T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİREYSEL MÜŞTERİLERİN KREDİ DEĞERLENDİRME SONUÇLARINI EN İYİ TAHMİN EDEN SCORECARD MODELİNİN OLUŞTURULMASI

YÜKSEK LİSANS TEZİ

Akın SÖYLEMEZ

İstatistik Anabilim Dalı

İstatistik Programı

Tez Danışmanı: Prof. Dr. Nalan CİNEMRE

NİSAN 2009

Akın SÖYLEMEZ tarafından hazırlanan BİREYSEL MÜŞTERİLERİN KREDİ DEĞERLENDİRME SONUÇLARINI EN İYİ TAHMİN EDEN SCORECARD MODELİNİN OLUŞTURULMASI adlı bu tezin ..... tezi olarak uygun olduğunu onaylarım.

.....
Tez Yöneticisi

Bu çalışma, jürimiz tarafından ..... Anabilim Dalında ..... tezi olarak kabul edilmiştir.

Başkan: : _____

Üye : _____

Üye : _____

Üye : _____

Üye : _____

Bu tez, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygundur.

ÖZET

Bu çalışmada amaç, son yıllarda sıklıkla kullanılan Veri Madenciliği (VM) tekniklerini kullanarak bir bankanın kredi değerlendirme sonuçlarını tahmin eden bir scorecard modeli kurmaktır.

Bu amaç doğrultusunda çalışmanın birinci bölümünde VM'nin tanımına, ikinci bölümünde VA'na, üçüncü bölümde VM fonksiyonlarına ve dördüncü bölümde VM süreç modellerine değinilmiştir. Beşinci bölümde ise SPSS Clementine programı üzerinden bir bankanın bireysel müşterilerinin kredi değerlendirme sonuçlarını en iyi tahmin eden scorecard modelinin oluşturulması ile ilgili uygulama yapılmıştır.

SUMMARY

In this study, the aim is using the data mining techniques which is used often nowadays, for setting up a scorecard model to predict credit evaluation process result for a bank.

In accordance with the purpose, in the first part of the study data mining definitions have been made, in the second part data warehouse has been told, in the third part data functions have been discussed and in the fourth part data mining process models have been dealt with. In the fifth part there has been a demonstration about setting up a scorecard model which predict the best scorecard model for credit evaluation process results for a bank's individual customer, using SPSS Clementine program.

ÖNSÖZ

Başta tez danışmanın Sn. Prof. Dr. Nalan Cinemre olmak üzere tüm hocalarıma, düzeltmelerde yardımcı olan Yrd. Doç. Dr. Semra ERPOLAT ve Arş. Gör. Elif Özge ÖZDAMAR'a, tezimi bitirmemde bana destek olan eşim Funda SÖYLEMEZ ve dostum Özgüç AKDAĞ'a tüm kalbimle teşekkür ederim.

Nisan 2009 Akın SÖYLEMEZ

İÇİNDEKİLER DİZİNİ

İÇİNDEKİLER DİZİNİ (Devam Ediyor)

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

BÖLÜM 1

VERİ MADENCİLİĞİ

1.1. VERİ MADENCİLİĞİNİN TANIMI

Veri madenciliği (VM) nin başlangıcı 19. yüzyıla dayanmaktadır. Bu yüzyılda Aristotle ve Bacon tarafından öne sürülen bilimsel metodolojik yaklaşımlar uzun yıllar kullanılmıştır. Bu yaklaşımların temeli; yüksek miktarlardaki veriyi toplama, geçmiş örnekleri araştırma ve elde edilen bulgulara göre varsayımlarda bulunmaya dayanmaktadır. Daha sonra Galileo tarafından, VM ile bulunan sonuçların bilim adamları tarafından da kontrol edilmesi gerektiği görüşü öne sürülmüştür. Bilimsel otoriteler tarafından da kabul gören bu yaklaşım, 20. yüzyıla gelindiğinde önemli bir rotasyona uğramıştır. Rotasyon sonucunda Galileo'nun öne sürdüğü bazı bilimsel yöntemlar tamamen tersine dönmüş ve "Doğrulayıcı Bilim" olarak adlandırılan, bir teorinin öncelikle varsayımla yapılabileceği, daha sonra deneysel verinin toplanabileceği yaklaşımı kabul görmeye başlamıştır.

VM, Galileo'nun bilimsel yöntem geleneğini takip eder ve daha sonraları 19. yüzyıldaki bilimsel metoda geri dönerek günümüze kadar gelir [10]. VM'nin bugün ulaştığı nokta ise neredeyse evlerde bile kullanılan bir terim haline gelmiş olmasıdır. En basitinden bir maç seyrederken sunucunun maç ile ilgili verdiği "15 maçtır yenilmeyen takımımız bugün tarihi bir zafere imza atmak üzere" ya da "futbol tarihimizde bir ilk ve takımımız Avrupa Şampiyonu" gibi bilgiler geçmişteki veriler ışığında ortaya çıkarıldığından aslında bir çeşit VM'dir.

VM, büyük hacimli veri yığınları içerisinden karar alabilmek için potansiyel olarak faydalı olabilecek, uygulanabilir ve anlamlı bilginin çıkartılmasıdır. VM geniş anlamda veri analiz tekniklerinin bütünüdür ve tek başına bir çözüm değildir. Mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminler yapmak için gerekli olan bilgileri elde etmeye yarayan bir araçtır. Ortaya çıkması hedeflenen bilgiler; üstü kapalı, çok net olmayan, önceden bilinmeyen, daha önce keşfedilmemiş ancak potansiyel olarak kullanışlı ve anlamlı kritik bilgilerdir.

VM döngüsü, veri yığınlarını elden geçirme ile başlayarak, analiz sonucunda ortaya çıkan sonuçların uzman gözüyle yorumlanması ile tamamlanır. VM çalışmalarının alt yapısının önemli bir bölümünü istatistik ve veritabanı uygulamaları oluşturur. VM büyük boyutlu VA'ların meydana çıkmasının bir sonucudur. 1960'lı yıllarda veriler elektronik ortamda toplanmaya ve geçmişteki veriler bilgisayar ile analiz edilmeye başlanmıştır. 1980'li yıllarda SQL ile verilerin dinamik ve anlık analiz edilmesine olanak sağlanmıştır. 1990'lı yıllara gelindiğinde toplanan verilerin hacimleri çok büyük boyutlara ulaşmış ve verilerin depolanması için VA'lar kullanılmaya başlanmıştır [1].

Günümüzde bilgisayarlar çok büyük miktardaki veriyi saklayabilmekte ve daha kısa sürede işleyebilmektedir. Ayrıca bilgisayarların ucuzlaması sayısal teknolojinin daha yaygın olarak kullanılmasını da sağlamıştır. Sayısal olarak toplanan ve uygun ortamlarda saklanan veri kütlelerinin değerlendirilip anlamlı bilgilere dönüşmesi için istatistik ve yapay zeka tekniklerinden yararlanılması, VM'ni ortaya çıkarmıştır. Bunun sonucunda da detaylı ve doğru bilgiye ulaşılmıştır.

Teknolojik gelişmeler ham verinin yeni fırsatlar üretmek üzere yönetim ve pazar ihtiyaçlarına yanıt verecek bilgiye dönüştürülmesini kolaylaştırmış bir anlamda kurumları VM üzerinde çalışmaya mecbur bırakmıştır. Aşağıda VM'nin çeşitli tanımlarına yer verilmiştir.

"VM, veri tabanı teknolojisi, makine öğrenmesi, desen tanıma, istatistik, görselleştirme gibi çok disiplinli bir bilim dalıdır [10].

"VM, önceleri bilinmeyen, geçerli ve etkin bilginin büyük veri tabanlarından çekilmesi ve daha sonra bu bilginin son iş kararını almak için kullanılmasını kapsayan bir süreçtir" [3].

"VM, aksi halde keşfedilemeyebilecek olan eğilimleri ve örüntüleri bulmak için çok miktarda verinin otomatikleştirilmiş analizidir" [9].

"VM, muazzam boyutlardaki veriden şirketlerin daha iyi kararlar almalarına yardımcı olup, pazarda rekabetçi olarak kalmalarını sağlayabilecek ilginç bilgileri keşfetme sürecidir" [4].

“VM, büyük hacimli veri içerisinden, anlamlı, gizli kalmış ve kuruluşun karar destek sistemi için faydalı olabilecek bilgilerin çıkartıldığı veri analiz tekniğidir” [19].

1.2. VERİ MADENCİLİĞİNİN ÖNEMİ

Verilen kararların doğruluğu karar veren kişinin yeteneklerine ve tecrübelerine olduğu kadar sahip olduğu bilginin yeterliliğine de bağlıdır. Bu nedenle artık bilgi, mal ve hizmetin yanında bir üçüncü faktör olarak değerlendirilmektedir. Bilginin yeterli olması bilgiyi oluşturan verilerin doğru depolanmasına, doğru işlenmesine ve doğru yorumlanmasına bağlıdır. Bunun yanında karar vericiler doğru kararları alabilmek için mümkün olduğunca çok veriyi depolamaya çaba göstermektedirler. Ancak verilerin toplanması çok zorlu bir süreç değildir. Asıl zor olan devamlı çoğalan ham veriyi anlamlı ve kullanılabilir bilgiye dönüştürebilmektir.

Veri madencisi, verileri inceler ve bu veriler içerisinde bulunan sayılar arasında anlamlı ilişkiler ortaya çıkarır. İlişkiler veri tabanlarında bulunur ve açık olarak belirtilmez. Çok miktardaki verinin hızlı bir şekilde toplanıp işlenebilmesi ve yönetilebilmesi VM ile mümkün olmaktadır. VM’nde görsel tekniklerin kullanımı büyük önem taşımaktadır. Çünkü karmaşık sonuçlar görsel olarak daha iyi anlaşılmaktadır.

VM ile şirketler önceden bilinmeyen bilgileri ortaya çıkararak karar verme süreçlerini iyileştirirler. VM teknikleri ile maliyetleri azaltmak, gelirleri artırmak, verimliliği artırmak, yeni fırsatlar ortaya çıkarmak, yeni keşifler yapmak, yoğun iş gücü gerektiren işleri otomatikleştirmek, sahtekarlıkları belirlemek ve müşteri deneyimini geliştirmek mümkündür.

1.3. VERİ MADENCİLİĞİNİN KULLANIM AMAÇLARI

Kış aylarında düşen yağış miktarı, en çok satılan bira markası, televizyon ratingleri, bir futbol maçının skoru, sedan araba alan müşterilerin yaşlarına kadar çevremizde yaşananların çok önemli bir kısmı sürekli kaydedilmektedir. Sayılardan oluşan bu kayıtlara ancak doğru şeklide bakıldığında bir anlam ifade etmektedir.

VM’nin asıl amacı, veri yığınlarından anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır. VM istatistiğin tersine tümdengelim

yaklaşımını kullanarak veriden fark edilmeyen ve öngörülmeyen bilgiyi çıkarmaya çalışır. VM'nin bazı kullanım amaçları aşağıda sıralandığı gibidir.

- Bir firma, kaybettiği müşteriler üzerinden analiz yaparak müşterilerini neden kaybettiğini bulup geri kazanmak ve mevcut müşterisini kaybetmemek için gerekli kararları öğrenebilir.
- Ürettiği ürün ya da verdiği hizmet ile ilgili hangi özelliklerin müşteri memnuniyetini artırdığını ortaya çıkartabilir.
- Bir banka, müşterilerinin kredi risklerini hesaplayarak hangi müşterinin kredisini sorunsuz olarak ödeyeceğini hangi müşterinin kredisini batıracağını tahmin edilebilir.
- Müşterilerin profillerine göre şirketin diğer ürünleri çapraz satış şeklinde müşterilere satılabilir.
- Piyasada oluşabilecek değişikliklere göre müşterinin nasıl davranacağı kestirilebilir.
- Müşteri segmentasyonu belirlenerek müşterilere çeşitli kampanyalar düzenlenebilir.
- Yeni bir ürün ile ilgili hedef kitlenin seçimi yapılabilir.
- Firmanın finansal yapısının, makro ekonomik değişmeler karşısındaki duyarlılığı ve oluşabilecek riskler belirlenebilir.
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik karlılık, pazar payı, ciro gibi karar vericeler için önemli olan finansal göstergelerin tahminleri yapılabilir.

1.4. VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

VM'nin uygulama alanlarını bilimsel ve iş dünyası olarak ikiye ayırmak mümkündür.

Bilimsel çalışmalarda VM gelişmiş veri toplama yöntemleri ile işlenmek üzere çok büyük miktarda verinin toplanmasına ihtiyaç duyulması, geleneksel yöntemlerin ham veriyi işlemede yetersiz kalması ve hipotez oluşturma, sınıflandırma, karar alma gibi bilimsel çalışma adımlarında bilim insanlarına destek sağlaması için kullanılır.

İş dünyasında VM'nin kullanılmasının temel nedeni, müşteriyi tanıyarak müşteri memnuniyeti sağlamak ve bu şekilde rekabet ortamında hızlı ve doğru kararlar alabilmektir. İş dünyasının her alanında VM büyük önem taşısa da aşağıda bilgiye dayalı yönetime en fazla ihtiyaç duyan sektörler ve bu sektörlerin ihtiyaç nedenleri sıralanmıştır.

Bankacılık sektörü: Farklı finansal göstergeler arasındaki gizli korelasyonların bulunması, kredi kartı dolandırıcıların tespit edilmesi, müşteri segmentasyonu, kredi taleplerinin değerlendirilmesi, usulsüzlük tespiti, risk analizleri ve risk yönetimi.

Sigorta sektörü: Yeni poliçe talep edecek müşterilerin tahmin edilmesi, sigorta dolandırıcılıklarının tespit edilmesi ve riskli müşteri tipinin belirlenmesi.

Telekomünikasyon sektörü: Pazar araştırması, müşteri segmentasyonu, satış gücü optimizasyonu, kampanya optimizasyonu ve çapraz satışlar.

Pazarlama ve perakende sektörü: Müşteri segmentasyonu, müşterilerin demografik özellikleri arasındaki bağlantıların kurulması, çeşitli pazarlama kampanyaları, mevcut müşterilerin elde tutulması, pazar sepeti analizi, çapraz satış analizleri, müşteri değerlendirme, müşteri ilişkileri yönetimi, çeşitli müşteri analizleri ve satış tahminleri.

İş dünyasında VM çalışmalarının büyük bir bölümü kaynakları daha verimli kullanmak, potansiyeli artırmak ve güvenliği sağlamak amacıyla kullanılmaktadır.

1.5. VERİ MADENCİLİĞİNDEKİ SORUNLAR

VM, girdi olarak ham veriyi sağlamak üzere veri tabanlarına dayanır. Bu da veri tabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda çeşitli sorunların yaşanmasına neden olur. VM'nde yaşanan diğer sorunlar ise verinin konu ile uyumsuzluğundan kaynaklanan sorunlardır. VM'nde karşılaşılan başlıca sorunları aşağıdaki alt başlıklar altında toplamak mümkündür.

Sınırlı Bilgi: Veri tabanları genel olarak VM dışındaki amaçlar için tasarlanmıştır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler veri tabanları içerisinde bulunmayabilir.

Eksik Değerler: Veri tabanlarındaki eksik ve yanlış bilgiden dolayı VM amacına tam olarak ulaşamayabilir. Bilgi yanlışlığı, ölçüm hatalarından, ya da öznel yaklaşımlardan kaynaklanabilir. Eksik değerler ise ölçümü yapılamamış veya daha sonradan kaybedilmiş değerler olup veri tabanında yer almaları durumunda aşağıda sıralanan yöntemlere başvurulabilir [7].

- **Eksik değerleri analizden çıkarmak:** Çok fazla sayıda eksik değerin olması durumunda kullanışlı olmayan bir yöntemdir.

- **Eksik değerleri el ile doldurmak:** Uzun zaman alan ve büyük veri tabanları için uygulanması zor bir yöntemdir.

- **Eksik değerler için global bir değer belirlemek:** Analiz sırasında belirlenen global değerin sıklıkla kullanılması ve hesaplamalara dahil edilmesi yanlış sonuçların elde edilmesine neden olabilir.

- **Eksik değer içeren veri için ortalama değer kullanmak:** Genellikle gelir bilgisi, kullandırılan kredi miktarı gibi değişkenlerin belirlenmesinde doğru sonuçlar veren bir yöntemdir.

- **En olası değer ile eksik veriyi tamamlamak:** Bu model için regresyon, karar ağaçları ya da yapay sinir ağaları kullanılabilir.

Gürültülü Değerler: Bir değişkendeki rastlantısal hata oranıdır. Verideki gürültüyü yok etmek için aşağıdaki yöntemler uygulanabilir [7].

- **Kutulama:** Bu yöntemde veriler komşu komşu değerlerine göre sıralanırlar. Sıralanmış bu veriler belirli sayıda kutulara konur. Ortalama değerlere göre, medyan değerine göre ya da limitlere göre gürültü değerleri düzeltilir. Eğer ortalama değere göre kutulama yapılacak ise gürültülü değer o kutu içindeki ortalama değer ile yer değiştirir.

- **Demetleme:** Birbirine benzer değerler gruplara ya da demetlere bölünerek her bir demetin sınır çizgisi belirlenir.

- **Regresyon:** Verinin bir fonksiyona sokularak o fonksiyon üzerinden yerleşmesi sağlanır ve böylece gürültülü değerler otomatik olarak elenmiş olur.

