Bir alt kırılım olan ilk 3 taksidinde 30 gün ve üzeri gecikmeye düşen müşterilerde ise bu oran yüzde 95,652'ye kadar yükselmektedir.İlk 3 taksidinde 30 gün ve üzeri gecikmeye düşmeyen müşterilerde yasal takip'e girme oranı yüzde 76,812'dir.

Şekil 5.7: CHAID algoritması karar ağacı düğüm 6, 7, 20, 21, 22, 23, 35, 36, 37, 42, 43, 44, 45, 46



CHAID modeli Clementine 12.0 programında şekil 5.8'de gösterildiği gibi modellenmiştir.

Şekil 5.8: Clementine 12.0'da CHAID algoritması modellemesi



5.2.3 C&RT (classification and regression tree)

1984 yılında Leo Breiman, Jerome Friedman, Richard Olshen ve Charles Stone tarafından geliştirilen bu algoritma bağımlı değişkenin kategorik olduğu durumlarda sınıflama,sürekli olduğu durumlarda ise regresyon amacı ile kullanılmakatadır. Bu algoritmada amaç doğruluğu en iyi olan modeli kurmaktır. Dallanma sürecinde değişken seçimi χ^2 ya da gini indeksi testi kullanılarak yapılır. Dallanma tüm durumları en iyi şekilde sınıflandırılıncaya ya da tahmin edilinceye kadar devam eder.ağacın yapısı yorumlanmayacak derecede karmaşık bir yapıya geldiğinde ya da önceden tanımlanmış ağaç derinliğine ulaşıldığında ya da diğer kriterlere ulaşıldığında büyüme durdurulabilir. Modelin doğruluğu bağımlı değişken kategorik olduğunda doğru tahmin edilen kayıtların oranı ile, sürekli olduğunda ise ortalama hata kareler ile ölçülür. Bu algoritmanın amacı yalnızca bir ağaç oluşturmak değil aynı zamanda sıralı iç içe her biri kendi içerisinde optimal olan budanmış ağaçlar oluşturmaktır.

5.2.3.1 C&RT Uygulaması ve Sonuçları

C&R karar ağacı modelinin performans ölçümleri aşağıdaki tabloda gösterildiği gibidir. 2832 adet kayıttan oluşan örneklemin rastgele belirlenmiş olan yüzde 66'sı eğitim kümesi olarak kullanılmış ve kurallar oluşturulmuştur.Verinin geri kalanı ise bu kuralların test edildiği test kümesi olarak kullanılmıştır. Doğru sınıflama oranı eğitim kümesinde yüzde 78,09 iken test kümesinde yüzde 74,88'e gerilemiştir.

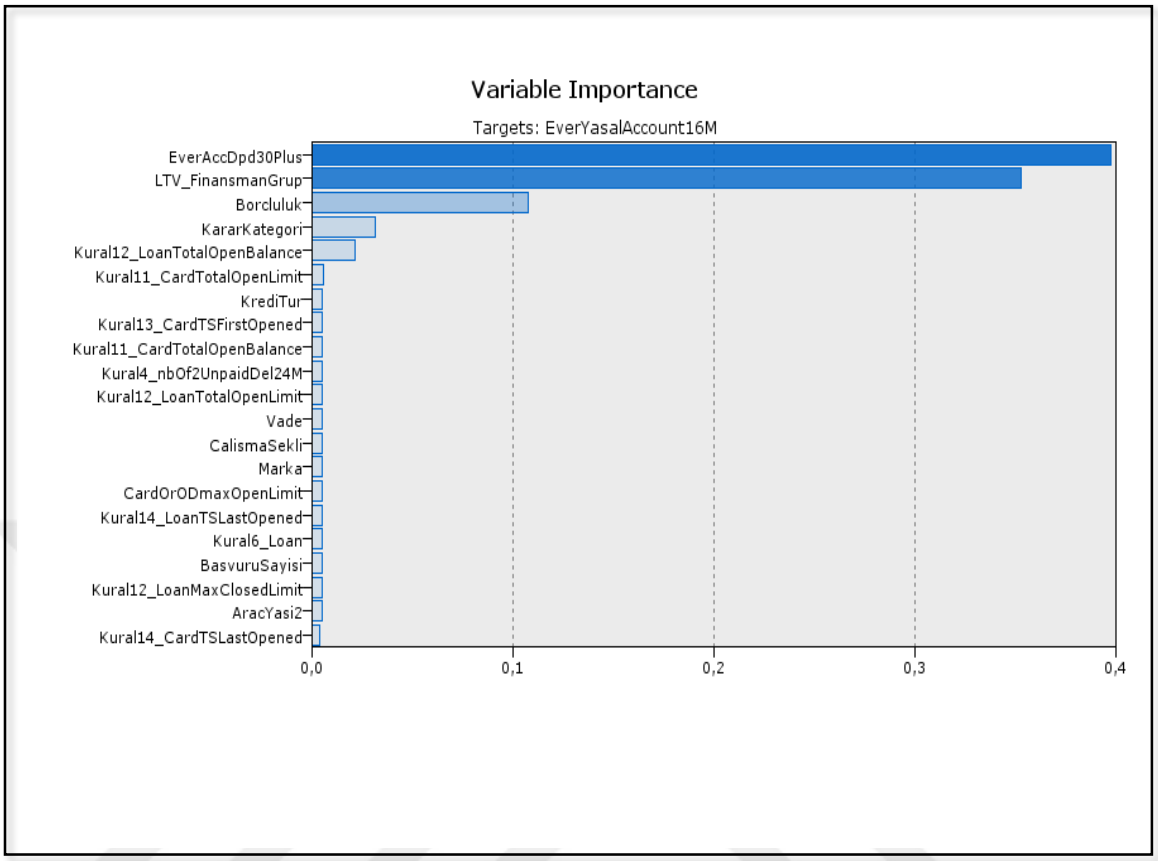
Tablo 5.9: C&RT algoritması denetimli sınıflandırma özeti

Sınıflandırma	**1_Eğitim**		**2_Test**	
Doğru	1422	78,09%	757	74,88%
Yanlış	399	21,91%	254	25,12%
Toplam	1821		1011	

*Kaynak:*Clementine 12.0

Müşterilerin ilk 16 ayda yasal takip'e girme eğilimlerine etki eden faktörler ve önem sıraları şekil 5.9'daki grafikte gösterildiği gibidir.