Anlamsız Veri: Veri tabanlarında gerekenden fazla anlamsız verinin olması sonuca ulaşmada VM'nin yetersiz kalmasına neden olabilir.

Veri Tabanlarının Büyük Boyutları: Sadece veri tabanlarının boyutu yüzünden VM yöntemlerinden herhangi birinin ham veri ile başarılı olma olasılığı yoktur. VM yöntemleri bu şekilde elde edilen sonuçların tüm veri tabanını temsil edeceğini umarak, veri tabanından bir örneğin çıkartılmasını gerektirebilir. Bir veri tabanının boyutunu küçültmek için aşağıda sıralanan iki yönteme başvurulur [8].

- **Veri alanından örnekleme:** Genellikle rastlantısal olarak kayıtlar seçilir ve veri setinden çıkartılır.

- **Özellik alanından örnekleme:** Her veri kaydının bazı özellikleri rastlantısal olarak seçilir ve veri setinden çıkartılır.

1.6. VERİ MADENCİLİĞİNİ ETKİLEYEN FAKTÖRLER

Bir veri küpündeki verilerin hepsi süreksiz olarak başlamıyorsa süreksiz olarak kabul edilir. Sürekli olarak başlayanlar ya da hem sürekli hem süreksiz olanlar ise süreksiz hale getirilir.

Sürekli bir veriyi süreksiz hale getirmenin en iyi yolunu bulabilmek kolay bir iş değildir. Veri tabanlarında bulunan veriler genellikle geleneksel istatistik varsayımlarına uymayan veriler olup karar vericinin mevcut durumu ifade eden sonuçları gösterebilmesinde önemli sorunlara neden olabilmektedir.

Veri küpünde yeralan verilerin analizinde kümeleme, sınıflandırma, regresyon gibi çok değişkenli yöntemler kullanılır. Bu yöntemlerin bazılarının uygulanmasında aşağıda sıralanan sorunlarla karşılaşılabilir [10].

- Veri içinde bilgiyi aramak genellikle bir bilgi varmış gibi gözükmesiyle sonuçlanır. Ancak aslında bilgi olarak görülen şeyler sadece rastlantısal dalgalanmalar olabilir.

- İstatistikte örnek boyutu genellikle küçüktür. VM'nde ise istatistiğin tersine genellikle çok büyüktür. Sonuç olarak VMnde hipotez testlerinin en yakın oran yöntemleri ve p-value önem derecesi en ufak etkileri bile anlamlı hale getirmeye yönelir. Bu yüzden Bayesian yöntemleri null hipotezi reddetme konusunda daha güvenli olduğu için çok daha tercih edilebilir. [2]

- Veri boyutu makinenin kapasitesini aşabilir.

- Data küpünün boyutsallığı arttıkça modelleme problemleri de artabilir.

- Yüksek boyutlu modellemelerde sayıca fazla bağımlı değişken olabilir. Fazla sayıda bağımlı değişken analizi zorlaştırabilir.

- Yüksek boyutlu olasılık tablolarını genel istatistikli modelleri kullanarak anlayabilmek ve analiz edebilmek için, mümkün olduğu kadar yüksek boyutu daha alt boyuta indirgeyebilmek ve bu sayede daha kontrol edilebilir hale getirebilmek gerekebilir.

BÖLÜM 2

VERİ AMBARI

1.1. VERİ AMBARI'NIN TANIMI

Veri Ambarı (VA), ilişkili verilerin sorgulandığı ve analizlerinin yapıldığı büyük ölçekli bir veri deposudur. Çok sayıda şube ve acenteleri bulunan günümüz şirketleri için müşteri verileri büyük önem taşımaktadır. Bu veriler genellikle çok büyük hacimli veriler olup Canlı Sistemler (CS) ve Karar Destek Sistemleri (KDS) ile işletilirler.

CS; firmanın karlılığının artırılması, firma politikalarının belirlenmesi, problemli noktaların tespit edilmesi ve aksiyonun sağlanması gibi önemli kararların verilmesinde şirket yöneticilerine yardım etmek amacıyla kullanılan sistemlerdir. Günümüzde üst yönetim için bu tip verilerin analizi "Balanced Scorecard" uygulamaları ile yapılmaktadır.

Veri tabanını canlı veri olarak da adlandırılan güncel verilerin oluşturduğu ve genellikle şirketlerin anlık gereksinimlerinin karşılanmasında kullanıldığı CS'de verilere mümkün olduğu kadar hızlı ve kısa bir sürede erişebilmek büyük önem taşımaktadır. Örneğin bir bankada batan kredilerin takibi önemli bir durum olup, kredi verilirken daha dikkatli olunması gerektiği bilgisi CS ile ilgili birimlere anlık olarak ulaştırılabilir. Canlı veriler anlık veri olduğu için kaynak veri olarak kullanılamamaktadır.

KDS ise CS aracılığıyla alınan kararların uygulanabilmesini sağlayan sistemlerdir. KDS'nde kullanılan veri güncel veri olmadığından veriye erişimin hızlı olması önemli değildir. Önemli olan husus sistemin iyi bir performansta çalışabilmesidir. Bu sistemler geçmişteki tüm verileri sakladıkları için veri hacimleri oldukça büyüktür. Bu kadar büyük hacimli veri içinde sorgulamalar yapmak KDS'in yavaş çalışmasına neden olmakta ve istenilen veriye ulaşılmasını zorlamaktadır.

KDS'nin bir çeşidi olan VA'lar CS tarafından beslenen büyük hacimli veri kümeleridir. Verilerin VA'lara aktarımları, şirketlerin ihtiyaçlarına göre anlık, günlük, haftalık hatta aylık bile olabilir. Veri aktarımları sırasında eski verilerin üzerine yeni verilerin aktarılamaması, ekleme ya da silme işlemlerinin yapılamaması, verilerin anlamsal bütünlüğü ve verinin güvenilirliği açısından çok önemlidir.

VA pahalı bir yatırım maliyeti olsa bile sonuç olarak getirisi (yararı) maliyetine oranla çok daha fazla olmaktadır. VA'lar bünyesinde data mart ve meta data adı verilen yapıları bulundurmaktadır.

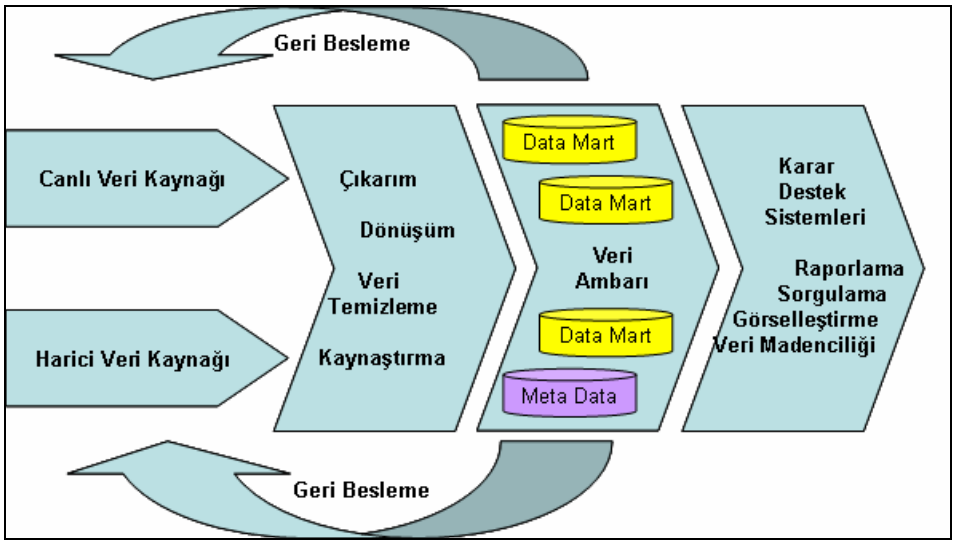
Data mart, küçük boyutlu (1-10 GB) bölümsel ambarlardır. VA'nın alt kümeleri olarak da adlandırılırlar. Şirketlerde farklı işler yapan bölümler için farklı data martlar kullanılır.

Meta data VA'ların en önemli bileşenlerinden biri olup verilerin tanımlandığı kısımdır. Meta data "veri hakkında veri" anlamındadır. Meta data her veri elementinin anlamını, hangi elementlerin hangileriyle nasıl ilişkili olduğunu ve kaynak verisi ile erişilecek veri gibi bilgileri içermektedir.

1.2. VERİ AMBARI MİMARİSİ

VA'lar oluşturulurken, kaynak olarak kullanılan CS'lerin veya harici veri kaynaklarının yönetimi bilgi kazancının artırımı acısından önem taşımaktadır. VA'larda büyük hacimli verilerin depolandığı göz önünde bulundurulursa, bunların tasarımı ve yönetilmesinin önemi daha kolay biçimde anlaşılır. Depolanacak verilerin VA'ya aktarım periyodu işletmenin ihtiyaçlarına göre belirlenip, günlük, haftalık veya aylık şekilde düzenlenebilir. Veri aktarım periyodu uzadıkça kayıtların güncelliğinin sağlanması zorlaşır. VA'nın mimarisinin gösterildiği Şekil 1.1'de de görüldüğü gibi VA tek bir canlı sistemi kaynak olarak kullanmamakta ayrıca kaynaklar arasında bir geri besleme mekanizması da sağlamaktadır. Bu mekanizma sayesinde VA kaynak sistemlerden ihtiyaç duyduğu istekleri elde edebilmektedir.

VA'lar ihtiyaç doğrultusunda alt veri depolarına yani data martlara aktarılabilmektedir. Genelde işletmeler departmanlar bazında alt veri kümeleri oluşturmaktadır.



Şekil 1.1. VA mimarisi.

Şekil 1.1.'de görüldüğü gibi, VA'lar üzerinden VM, raporlama, sorgulamalar, görselleştirme ve alt veri depoları oluşturulabilmektedir. VA'ların farklı kaynaklardan beslendiği göz önünde bulundurulursa, bu verilerin VA'lara aktarımından önce birtakım düzenlemeler veya dönüşümler yapılması gerekmektedir. Veri aktarımı sırasında, mevcut verilerin tekrar aktarılması, bazı verilerin güncellenmesi veya silinmesi söz konusu olabilmektedir. Bu düzenlemeler sonucunda verilerin uygun bir yapıda bütünlüğü sağlanıp daha sonra VA'ya aktarımı gerçekleşecektir. Ham verilerden KDS'ye kaynak olabilecek verilerin kazanımı için uygulanması gereken işlemler; çıkarım, dönüşüm, temizleme, kaynaştırma şekilde sıralanabilmektedir [19].

Çıkarım, VA'lara kaynak olabilecek farklı yapılara sahip verilerin elde edilmesi için gerekli adımdır. Dönüşüm; kaynaklardan sağlanan ham verilerin KDS için anlamlı hale getirilmesidir.

Veri temizleme, kaynak verilerde bulunan eksik veya hatalı verilerin anlamlaştırılmasıdır. Örneğin eksik verilere uygun değerlerin atanması, tutarsızlığı ortadan kaldıracak ve böylece daha doğru ve güvenilir sonuçlar elde edilmesini sağlayacaktır.

Kaynaştırma, farklı özelliklere sahip veri kaynaklarından sağlanan verilerin belirli bir tablo yapısı altında birleştirilmesidir. Kaynaklar veri dosyaları, veri tabanları, veri küpleri olabilir.

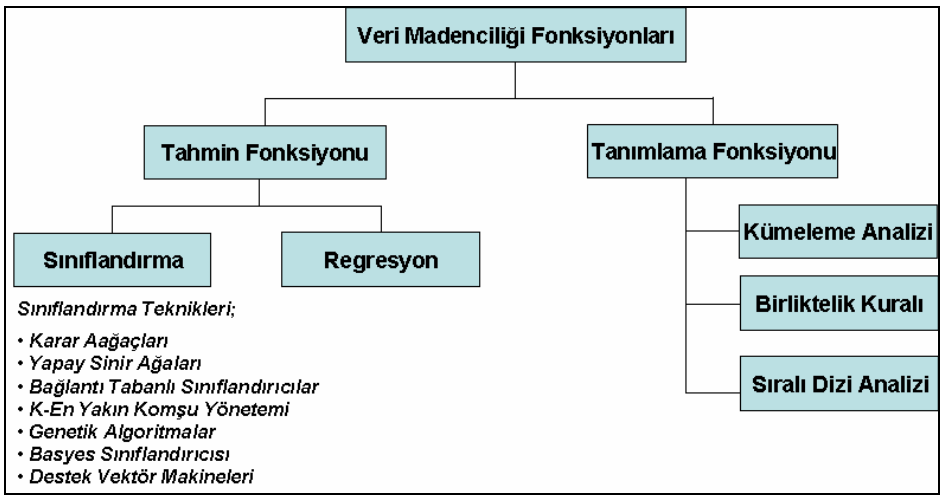
BÖLÜM 3

VERİ MADENCİLİĞİ FONKSİYONLARI

3.1. FONKSİYONLARIN GENEL ÖZELLİKLERİ

VM fonksiyonları tahmin fonksiyonları ve tanımlama fonksiyonları olmak üzere iki ana kategoride incelenir.

VM'de tahmin, ciddi ve güçlü bir hedef iken bilgi keşfi daha zayıf bir yaklaşım olup genellikle tahmine ön bilgi oluşturmak amacıyla kullanılır. Tanımlama ise bulguların tahmin edilmesinin sağlanmasıdır. VM'de kullanılan tahmin fonksiyonları sınıflandırma ve regresyon analizleri; tanımlama fonksiyonları ise kümeleme, birliktelik ve sıralama analizleri olarak incelenebilir. Şekil 3.1.'de VM fonksiyonları ayrıntılı biçimde gösterilmiştir. Gerçek problemlerin sıralanan bu analizlerden hangisiyle çözüleceğinin belirlenmesi için kategorize edilmesi, VM ile ilgili yapılması gereken ilk adımdır [10].



Şekil 3.1. VM fonksiyonları.

3.2. TAHMİN FONKSİYONLARI

Geçmiş verilerden yararlanarak, gelecek ile ilgili bir sonucu tahmin etmek için kullanılan fonksiyonlardır. Diğer bir değişle, yeni bir nesnenin niteliklerini inceleme ve bu nesneyi önceden tanımlanmış bir sınıfa almaktır. Tahmin modellerinde olası sonucu öngörmeye yarayan faktörler ve sonuç yer alır. Model kurulurken geçmiş

deneyimlerden, faktörlerin aldığı değerlere göre elde edilen sonuçlar girdi olarak kullanılır. Beklenen sonuç kategorik değer ya da nümerik değerdir.

Tahmin edilen sonuçların kalitesi tahmin edilen sonuç kadar önemlidir. Çoğunlukla tahmin edilen sonuç ile birlikte, bu sonucun kalitesine yönelik güvenlik aralığı gibi değerler de belirlenir. VMnde tahmin fonksiyonları sınıflandırma ve regresyon analizi olarak iki kategoride incelenebilir.

3.2.1. SINIFLANDIRMA

En temel VM fonksiyonlarından biri olan sınıflandırma, kategorik sonuçları tahmin etmede kullanılır. Sınıflandırma tekniklerinden biri ile model oluşturmak, sonuçları önceden bilinen durumlar ve bu durumlar ile ilgili faktörlerin aldığı değerlerin bilinmesi ile mümkündür. Bu değerler eğitim verisi olarak adlandırılır. Eğitim verisinin yanında test verileri de kullanılmaktadır. Eğitim kümesi modelin oluşturulmasında test kümesi ise modelin doğrulanmasında kullanılır. Hastalık teşhisi, kredi başvuru değerlendirmesi, insan davranışlarının incelenmesi, sınıflandırma analizlerinin kullanıldığı alanlardan birkaçıdır [1]. Sınıflandırma teknikleri aşağıda kısaca açıklanmıştır.

3.2.1.1. Karar Ağaçları

Karar ağaçları, sınıflandırma ve tahminleme yapmada kullanılan popüler bir VM tekniğidir. Karar ağaçlarının yapay sinir ağlarından daha avantajlı tarafı kullanıcının kolayca anlayabileceği şekilde ifade edilebilmesi ve ortaya kurallar çıkartabilmesidir. Karar ağaçları, ağaca benzer bir sınıflandırıcı yapıya sahip olup ağaçtaki her yaprak ya bir karar düğümünü ya da kararı temsil eder [7].

Morgan ve Sonquist tarafından University of Michigan'da 1970'li yılların başlarında kullaılan Automatic Interaction Detector-AID, karar ağacı temelli ilk algoritma ve yazılımdır. AID tekniği en kuvvetli ve en iyi tahmini gerçekleştirebilmek için bağımlı ve bağımsız değişkenler arasında mümkün olan bütün ilişkilerin incelenmesine dayanmaktadır. Ancak AID'in bağımlı ve bağımsız değişkenler arasındaki ilişkilerin tanımlanmasında aşırı saldırgan davrandığı ve bunun sonucunda anlamlı ve anlamsız ilişkileri ayırt edemediği yönünde Einhorn başta olmak üzere bir çok araştırmacı tarafından çeşitli yayınlar yapılmıştır.

İlk temelleri AID yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sınıflandırılmıştır. Geliştirilen bu algoritmalar aşağıda sıralandığı gibidir.

- CHAID (Chi-Squared Automatic Interaction Detector; G.V. Kass; 1980),
- C&RT (Classification and Regression Trees; Breiman, Friedman, Olshen ve Stone; 1984),
- ID3 (Quinlan; 1986),
- Exhaustive CHAID (Biggs, de Ville ve Suen; 1991),
- C4.5 (Quinlan; 1993),
- MARS (Multivariate Adaptive Regression Splines; Friedman),
- QUEST (Quick, Unbiased, Efficient Statistical Tree; Loh ve Shih, 1997),
- C5.0 (Quinlan),
- SLIQ (Supervised Leaming in Quest; Mehta, Agarwal veve Rissanen),
- SPRINT (Scalable Parallelizable Induction of Decision Trees; Shafer, Agrawal ve Mehta) başlıcalarıdır.

Karar ağacı kurgulanmasının, yorumlanmasının ve veri tabanları ile entegrasyonunun kolaylığı nedeni ile en yangın kullanılan sınıflandırma tekniklerinden biridir. Güvenirliliklerinin iyi olması da başka bir tercih edilme sebebidir. Karar ağaçlarının hedefi bağımlı değişkendeki farklılıkları maksimize edecek şekilde veriyi sıralı bir şekilde parçalara ayırmaktır. Sınıflandırma ağacı olarak da adlandırılmaktadırlar [10].

İstatistiksel yöntemlerde veya yapay sinir ağalarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçlarında ise ağaç oluşturulduktan sonra kökten yaprağa doğru inilerek kuralların yazılması mümkündür. Sonradan başka bir teknik kullanılacak olsa bile karar ağacı yöntemi ile veri hakkında kısa bir ön çalışma yaparak önemli değişkenler ve yaklaşık kurallar konusunda karar vericiye bilgi

verilebilir. Karar ağacı tekniği kullanılarak verinin sınıflandırılması öğrenme, sınıflama ve uygulama olmak üzere üç aşamadan oluşur [10].

Öğrenme: Önceden sonuçları bilinen verilerden model oluşturmaktır.

Sınıflama: Yeni bir veri setinin (test verisi) modele uygulanıp karar ağacının doğruluğunun sınanmasıdır. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflandırmanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır.

Uygulama: Elde edilen doğruluk kabul edilebilir oranda ise, karar ağacının yeni verilerin sınıflandırılması amacıyla kullanılmasıdır.

Aşağıda tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağacı analizlerin yaygın olarak kullanıldığı alanlar sıralanmıştır.

- Belirli bir sınıfın muhatemel üyesi olacak elemanların belirlenmesi,
- Olayların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Parametrik modellerin oluşturulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikliye dönüştürülmesi.

Karar ağacının kullanıldığı uygulamalardan bazıları aşağıda sıralanmıştır.

- Demografik gruplardan hangilerinin mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi (Direct Mail),
- Müşterilere kredi geçmişlerini kullanarak kredi verilmesi (Credit Scoring),

- Geçmişte işletmeye en faydalı olan müşterilerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi,
- Üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesidir.

Karar ağaçlarının güçlü olduğu noktalar aşağıda yer almaktadır.

- Üretilen sonuçlara kolayca ulaşılabilir,
- Denetimli öğrenme için kullanılan bir tekniktir,
- Sonuçlar kurallara dönüştürülebilir,
- Çok sayıda işlem yapılmasına gerek duymadan sınıflandırma işlemini gerçekleştirebilir,
- Hem kategorik (nominal / ordinal) hem de sayısal veriler üzerinde işlem yapabilmektedir,
- Karar ağaçları, sınıflandırma ve tahmin problemleri için hangi değişkenlerin daha önemli olduğunu açıkça ifade etmektedir.

Karar ağacının öngörü için kullandığı çalışmalarda, öngörü yapılacak değişkenin sürekli değerler alması durumunda uygun sonuçlar üretilememesi karar ağacının zayıf olduğu yöndür.

3.2.1.2. Yapay Sinir Ağları

Yapay sinir ağları, insan beyninin işleme mantığı temel alınarak nöronların matematiksel olarak modellenmesidir. Bu yöntem ile kurulan model kontrol edilmekte ve öğrenme faaliyeti ile model geliştirilmektedir. Süreç davranış biçimlerini anlamak ve hatayı en aza indirmek üzerine kuruludur. Bilgiyi almak ve daha sonra her uygulamadan bir ders çıkartmak gibi de düşünülebilir. Yapay sinir ağları istatistiksel yöntemler gibi veri hakkında parametrik bir model öngörmez.

Doğru sınıflandırma sağlayan ve doğru sonuçlar veren bir yöntem olmak ile birilikte en önemli dezavantajı, öğrenme süresinin uzun olması ve çıkan sonuçlarının ifade edilmesinin ya da tanımlanmasının zor olmasıdır.

3.2.1.3. Bağlantı Tabanlı Sınıflandırıcılar

Bağlantı tabanlı sınıflandırıcılar, veri içindeki bağları ortaya çıkartmak için kullanılan yöntemdir. Genelde ürün ve müşteri arasındaki bağların ortaya çıkartıldığı market sepeti analizlerinde, hedefe yönelik pazarlama alanlarında veya stok fiyat değişimlerinde kullanılır [7].

3.2.1.4. k-En Yakın Komşu Yöntemi

Demetleme yöntemi olarak da bilinen k-en yakın komşu yönteminde k değeri, komşu olan kayıtların sayısını simgeler. Yöntem verilen N adet prototip örüntüye ve bunların doğru sınıflandırılmasına göre, sınıflandırılmamış olan bir örüntüyü en yakın komşu gruba bağlar. Sınıflandırmanın doğruluğu oranı k değerinin artmasına bağlı olarak artış gösterir [7].

3.2.1.5. Genetik Algoritmalar

Genetik algoritmalar evrimsel gelişimi taklit ederek optimum sonuçlara ulaşmayı sağlayan yapılardır. Bu algoritmalarda işlem elemanları olarak çaprazlama, mutasyon ve seçme operatörleri kullanılır. Algoritmaların işleyişi yüksek uygunluğa sahip çözümlerin seçilip mutasyona uğratılarak daha yüksek uygunlukta çözümler üretmek üzere yeniden kullanılması şeklindedir. Genetik algoritmalar genellikle iş tarifelerinde ve motor dizaynlarında kullanılır [7].

3.2.1.6. Bayes Sınıflandırıcısı

Bayes sınıflandırıcısı istatistiksel bir sınıflandırıcı olup eldeki verilerin, belirlenmiş olan sınıflara ait olma olasılıklarını öngörür. İstatistikteki Bayes teorimeine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelini oluşturarak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçların elde edilmesine olanak sağlar. Bayes sınıflandırıcısı, belirsizlik taşıyan durumlarda karar verme konusunda çok kullanışlıdır. Dezavantajı ise değişkenler arasındaki ilişkinin modellenemiyor olması ve değişkenlerin birbirlerinden tamamen bağımsız olduklarının varsayımıdır [1].

3.2.1.7. Destek Vektör Makineleri

Destek Vektör Makineleri (Support Vector Machine, SVM), doğrusal olmayan bir şekilde ayrılabilen öbekler için optimal hiper-düzlemi bulmaya çalışan yöntemdir. Bu yüzden SVM'lerin VM'deki uygulamaları özellikle sınıflandırma üzerinde olmuştur. Elde edilen sonuçlar bu yöntemin sınıflandırmada oldukça başarılı olduğunu göstermiştir.

SVM modeli; öğrenme, sınıflandırma, kümeleme, yoğunluk tahmini ve veriden regresyon kuralları üretmek için kullanılan bir eğitme algoritmasıdır. Temeli ilk olarak 1960'lı yıllarda V. Vapnik tarafından atılmış olan SVM modeli, 1995'li yıllarda daha çok sınıflandırmaya yönelik olarak kullanılmıştır. Son zamanlardaki kullanımı ise daha çok regresyon ile yoğunluk fonksiyonu tahminli teknikler üzerinde olmuştur [6].

Oldukça yeni bir öğrenme algoritması olan ve çeşitli alanlarda başarılı sonuçlar veren SVM; kredi derecelendirme, zaman serileri tahmini ve sigorta tazminat talebinde hile tespiti gibi çeşitli finansal uygulamalarda da etkili olmaktadır [19].

SVM, firmalara yönelik çalışmalar ağırlıkta olmak üzere, iflas tahmininde de kullanılmıştır. Bu çalışmalarda, SVM'nin diskriminant analizi, lojistik regresyon ve geri beslemeli sinir ağlarından daha iyi sonuç verdiği gözlemlenmiştir [15]. Literatürde artık tek başına istatistiksel yöntemler kullanmak yerine bu tekniklerle birlikte sinir ağları, yöneylem araştırması, bulanık mantık ve SVM gibi yöntemler de kullanılarak daha etkili sonuçlar elde edilmektedir.

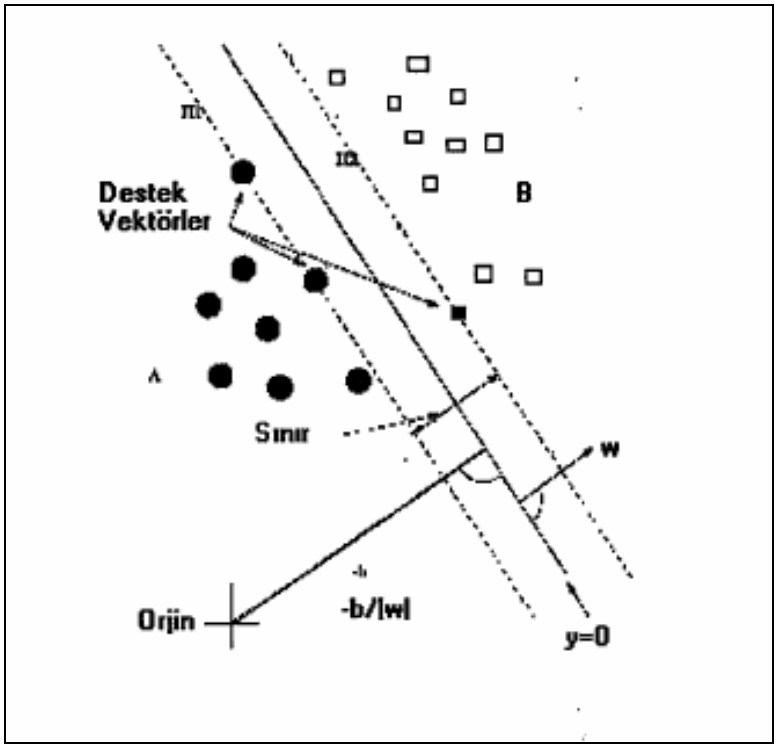
SVM, istatistiksel öğrenme teorisi üzerine inşa edilmiş bir yöntemdir. Modelde eğitim kümesindeki hata ile VC (Vapnik-Chervonenkis) boyutuna göre ifade edilen hipotez uzayının karmaşıklığını azaltan bir çözümün bulunduğu savunulmaktadır. Bu bakımdan, SVM ile bulunan fonksiyon, veriye yakınlık ve çözümün karmaşıklığı arasındaki geçiştir.

SVM'de amaç, veri noktalarını mümkün olduğu kadar iyi sınıflandıran ve mümkün olduğu kadarıyla iki sıfır noktaya ayıran optimum ayırıcı düzlemin bulunmasıdır. Yani iki sınıf arasındaki uzaklığın maksimum temel taşları olduğu durumun bulunmasıdır. Bu sınıflandırma mantığının temel taşlarını ise her iki sınıfın uç noktalarında bulunan ve eğitme örneklerinin arasından seçilen destek vektörleri

oluşturmaktadır. Ayırıcı düzlemin optimum olması genelleme yeteneğinin de maksimum düzeyde olmasını sağlayacaktır. Fakat verileri iki sınıfa ayıracak çok sayıda ayırıcı düzlemin olması bu düzlemler üzerinde bazı ön işlemlerin yapılmasını gerektirmektedir. Öncelikle ayırıcı düzlem sayısı sınıflar arası mesafe ölçüsüyle sınıf genişliği ölçüsünün çarpımı 1 alınarak sınıflandırılır. Sonra bu ayırıcı düzlemler arasında optimum olanının bulunması gerekir. Optimum ayırıcı düzlem, her iki sınıfın en uç verileri arasındaki mesafenin (iki sınıfın destek vektörleri arasındaki mesafe) maksimum olduğu durumu sağlayan ayırıcı düzlemdir. Bu ayırıcı düzlem bahsedilen aralığın tam ortasından geçmektedir.

SVM, optimum ayırıcı düzlemi bulurken önce optimizasyon problemini formülleştirir daha sonra ikinci dereceli programlama adı verilen yöntemleri kullanarak problemi çözer. SVM öğrenme yöntemi teorik olarak VC boyutu ve risk minimizasyonu prensiplerine uygun olarak tasarlanmıştır.

Matematiksel olarak SVM, dağılımdan bağımsız formül yapısı üzerine kuruludur. SVM tekniğinde önce eğitim verileri alınıp SVM eğitilerek sınıflandırıcı model oluşturulur. Daha sonra çıkış değerini önceden bildirdiğimiz test verileri için sistemin hesaplayacağı çıkış değerleri belirlenir. Bu iki değer arasındaki farklılık oranına göre SVM'nin sınıflandırma performansı değerlendirilir.



Şekil 3.2. SVM sınıflandırıcısı.

Şekil 3.2'de SVM sınıflandırıcısının yapısı gösterilmektedir. Amaç optimum ayırıcı düzlem vektörünün bulunmasıdır. Giriş verileri (+1,-1) değerlerini alan iki sınıfa ayrılır. Destek vektörleri bulunduğu sınıfın orijine en yakın olan elemanlarıdır. Ayırıcı düzlemin optimum olması için destek vektörleriyle ayırıcı düzlem arasındaki mesafenin maksimum olması gerekir [6].

SVM modelinde kernel özelliği yeralmaktadır. Clementine programında model için belirlenen kernel seçeneği lineer, polynomial ve simple (kernel'in kullanılmadığı durum) olmak üzere üç ayrı tiptedir. Farklı kernel tipleri, ayırıcının farklı yollarla hesaplanabilmesini sağlar. Bu nedenle genellikle uygulamalarda farklı kernel tiplerinin denenmesi önerilir. Kavramsal olarak kernel fonksiyonları orijinal veriyi daha yüksek bir boyut uzayına taşır ve girdi veri kümesini, dönüştürülmüş uzayda lineer olarak ayrılabilir hale getirir. Kernel fonksiyonlarının seçimi büyük ölçüde uygulamaya bağlıdır. Uygulamada kullanılan veri setinden hassas sonuçların alınabilmesi için kernel özelliği lineer, polynomial ve simple olduğu durumlardaki SVM modelleri ele alınmıştır.

3.2.2. REGRESYON ANALİZİ

Regresyon analizi bir ya da daha çok değişkenin başka değişkenler cinsinden tahmin edilmesini sağlayacak ilişkiler bulmak ve bunları tanımlamaktır. Regresyon analizinin temelinde gözlenen bir olay değerlendirilirken bu olayın hangi olaylardan etkilendiğinin belirlenmesi yatmaktadır. Bu olaylar bir ya da birden çok olabileceği gibi etki düzeyleri farklı seviyelerde de olabilirler. Regresyon analizinde verilerin matematiksel bir fonksiyon şeklinde tanımlanması gerekmektedir. Oluşturulan matematiksel fonksiyon bir bağımlı değişken ve bir veya birden çok bağımsız değişkenden oluşabilir. Değişkenler sayılabilir veya ölçülebilir nitelikte olabilir. Tek değişkenli modeller basit doğrusal regresyon, birden fazla bağımsız değişkenli modeller ise çoklu regresyon modelleri olarak adlandırılır [1]. Basit doğrusal regresyon ise kendi içinde doğrusal ve doğrusal olmayan olmak üzere ikiye ayrılır.

Basit doğrusal regresyon iki değişken arasındaki ilişkiyi tanımlamayı amaçlar. Yöntemde amaç, eldeki verileri kullanarak bir doğru oluşturmaktır. Doğru tüm veri noktalarından tahmin edilen eğriye olan uzaklığın kareleri minimize edilmek

suretiyle optimize edilir. Doğru elde edildikten sonra iki değişken arasındaki ilişkinin gücü R-kare değeri ile ölçülür.

Doğrusal olmayan regresyon analizi ise bağımlı ve bağımsız değişkenler arasındaki ilişkinin doğrusal olmadığı durumlarda kullanılan yöntemdir. Yöntemde daha iyi bir uyumun elde edilebilmesi için çeşitli algoritmalar kullanılarak bağımsız değişkenler değişime uğratılır.

Birden fazla bağımsız değişkenin bağımlı değişken üzerindeki etkisi incelenmek istendiğinde ise çoklu regresyon adı verilen yöntem kullanılır.

3.2.2.1. Loess (Lowess) Regresyon Analizi

Loess lokal olarak tahmin edilmiş saçılım grafiği düzlemesi, Lowess ise lokal olarak ağırlıklandırılmış saçılım grafiği düzlemesi anlamına gelmektedir. Loess parametrik olmayan regresyon tekniklerinin en esnek olanlarından biridir. Düzeltme/düzleme (smoothing) fonksiyonunu kullanarak gürültüyü azaltıp değişkenler arasındaki ilişkiyi en düşük düzeyde kabul ederek genel örüntüleri yakalamaya çalışır. Loess genel olarak trendleri gözlemlemenin zor olduğu çok büyük veri setlerinde iki değişken arasındaki ilişkiyi görsel olarak göstermek için kullanılır [5].