Şekil 5.9: C&RT algoritması değişken önem sıralaması

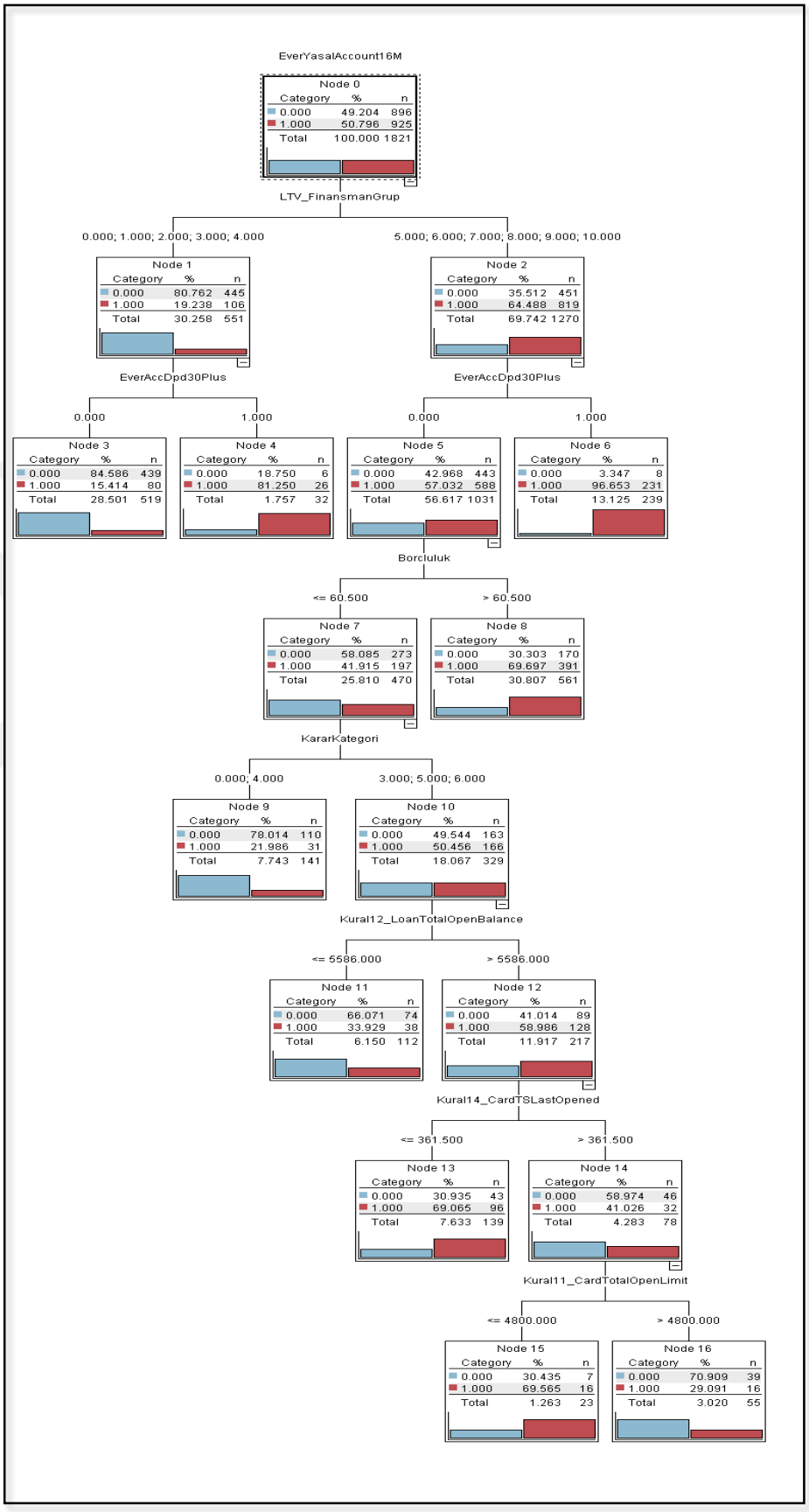


C&R algoritmasına göre müşterinin yasal takibe girme eğilimini belirleyen en önemli faktör yüzde 40'lık önem oranına sahip olan müşterinin krediyi ödemeye başladıktan sonra ilk 3 taksidinde 30 gün ve üstü gecikmeye düşüp düşmediğinin bilgisinin tutulduğu "EverAccDpd30Plus" değişkenidir. En önemli 2.faktör müşterinin alacağı taşıt fiyatının yüzde kaçı için kredi talep ettiğinin bilgisinin tutulduğu "LTV" değişkenidir ve önem oranı yüzde 35'tir. Müşterinin kredi kartı borçluluğunu gösteren "borçluluk" değişkeni yüzde 11'lik önem oranı ile müşterinin yasal takip'e girme eğilimini etkileyen en önemli 3. faktördür. En önemli 4. Faktör ise başvurunun sistemde aldığı karar bilgisinin tutulduğu "Karar Kategori" değişkenidir ve önem oranı yüzde 4'tür. Müşterinin güncel açık kredi borcu bilgisinin tutulduğu "LoanTotalOpenBalance" değişkeni 5. Önemli faktör olarak grafikte yeri almaktadır ve önem oranı yüzde 2'dir. Yüzde 1'den daha düşük önem derecesine sahip diğer değişkenler ise sırasıyla şekil 5.9'daki grafikte gösterildiği gibidir.

C&RT algoritması karar ağacı kural setini yazmak gerekirse aşağıdaki şekilde yazılabilir.

```
LTV_FinansmanGrup in [ 0.000 1.000 2.000 3.000 4.000 ] [ Mode: 0 ] (551)
    EverAccDpd30Plus in [ 0.000 ] [ Mode: 0 ] => 0,0 (519; 0,846)
    EverAccDpd30Plus in [ 1.000 ] [ Mode: 1 ] => 1,0 (32; 0,812)
LTV_FinansmanGrup in [ 5.000 6.000 7.000 8.000 9.000 10.000 ] [ Mode: 1 ] (1.270)
    EverAccDpd30Plus in [ 0.000 ] [ Mode: 1 ] (1.031)
        Borcluluk <= 60,500 [ Mode: 0 ] (470)
            KararKategori in [ 0.000 4.000 ] [ Mode: 0 ] => 0,0 (141; 0,78)
            KararKategori in [ 3.000 5.000 6.000 ] [ Mode: 1 ] (329)
                Kural12_LoanTotalOpenBalance <= 5.586 [ Mode: 0 ] => 0,0 (112; 0,661)
                Kural12_LoanTotalOpenBalance > 5.586 [ Mode: 1 ] (217)
                    Kural14_CardTSLastOpened <= 361,500 [ Mode: 1 ] => 1,0 (139; 0,691)
                    Kural14_CardTSLastOpened > 361,500 [ Mode: 0 ] (78)
                        Kural11_CardTotalOpenLimit <= 4.800 [ Mode: 1 ] => 1,0 (23; ,696)
                        Kural11_CardTotalOpenLimit > 4.800 [ Mode: 0 ] => 0,0 (55; ,709)
        Borcluluk > 60,500 [ Mode: 1 ] => 1,0 (561; 0,697)
    EverAccDpd30Plus in [ 1.000 ] [ Mode: 1 ] => 1,0 (239; 0,967)
```

Şekil 5.10: C&RT algoritması karar ağacı



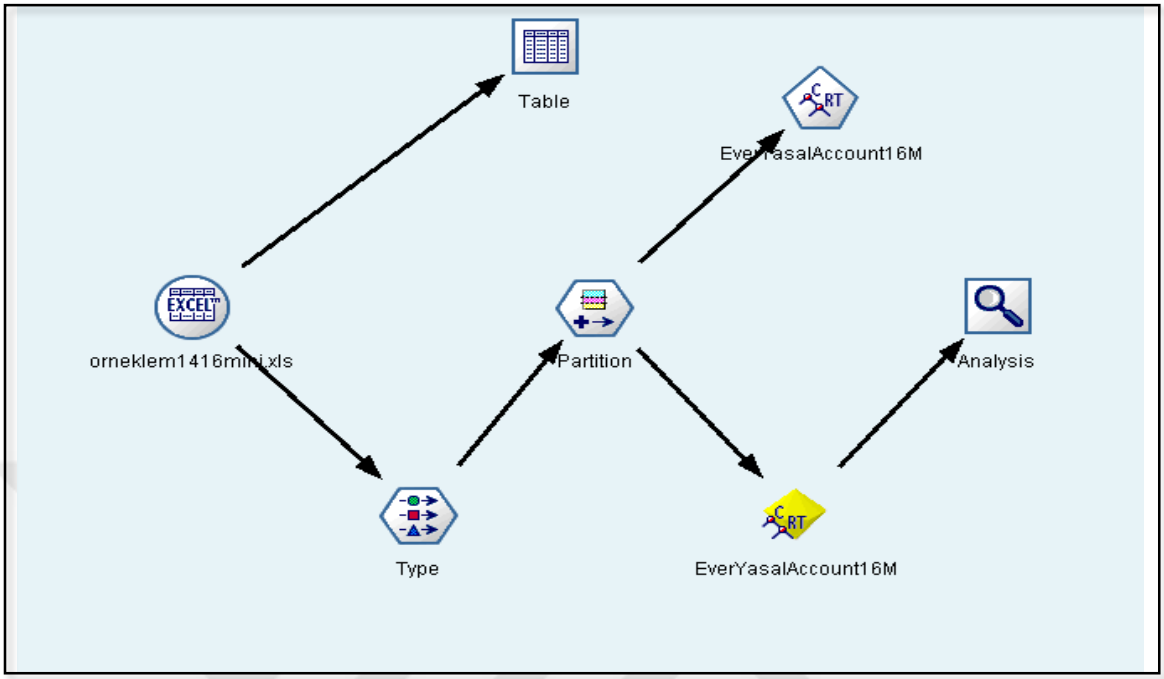
Şekil 5.10'da görüldüğü gibi elde edilen karar ağacı 7 seviye ve 16 alt düğümden oluşmaktadır. Karar ağacı incelendiğinde müşterilerin yasal takip'e girip girmediğini gösteren bağımlı değişken ilk önce "LTV" faktörü üzerinden 2 dala ayrılmıştır. Bu 2 dal daha sonra "EverAccDpd30Plus" faktörü ile alt dallara bölünmüştür.Bu alt dallar ise kendi içlerinde "Borçluluk", "KararKategori", "LoanTotalOpenBalance", "CardTSLastOpened" ve "CardTotalOpenLimit" faktörleri ile sonraki kırılımlarını gerçekleştirmişlerdir.

Bu tez kapsamında yasal takibe girme eğilimi olan müşterilerin özellikleri ortaya konulmaya çalışıldığından yasal takibe girme oranı üzerinde etkisi olan düğümlerin yorumu üzerinde yoğunlaşılmıştır.

En başta tüm müşteriler içerisinde yaklaşık yüzde 51 yasal takip'e gitme oranı var iken bu oran LTV'Sİ yüzde 70'den fazla olan müşterilerin seçildiği düğüm 2'de yüzde 64,488'e çıkmaktadır. Bir alt kırılım olan ilk 3 taksidinde 30 gün ve üzeri gecikmeye düşen müşterilerde ise bu oran yüzde 96,653'e kadar yükselmektedir. İlk 3 taksidinde 30 gün ve üzeri gecikmeye düşmeyen müşterilerde ise borçluluğu yüzde 60,5'ten fazla olan müşterilerin yasal takip'e girme oranları yüzde 69,697'dir.

C&RT modeli Clementine 12.0 programında şekil 5.11'de gösterildiği gibi modellenmiştir.

Şekil 5.11: Clementine 12.0'da C&RT algoritması modellemesi



5.2.4 QUEST (quick, unbiased, efficient statistical tree)

1997 yılında Loh&Shih tarafından geliştirilen bu algoritma ikili karar ağacı yapısı kullanan bir sınıflama algoritmasıdır. Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki ANOVA F testi, Levene's Test ya da Pearson's χ^2 testi ile hesaplanır.kayıp veriler için vekil değişken oluşturulur.en yüksek bağlantılı değişken için sınıf oluşturulur ve bu şekilde büyüme devam eder.ağacı budama için on katlı çapraz doğrulama (ten fold cross validation) adı verilen bir metod kullanılır.ağacın boyutunu küçültmek için ise otomatik maliyet-karmaşıklık budama (automatic cost-complexity prunning) tekniği kullanılabilir.

5.2.4.1 QUEST algoritması uygulaması ve sonuçları

Quest karar ağacı modelinin performans ölçümleri tablo5.10'da gösterildiği gibidir. 2832 adet kayıttan oluşan örneklemin rastgele belirlenmiş olan yüzde 66'sı eğitim kümesi olarak kullanılmış ve kurallar oluşturulmuştur.Verinin geri kalanı ise bu kuralların test edildiği test kümesi olarak kullanılmıştır. Doğru sınıflama oranı eğitim kümesinde yüzde 76,88 iken test kümesinde yüzde 74,58'e gerilemiştir.

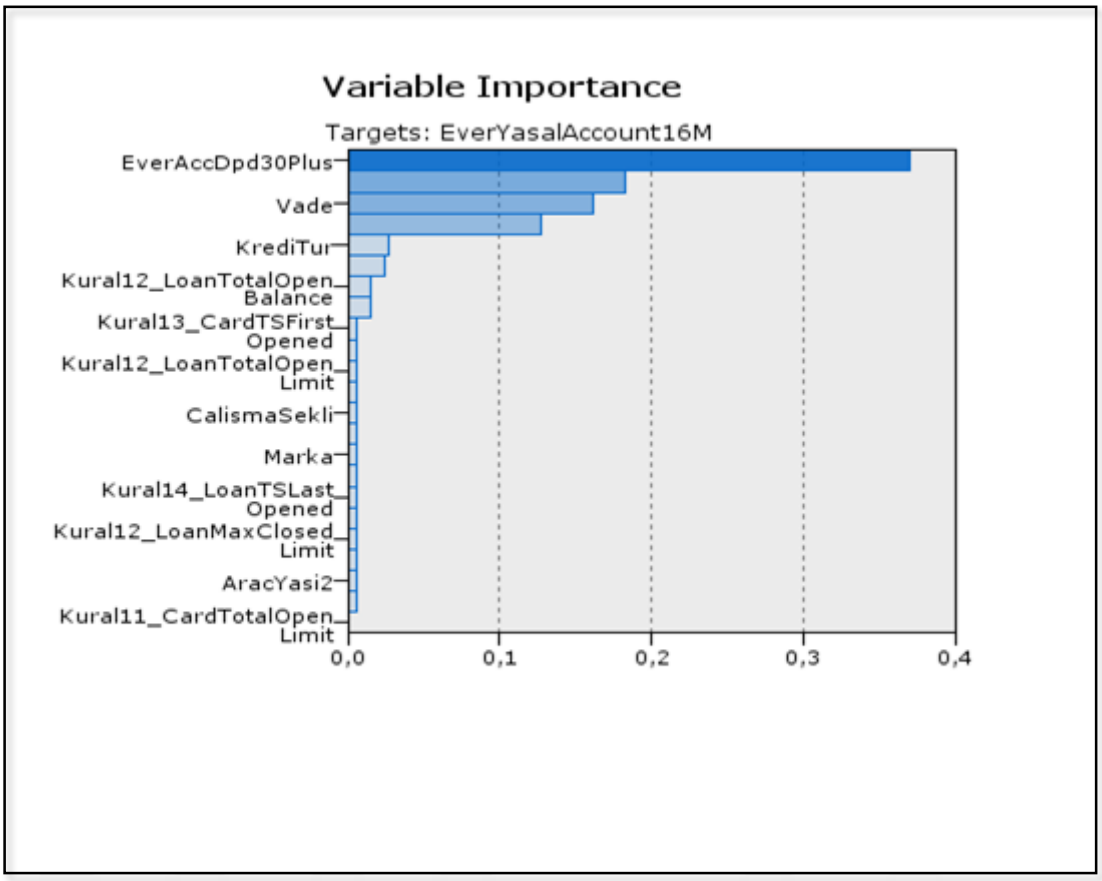
Tablo 5.10: QUEST algoritması denetimli sınıflandırma özeti

Sınıflandırma	1_Eğitim		2_Test	
Doğru	1400	76,88%	754	74,58%
Yanlış	421	23,12%	257	25,42%
Toplam	1821		1011	

*Kaynak:*Clementine 12.0

Müşterilerin ilk 16 ayda yasal takip'e girme eğilimlerine etki eden faktörler ve önem sıraları şekil 5.12'deki grafikte gösterildiği gibidir.

Şekil 5.12: QUEST algoritması değişken önem sıralama



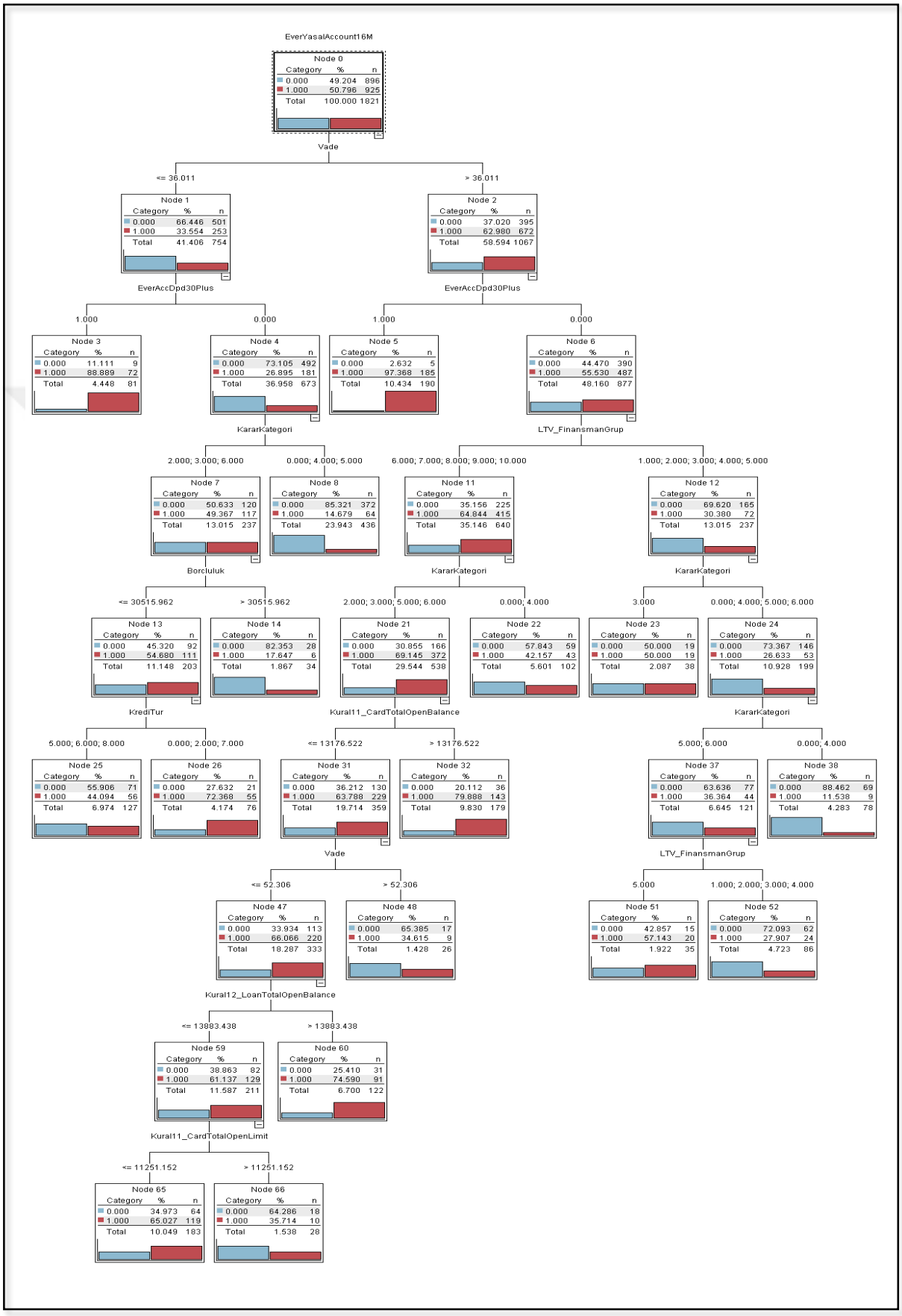
Quest algoritmasına göre müşterinin yasal takiBE girme eğilimini belirleyen en önemli faktör yüzde 37'lik önem oranına sahip olan müşterinin krediyi ödemeye başladıktan sonra ilk 3 taksidinde 30 gün ve üstü gecikmeye düşüp düşmediğinin bilgisinin tutulduğu "EverAccDpd30Plus" değişkenidir. En önemli 2. Faktör ise başvurunun sistemde aldığı karar bilgisinin tutulduğu "Karar Kategori" değişkenidir ve önem oranı yüzde 18'dir. yüzde 16'lık önem oranı ile "vade" değişkeni en önemli 3. Faktör olurken müşterinin alacağı taşıt fiyatının yüzde kaçı için kredi talep ettiğinin bilgisinin tutulduğu "LTV" değişkeni en önemli 4. faktördür ve önem oranı yüzde 14'tür. Yüzde 4'lük önem oranı ile "KrediTur" değişkeni 5. Sırada yer alırken müşterinin güncel toplam açık kart borcunu gösteren "CardTotalOpenBalance" değişkeni 6. Sırada yer almaktadır ve önem oranı yüzde 3,8'dir. Müşterinin güncel toplam açık kredi borcunu gösteren "LoanTotalOpenBalance" değişkeni ise yüzde 3,4'lük önem oranı ile 7. sırada yer almaktadır. Müşterinin kredi kartı borçluluğunu gösteren "borçluluk" değişkeni ise yüzde 3,3'lük önem oranı ile müşterinin yasal takip'e girme eğilimini etkileyen 8.

faktördür. Yüzde 1'den daha düşük önem derecesine sahip diğer değişkenler ise sırasıyla şekil 5.12'de gösterildiği gibidir.