Loess regresyon analizinde, X ekseninde belli bir genişlikte olan noktalar tahmin edilmiş noktalara karşı gelecek şekilde seçilir ve verinin alt kümesinde düşük dereceli bir polinomsal denklem oluşturur. Analizde tahmin edilmiş değerlere daha yakın noktalara daha çok ağırlık değeri verilmektedir. Bu sonuç denklemi ilgili nokta için tahmin etmede kullanılır. Daha sonra veri bir birim sağa kaydırılır ve süreç ikinci nokta için yeni bir tahminin oluşturulması şeklinde devam eder. Elde edilen sonuç noktaları bir çizgi ile birleştirilir. Bant aralığı ne kadar küçükse o kadar az nokta kullanılıyor demektir ve final çizgisi de o kadar az düzleştirilir.

Loess regresyon analizin avantajları aşağıdaki ekilde sıralanabilir.

- Basit ve esnek olması
- Değişkenler arası ilişki anlamında kabullerin olmaması
- Kompleks ilişkileri görsel olarak gösterebilme özelliği

- Kullancıların gerektiğinde yeni değerleri fit etmesi ve modeli güncelleyebilmesi.

Loess regresyon analizin dezavantajları ise şöyle sıralanabilir:

- Çok yoğun datasetlerine ihtiyaç duyması
- Hazır bir formulasyonu olmaması dataların transportunun zor olması
- Karmaşık işlemlere sahip olmasına rağmen bilgisayar programları tarafından kolaylıkla hesaplanabiliyor olması.

3.2.2.2. Lojistik Regresyon Analizi

Lineer regresyona benzeyen lojistik regresyonda bağımlı değişkenin kesikli veya kategorik olması (sürekli olmaması) en önemli farklılığıdır. Bu fark özellikle bir teklife yanıt ya da çalışmada olduğu gibi, bir banka müşterisinin aldığı krediyi batırıp batırmadığının araştırılmaya çalışıldığı kesikli aksiyonları belirlemeye yönelik sınıflandırma modellerinde önem kazanmaktadır. Sınıflandırma analizlerinde doğrusal regresyonun kullanılması mümkün değildir. Bunun yerine kullanılan lojistik regresyonda, çok değişkenli normal dağılım varsayımına ihtiyaç duyulmadığından bu tür uygulamalarda avantaj sağlamaktadır. Lojistik regresyon ile bağımsız değişkenler kullanılarak ikili çıktısı olan bağımlı değişkenin istenilen durumunun gerçekleşme olasılığı hesaplanır. Regresyon yapabilmek için bağımlı değişken sürekli değere dönüştürülür. Bu değer beklenen olayın olma olasılığıdır [10].

3.3. TANIMLAMA FONKSİYONLARI

Tanımlama fonksiyonlarının amaçları belli bir hedefi tahmin etmek değil, veri setinde yer alan veriler arasındaki ilişkileri, bağlantıları ve davranışları bulmaktır. Mevcut verileri yorumlayarak davranış biçimleri ile ilgili tespitler yapmayı ve bu davranış biçimlerini gösteren alt veri setlerinin özelliklerini tanımlamayı hedefler. Bu sayede tanımı bilinen yeni bir verinin yapıya katıldıktan sonra ne şekilde hareket edeceği konusunda karar almaya destek olur.

3.3.1. KÜMELEME ANALİZİ

Kümeleme analizinde amaç, nesnelerin belirlenen özellikler doğrultusunda kümelere ayrılmalarını sağlamaktır. Benzer işleve sahip olan kümeleme analizini sınıflandırma analizlerinden ayıran en önemli fark, kümeleme analizinde önceden tanımlanmış sınıfların olmayışıdır [7].

Kümeleme işlemi çoğunlukla VM'nin diğer yöntemleri veya modelleme çeşitleri için bir ön analiz olarak kullanılır. Örneğin pazarda müşterilerin segmentasyonunu belirlerken kümeleme ilk adım olabilir.

3.3.2. BİRLİKTELİK KURALI

Büyük veri kümelerinde yeralan farklı veriler arasındaki birliktelik ilişkilerinin bulunabilmesi için yapılan bir analizdir. Analizde amaç, belirli bir veri kümesinde yüksek sıklıkta birlikte görünen özellik değerlerine ait ilişkilerin keşfedilmesidir. Analiz sonucunda elde edilen birliktelik kuralları ile şirketlerin daha verimli karar almaları sağlanmaktadır.

Birliktelik kuralının en klasik örneği, müşterilerin beraber satın aldığı ürünlerin belirlenmesini sağlayan sepet analizidir. Analizde amaç ürünler arasındaki negatif ya da pozitif korelasyonları bularak müşterilerin satın alma alışkanlıklarını ortaya çıkarmaktır. Örneğin, çocuk bezi satın alan müşterilerin %30'u süt de almaktadır. Bu verilere sahip olan marketler, birlikte satılan ürünleri yakın raflara koyarak ya da ürün kataloglarında bu ürünlerin birlikte görülmesini sağlayarak müşteriler için cazip ürün paketleri oluşturmak suretiyle satışlarını artırabilirler.

3.3.3. SIRALI DİZİ ANALİZİ

Gözlem sonuçlarını zaman ve mekan özelliklerine göre sıralanmış biçimde gösteren yapılara sayı dizileri, bu dizilerdeki trendleri ve döngüleri anlamak için kullanılan analize ise sıralı dizi analizi denir. Sıralı dizi analizinde ilişki kayıtları incelenerek zaman içinde sıkça rastlanan trendler ve benzer trendler bulunur. Bu trendler daha sonra veri içindeki ilişkileri tanımlamak için kullanılır. Örneğin, aldığı kredinin ilk üç taksitinin iki veya daha fazlasını geç ödemiş olan müşterilerin %70 olasılıkla kanuni takibe gitmesi, bu analiz sonucunda ortaya çıkar.

Seriler özelliklerine göre zaman serisi, mekan serisi, bölünme serisi ve bileşik seri olmak üzere dört grupta incelenebilir [10]. Seride gözlem sonuçlarının zamana göre sınıflandırılması zaman serilerini, mekana göre sıralanması mekan serilerini, kriterlere göre sıralanması frekans serilerini, iki ya da daha fazla özelliğe göre bir arada gösterilmesi ise bileşik serileri oluşturur.

BÖLÜM 4

VERİ MADENCİLİĞİ SÜREÇ MODELLERİ

4.1. SÜREÇ MODELLERİNE GİRİŞ

VM'nde süreç modelleri akademik, endüstriyel ve karışık süreç modelleri olmak üzere üç grupta incelenir. Aşağıdaki alt kesimlerde süreç modellerinin ayrıntıları üzerinde durulmuştur

4.2. AKADEMİK SÜREÇ MODELİ

Bilgi keşfi modellerinin geliştirilmesine yönelik ilk çalışmalar akademik alanda başlamıştır. VM alanının şekillendiği 90'lı yılların ortasında araştırmacılar, VM yöntemlerini karmaşık bilgi keşfi sürecini kullanan kişilere yol göstermesi amacıyla çok aşamalı prosedürlerden oluşturmuşlardır. Araştırıcıların yöntemleri çok aşamalı prosedürler halinde geliştirmelerinin nedeni, bilgi keşifi modellerinin uygulama alanın yürütülmesine yardım edecek aktivite düzenini oluşturmaktır.

1996 yılında Anand ve Buchner tarafından sekiz ve 1998 yılında Fayyad tarafından dokuz aşamadan oluşan iki akademik süreç modeli geliştirilmiştir. Aşağıda bu modellerden araştırma alanında öncü olduğu kabul edilen Fayyad'ın yöntemi verilmiştir.

Uygulama alanın belirlenmesi ve geliştirilmesi aşaması: Konu ile ilgili ön bilgilerin elde edildiği ve son kullanıcıların hedeflerinin öğrenildiği aşamadır.

Hedef veri setinin yaratılması aşaması: Veri madencisinin, keşif sürecinde kullanılacak değişkenlerin alt kümelerini belirlediği ve örneklemini seçtiği aşamadır. Bu aşamada alt kümeler çoğunlukla mevcut verinin belirlenen hedef doğrultusunda sorgulanması ile belirlenir.

Veri temizleme ve ön işleme aşaması: Bu aşamada dışa düşen veriler giderilir, verideki gürültü ve eksik değerlere atama yapılır ve zaman serisinden doğan bilgi ve değişiklikler dikkate alınır.

Veri indirgeme ve gözlemleme aşaması: Boyut indirgeme ve dönüştürme yöntemleri ile veri setindeki değişkenlerin gerekli ve gereksiz değişkenler olmak üzere ayrıştırıldığı aşamadır.

VM modelinin seçimi aşaması: Veri madencisinin, birinci aşamada belirlenen hedeflere ulaşabilmesi için VM yöntemlerinden (sınıflandırma, regresyon, kümeleme vb.) hangisi ya da hangilerinin kullanımlarının uygun olacağını belirlediği aşamadır.

VM algoritmasının seçilmesi aşaması: Veri madencisinin veri setindeki örüntüleri tespit edecek yöntemleri seçtiği ve bu yöntemlara uygun parametreleri tanımladığı aşamadır.

VM aşaması: Veri gösterim biçimlerinin (sınıflandırma kuralları, karar ağaçları, regresyon modelleri, eğilimler vb.) oluşturulduğu aşamadır.

Tespit edilen örüntülerin yorumlanması aşaması: Belirlenen modellerde kullanılacak verinin görselleştirilmesinin yapıldığı aşamadırr.

Keşfedilen bilginin konsolide edilmesi aşaması: Keşfedilen bilginin oluşturulan bir performans sistemine uygulanması aşamasıdır. Bu aşamada keşfedilen bilgili ile ilgili dokümantasyon ve raporlama gerçekleştirilir. Ayrıca konu ile ilgili önceki inanışların mevcut durum ile olan benzerlik ve farklılıkları tespit edilir.

Akademik süreç modeli, endüstriyel konularla ilgilenmese bile kolaylıkla endüstriyel alanlara uygulanabilir [4].

4.2. ENDÜSTRİ SÜREÇ MODELİ

Süreç modellerineki akademik çalışmaları kısa bir zaman sonra endüstri alanındaki çalışmalar takip etmiştir. Bu dönemde geçmiş endüstriyel tecrübesi olan bireylerce önerilen yaklaşımların yanı sıra büyük endüstriyel konsorsiyumlara kadar geniş bir yelpaze tarafından önerilen farklı yaklaşımları da içeren modeller geliştirilmiştir.

Endüstri süreç modelleri iki önemli modeli içermektedir. Modellerden biri, IBM firmasının desteği ile Cabena tarafından geliştirilen beş aşamalı model iken diğeri ise 1990'larda Avrupalı dört şirketlerce oluşturulmuş büyük bir konsorsiyum tarafından gerçekleştiren altı aşamalı endüstriyel CRISP-DM (Cross-Industry Standard Process for Data Mining/Endüstürüler arası VM için standart süreç) modelidir. CRISP-DM

süreç modeli zamanla öncü model konumuna gelmiştir. Konsorsiyumda bulunan şirketlerin ortak özelliği veri ve gözlem çalışmaları üzerinde uzmanlaşmış olmalarıdır. Bu modelin geliştirilmesi sürecinde önemli bir endüstriyel destek sağlanmıştır. Aynı zamanda Avrupa Biriliği tarafından geliştirilmiş olan ESPRIT programınca da mali destek görmüştür. Bu model hali hazırda üç yüzden fazla kurumsal kullanıcı ve servis sağlayıcıya sahiptir. CRIPS-DM modelini oluşturan altı aşama aşağıda özetlenmiştir.

İşin anlaşılması aşaması: Bu aşamada şirket gözü ile hedeflerin ve gereksinimlerin anlaşılmasına odaklanılmaktadır. Bu aşamanın bulguları aynı zamanda VM modelini tanımlanmakta ve amaçlara ulaşmayı sağlayacak ön proje planı için VM modeline çevrilmektedir. İşin anlaşılması aşaması aşağıda sıralanan alt aşamalara ayrılır.

- İş hedeflerinin belirlenmesi
- Durumun tespit edilmesi
- VM hedeflerinin belirlenmesi
- Proje planının yaratılması

Verinin anlaşılması aşaması: Bu aşama verinin toplanması ve verinin anlaşılması ile başlar; veri kalite sorunlarının tanımlanması, verinin ön gözlemlerinin yapılması ve önemli veri alt kümelerinin tespit edilmesi ile devam eder. Aşağıda verinin anlaşılması aşamasının alt aşamaları sıralanmıştır.

- Başlangıç verinin toplanması
- Verinin tanımlanması
- Veride ön gözlem gerçekleştirmesi
- Veri kalitesinin doğrulanması

Veri ön işleme aşaması: Nihai veri setinin oluşturulması için yapılan tüm işlemleri içeren aşamadır. Tablo, kayıt, özellik seçimi, veri temizleme, yeni özelliklerin yaratılması ve veri dönüştürme işlemlerinin gerçekleştirildiği aşama sonunda VM

programında kullanılacak nihai veri elde edilmiş olur. Veri ön işleme aşamasını oluştural alt aşamalar aşağıda sıralandığı gibidir.

- Veri seçimi
- Veri temizleme
- Veri setinin kurulması
- Veri entegrasyonu
- Veri alt kümelerinin biçimlendirilmesi

Modelleme aşaması: Modelleme, daha çok aynı VM problemi için farklı yöntemlar geliştirilmesi ve sonuçları optimize eden parametrelerin kalibre edilmesi süreçlerini içermektedir. Bu süreçler bazı yöntemlerin belirli tipte girdi verileri gerektirmesinden dolayı çoğunlukla döngüsel bir yapı içermektedirler. Modellemenin alt aşamaları aşağıda sıralandığı gibidir.

- Modelleme tekniklerinin seçilmesi
- Test tasarımlarının yaratılması
- Modellerin geliştirilmesi
- Geliştirilen modellerin değerlenmesi

Değerlendirme aşaması: Yüksek kaliteli veri analizini sağlamak amacı ile çeşitli modeller kurulduktan sonra bu modellerin belirlenen iş hedeflerini gerçekleştirmedeki başarısının değerlendirildiği aşamadır. Aşamanın işleyişi modellerin kurulmasında kullanılan şemaların tekrar gözden geçirierek kurulan modellerin iş hedeflerini karşılama noktasındaki yeterliliğinin tespit edilmesi şeklindedir. Değerlendirme aşamasının sonunda VM modelinden elde edilecek sonuçların hedeflerimize uygunluğu konusunda karar sahibi olunur. Aşama aşağıda sıralanan alt aşamalara ayrılır.

- Sonuçların değerlendirilmesi
- Sürecin gözden geçirilmesi

- Bir sonraki basamağın belirlenmesidir

Konumlama aşaması: Bu aşmada keşfedilen bilgi son kullanıcının kullanımına uygun şekilde organize edilip sunulur. Konumlama aşaması gereksinimlere bağlı olarak raporlama gibi kolay bir aşama olabileceği gibi bilgi keşif modeli sürecinin yenilenmesi gibi zor bir aşamada olabilir. Bu aşmanın alt aşamaları aşağıdaki gibidir.

- Plan konumlandırma,
- Plan görüntüleme ve bakım
- Son raporun oluşturulması
- Sürecin alt aşamalarının gözden geçirilmesidir.

4.3. KARIŞIK SÜREÇ MODELİ

Akademik ve endüstriyel modellerin kurulmasını izleyen süreçte her iki yöntemin yaklaşımlarını da kullanan karışık yani melez modeller ortaya çıkmıştır. Bunlara örnek olarak Cios tarafından geliştirilen bir altı aşamalı bilgi keşfi modeli gösterilebilir. Bu modelin geliştirilmesinde CRISP-DM modeli baz alınmış ve akademik gereksinimlere göre model uyarlanmıştır. CRISP-DM modeli ile karışık süreç modeli arasındaki temel farklılıklar aşağıdaki gibidir.

- Karışık süreç modelinde süreç basamakları daha genel ve araştırma odaklıdır
- Modelleme aşaması yerine VM aşaması getirilmiştir
- Karışık modelde bir takım dışsal geri bildirim mekanizmaları uygulanmıştır
- Belirli bir alana yönelik keşfedilen bilgiler başka bir alanda kullanılabilinmektedir.

Karışık süreç modelini oluşturan altı aşama aşağıda sıralanmıştır.

Problem alanının anlaşılması aşaması: Bu aşamada uzmanlarla işbirliği yapılarak kısıtları içerecek şekilde problemin tanımı, projenin hedefleri, soruna ilişkin mevcut çözüm yaklaşımları ve bu kullanılan terminolojinin öğrenilmesi işlemleri

gerçekleştirilir. Daha sonra proje hedefleri VM hedeflerine çevrilir ve sonraki aşamalarda kullanılmak üzere bir başlangıç VM modeli seçilir.

Verinin anlaşılması aşaması: Örneklem verisinin toplanarak veriye ilişkin biçim ve boyut gibi bazı özelliklerin belirlendiği aşamadır. Bu aşamada eldeki mevcut veri seti kullanılabilirlik, eksik ya da kayıp veriler ile gereksiz verilerin olup olmadığı açısından incelenir. İnceleme sonucunda veri setinin VM hedeflerine yönelik bir veri seti olduğunun doğrulanması gerekir.

Verinin hazırlanması aşaması: Bu aşamada takip eden VM sürecinde hangi verilerin girdi olarak kullanılacağı tespit edilir. Veri bütünlüğünün kontrol edilmesi, gürültünün ve kayıp verilerin giderilmesi gibi işlemleri gerçekleştirmede kullanılan örnekleme, korelasyon ve anlamlılık testleri ve çeşitli veri temizleme yöntemleri kullanılır. Temizleme işleminin ardından verinin seçimi ve ayıklanması işlemleri ile veri setinin boyutu indirgenir, veri setine ilişkin yeni özellikler türetilir ve özetlemeler yapılır. Sonuç olarak elde edilen veri, VM girdi gereksinimlerini karşılayacak düzeye getirilir.

VM aşaması: Bu aşamada çeşitli VM modelleri kullanılarak işlenmiş veriden bilgi üretilir.

Keşfedilen bilgilerin değerlendirilmesi aşaması: Değerlendirme aşamasında uzman yorumu ile sonuçların anlaşılması, elde edilen sonuçların yeni ve kayda değer bir bilgi içerip içermediği ve keşfedilen bilginin olası etkileri tespit edilir. Onaylanmış modeller ayrılarak tüm modelin iyileştirilmesi için yapılması gereken alternatif aksiyonlar belirlenir. Ayrıca süreçteki hatalara ilişkin bir liste oluşturulur.

Keşfedilen bilginin kullanılması aşaması: Keşfedilen bilginin nerede ve nasıl kullanılacağına ilişkin bir planlamanın yapıldığı aşamadır. Mevcut konunun uygulama alanı dışındaki alanlarda da kullanılabilmesi için çeşitli düzenlemeler yapılıp keşfedilen bilginin etkisini incelemek üzere bir plan geliştirilir. Daha sonra tüm projenin belgelenmesi sağlanıp elde edilen bilgiler uygulanmaya konur.

BÖLÜM 5

BİREYSEL MÜŞTERİLERİN KREDİ DEĞERLENDİRİLMESİNE YÖNELİK SCORECARD UYGULAMASI

5.1. BANKACILIK VE FİNANS SEKTÖRÜNDE VERİ MADENCİLİĞİ

Bilgi teknolojisi sadece servis kalitesini geliştirmek için değil bunun dışında rekabet avantajı kazanmak için de kapsamlı olarak kullanılmaktadır. Bankacılık endüstrisi, müşterileri hakkında sahip oldukları bilgilerin öneminin farkına varmıştır. Bankalar tarafından yıllardır toplanan yüksek miktardaki veriler, manuel olarak incelenemeyecek bir hal almış ve VM teknolojileri için büyük bir fırsat olmuştur. Aşağıda bankacılık ve finans endüstrilerinde VM'nin kullanıldığı alanlardan bazıları sıralanmıştır.

- Müşteri kredi başvurusu değerlendirme (Scorecard modelleri)
- Farklı finansal göstergeler arasındaki saklı korelasyonların bulunması
- Müşteri segmentasyonu
- Sadık müşterilerin belirlenmesi
- Bankadan ayrılacak olan müşterilerin öngörülmesi
- Müşteri karlılığı
- Kampanya yönetimi
- Çapraz / dikey satış
- Ürün yönetimi
- Fiyatlandırma
- Risk yönetimi

Bank of America VM'ni, hangi müşterilerin hangi ürünü kullandığını tespit etmek ve böylece müşterilere doğru ürün ve servisi önerebilmek için kullanmıştır [10].

New York'taki Chase Manhattan Bankası'nın müşterileri rakip bankalara gitmeye başlayınca müşteri hesaplarını analiz ederek kendi hesap gereksinimlerinde değişiklikler yapabilmek için VM'ni kullanmış ve bu sayede karlı müşterilerini elde tutmayı başarmıştır [9].

Türkiye'deki büyük bankalar da VM tekniklerini bir çok alanda kullanmaktadır. Özellikle müşterilere verilecek krediler için scorecard modelleri geliştirmede VM sık kulanılmaktadır. Ayrıca bankacılık ürünlerinin satışlarını telefonda yapabilmek için müşteri profilleri VM yöntemleri ile analiz edilerek doğru müşteriye doğru ürünün satılması sağlanmaktadır.

5.2. SPSS CLEMENTINE

Clementine, ilk olarak 1989 yılında Dr. Alan Montgomery ve beş meslektaşı tarafından kurulmuştur. Başlangıçta hedefi teknik uzmanlara ihtiyaç duymadan çevreyle entegre VM operasyonlarının iş adamları tarafından, anlaşılabilir hale getirmekti. Clementine 1992 yılında ürün araştırmalarına başladı ve ilk ürününü 1994 yılında pazara sundu. Clementine, iş zekasında makine kullanımını getiren ilk ürünlerden biri olmakla birlikte uzman olmayan kişiler tarafından da anlaşılabilir bir ara yüze sahiptir. Sonuç olarak pazardaki lider konumunu yakalamış ve günümüze kadar getirmiştir. SPSS 1998 yılında ISL firması tarafından devralınınca Clementine için yeni olanaklar açılmıştır [10].

5.2.1. SPSS CLEMENTINE'DE KULLANILAN MODELLER

Modelleme ve görüntüleme SPSS Clementine'nin kalbidir. CRIPS-DM süreç modeli ile desteklenen SPSS Clementine, istatistiksel model ve teknikleri bir süreç olarak gösteren ve son kullanıcıya hitap eden bir programdır [10]. SPSS Clementine programında kullanılan modeller aşağıda sıralandığı gibidir.

- Otomasyon uygulayan modeller; Binnary Classifer, Numeric Predictor, Time Series.

- Sınıflandırma modelleri; C&R Tree, QUEST, CHAID, Decision List, Regression, PCA/Factor, Neural Network, C5.0, Feature Selection, Discriminant, Logistic, GenLin, Cox, SVM, Bayes Network, SLRM.
- Birliktelik modelleri; GRI, Apriori, Carma, Sequence.
- Bölümleme modelleri; K-Means, Kohonen Two Step, Anomaly.

5.2.2. SPSS CLEMENTINE'NIN AVANTAJLARI

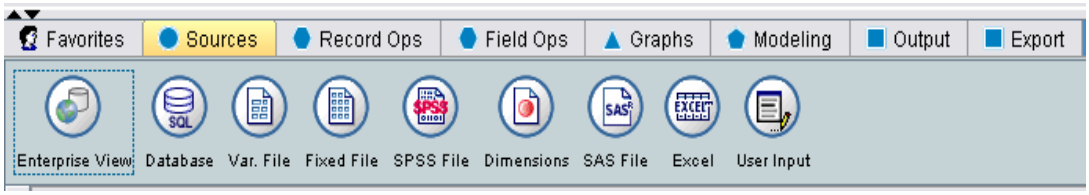
SPSS Clementine programının avantajları aşağıdaki şekilde sıralanabilir [15].

- VM teknikleri çok gelişmiştir
- En iyi modeli bulan araçlara sahiptir
- Bir model algoritmasının sonuçları bir başka model algoritmasında girdi olarak kullanılabilir
- Karmaşık değişken süreçler için yazı dili (script) mevcuttur
- Kullanımı kolaydır

5.2.3. SPSS CLEMENTINE'DA KULLANILAN MODÜLLER

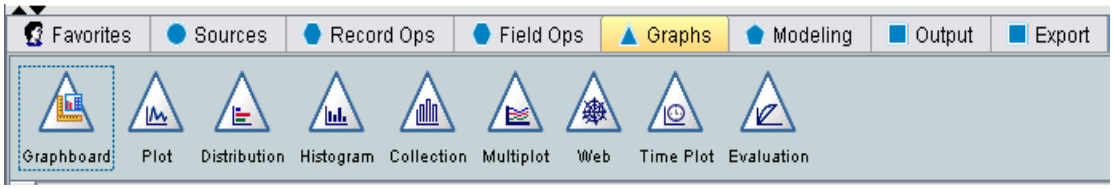
Veri işleme yani veri madenciliği programı olan SPSS Clementine programı oldukça gelişmiş modelleme ve grafik araçlarına sahiptir.

Şekil 5.1'de Clementine programının "Kaynaklar" kısmının yeraldığı görüntü bulunmaktadır. Buradaki seçeneklere göre veriler istenilen formatta Clementine'e ortamına aktarılabilir.



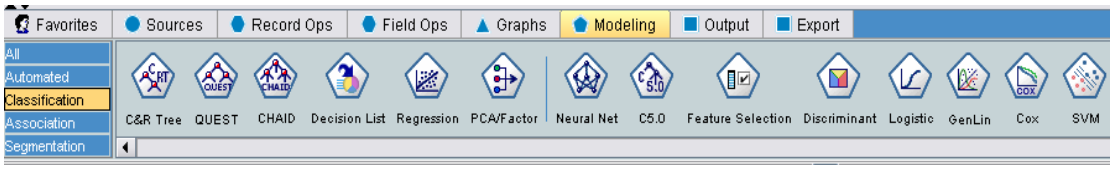
Şekil 5.1. Clementine programının "Kaynaklar" kısmı.

Şekil 5.2'de Clementine programının "Grafik" kısmının yeraldığı görüntü bulunmaktadır. Oldukça gelişmiş olan bu kısımda yer alan grafik modülleri ile veri seti grafiğe dönüştürülebilir.



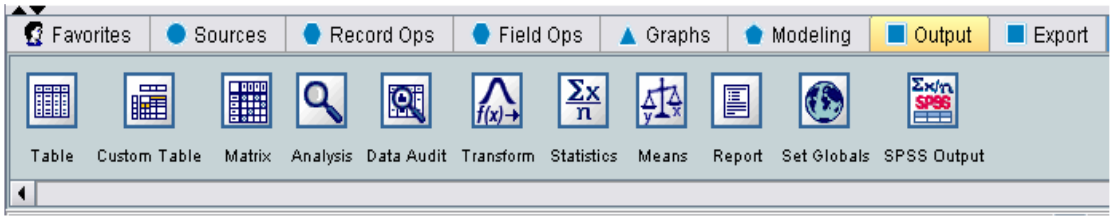
Şekil 5.2. Clementine programının "Grafik" kısmı.

Clementine programının "Model" kısmının yeraldığı görüntü bulunmaktadır. Oldukça gelişmiş olan bu kısımda bir çok model kullanılarak veri seti analiz edilebilir. Şekil 5.3'de sınıflandırma amacıyla kullanılan birkaç analiz modülü yer almaktadır.



Şekil 5.3. Clementine programının "Model" kısmı.

Şekil 5.4'te yeralan Clementine programının "Çıktılar" kısmı, yapılan analizler sonucunda elde edilen değerlerin sunumu ile ilgili araçlar sunar.



Şekil 5.4. Clementine programının "Çıktılar" kısmı.

5.3. SCORECARD UYGULAMASI

Bankacılık ve finans sektöründe en iyi istatistiksel araştırma modellerini kullanan programlar ile kredi scorecard analizi yapan firmaların sayısı gittikçe artmaktadır [10].

Bir bankanın bireysel müşterilerine ait kredi değerlendirmesinin yapıldığı bu çalışmada, kullanılacak en iyi scorecard modelinin belirlenmesinde endüstriyel süreç

modeli kullanılmıştır. Modeli oluşturan aşamalar izleyen alt bölümlerde ayrıntılandırılmıştır.

5.3.1. İŞİN ANLAŞILMASI AŞAMASI

Son zamanlarda bankalar kredi vermek için büyük rekabet içine girmişlerdir. Bankaların yurtdışı bankalar ile birleşmeleri bankaların kredibilitesini artırmış ve yurtdışındaki mevduat Türkiye'ye girmiştir. Yabancı bankaların topladığı mevduatı, Türk bankaları yurt içindeki müşterilerine kredi olarak satmaktadır. Bu nedenle bankalar kredi verirken çok dikkatli olmalı ve risklerini minimize edebilecek yöntemleri kullanmalıdırlar. Bu yöntemlardan en çok kullanılanı ise scorecard yöntemidur.

Çalışmanın amacı, kredi kullanmak isteyen bireysel müşteriler için banka açısından riski minimize edecek bir model oluşturmaktır.

5.3.2. VERİ'NİN ANLAŞILMASI AŞAMASI

Uygulamada kullanılan veri seti, bir bankanın bireysel müşterilerine ait olup 1000 adet bireysel müşteri ve bu müşteriler ile ilgili 22 değişkenden oluşmaktadır. Bu değişkenler müşterilerin demografik özellikleri ile kullandıkları kredinin özelliklerini belirtmektedir. Değişkenlerden 7'si sürekli 15'i ise kesikli türde olup, kesikli verilerden oluşan değişkenlerin 3 tanesi binominal veri içermektedir. 1000 müşteriden oluşan verinin 22 değişkeni ve bu değişkenlerin açıklamaları Tablo 5.1'deki gibidir.

Tablo 5.1. Değişkenlerin özellikleri.

Değişken Adı	Değişken türü	Açıklama
Yaş	Sürekli veri	Kredi kullananların yaşları
Kredi miktarı	Sürekli veri	Talep edilen kredi miktarı TL
Vadeli hesap bakiyesi	Nominal ya da ordinal veri (kategorik veri)	Bugünkü vadeli hesap bakiyesi: 1 = 0 TL'den az 2 = 0 ile 200 TL arası 3 = 200 TL'den fazla 4 = hesabı yok
Kefil	Nominal veri (kategorik veri)	Kefil durumu: 1 = kefili yok 2 = co-applicantri 3 = kefili var

Tablo 5.1. Değişkenlerin özellikleri (Devam).

Bakmakla yükümlü olduğu kişi sayısı	Sürekli veri	Bakmakla yükümlü olduğu kişi sayısı
Kredinin süresi	Sürekli veri	Kredinin aylık taksit sayısı
İş tecrübesi	Ordinal veri (kategorik veri)	Şu ana kadarki iş tecrübesi: 1 = işsiz 2 = 1 yıldan az 3 = 1 yıl ile 4 yıl arası 4 = 4 yıl ile 7 yıl arası 5 = 7 yıldan fazla
Bankadaki hesap sayısı	Sürekli veri	Bankada bulunan hesap sayısı
İşletme sahibi	Binominal veri	İşletme sahibi mi: 1 = işletmede çalışan 2 = işletme sahibi
Kredi geçmişi	Ordinal veri (kategorik veri)	Kredi tarihçesi: 0 = hiç kredi almamış 1 = bankamızdan aldığı tüm kredileri tam olarak zamanında geri ödemiş 2 = şu ana kadar aldığı tüm kredileri zamanında ödemiş 3 = aldığı kredileri gecikmeli ödemiş 4 = kritik durum / diğer bankalardan aldığı krediyi ödememiş
Oturduğu ev	Nominal veri (kategorik veri)	Oturduğu evin durumu: 1 = kiralık 2 = kendi evi 3 = para vermeden oturuyor
Gelir	Sürekli veri	Vergiler düştükten sonra aylık kazanılan net gelir
İşi	Ordinal veri (kategorik veri)	İşi: 1 = işsiz 2 = vasıfsız işçi 3 = vasıflı işçi / memur 4 = yönetici / iş sahibi / yüksek vasıflı işçi
Medeni durumu	Nominal veri (kategorik veri)	Medeni durumu ve cinsiyeti: 1 = evli ya da boşanmış, dul 2 = bekar
Cinsiyeti	Nominal veri (kategorik veri)	Cinsiyeti: 1 = erkek 2 = kadın
Diğer kredileri	Nominal ya da ordinal veri (kategorik veri)	Diğer kredileri: 1 = başka bir bankaya kredi ödemesi var 2 = mağzaya kredi ödemesi var 3 = kredi ödemesi yok

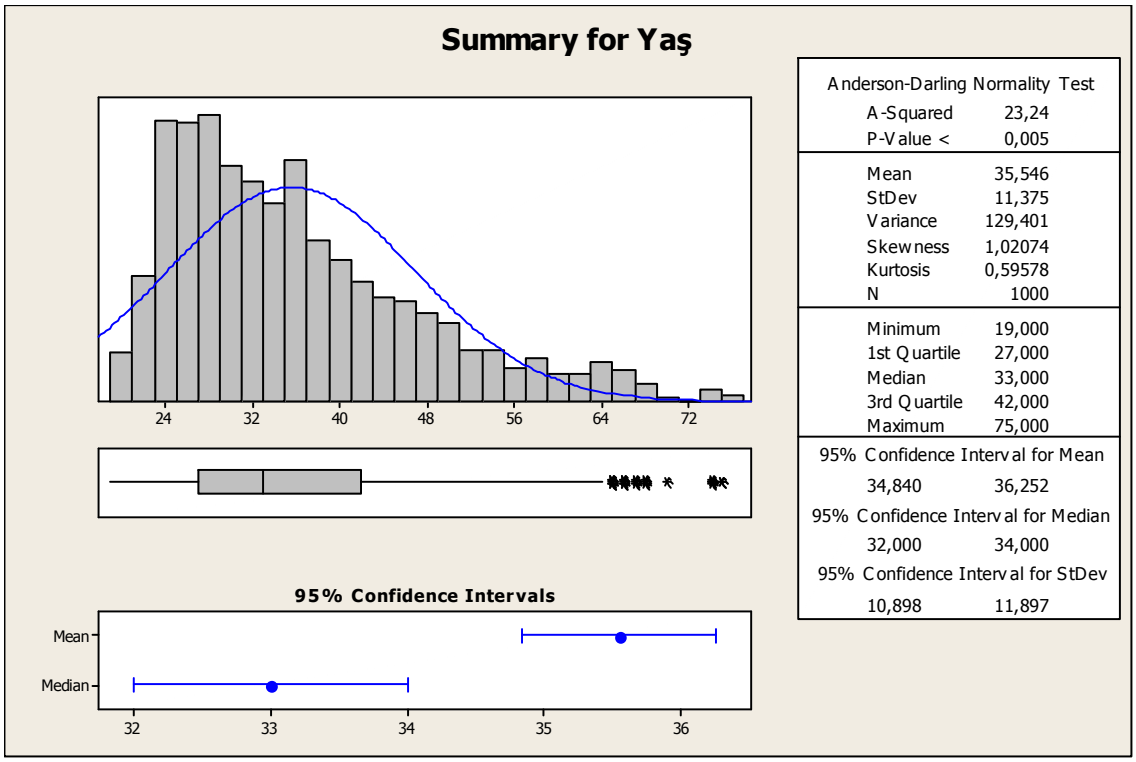
Tablo 5.1. Değişkenlerin özellikleri (Devam).