QUEST algoritması karar ağacı kural setini yazmak gerekirse aşağıdaki şekilde yazılabilir.

```
Vade <= 36,011 [ Mode: 0 ] (754)
  EverAccDpd30Plus = 1 [ Mode: 1 ] => 1,0 (81; 0,889)
    EverAccDpd30Plus = 0 [ Mode: 0 ] (673)
          KararKategori = 2 or KararKategori = 3 or KararKategori = 6 [ Mode: 0 ] (237)
              Borcluluk <= 30515,962 [ Mode: 1 ] (203)
                 KrediTur = 5 or KrediTur = 6 or KrediTur = 8 [ Mode: 0 ] => 0,0 (127; 0,559)
                 KrediTur = 0 or KrediTur = 2 or KrediTur = 7 [ Mode: 1 ] => 1,0 (76; 0,724)
              Borcluluk > 30515,962 [ Mode: 0 ] => 0,0 (34; 0,824)
        KararKategori = 0 or KararKategori = 4 or KararKategori = 5 [ Mode: 0 ] => 0,0 (436; 0,853)
Vade > 36,011 [ Mode: 1 ] (1.067)
  EverAccDpd30Plus = 1 [ Mode: 1 ] => 1,0 (190; 0,974)
    EverAccDpd30Plus = 0 [ Mode: 1 ] (877)
                                          LTV_FinansmanGrup = 6 or LTV_FinansmanGrup = 7 or
                                          LTV_FinansmanGrup = 8 or LTV_FinansmanGrup = 9 or
                                                  LTV_FinansmanGrup = 10 [ Mode: 1 ] (640)
            KararKategori = 2 or KararKategori = 3 or KararKategori = 5 or KararKategori = 6 [ Mode: 1 ] (538)
                Kural11_CardTotalOpenBalance <= 13176,522 [ Mode: 1 ] (359)
                    Vade <= 52,306 [ Mode: 1 ] (333)
                       Kural12_LoanTotalOpenBalance <= 13883,438 [ Mode: 1 ] (211)
                          Kural11_CardTotalOpenLimit <= 11251,152 [ Mode: 1 ] => 1,0 (183; ,65
                          Kural11_CardTotalOpenLimit > 11251,152 [ Mode: 0 ] => 0,0 (28; 0,643
                       Kural12_LoanTotalOpenBalance > 13883,438 [ Mode: 1 ] => 1,0 (122; 0,746)
                    Vade > 52,306 [ Mode: 0 ] => 0,0 (26; 0,654)
                Kural11_CardTotalOpenBalance > 13176,522 [ Mode: 1 ] => 1,0 (179; 0,799)
            KararKategori = 0 or KararKategori = 4 [ Mode: 0 ] => 0,0 (102; 0,578)
                LTV_FinansmanGrup = 1 or LTV_FinansmanGrup = 2 or LTV_FinansmanGrup = 3 or LTV_FinansmanGrup = 4 or
                                                  LTV_FinansmanGrup = 5 [ Mode: 0 ] (237)
             KararKategori = 3 [ Mode: 1 ] => 0,0 (38; 0,5)
             KararKategori = 0 or KararKategori = 4 or KararKategori = 5 or KararKategori = 6 [ Mode: 0 ] (199)
                KararKategori = 5 or KararKategori = 6 [ Mode: 0 ] (121)
                       LTV_FinansmanGrup = 5 [ Mode: 1 ] => 1,0 (35; 0,571)
                                LTV_FinansmanGrup = 1 or LTV_FinansmanGrup = 2 or LTV_FinansmanGrup = 3 or
                                          LTV_FinansmanGrup = 4 [ Mode: 0 ] => 0,0 (86; 0,721)
                KararKategori = 0 or KararKategori = 4 [ Mode: 0 ] => 0,0 (78; ,885)
```

Şekil 5.13: QUEST algoritması karar ağacı



Şekil 5.13'te görüldüğü gibi elde edilen karar ağacı 8 seviye ve 30 alt düğümden oluşmaktadır. Karar ağacı incelendiğinde müşterilerin yasal takip'e girip girmediğini gösteren bağımlı değişken ilk önce "vade" faktörü üzerinden 2 dala ayrılmıştır. Bu 2 dal daha sonra "EverAccDpd30Plus" faktörü ile alt dallara bölünmüştür.Bu alt dallar ise kendi içlerinde "kararKategori", "LTV", "Borçluluk", "KrediTür", "CardTotalOpenBalance", "LoanTotalOpenBalance" ve "CardTotalOpenLimit" faktörleri ile sonraki kırılımlarını gerçekleştirmişlerdir.

Bu tez kapsamında yasal takibe girme eğilimi olan müşterilerin özellikleri ortaya konulmaya çalışıldığından yasal takip'e girme oranı üzerinde etkisi olan düğümlerin yorumu üzerinde yoğunlaşılmıştır.

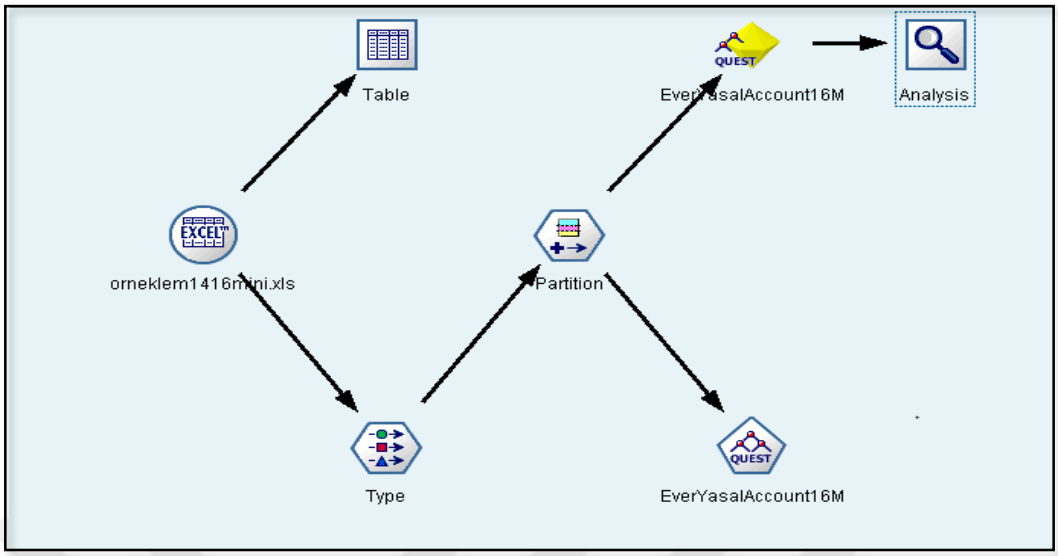
En başta tüm müşteriler içerisinde yaklaşık yüzde 51 yasal takip'e gitme oranı var iken bu oran vadesi 36 aydan fazla olan müşterilerin seçildiği düğüm 2'de yüzde 62,980'e çıkmaktadır. Bir alt kırılım olan ilk 3 taksidinde 30 gün ve üzeri gecikmeye düşen müşterilerde ise bu oran yüzde 97,368'e kadar yükselmektedir.

İlk 3 taksidinde 30 gün ve üzeri gecikmeye düşmeyen müşterilerde ise LTV'si yüzde 71-95 aralığında olan müşterilerde en başta yüzde 51 olan oran yüzde 64,844 seviyelerine yükselmektedir. Bu müşterilerden karar kategorisi oto red,oto red istisnası,oto onay istisnası ve gri alan olan ve toplam açık kredi kartı borcu 13.177 TL'den fazla olan müşterilerin yasal takip'e girme oranı yüzde 79,888'dir.

Karar kategorisi oto red,oto red istisnası,oto onay istisnası ve gri alan olan , toplam açık kredi kartı borcu 13.177 TL'ye eşit ve daha az olan ,vadesi 52 aya eşit ve daha az olan ve toplam açık kredi borcu 13.883 TL'den fazla olan müşterilerin ise yasal takip'e girme oranı yüzde 74,590'dır.

QUEST modeli Clementine 12.0 programında aşağıdaki şekil 5.14'te gösterildiği gibi modellenmiştir.

Şekil 5.14: Clementine 12.0'da QUEST algoritması modellemesi



5.2.5 Karar Listesi

Tablo 5.11'de karar listesi algoritmasından çıkan kural setleri görülmektedir.

Tablo 5.11: Karar listesi sınıflandırma özeti

no	Segment	Skor	Kapsam(n)	Sıklık	Olasılık
Kalan Dahil Tüm Segmentler			**1821**	**925**	**50,80%**
1	**KararKategori,Kural14_CardTSLastOpened** KararKategori=3 and Kural14_CardTSLastOpened > 0 and Kural14_CardTSLastOpened <= 144	1.0	63	55	87,30%
2	**LTV_FinansmanGrup,EverAccDpd30Plus** LTV_FinansmanGrup=7 and EverAccDpd30Plus=1	1.0	88	86	97,73%
3	**Kural12_LoanTotalOpenBalance,KararKategori** Kural12_LoanTotalOpenBalance > 26.971 and KararKategori=3	1.0	63	53	84,13%
4	**KararKategori,EverAccDpd30Plus** KararKategori=6 and EverAccDpd30Plus=1	1.0	54	50	92,59%
5	**EverAccDpd30Plus** EverAccDpd30Plus=1	1.0	106	98	92,45%
6	**KararKategori,Kural11_CardTotalOpenBalance** KararKategori=6 and	1.0	67	53	79,10%

	Kural11_CardTotalOpenBalance> 30.955				
7	**KararKategori,Kural11_CardTotalOpenBalance** KararKategori=6 and Kural11_CardTotalOpenBalance > 16.040 and Kural11_CardTotalOpenBalance <= 30.955	1.0	55	43	78,18%
8	**Borcluluk** Borcluluk > 92 and Borcluluk <= 100	1.0	115	70	60,87%
9	**Kural12_LoanTotalOpenBalance** Kural12_LoanTotalOpenBalance > 44.800	1.0	68	41	60,29%
10	**KararKategori,LTV_FinansmanGrup** KararKategori=6 and LTV_FinansmanGrup=7	1.0	69	50	72,46%
Kalan			**1073**	**326**	**30,38%**

*Kaynak:*Clementine 12.0

Bu tez kapsamında uygulanan karar listesi çalışmasında toplamda 2832 adet kayıt içeren örneklemden rastgele yüzde 66'sı eğitim grubu olarak seçilmiştir. Ve bu seçilen 1821 kayıt ile karar listesi oluşturulmuştur.

Tablo 5.11'de ilk satırda 1921 kaydın 925 tanesinin ilk 16 ayında yasal takip'e girdiği görülmektedir. En son satırda ise oluşturulan 10 adet kurallar bütünü ile bu 1921 kaydın 848 tanesine ulaşıldığı 1073 tanesine ulaşılamadığı, 925 kaydın da 599 tanesine ulaşıldığı 326 tanesine ulaşılamadığı gösterilmektedir.