Varlıkları	Nominal ya da ordinal veri (kategorik veri)	Kredi için varlık durumu: 1 = taşınmaz gayrimenkulü var 2 = taşınmaz gayrimenkulü yok, ev almış ve ödemeye devam ediyor / hayat sigortası var 3 = gayrimenkul ve birikmiş parası yok, arabası var 4 = herhangi bir varlığı yok / bilinmiyor
Kredi alma nedeni	Nominal veri (kategorik veri)	Kerdi başvuru sebebi: 0 = yeni bir araba 1 = arabasında kullanacak 2 = beyaz eşya ya da mobilya 3 = elektronik eşya 4 = elekronik ev aletleri 5 = tamirat 6 = eğitim 7 = tatil 8 = mesleki eğitim 9 = iş 10 = diğer
Oturma süresi	Sürekli veri	Geçerli adresteki oturma süresi
Para birikimi	Nominal ya da ordinal veri (kategorik veri)	Para birikimi: 1 = 100 TL'den az 2 = 100 TL ile 500 TL arasında 3 = 100 TL ile 1000 TL arasında 4 = 1000 TL'den fazla 5 = herhangi bir birikimi yok / bilinmiyor
Telefon	Binominal	Bankada kayıtlı telefonu var mı: 1 = yok 2 = evet müşteri kaydının altında telefonu var
Değerlendirme	Binominal	Kredi rating durumu (iyi veya kötü)

Değerlendirme değişkeni müşterinin aldığı krediyi batırıp batırmadığını gösteren ve modele temel olan değişkendir.

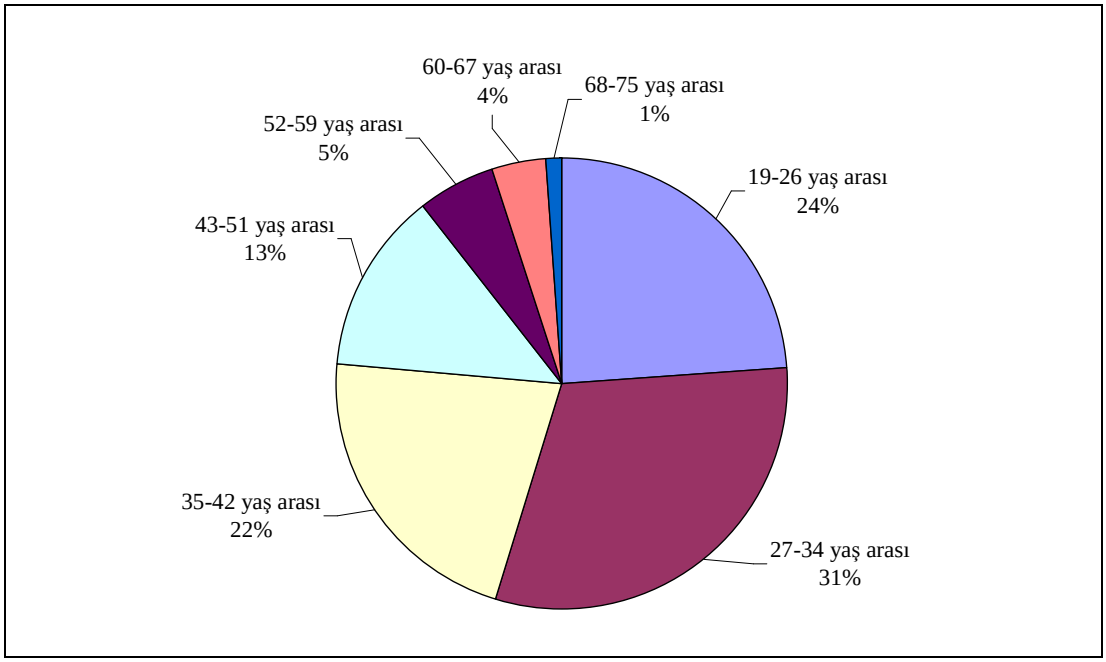
Aşağıda yaş, gelir düzeyi, cinsiyet,müşterinin varlıkları, kredi alma nedeni, kredi geçmişi değişkenlerine ilişkin bazı istatistiksel analizlere yer verilmiştir.

Yaş: Yapılan analiz sonucunda kredi alan 1000 müşterinin yaş dağılımlarına ilişkin elde edilen istatistiksel bilgiler Şekil 5.5'de gösterildiği gibidir. Buna göre müşterilerin yaş ortalamasının 35,546; standart sapmasının ise 11,375 olduğu ve yaşların normal dağılıma uymadığı görülmektedir.



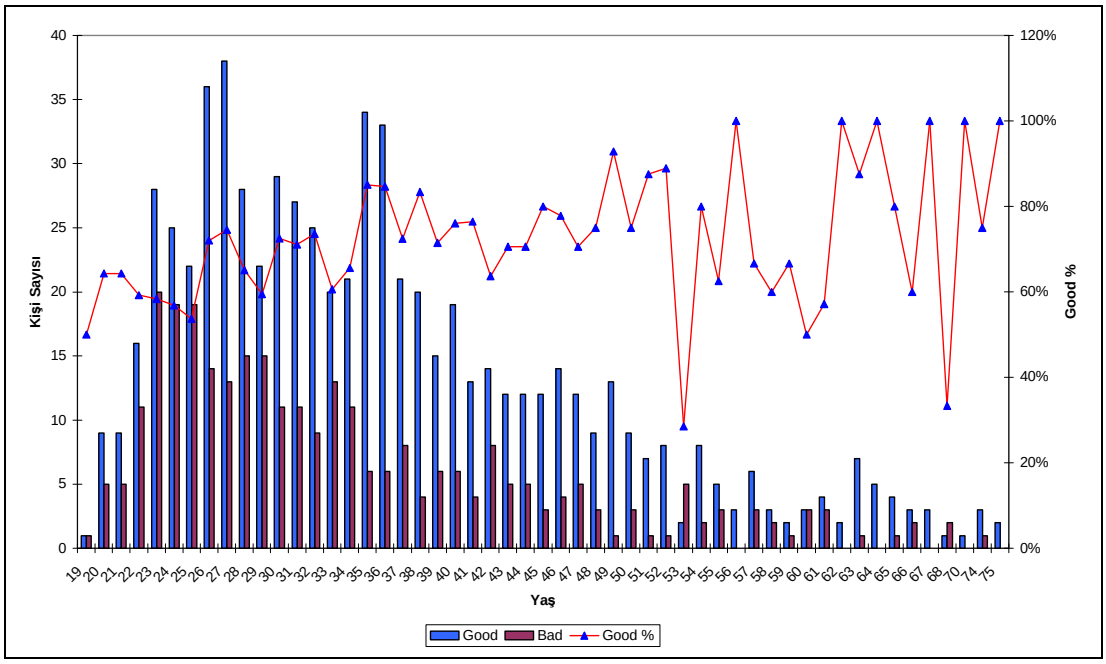
Şekil 5.5. Yaş dağılımının istatistiksel göstergeleri.

Şekil 5.6'da yaşların grafiksel dağılımı yeralmaktadır. Buna göre kredi kullanan müşterilerin %24'ünün 19 ile 26 yaşları arasında, %31'inin ise 27 ile 31 yaşları arasında olduğu görülmektedir.



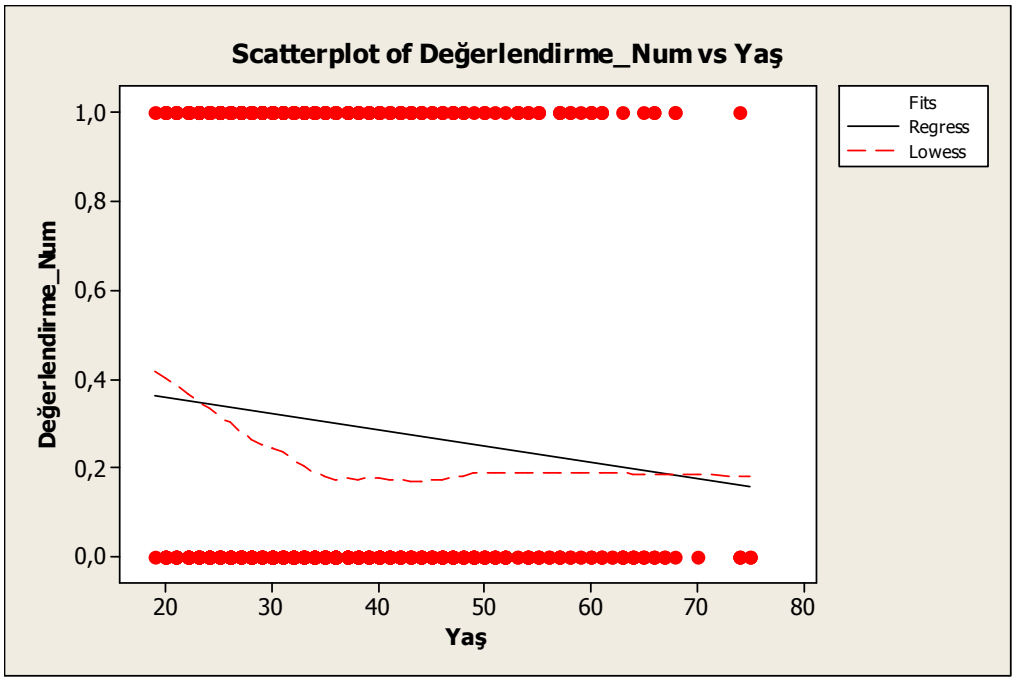
Şekil 5.6. Yaş dağılımı.

Yaş'a göre kredi değerlendirme sonuçlarının grafiksel olarak gösterildiği Şekil 5.7'den krediyi batırma oranı %71 olan müşterilerin yaşlarının 53 olduğu görülmektedir. En fazla kredi kullanan kişilerin yaşı ise 27 olup krediyi batırma oranları %25'tir.



Şekil 5.7.Yaş'a göre kredi değerlendirme sonuçları ile krediyi batırmama oranı.

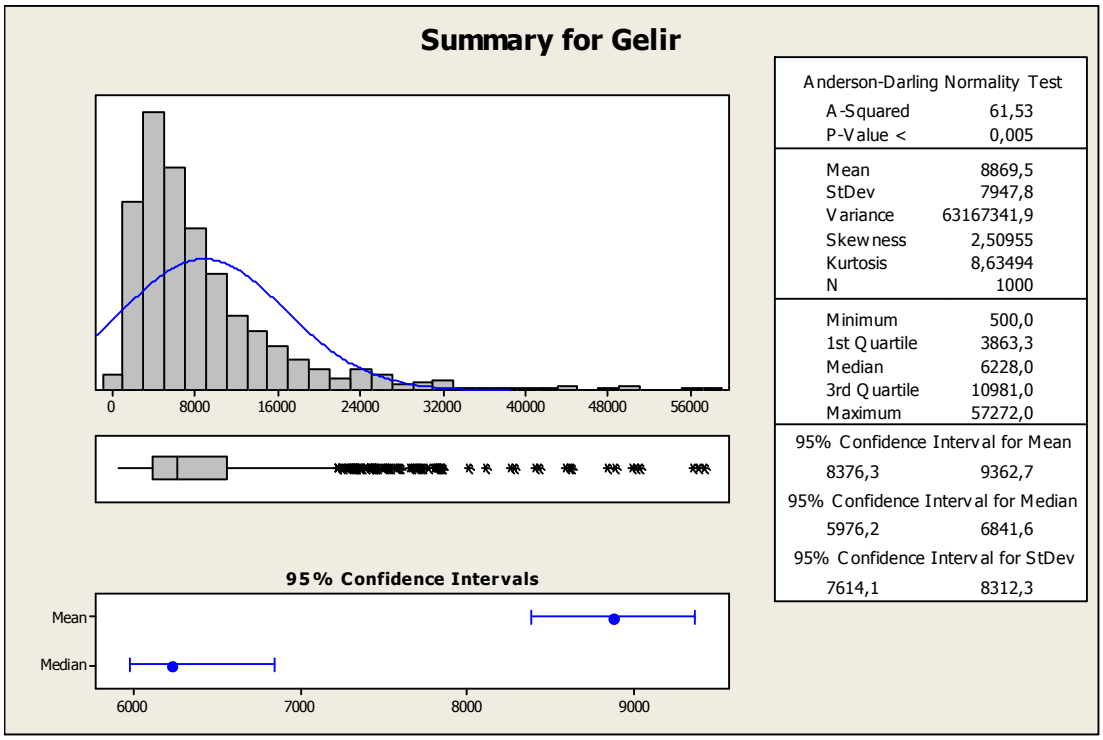
Yaş ile kredi değerlendirme sonuçları arasındaki ilişki lowess regresyonu ile de gösterilebilir.



Şekil 5.8. Yaş ile kredi değerlendirme sonuçları arasındaki lowess regresyonu.

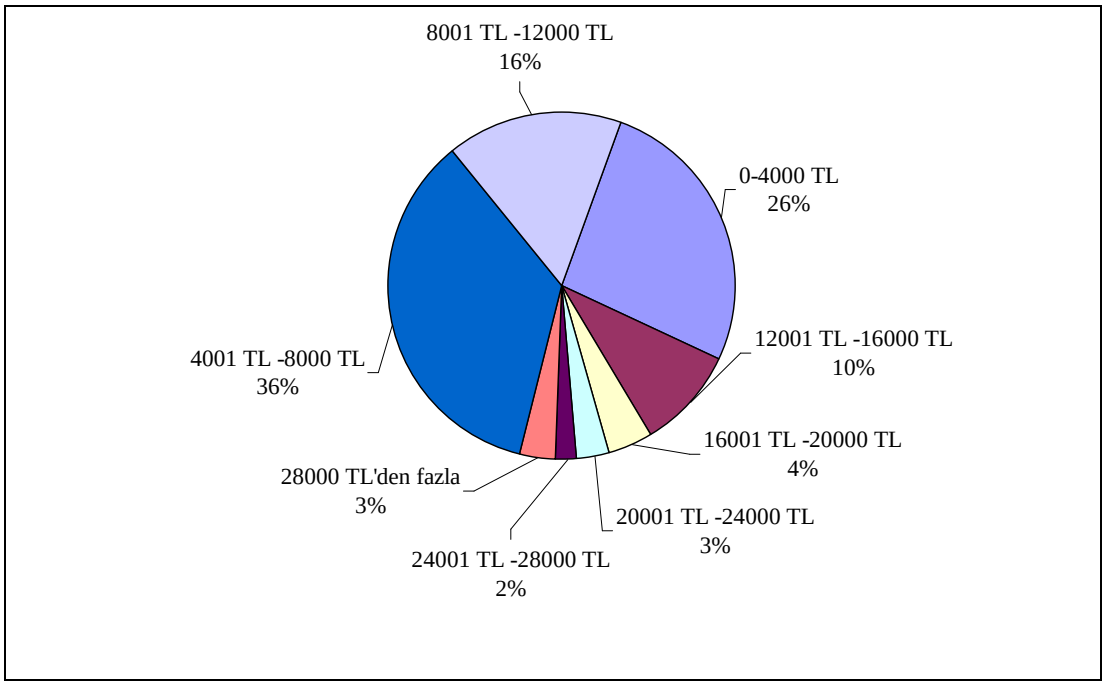
Şekil 5.8'de yeralan yaş ile kredi değerlendirme sonuçları arasındaki lowess regresyonundan, 19 ile 35 yaş aralığı için yaş arttıkça kredinin geri ödenmesinde sorun azalırken 35 yaş sonrasında ise yaşın artması kredi sonucunu etkilememektedir.

Gelir Düzeyi: Kredi alan 1000 müşterinin gelir düzeylerine ilişkin elde edilen istatistiksel bilgiler Şekil 5.9'de gösterilmiştir. Buna göre, kredi alan müşterilerin gelir düzeylerinin ortalaması 8.869,536 TL olup standart sapması ise 7.947,788'dir. Müşterilerin gelir düzeylerinin normal dağılımlı olmadıkları da görülmektedir.



Şekil 5.9. Gelir dağılımının istatistiksel göstergeleri.

Şekil 5.10'dan kredi alan müşterilerin %36'sının gelir düzeyi 4001 ile 8000 TL arasında olup, 4000 TL altında gelire sahip olanların oranı ise %26'dır. Kredi alanların %16'sının gelir düzeyleri ise 8001 ile 12000 arasındadır.



Şekil 5.10. Gelir dağılımı.

Tablo 5.2'de gelir ile yaş değişkenlerinin çapraz tablosu yeralmaktadır. Buna göre 4001-8000 gelir grubuna girip kredi kullanan kişi sayısı 352 olup, bunların %31'inin 27-34 yaşları arasında olduğu görülmektedir.