Karar listesini yorumlanacak olursa;

i. Karar Kategorisi 3 olan ve CardTSLastOpened değişkeni 0 günden büyük 144 günden küçük olan 63 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 87,30'dur ve bu da 55 müşteriye karşılık gelmektedir.
ii. LTV'si 7 olan ve EverAccDpD30Plus değişkeni 1 olan 88 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 97,73'tür ve bu da 86 müşteriye karşılık gelmektedir.
iii. LoanTotalOpenBalance'ı 26,971 TL'den büyük olan ve karar kategorisi 3 olan 63 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 84,13'tür ve bu da 53 müşteriye karşılık gelmektedir.
iv. Karar Kategorisi 6 olan ve EverAccDpD30Plus değişkeni 1 olan 54 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 92,59'dur ve bu da 50 müşteriye karşılık gelmektedir.
v. EverAccDpD30Plus değişkeni 1 olan 106 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 92,45'tir ve bu da 98 müşteriye karşılık gelmektedir.

vi. Karar Kategorisi 6 olan ve CardTotalOpenBalance'ı 30.955 TL'den büyük olan 67 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 79,10'dur ve bu da 53 müşteriye karşılık gelmektedir.

vii. Karar Kategorisi 6 olan ve CardTotalOpenBalance'ı 16.040 TL'den büyük 30.955 TL'den küçük olan 55 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 78,18'dir ve bu da 43 müşteriye karşılık gelmektedir.

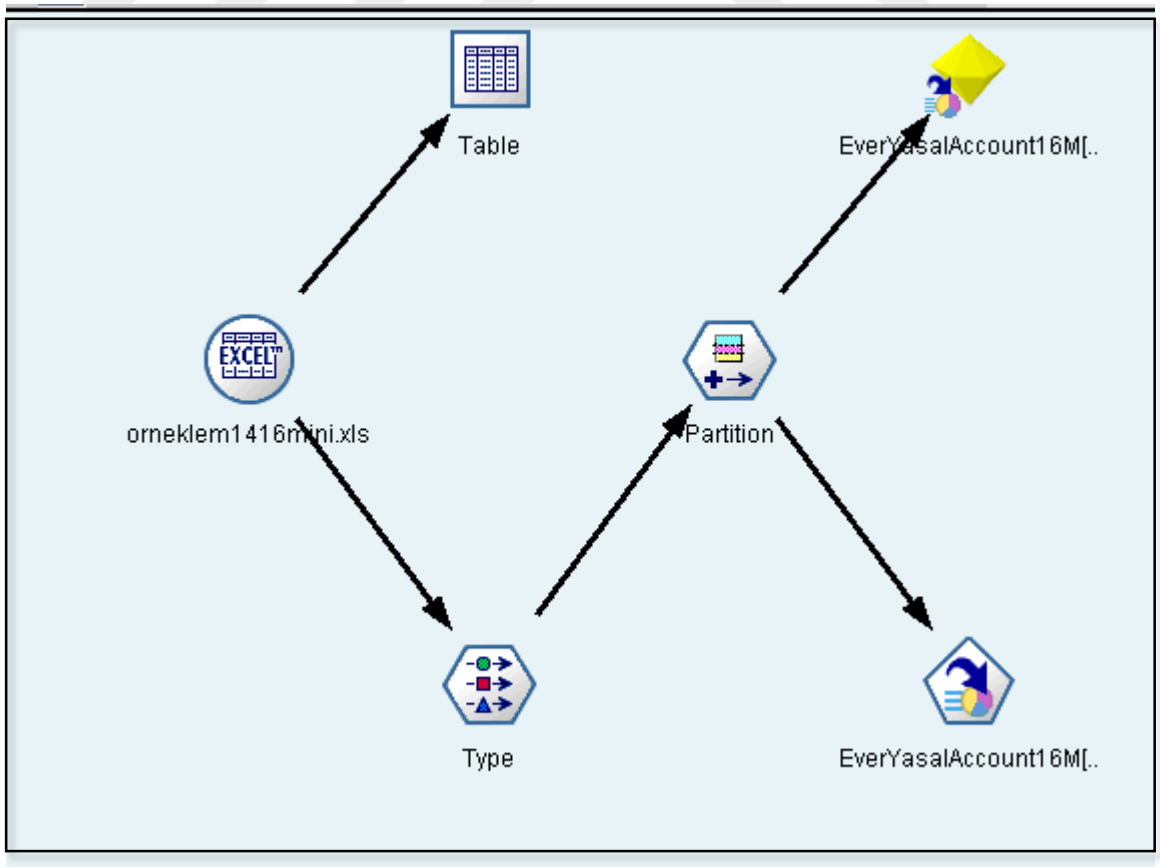
viii. Borçluluğu yüzde 92 ile yüzde 100 arasında olan 115 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 60,87'dir ve bu da 70 müşteriye karşılık gelmektedir.

ix. Loan Total Open Balance'ı 44.800 TL'nin üzerinde olan olan 68 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 60,29'dur ve bu da 41 müşteriye karşılık gelmektedir.

x. Karar Kategorisi 6 olan ve LTV'si 7 olan 69 müşterinin ilk 16 ayında yasal takip'e girme olasılığı yüzde 72,46'dır ve bu da 50 müşteriye karşılık gelmektedir.

Karar Listesi Clementine 12.0 programında şekil 5.15'te gösterildiği gibi modellenmiştir.

Şekil 5.15: Clementine 12.0'da karar listesi algoritması modellemesi



5.3 BAYES TABANLI AĞLAR

1985 yılında Judea Pearl tarafından geliştirilen bu algoritma bayes teoremini baz almaktadır.Bu algoritmada düğümler rastlantı değişkenlerini düğümler arasındaki bağlar ise rastlantı değişkenleri arasındaki olasılıksal bağımlılık durumlarını gösterir.Bu bağımlılıklar istatistiksel yöntemlerden yararlanılarak hesaplanmaktadır.değişkenler arasındaki ağların yönlendirilmemiş olduğu modeller Markov ağları , yönlendirilmiş olduğu modeller ise bayesci ağlar olarak adlandırılmakatadır.Bayesci ağlarda değişkenler arasında koşullu bağımsızlık özelliği geçerlidir. Koşullu bağımsızlık bir değişkenin bağlı olduğu üst değişkenlerinin durumları bilindiğinde üst değişkenin bağlı olmadığı diğer değişkenlerden bağımsız olması anlamına gelmektedir.bayesci ağ algoritmasında amaç doğruluğu en yüksek modeli elde etmektir.bayesci ağ modelinin bağımlılık yapısı ve koşullu olasılık fonksiyonları veriden öğrenilerek elde edilir.bu öğrenmenin iki farklı çeşidi vardır. Birincisi modelde yer alan bağlantıları belirleyen yapısal öğrenme ikincisi ise koşullu olasılık fonksiyonlarının parametrik yapısını öğrenme yani parametre öğrenimidir.

5.3.1 Naive Bayesian Algoritması

Bu algoritma adını 1701-1761 yılları arasında yaşamış ingiliz matematikçi thomas bayes'ten alır. Bu yöntem pratik ve en düşük hata oranına sahip olması özellikleri ile sınıflandırma yöntemleri arasında en sık kullanılanıdır. Eldeki sınıflardan hangisine ait olduğu bilinmeyen bir veri örneğinin mevcut sınıflardan hangisine ait olma olasılığı en yüksek ise veri o sınıfa atanır. Burada bilinmeyen her veri birbirinden bağımsızdır ve hepsi aynı derecede önemlidir.bir veri başka bir veri hakkında bilgi içermez.

5.4 DİĞER SINIFLANDIRMA YÖNTEMLERİ

Tezin bu bölümünde WEKA ile yapılmış olan sınıflandırma sonuçları yer almaktadır.WEKA "waikato Environment for knowledge analysis" kelimelerinin baş

harflerinden oluşan ve waikato üniversitesinde java dilinde geliştirilmiş olan, makine öğrenimi algoritmalarını içeren yazılımdır.

WEKA'da temel olaraka yapılan 3 veri madenciliği işlemi vardır.

i. Sınıflandırma
ii. Kümeleme
iii. Birliktelik

Bunların yanında ;

i. Veri ön işleme
ii. Görselleme gibi veri kümeleri üzerinde ön ve son işlemler de yapılabilmektedir.

WEKA'da veri madenciliği işlemi yapmak için veriler kendisine özgü geliştirilmiş olan ARFF (Attribute Relationship file format) dosya yapısı ile programa entegre edilmelidir. ARFF bir metin dosyası olup dosyanın ilk satırında dosyadaki ilişki tipi (relation) , İkinci satırından itibaren veri kümesindeki değişkenler (attributes) tutulmaktadır.Değişkenlerin bitiminde ise veri setindeki her bir kayıt satır satır yazılmaktadır.Her bir kaydın her bir değişkene karşılık gelen değerleri ise virgülle ayrılmaktadır.

5.4.1 Model Başarı Ölçütleri

Model başarı ölçütler veriyi sınıflamada kullanılan modelin veriyi ne kadar doğru sınıflandırdığını ölçmede kullanılan ölçütlerdir.

5.4.1.1 Verinin iki alt örnekleme ayrılması (Hold out yöntemi)

Bu yöntem sınıflandırılacak verinin rastgele olarak en az iki alt kümeye ayrılmasını baz alır. Modelin kurulduğu küme eğitim(training set) kümesi, kurulan modelin sonuçlarının test edildiği küme ise test (testing set)kümesidir. Modelin doğruluk değeri (accuracy) test kümesi ile belirlenmektedir. Bir modelin doğruluğunun belirlenmesindeki en temel yöntem basit geçerlilik'tir.(simple validation)basit geçerlilik yönteminde verilerin yüzde 5 ile yüzde 34'ü arasındak bir kısım test için ayrılır ve kalan kısmı üzerinde

algoritmanın geliştirilmesi sağlanır.Bu çalışmada eğitim kümesi tüm verilerin yüzde 66'sını kapsamaktadır, test kümesi ise yüzde 34'ünü kapsamaktadır.

Hata Oranı: Yanlış olarak sınıflanan gözlem sayısının toplam gözlem sayısına oranıdır.

Doğruluk Oranı: doğru olarak sınıflanan gözlem sayısının toplam gözlem sayısına oranıdır.

5.4.1.2 Genel doğruluk (overall accuracy)

Bağımlı değişkenin iki kategorili olduğu durumda sınıflama tablosu tablo 5.12'deki gibi gösterilmektedir.

Tablo 5.12: Hata Matrisi

		Gerçek durum	
		1	**0**
Tahmin Edilen	**1**	Doğru Pozitif / True Positive (TP)	Yanlış Pozitif/False Positive (FP)
	0	Yanlış Negatif / False Negative (FN)	Doğru Negatif / True Negative (TN)

Bu durumda $N \Rightarrow Örneklem\ Büyüklüğü$ olmak üzere aşağıdaki eşitlik kullanılarak hesaplanmaktadır.

$$Genel\ Doğruluk\ (GD) = \frac{TP+TN}{N} \quad (5.22)$$

5.4.1.3 Dengeli doğruluk (balanced accuracy)

Bağımlı değişken kategorilerinin dengesiz olduğu durumlarda önerilen bir model başarı ölçütüdür. Aşağıdaki eşitlik kullanılarak hesaplanmaktadır.

$$Dengeli\ Doğruluk\ (DD) = \frac{{}^{TP}/_{(TP+FN)} + {}^{TN}/_{(FP+TN)}}{2} \quad (5.23)$$

5.4.1.4 Duyarlılık (sensitivity),recall,hit rate,true positive rate

Gerçek pozitifler arasında, modelin pozitifleri doğru bir şekilde sınıflama başarısıdır. Aşağıdaki eşitlik kullanılarak hesaplanmaktadır.

$$Duyarlılık(DUY) = \frac{TP}{TP+FN} \tag{5.24}$$

5.4.1.5 Seçicilik (specifity), true negative rate

Gerçek negatifler arasında, modelin negatifleri doğru bir şekilde sınıflama başarısıdır. Aşağıdaki eşitlik kullanılarak hesaplanmaktadır.

$$Seçicilik(SEC) = \frac{TN}{TN+FP} \tag{5.25}$$

5.4.1.6 Hassasiyet, precision, positive predictive value

$$Hassasiyet = \frac{TP}{TP+FP} \tag{5.26}$$

5.4.1.7 Negatif tahmin değeri, negative predictive value

$$Negatif\ Tahmin\ Değeri = \frac{TN}{TN+FN} \tag{5.27}$$

5.4.1.8 Yanlış pozitif oranı, false positive rate

$$Yanlış\ Pozitif\ Oranı = \frac{FP}{TN+FP} \tag{5.28}$$

5.4.1.9 Yanlış negatif oranı, false negative rate, miss rate

$$Yanlış\ Pozitif\ Oranı = \frac{FN}{FN+TP} \tag{5.29}$$

5.4.1.10 Matthews korelasyon katsayısı (matthews correlation coefficient)

Bağımlı değişkenin 2 kategorili olduğu durumlarda kullanılan bir model başarı ölçütüdür. Bağımlı değişkene ait prevelans değerinin dengesiz olduğu durumlarda diğer başarı ölçütlerine göre daha doğru sonuç vermesi avantajlı yönüdür. ±1 aralığında değer almaktadır. +1 değeri en iyi durumu, 0 değeri rastgele bir tahmin yapıldığını,-1 değeri ise en kötü durumu yani yanlış tahmin yapıldığını göstermektedir. Phi katsayısı olarakta bilinen bu katsayı aşağıdaki eşitlik kullanılarak hesaplanmaktadır.

$$MKK = \frac{(TP \times TN)+(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5.30)$$

5.4.1.11 F-ölçütü (F-measure)

Duyarlılık ve seçicilik ölçütlerinin tek başlarına yeterli olmadıkları düşünülerek elde edilmiştir.Doğru negatif hücresini dikkate almaz ve aşağıdaki eşitlik kullanılarak hesaplanmaktadır.

$$F = \frac{2TP}{2TP+FP+FN} \quad (5.31)$$

5.4.1.12 ROC eğrisi (receiver operating characteric)

ROC eğrisi modeller arasında güvenilir bir karşılaştırma yapmaya olanak sağlar.İkili sınıflandırma algoritmalarında hassasiyetin kesinliğe oranı ile ortaya çıkmaktadır. Başka bir ifadeyle doğru pozitiflerin, yanlış pozitiflere olan oranı ROC eğrisi ile ifade edilmektedir. Kesinlik ve hassasiyet arasındaki dengeyi değerlendirmek için kullanılan ROC eğrisinin altında kalan alan ROC puanı olarak tanımlanır. ROC puanı 1'e ne kadar yakın ise pozitiflerin negatiflerden o kadar mükemmel bir şekilde ayrıldığı anlamına gelir, ROC puanı 0 iken ise herhangi bir pozitifin bulunamadığı sonucu çıkmaktadır.Eğri altındaki alanın yorumlanmasında tablo 5.13'te gösterilen derecelendirme kullanılabilir.

Tablo 5.13: ROC Eğrisi Derecelendirme bilgileri

ROC Puanı	Derecelendirme
0.90-1.00	Mükemmel
0.80-0.90	İyi
0.70-0.80	Orta
0.60-0.70	Zayıf
0.50-0.60	Başarısız

5.4.2 WEKA'da Uygulanan Algoritmalar ve Sonuçları

8 ana sınıflandırma yöntemi altında toplam 49 adet algoritma'ya ait sınıflandırma başarı ölçütleri en yüksek doğru sınıflama oranına sahip olandan en düşük doğru sınıflama sahip oranına sahip olana doğru sıralanarak tablo 5.13'te verilmiştir.

Tablo 5.14: WEKA'da uygulanan algoritmalar ve performans özet bilgileri

	Correctly Classified Instances	Incorrectly Classified Instances	Root mean squared error	TP Rate	FP Rate	Precision	Recall	F Measure	ROC Area	Classifier
Random SubSpace	78%	22%	0,402	0,78	0,22	0,78	0,78	0,78	0,85	Meta
Threshold Selector	77%	23%	0,3906	0,77	0,23	0,77	0,77	0,77	0,86	Meta
Classif. Via Regression	77%	23%	0,3979	0,77	0,23	0,77	0,77	0,77	0,86	Meta
Logistic	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	Functions
MultiClass Classifier	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	Meta
LAD Tree	77%	23%	0,4043	0,77	0,23	0,77	0,77	0,77	0,84	Trees
Rotation Forest	76%	24%	0,4006	0,76	0,24	0,77	0,76	0,76	0,84	Meta
ADTree	76%	24%	0,4118	0,76	0,24	0,76	0,76	0,76	0,84	Trees
Random Forest	76%	24%	0,39	0,76	0,24	0,77	0,76	0,76	0,86	Trees

Decorate	76%	24%	0,4041	0,76	0,24	0,76	0,76	0,76	0,84	Meta
SPegasos	76%	24%	0,4887	0,76	0,24	0,76	0,76	0,76	0,76	Functions
LMT	76%	24%	0,3949	0,76	0,24	0,76	0,76	0,76	0,85	Trees
Bagging	76%	24%	0,4039	0,76	0,24	0,76	0,76	0,76	0,84	Meta
LogitBoost	76%	24%	0,4062	0,76	0,24	0,76	0,76	0,76	0,84	Meta
Dagging	75%	25%	0,4285	0,75	0,25	0,75	0,75	0,75	0,83	Meta
SMO	75%	25%	0,5013	0,75	0,25	0,75	0,75	0,75	0,75	Functions
AdaBoost M1	75%	25%	0,4156	0,75	0,25	0,75	0,75	0,75	0,83	Meta
Random Committee	75%	25%	0,4107	0,75	0,25	0,75	0,75	0,75	0,83	Meta
DTNB	74%	26%	0,4167	0,75	0,26	0,75	0,75	0,74	0,83	Rules
Filtered Classifier	74%	26%	0,4499	0,74	0,26	0,74	0,74	0,74	0,78	Meta
Attribute Selected Classifier	74%	26%	0,4483	0,74	0,26	0,74	0,74	0,74	0,78	Meta
J48graft	74%	26%	0,4848	0,74	0,26	0,74	0,74	0,74	0,72	Trees
NBTree	74%	26%	0,4487	0,74	0,27	0,74	0,74	0,74	0,82	Trees
REPTree	74%	26%	0,4367	0,74	0,27	0,74	0,74	0,74	0,79	Trees
SimpleCart	74%	26%	0,4304	0,74	0,27	0,74	0,74	0,74	0,79	Trees
Decision Table	74%	26%	0,4174	0,74	0,27	0,74	0,74	0,74	0,82	Rules
FT	73%	27%	0,4777	0,73	0,27	0,73	0,73	0,73	0,77	Trees
END	73%	27%	0,4876	0,73	0,27	0,73	0,73	0,73	0,72	Meta
Ordinal Class Classifier	73%	27%	0,4876	0,73	0,27	0,73	0,73	0,73	0,72	Meta
J48	73%	27%	0,4876	0,73	0,27	0,73	0,73	0,73	0,72	Trees
Multilayer Perceptron	73%	27%	0,4629	0,73	0,27	0,73	0,73	0,73	0,80	Functions
Ridor	72%	28%	0,5266	0,72	0,29	0,74	0,72	0,72	0,72	Rules
Naive Bayes	72%	28%	0,4721	0,72	0,28	0,73	0,72	0,72	0,80	Bayes
JRip	72%	28%	0,4406	0,72	0,28	0,72	0,72	0,72	0,77	Rules
PART	72%	28%	0,49	0,72	0,29	0,72	0,72	0,72	0,75	Rules
BFTree	71%	29%	0,5069	0,71	0,29	0,71	0,71	0,71	0,70	Trees
RBF Network	71%	29%	0,4446	0,71	0,28	0,73	0,71	0,71	0,81	Functions
MultiBoost AB	71%	29%	0,4895	0,71	0,29	0,71	0,71	0,71	0,79	Meta
Conjunctive Rule	70%	30%	0,4528	0,70	0,32	0,74	0,70	0,68	0,69	Rules
LWL	69%	31%	0,4444	0,69	0,32	0,72	0,69	0,68	0,78	Lazy

Raced Incremental LogitBoost	69%	31%	0,4605	0,69	0,32	0,70	0,69	0,69	0,69	Meta
OneR	69%	31%	0,5553	0,69	0,32	0,70	0,69	0,69	0,69	Rules
Decision Stump	69%	31%	0,4616	0,69	0,32	0,70	0,69	0,69	0,69	Trees
Voted Perceptron	69%	31%	0,5606	0,69	0,32	0,69	0,69	0,69	0,70	Functions
IB1	68%	32%	0,5683	0,68	0,32	0,68	0,68	0,68	0,68	Lazy
IBk	68%	32%	0,568	0,68	0,32	0,68	0,68	0,68	0,68	Lazy
Random Tree	67%	33%	0,5746	0,67	0,33	0,67	0,67	0,67	0,67	Trees
KStar	61%	39%	0,6008	0,61	0,39	0,61	0,61	0,61	0,66	Lazy
VFI	61%	39%	0,4985	0,61	0,41	0,69	0,61	0,56	0,74	Misc

*Kaynak:*WEKA

Her bir ana sınıfın en yüksek doğru sınıflama oranına sahip algoritmaları seçildiğinde ilk 15 algoritmayı tablo 5.15'te gösterildiği şekilde sıralamak mümkündür.