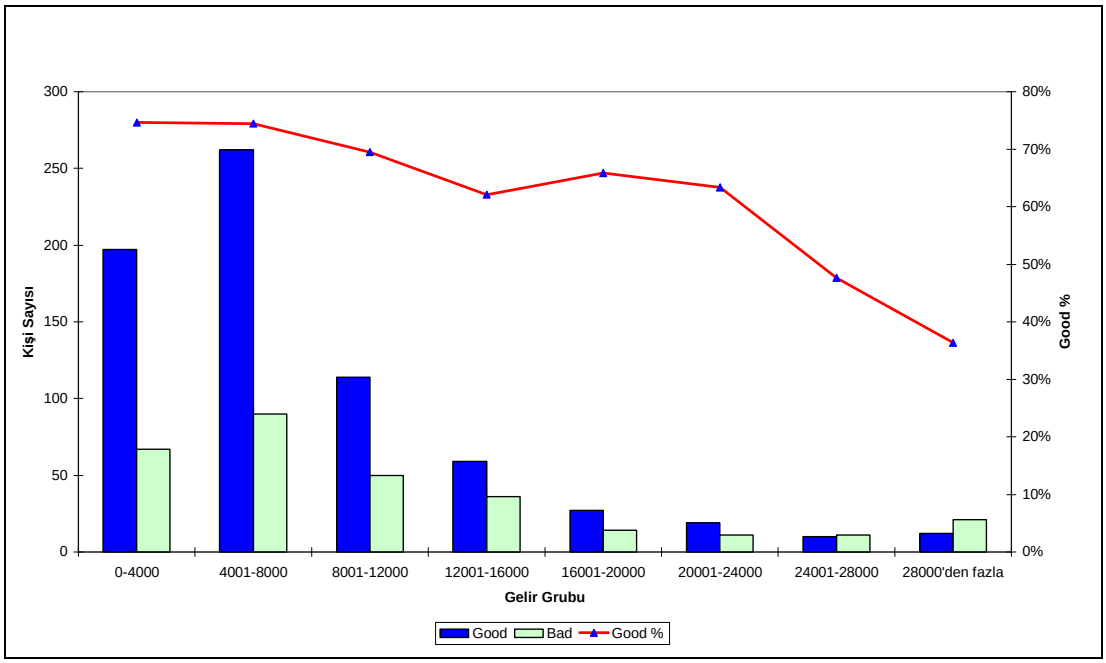
Tablo 5.2. Gelire göre yaş grupları tablosu.

Gelir (TL)		**Yaş Grupları**							
		19-26	**27-34**	**35-42**	**43-51**	**52-59**	**60-67**	**68-75**	**Toplam**
0-4000	**Adet**	80	70	54	32	13	12	3	264
	% Satır	30,3	26,52	20,45	12,12	4,92	4,55	1,14	100
	% Sütun	33,33	22,73	24,88	24,62	24,07	29,27	30	26,4
4001-8000	**Adet**	89	109	70	48	18	17	1	352
	% Satır	25,28	30,97	19,89	13,64	5,11	4,83	0,28	100
	% Sütun	37,08	35,39	32,26	36,92	33,33	41,46	10	35,2
8001-12000	**Adet**	31	61	37	18	9	7	164	327
	% Satır	18,9	37,2	22,56	10,98	5,49	4,27	0,61	100
	% Sütun	12,92	19,81	17,05	13,85	16,67	17,07	10	16,4
12001-16000	**Adet**	15	27	23	17	6	2	5	95
	% Satır	15,79	28,42	24,21	17,89	6,32	2,11	5,26	100
	% Sütun	6,25	8,77	10,6	13,08	11,11	4,88	50	9,5
16001-20000	**Adet**	10	16	11	3	1	0	0	41
	% Satır	24,39	39,02	26,83	7,32	2,44	0	0	100
	% Sütun	4,17	5,19	5,07	2,31	1,85	0	0	4,1

Tablo 5.2. Gelire göre yaş grupları tablosu (Devam).

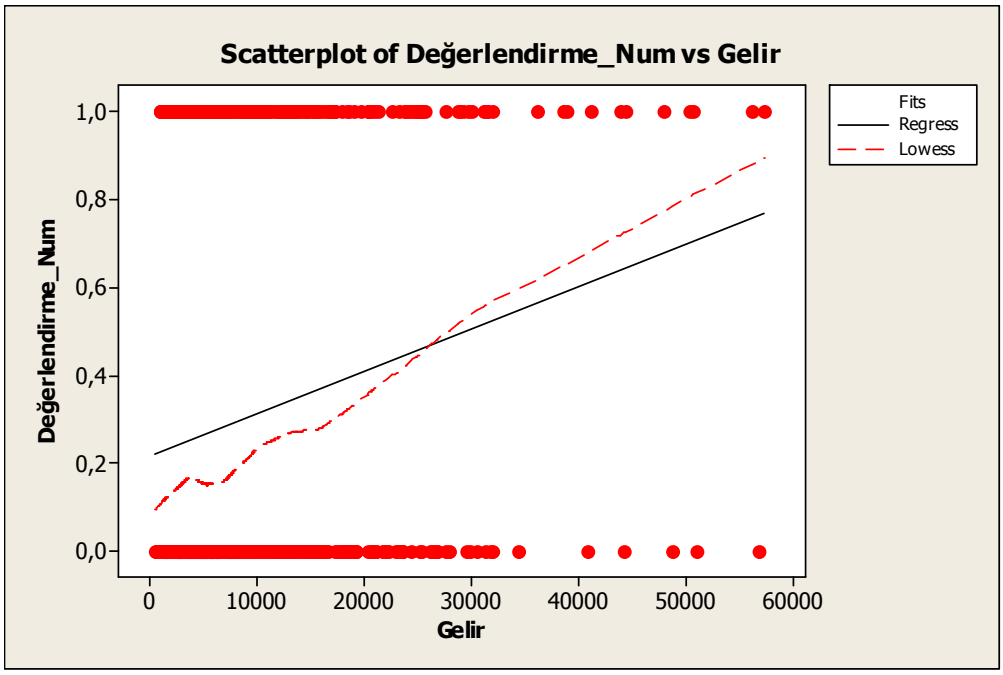
20001-24000	**Adet**	6	10	9	3	1	1	0	30
	% Satır	20	33,33	30	10	3,33	3,33	0	100
	% Sütun	2,5	3,25	4,15	2,31	1,85	2,44	0	3
24001-28000	**Adet**	2	8	4	4	2	1	0	21
	% Satır	9,52	38,1	19,05	19,05	9,52	4,76	0	100
	% Sütun	0,83	2,6	1,84	3,08	3,7	2,44	0	2,1
28000 fazla	**Adet**	7	7	9	5	4	1	0	33
	% Satır	21,21	21,21	27,27	15,15	12,12	3,03	0	100
	% Sütun	2,92	2,27	4,15	3,85	7,41	2,44	0	3,1
Toplam	**Adet**	240	308	217	130	54	41	10	1000
	% Satır	24	30,8	21,7	13	5,4	4,1	1	100
	% Sütun	100	100	100	100	100	100	100	100

Gelir gruplarının kredi değerlendirmesi Şekil 5.11'de gösterilmiştir. Buna göre gelir grubu artıkça kredinin geri ödenme yüzdesinin düştüğü gözlemlenmektedir. Genel olarak tüm yaş gruplarına bakıldığında, ortalama gelir düzeyinin yüksek olduğu durumda kredi ödemesinde sorun yaşandığı daha düşük gelir düzeylerinde ise kredi ödemesinin zamanında ve sorunsuz olarak yapıldığı görülmektedir. Bu durum, kişinin gelir düzeyi yükseldikçe yaşam standartlarının da yükseldiği ve dolayısıyla yaptığı harcamaların da arttığını göstermektedir.



Şekil 5.11. Gelir gruplarına göre kredi değerlendirme sonuçları.

Gelir ile kredi değerlendirme sonuçları arasındaki ilişkinin loess regresyonu analizi Şekil 5.12'deki gibidir. Analiz sonucuna göre gelir arttıkça kredinin ödenmesi ile ilgili sorun yaşandığı görülmektedir.



Şekil 5.12. Gelir ile kredi değerlendirme sonuçları arasındaki lowess regresyonu.

Cinsiyet: Kredi alan müşterilerin %69'u erkek, %31'i bayandır. Yaş gruplarına göre cinsiyet dağılımı Tablo 5.3'deki gibidir.

Tablo 5.3. Cinsiyete göre yaş grupları.

Yaş		Cinsiyet		Toplam
		Erkek	Kadın	
19-26	Adet	115	125	240
	% Satır	47,92	52,08	100
	% Sütun	16,67	40,32	24
27-34	Adet	220	88	308
	% Satır	71,43	28,57	100
	% Sütun	31,88	28,39	30,8
35-42	Adet	176	41	217
	% Satır	81,11	18,89	100
	% Sütun	25,51	13,23	21,7
43-51	Adet	108	22	130
	% Satır	83,08	16,92	100
	% Sütun	15,65	7,1	13
52-59	Adet	32	22	54
	% Satır	59,26	40,74	100
	% Sütun	4,64	7,1	5,4

Tablo 5.3. Cinsiyete göre yaş grupları (Devam).

60-67	**Adet**	31	10	41
	% Satır	75,61	24,39	100
	% Sütun	4,49	3,23	4,1
68-75	**Adet**	8	2	10
	% Satır	80	20	100
	% Sütun	1,16	0,65	1
Toplam	**Adet**	690	310	1000
	% Satır	69	31	100
	% Sütun	100	100	100

Tablo 5.4'te cinsiyete göre kredi değerlendirmesi sonuçları yeralmaktadır. Tablodan erkeklerin aldıkları krediyi batırma oranı %28 iken bayanlarda bu oran %35'e yükselmektedir.

Tablo 5.4. Cinsiyete göre kredi değerlendirmesi.

Cinsiyet		**Değerlendirme**		**Toplam**
		Kötü	**İyi**	
Erkek	**Adet**	191	499	690
	% Satır	27,68	72,32	100
	% Sütun	63,67	71,29	69
Kadın	**Adet**	109	201	310
	% Satır	35,16	64,84	100
	% Sütun	36,33	28,71	31
Toplam	**Adet**	300	700	1000
	% Satır	30	70	100
	% Sütun	100	100	100

Kredi batırma oranın istatistiksel olarak anlamlı olup olmadığının sınanması için Chi-kare testi yapılmıştır. Test sonucu 5,699 olup kredi değerlendirmeleri açısından erkekler ile bayanlar arasında istatistiksel olarak anlamlı bir farklılığın olduğu 0,05 yanılma düzeyinde gözlemlenmiştir.

Müşterilerin kredi geçmişinin cinsiyete göre değerlendirilmesi Tablo 5.5'deki gibidir. Tablodan bayanların %60'ının, erkeklerin ise %50'sinin kredisini zamanında ödediği görülmektedir.

Tablo 5.5. Cinsiyete göre kredi geçmişi.

Kredi Geçmişi		Cinsiyet		
		Erkek	Kadın	Toplam
Banka kredisini geç ödememiş	Adet	32	17	49
	% Satır	65,31	34,69	100
	% Sütun	4,64	5,48	4,9
Hiç kredi almamış	Adet	28	12	40
	% Satır	70	30	100
	% Sütun	4,06	3,87	4
Krediyi geç ödememiş	Adet	344	186	530
	% Satır	64,91	35,09	100
	% Sütun	49,86	60	53
Krediyi geç ödemiş	Adet	72	16	88
	% Satır	81,82	18,18	100
	% Sütun	10,43	5,16	8,8
Krediyi ödememiş	Adet	214	79	293
	% Satır	73,04	26,96	100
	% Sütun	31,01	25,48	29,3
Toplam	Adet	690	310	1000
	% Satır	69	31	100
	% Sütun	100	100	100

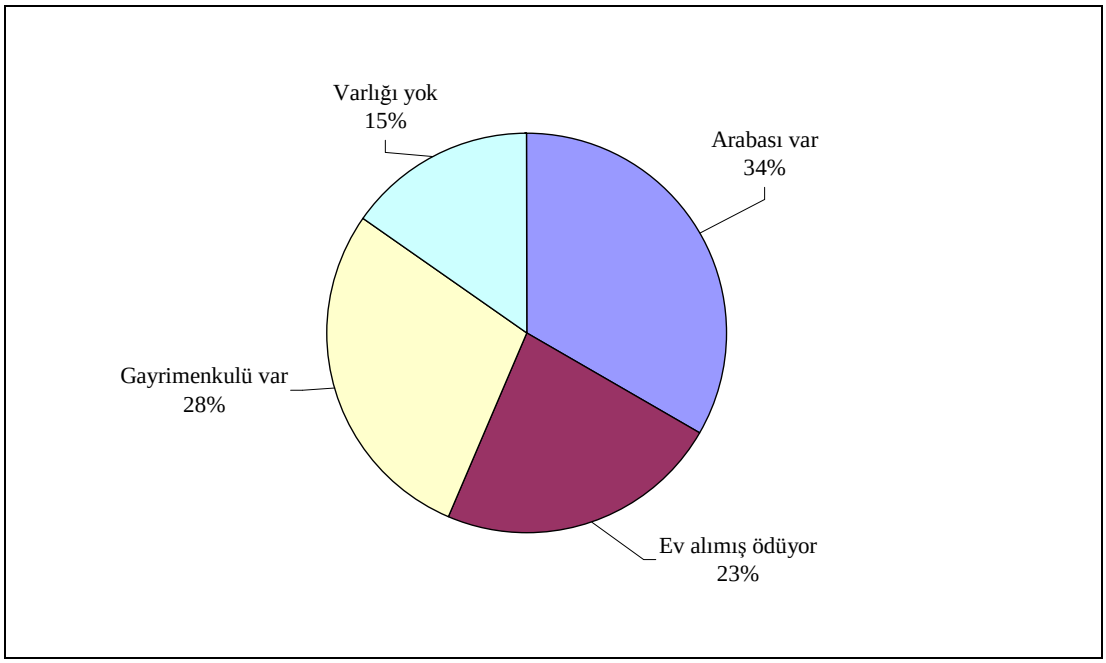
Cinsiyete göre müşterilerin kredi geçmişinin istatistiksel olarak anlamlı olup olmadığının sınanması için yapılan Chi-kare testi sonucunda test istatistiği 13,478 olarak hesaplanmıştır. Buna göre erkeler ile bayanlar arasında kredi geçmişi açısından farklılığın olduğu 0,05 yanılma düzeyinde gözlemlenmiştir.

Cinsiyete göre gelir grupları dağılımı Tablo 5.6'daki gibidir. Tablodan kredi kullananların %35,20'sinin 4001-8000 gelir grubunda olduğu ve bu grupta yeralan müşterilerin %64,77'sinin erkek, geri kalan %35,23'ünün ise kadın olduğu görülmektedir. Kredi kullanan bayanlar arasından bu gelir grubuna ait olan kişi sayısı 124 olup, tüm bayanların %40'nı oluşturmaktadırlar.

Tablo 5.6. Cinsiyete göre gelir grubları.

Gelir (TL)		Cinsiyet		Toplam
		Erkek	Kadın	
0-4000	Adet	162	102	264
	% Satır	61,36	38,64	100
	% Sütun	23,48	32,9	26,4
4001-8000	Adet	228	124	352
	% Satır	64,77	35,23	100
	% Sütun	33,04	40	35,2
8001-12000	Adet	123	41	164
	% Satır	75	25	100
	% Sütun	17,83	13,23	16,4
12001-16000	Adet	75	20	95
	% Satır	78,95	21,05	100
	% Sütun	10,87	6,45	9,5
16001-20000	Adet	34	7	41
	% Satır	82,93	17,07	100
	% Sütun	4,93	2,26	4,1
20001-24000	Adet	26	4	30
	% Satır	86,67	13,33	100
	% Sütun	3,77	1,29	3
24001-28000	Adet	17	4	21
	% Satır	80,95	19,05	100
	% Sütun	2,46	1,29	2,1
28000 fazla	Adet	25	8	33
	% Satır	75,76	24,24	100
	% Sütun	3,62	2,58	3,3
Toplam	Adet	690	310	1000
	% Satır	69	31	100
	% Sütun	100	100	100

Müşteri Varlıkları: Çalışmada kredi alan müşterilerin varlıkları, teminatları incelenmiştir. Şekil 5.13'de kredi kullanan müşterilerin varlık durumlarının grafiği yeralmaktadır. Buna göre müşterilerden taşınmaz bir gayrimenkule sahip olanların %28, ev almış ve ödemeye devam edenlerin %23, arabası olanların %34, herhangi bir varlığı olmayanların ise %15 oranında olduğu görülmektedir.



Şekil 5.13. Kredi kullanan müşterilerin varlık durumlarının dağılımı.