Tablo 5.15: WEKA'da uygulanan doğru sınıflandırma oranı en yüksek 15 algoritma ve performans özet bilgileri

	Correctly Classified Instances	**Incorrectly Classified Instances**	**Root mean squared error**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Classifier**
RandomSub Space	78%	22%	0,402	0,78	0,22	0,78	0,78	0,78	0,85	Meta
Threshold Selector	77%	23%	0,3906	0,77	0,23	0,77	0,77	0,77	0,86	Meta
Classification Via Regression	77%	23%	0,3979	0,77	0,23	0,77	0,77	0,77	0,86	Meta
Logistic	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	Functions
MultiClass Classifier	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	Meta
LADTree	77%	23%	0,4043	0,77	0,23	0,77	0,77	0,77	0,84	Trees
Rotation Forest	76%	24%	0,4006	0,76	0,24	0,77	0,76	0,76	0,84	Meta
ADTree	76%	24%	0,4118	0,76	0,24	0,76	0,76	0,76	0,84	Trees
Random Forest	76%	24%	0,39	0,76	0,24	0,77	0,76	0,76	0,86	Trees
Decorate	76%	24%	0,4041	0,76	0,24	0,76	0,76	0,76	0,84	Meta
SPegasos	76%	24%	0,4887	0,76	0,24	0,76	0,76	0,76	0,76	Functions

DTNB	74%	26%	0,4167	0,75	0,26	0,75	0,75	0,74	0,83	Rules
Naive Bayes	72%	28%	0,4721	0,72	0,28	0,73	0,72	0,72	0,80	Bayes
LWL	69%	31%	0,4444	0,69	0,32	0,72	0,69	0,68	0,78	Lazy
VFI	61%	39%	0,4985	0,61	0,41	0,69	0,61	0,56	0,74	Misc

*Kaynak:*WEKA

Bu 15 algoritmanın bağımlı değişkenin kategorisine göre göstermiş oldukları performans ise tablo 5.16'da gösterilmektedir.

Tablo 5.16: WEKA'da uygulanan doğru sınıflandırma oranı en yüksek 15 algoritma ve bağımlı değişkenin kategorisine göre göstermiş oldukları performans özet bilgileri

	Class	Correctly Classified Instances	Incorrectly Classified Instances	Root mean squared error	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Classifier
RandomSub Space	0				0,76	0,21	0,78	0,76	0,77	0,85	Meta
	1				0,79	0,24	0,78	0,79	0,79	0,85	
	Weighted Avg.	78%	22%	0,402	0,78	0,22	0,78	0,78	0,78	0,85	
Threshold Selector	0				0,73	0,19	0,79	0,73	0,76	0,86	Meta
	1				0,81	0,27	0,76	0,81	0,79	0,86	
	Weighted Avg.	77%	23%	0,3906	0,77	0,23	0,77	0,77	0,77	0,86	
Classification Via Regression	0				0,75	0,21	0,77	0,75	0,76	0,86	Meta
	1				0,79	0,25	0,77	0,79	0,78	0,86	
	Weighted Avg.	77%	23%	0,3979	0,77	0,23	0,77	0,77	0,77	0,86	
Logistic	0				0,76	0,23	0,76	0,76	0,76	0,86	Functions
	1				0,77	0,24	0,77	0,77	0,77	0,86	
	Weighted Avg.	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	
MultiClass Classifier	0				0,76	0,23	0,76	0,76	0,76	0,86	Meta
	1				0,77	0,24	0,77	0,77	0,77	0,86	
	Weighted Avg.	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	
LADTree	0				0,78	0,25	0,75	0,78	0,76	0,84	Trees
	1				0,75	0,22	0,78	0,75	0,77	0,84	
	Weighted Avg.	77%	23%	0,4043	0,77	0,23	0,77	0,77	0,77	0,84	
Rotation	0				0,72	0,19	0,78	0,72	0,75	0,84	Meta

Forest	1				0,81	0,28	0,75	0,81	0,78	0,84	
	Weighted Avg.	76%	24%	0,4006	0,76	0,24	0,77	0,76	0,76	0,84	
ADTree	0				0,76	0,23	0,76	0,76	0,76	0,84	Trees
	1				0,77	0,24	0,77	0,77	0,77	0,84	
	Weighted Avg.	76%	24%	0,4118	0,76	0,24	0,76	0,76	0,76	0,84	
Random Forest	0				0,72	0,20	0,78	0,72	0,75	0,86	Trees
	1				0,80	0,28	0,75	0,80	0,78	0,86	
	Weighted Avg.	76%	24%	0,39	0,76	0,24	0,77	0,76	0,76	0,86	
Decorate	0				0,74	0,22	0,76	0,74	0,75	0,84	Meta
	1				0,78	0,26	0,76	0,78	0,77	0,84	
	Weighted Avg.	76%	24%	0,4041	0,76	0,24	0,76	0,76	0,76	0,84	
SPegasos	0				0,72	0,20	0,77	0,72	0,75	0,76	Functions
	1				0,80	0,28	0,75	0,80	0,78	0,76	
	Weighted Avg.	76%	24%	0,4887	0,76	0,24	0,76	0,76	0,76	0,76	
DTNB	0				0,69	0,20	0,76	0,69	0,72	0,83	Rules
	1				0,80	0,31	0,73	0,80	0,76	0,83	
	Weighted Avg.	74%	26%	0,4167	0,75	0,26	0,75	0,75	0,74	0,83	
Naive Bayes	0				0,82	0,37	0,68	0,82	0,74	0,80	Bayes
	1				0,63	0,18	0,78	0,63	0,70	0,80	
	Weighted Avg.	72%	28%	0,4721	0,72	0,28	0,73	0,72	0,72	0,80	
LWL	0				0,50	0,13	0,79	0,50	0,61	0,78	Lazy
	1				0,87	0,50	0,65	0,87	0,74	0,78	
	Weighted Avg.	69%	31%	0,4444	0,69	0,32	0,72	0,69	0,68	0,78	
VFI	0				0,26	0,06	0,82	0,26	0,40	0,74	Misc
	1				0,94	0,74	0,58	0,94	0,72	0,74	
	Weighted Avg.	61%	39%	0,4985	0,61	0,41	0,69	0,61	0,56	0,74	

*Kaynak:*WEKA

6.KARŞILAŞTIRMA VE SONUÇ

Bu tez kapsamında, günümüzde anlamlı ve nitelikle bilgiye ulaşmanın en temel aracı olarak kabul edilen çeşitli veri madenciliği yöntemleri ile bireysel araç kredilerinin yasal takibe girme durumları hakkında tahmin modellerinin oluşturulması amaçlanmıştır. Bu bağlamda özel bir tüketici finansman şirketinin veritabanından 01.01.2010 – 31.12.2013 tarihleri arasına ait tüm başvuru verileri alınmıştır. Tüm başvurulardan bireysel ve kredileşmiş başvurular seçilerek alınan data tez konusuna uygun olacak şekilde indirgenmiştir. İndirgenen datadan eksik, hatalı, kayıp ve aykırı değerlerin kontrolü yapılıp data temizlenerek kullanıma uygun hale getirilmiştir. Uygulanan her bir model öncesi veri seti tekrar gözden geçirilip gerektiği durumlarda ihtiyaca uygun olacak şekilde değişken indirgemesi, bütünleştirilmesi ya da veri dönüştürülmesi yapılmıştır.Bu tez kapsamında tahmin edici veri madenciliği modelleri denetimli öğrenme tekniği ile uygulanmıştır. Kullanılan başlıca sınıflandırma yöntemleri ise ikili lojistik regresyon analizi, CHAID,C&RT,QUEST karar ağacı algoritmaları, karar listesi algoritmalarıdır.Datada ilk önce uygulanan lojistik regresyon analizi ile başlangıçta elde bulunan 66 adet değişkenden anlamlı olmayan 40 tanesi çıkartılmıştır. Ve diğer tüm algoritmalar rastgele seçilen ve toplam 2832 adet müşteriye ait 26 adet değişkenden oluşan örneklemin yüzde altmışaltısını temsil eden eğitim kümesi üzerinde geliştirilmiş ve geri kalan test kümesi üzerinde test edilmiştir.

Tablo 6.1: Clementine 12.0'da geliştirilen algoritmalar ve sınıflandırma oranları

	Doğru Sınıflandırma Oranları	
	Eğitim Kümesi	**Test Kümesi**
Lojistik Regresyon	80,07%	78,73%
C&RT	78,09%	74,88%
CHAID	77,21%	72,40%
QUEST	76,88%	74,58%
N	**1821**	**1011**

Eğitim kümesinde de test kümesinde de en yüksek doğru sınıflandırma oranına sahip olan lojistik regresyon analizinden çıkan sonuçlara göre tahsisçinin manuel olarak karar verdiği (karar kategorisi oto onay istisnası ve gri alan olan ya da çok nadirde olsa politika kurallarının ezildiği başvurular) otomatik red ve onay kurallarının yeterli gelmediği durumlarda tahsisçinin deneyimi ve yorumu ile ikinci kez değerlendirilen başvuruların yasal takibe girme oranlarında ciddi bir azalma olduğu gözlenmektedir. Bu da sistemsel olarak karar mekanizmasında var olan mevcut kurallara ek kurallar getirilmesinin doğru olacağını göstermektedir. Müşterilerin talep ettikleri kredi oranı azaldıkça yasal takibe girme oranlarında önemli bir düşüş olduğu görülmüştür. Bu da müşterinin sahip olduğu nakit miktarının fazlalığı dolayısı ile ekonomik durumunun iyi olması ile doğrudan ilişkilidir. Benzer şekilde bir müşterinin diğer mevcut kredi borçlanmaları arttıkça yasal takibe girme oranı da hızla artmaktadır. Öte yandan şubelerin az ancak müşteri potansiyelinin fazla olduğu şirketin genel merkezinden uzak İç Anadolu, Ege ve Doğu Anadolu gibi bölgelerde müşterilerin yasal takibe girme oranlarının göreceli olarak yüksek olduğu gözlemlenmiştir. Bunun sebebi olarak bu bölgelerde bulunan az sayıda şube ile satış hedeflerinin gerçekleştirilmesi için yapılan baskılar sonucu daha esnek davranan tahsisçi etkisi ya da manipule edilmeye müsait ve satış temsilcisi tarafından başvuru esnasında girilen demografik bilgiler gösterilebilir.

İkinci ve üçüncü en yüksek doğru sınıflama oranına sahip C&RT ve CHAID algoritmalarına göre karar ağacında ilk dallanmaya sebep olan ve bir müşterinin yasal takibe girmesine sebep olan en önemli etken müşterinin talep ettiği kredi oranıdır. İkinci en büyük etken ise müşterilerin aldıkları araç kredisini öderken ilk 3 taksidini ödedikleri dönemde 30 gün ve üzeri gecikme yaşamalarıdır. C&RT algoritmasına göre talep ettiği kredi oranı yüzde yetmiş ve üzeri olup ilk 3 taksit döneminde 30 gün ve üzeri gecikme yaşayan müşterilerin yaklaşık olarak yüzde doksanyedisi yasal takibe girmektedir. Bu oran CHAID algoritmasında yaklaşık yüzde doksansekizdir. Burada dikkat çeken nokta lojistik regresyonda tek başına yasal takibe girme etkisi oldukça düşük olan ilk 3 taksit ödeme döneminde 30 gün ve üzeri gecikmeye düşme bilgisinin kredi borçlanma oranı ile birlikte ele alındığında oldukça etkili olduğunun tespit edilmesidir.

Dördüncü en yüksek sınıflama oranına sahip olan QUEST algoritması ele alındığında karar ağacında ilk dallanmaya sebep olan ve bir müşterinin yasal takibe girmesine sebep olan en önemli etken müşterinin kredi alırken talep ettiği vadedir. İkinci en büyük etken ise diğer karar ağacı algoritmalarında da olduğu gibi müşterilerin aldıkları araç kredisini öderken ilk 3 taksidini ödedikleri dönemde 30 gün ve üzeri gecikme yaşamalarıdır. Buna göre vadesi 36 aydan yüksek olan müşterilerden ilk 3 taksit döneminde 30 gün ve üzeri gecikme yaşayan müşterilerin yaklaşık olarak yüzde doksanyedisi yasal takibe girmektedir.

Bir başka algoritma olarak geliştirilen karar listesinde bu 3 karar ağacı algoritmasında ortaya çıkan kural setlerinden daha farklı olan ve bir müşterinin yasal takibe girmesi üzerinde etkin rol oynayan bir takım kurallar daha tespit edilmiştir. Bunlardan en etkin olanı otomatik olarak karar verilemeyen ve tahsisçinin görüşüne sunulan başvurulardan ilk 3 taksit döneminde 30 gün ve üzeri gecikme yaşayan müşterilerdir. Bu tür müşterilerin yaklaşık olarak yüzde doksanüçü yasal takibe girmektedir. Bir diğer etkin kural ise oto red istisnası olan ve kredi geçmişi iki yıldan az olan müşterilerdir. Bu tür müşterilerin yaklaşık olarak yüzde seksenyedisi yasal takibe girmektedir. Aynı şekilde oto red istinası olan ve açık kredi borcu yirmialtıbindokuzyüzyetmişbir TL'den yüksek olan müşterilerin yasal takibe gitme oranı yüzde seksendörttür.

Bir veri madenciliği yazılımı olan Clementine 12.0'da geliştirilen bu algoritmaların yanısıra WEKA'da uygulanan bir çok algoritma ile de seçilen örneklem performans açısından değerlendirilmiştir. Bu algoritmalardan en yüksek doğru sınıflandırma derecesine sahip olan beş algoritma ve performans değerleri tablo 6.2'de gösterildiği gibidir.

Tablo 6.2: WEKA'da uygulanan algoritmalar ve performans özet bilgileri

	Correctly Classified Instances	Incorrectly Classified Instances	Root mean squared error	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Classifier
RandomSub Space	78%	22%	0,402	0,78	0,22	0,78	0,78	0,78	0,85	Meta
Threshold Selector	77%	23%	0,3906	0,77	0,23	0,77	0,77	0,77	0,86	Meta
Class. Via Regression	77%	23%	0,3979	0,77	0,23	0,77	0,77	0,77	0,86	Meta
Logistic	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	Functions
MultiClass Classifier	77%	23%	0,3909	0,77	0,23	0,77	0,77	0,77	0,86	Meta

Gerçek veriler ile geliştirilen modeller için yüzde yetmişbeş ve üstü doğru sınıflandırma oranı iyi bir göstergedir. Aynı zamanda kazanç eğrisinin yüzde seksenbeş ve daha yüksek değerlerde olması verilerin birbirinden iyi ayrıldığını ve veri setinin iyi modellendiğini göstermektedir.

En iyi sınıflandırma oranını veren ikili lojistik regresyon ve benzer dallanmalar ile onu destekleyen karar ağaçları algoritmaları sonuçları göz önünde bulundurulduğunda ortalama kredi tutarı≈ 30.000 TL olan bu şirketin 36 aylık datasında öngörüp önlem alabileceği kredi tutarı 22.440.000 TL'dir. Bu rakamda ortalama bir ayda finansman sağlamış olduğu toplam kredi tutarına eşdeğerdir.Öte yandan bu algoritmaların başvuru aşamasında çalışması durumunda ise 12.180.000 TL kadar kesin riskli tutardan kaçınılmış olur.

KAYNAKÇA

Kitaplar

Agresti, A.,1996. *An introduction to categorical data analysis*. New York: John Wiley and sons.

Akkaya, Ş.& Pazarlıoğlu, V.,2000. *Ekonometri 1*. İzmir:Anadolu Matbaacılık

Akkaya, Ş.& Pazarlıoğlu, V.,1998. *Ekonometri 2*. İzmir:Anadolu Matbaacılık

Alpaydın, E.,2004. *Introduction to machine learning*. Cambridge:MIT Press

Edelsteın, H.,2000. *Mining large database - a case study**, two crows corporation***.

Evgeny, A. & Elena, P.,2010. *Applying chaıd for logistic regression diagnostics and classification accuracy improvement, journal of targeting, measurement and analysis for marketing*

Finlay, S.,2014. *Predictive analytics, data mining and big data. myths, misconceptions and methods*. Basingstoke: Palgrave Macmillian. p. 237.

Fukunaga, K.,1990. *Introduction to statistical pattern recognition*. San Diego:CA,Academic

Gujarati, D.N.,2001, *Basic econometrics.* New York: McGraw-Hill

Han, J. & Kamber, M. 2006. *Data mining: concepts and techniques*.Second Edition. USA:Elsevier.

Han, J.K., *Morgan, M.,* 2000. *Data mining concepts and techniques*. San Francisco, USA: *Kaufmann Publishers*

Hawkins, D.M. & Kass, G.V.,1982. *Automatic interaction detection*. Cambridge: Cambridge University Press

Hosmer, D.W.& Lemeshow, S.,2000. *Applied logistic regression*.USA:Wiley.

Magidson, J.,1994. *The CHAID approach to segmentation modeling: chi-squared automatic interaction detection*. Oxford: Blackwell

Maimon, O. & Rokach, L.,2005. *Data mining and knowledge discovery handbook*. Ramat-Aviv:Springer Press

Tatlıdil, H., 2002. *Uygulamalı çok değişkenli istatistiksel analiz*. Ankara: Ziraat Matbaacılık

Olson, D.L.& Delen, D. 2008. *Advanced data mining techniques*. Berlin:Springer Publishing.

Yaralıoğlu, K.,2004.*Uygulama karar destek yöntemleri*. 9. Baskı. İstanbul:Akse Basım.

Zhu, X.Q.& Davidson, I., 2007. *Knowledge discovery and data mining: challenges and realities*. New York: Hershey

Süreli Yayınlar

Akpınar, H., 2000.Veri tabanlarında bilgi keşfi ve veri madenciliği. *İ.Ü. İşletme Fakültesi Dergisi*, **29**(1)

Belson, W. A.,1959. Matching and prediction on the principle of biological classification. *Applied Statistics*, **1**(8), ss. 65–75.

Bircan, H. & Karagöz, Y.,2004. Lojistik regresyon analizi: tıp verileri üzerinde bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*,**1**(9)

Cox, A.L., 2002. Data mining and causal modelling of customer behaviours.*Telecommunication Systems*. **1**(21), s. 356.

Sever, H. & Oguz B., 2002. Veritabanlarında bilgi keşfine formal bir yaklaşım: kısım I eşleştirme sorguları ve algoritmalar. *Bilgi Dünyası*. **3**(2),ss.173-204.

Sever, H. & Oguz B., 2002. Veritabanlarında bilgi keşfine formal bir yaklaşım: kısım II eşleştirme sorgularının biçimsel kavram analizi ile modellenmesi. *Bilgi Dünyası*, **4**(1),ss.15-44

Shearer C.,2000.The CRISP-DM model: the new blueprint for data mining.*Data Warehousing* **4**(2),ss.12-16.

Kurgan, L. & Musilek P., 2006. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*.**1**(21),ss.1- 24.

Othman, M.F.& Yau, T.M.S.,2007. Comparison of different classification techniques using weka for breast cancer. *3rd Kuala Lumpur International Conference on Biomedical Engineering IFMBE Proceedings.* **15**(13),ss. 520-523.