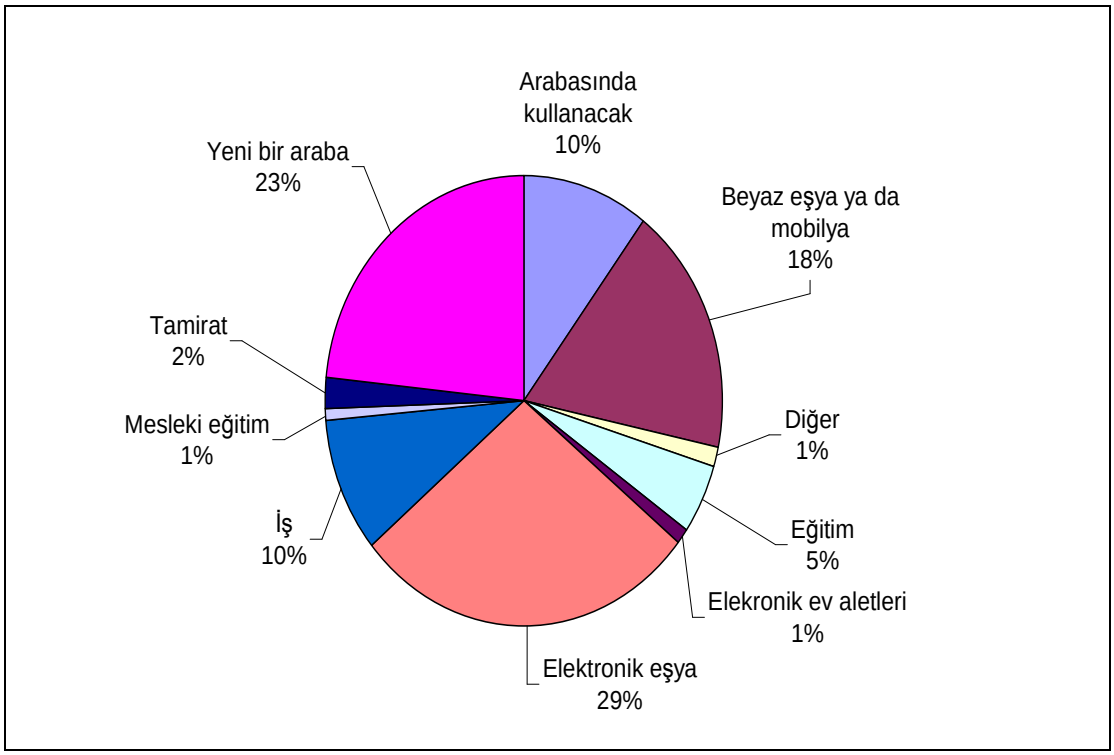
Tablo 5.7'de müşterilerin varlıklarına göre kredinin geri ödenmesi arasındaki ilişki yeralmaktadır. Tabloya göre varlıklı müşterilerin krediyi geri ödeme yüzdeleri varlığı olmayan müşterilerin geri ödeme yüzdelerinden fazladır.

Tablo 5.7. Müşterilerin varlıklarına göre kredi değerlendirmesi.

Müşteri Varlıkları		**Değerlendirme**		**Toplam**
		Kötü	**İyi**	
Arabası var	**Adet**	102	230	332
	% Satır	30,72	69,28	100
	% Sütun	34	32,86	33,2
Ev almış ödüyor	**Adet**	71	161	232
	% Satır	30,6	69,4	100
	% Sütun	23,67	23	23,2
Gayrimenkulu var	**Adet**	60	222	282
	% Satır	21,28	78,72	100
	% Sütun	20	31,71	28,2
Varlığı yok	**Adet**	67	87	154
	% Satır	43,51	56,49	100
	% Sütun	22,33	12,43	15,4
Toplam	**Adet**	300	700	1000
	% Satır	30	70	100
	% Sütun	100	100	100

Müşterilerin varlıklarına göre kredinin geri ödenmesi arasındaki ilişkinin istatistiksel olarak anlamlı olup olmadığının sınanması için yapılan Chi-kare testi sonucunda test istatistiği 23,720 olarak hesaplanmıştır. Bu durum varlıklı müşteriler ile varlıksız müşterilerin kredi geri ödemeleri arasında fark anlamlı bir farklılığın olduğunu 0,05 yanılma düzeyinde göstermektedir.

Kredi Alma Nedeni: Yapılan analiz sonucunda müşterilerin %29'unun elektronik eşya almak için, %23'ünün yeni bir araba için, %18'inin beyaz eşya ya da mobilya almak için kredi aldıkları belirlenmiştir. Kredi alma nedenlerinin grafiksel olarak gösterimi Şekil 5.14'deki gibidir.



Şekil 5.14. Müşterilerin kredi alma nedenlerinin dağılımı.

Kredi kullanan 300 kadının kredi alma nedenleri ile aldıkları krediyi geri ödemeleri durumlarının birlikte incelendiği Tablo 5.8'e göre, toplam alınan krediler içinde en fazla alınan kredi %27,41'lik oranla beyaz eşya-mobilya kredisi, daha sonra %22,58 oranla yeni bir araba kredisidir. Bayanlar tarafından alınan kredilerin %35,16'sında sorun yaşanmış ve en fazla sorun ise yeni bir araba kredisi alan müşterilerde ortaya çıkmıştır.

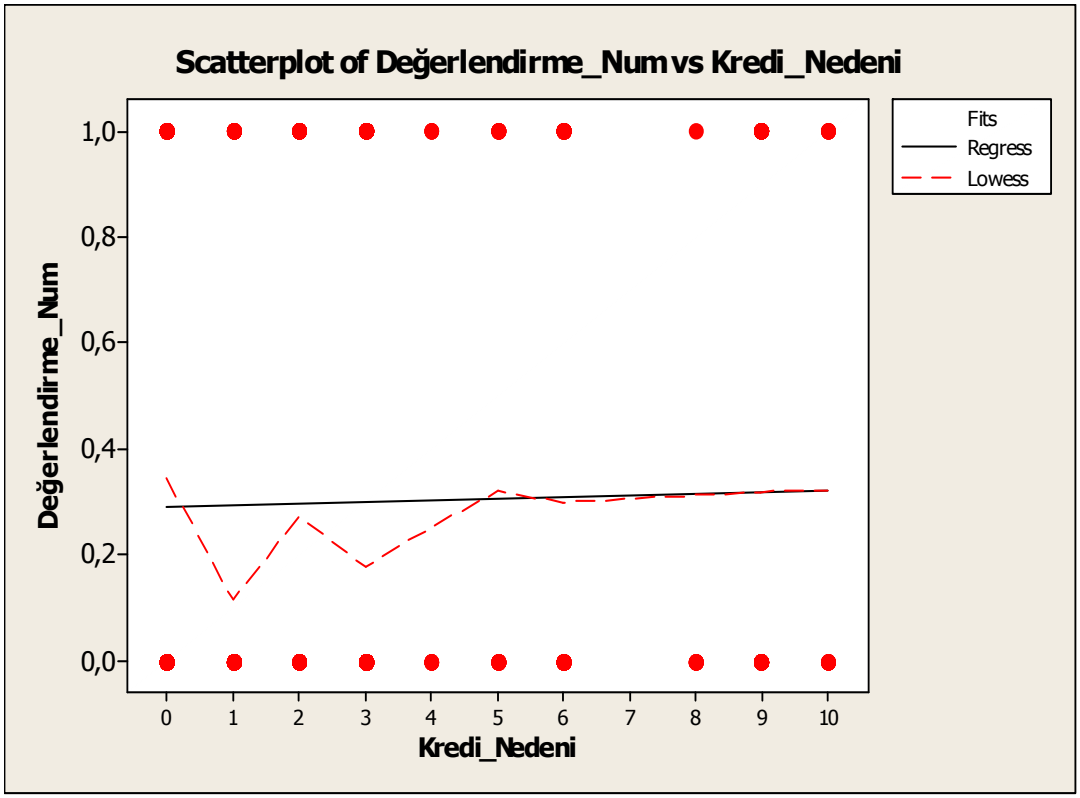
Tablo 5.8. Kadınların kullandıkları kredilerin nedenlerine göre kredi değerlendirmesi

Kredi Nedeni		**Değerlendirme**		
		Kötü	**İyi**	**Toplam**
Arabasında kullanacak	**Adet**	8	16	24
	% Satır	33,33	66,67	100
	% Sütun	7,339	7,96	7,742
Beyaz eşya-mobilya	**Adet**	28	46	74
	% Satır	37,84	62,16	100
	% Sütun	25,688	22,886	23,871
Diğer	**Adet**	2	1	3
	% Satır	66,67	33,33	100
	% Sütun	1,835	0,498	0,968
Eğitim	**Adet**	9	12	21
	% Satır	42,86	57,14	100
	% Sütun	8,257	5,97	6,774
Elektronik ev aletleri	**Adet**	2	4	6
	% Satır	33,33	66,67	100
	% Sütun	1,835	1,99	1,935
Elektronik eşya	**Adet**	19	66	85
	% Satır	22,35	77,65	100
	% Sütun	17,431	32,836	27,419
İş	**Adet**	7	12	19
	% Satır	36,84	63,16	100
	% Sütun	6,422	5,97	6,129
Mesleki eğitim	**Adet**	0	3	3
	% Satır	0	100	100
	% Sütun	0	1,493	0,968
Tamirat	**Adet**	2	3	5
	% Satır	40	60	100
	% Sütun	1,835	1,493	1,613
Yeni bir araba	**Adet**	32	38	70
	% Satır	45,71	54,29	100
	% Sütun	29,358	18,905	22,581
Toplam	**Adet**	109	201	310
	% Satır	35,16	64,84	100
	% Sütun	100	100	100

Kadınların kredi alma nedenleri ile aldıkları krediyi geri ödemeleri arasındaki ilişkinin istatistiksel olarak anlamlı olup olmadığının sınanması için yapılan Chi-kare testi sonucunda test istatistiği 13,366 olarak hesaplanmıştır. Bu durum bayanların

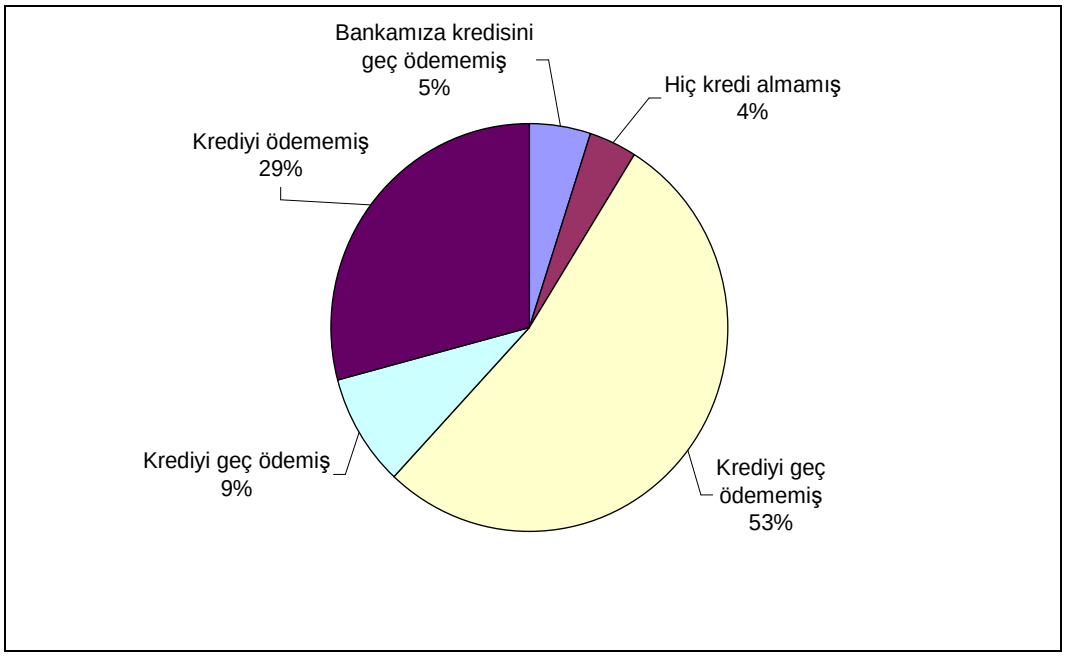
batırdıkları kredilerin krediyi alış nedeni ile ilgili olmadığının göstergesi olarak yorumlanabilir.

Kredi alma nedeni ile kredi değerlendirme sonuçları arasındaki ilişkinin Lowess eğrisi ile gösterimi Şekil 5.15'deki gibidir.



Şekil 5.15. Kredi nedeni ile kredi değerlendirme sonuçları arasındaki loess regresyonu.

Kredi Geçmişi: Müşterilerin kredi geçmişlerinin grafiksel olarak gösterimi Şekil 5.16'da olduğu gibidir. Buna göre kredi alanların %53'ünü diğer bankalardan aldığı kredilerin taksitlerini tam zamanında yatırmış ve tüm kredisini ödemiş olanlar, %5'ini çalışmada ele alınan bankadan kredi alıp kredi taksitlerini tam zamanında ve eksiksiz olarak ödemiş olanlar, %29'unu kritik düzeyde müşterilerin oluşturduğu diğer bankalara borcu olanlar ile aldıkları krediyi ödeyememiş olanlar, %9'unu taksit ödemelerini geciktirmiş olanlar ve %4'ünü ise hiç kredi almamış olanlar oluşturmaktadır.



Şekil 5.16. Müşterilerin kredi geçmişlerinin dağılımı.

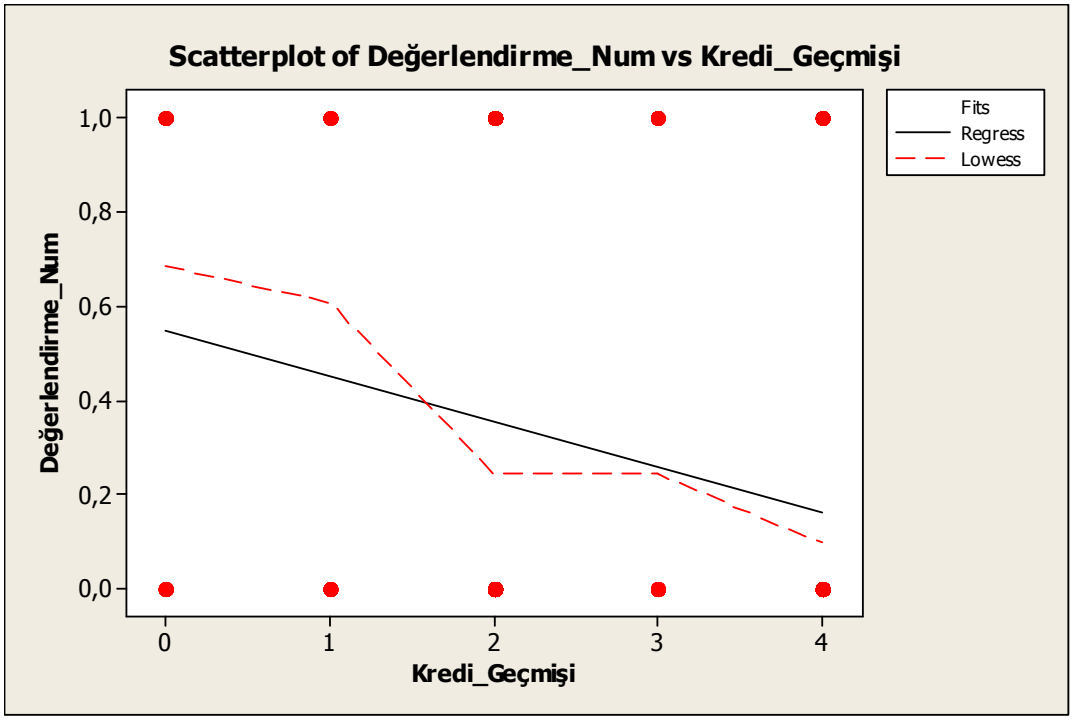
Kredi geçmişi ile bankadan alınan kredinin geri ödemesinin incelendiği Tablo 5.9'a göre, daha önce kredisini ödememiş müşterilerin %17'si, daha önce hiç kredi almamış müşterilerin ise %62'si aldığı krediyi batırmıştır.

Tablo 5.9. Müşterilerin kredi geçmişlerine göre kredi değerlendirme tablosu

Kredi Geçmişi		**Değerlendirme**		**Toplam**
		Kötü	**İyi**	
Banka kredisini geç ödememiş	**Adet**	28	21	49
	% Satır	57,14	42,86	100
	% Sütun	9,33	3	4,9
Hiç kredi almamış	**Adet**	25	15	40
	% Satır	62,5	37,5	100
	% Sütun	8,33	2,14	4
Krediyi geç ödememiş	**Adet**	169	361	530
	% Satır	31,89	68,11	100
	% Sütun	56,33	51,57	53
Krediyi geç ödemiş	**Adet**	28	60	88
	% Satır	31,82	68,18	100
	% Sütun	9,33	8,57	8,8
Krediyi ödememiş	**Adet**	50	243	293
	% Satır	17,06	82,94	100
	% Sütun	16,67	34,71	29,3
Toplam	**Adet**	300	700	1000
	% Satır	30	70	100
	% Sütun	100	100	100

Kredi geçmişi ile bankadan alınan kredinin geri ödemesi arasındaki ilişkinin istatistiksel olarak anlamlı olup olmadığının sınanması için yapılan Chi-kare testi sonucunda test istatistiği 61,691 olarak hesaplanmıştır. Bu durum batan kredilerin müşterilerin kredi geçmişi ile ilgili olduğunu 0,05 yanılma düzeyinde göstermektedir.

Yapılan Lowess regresyon analizi sonucunda ise kredi geçmişi ile kredi sonuçları arasında ters bir ilişki olduğu, yani kredi geçmişi kötüye doğru gittikçe kredinin sorunsuz olarak ödendiği görülmektedir. Bu durum Şekil 5.17'de gösterilmiştir.



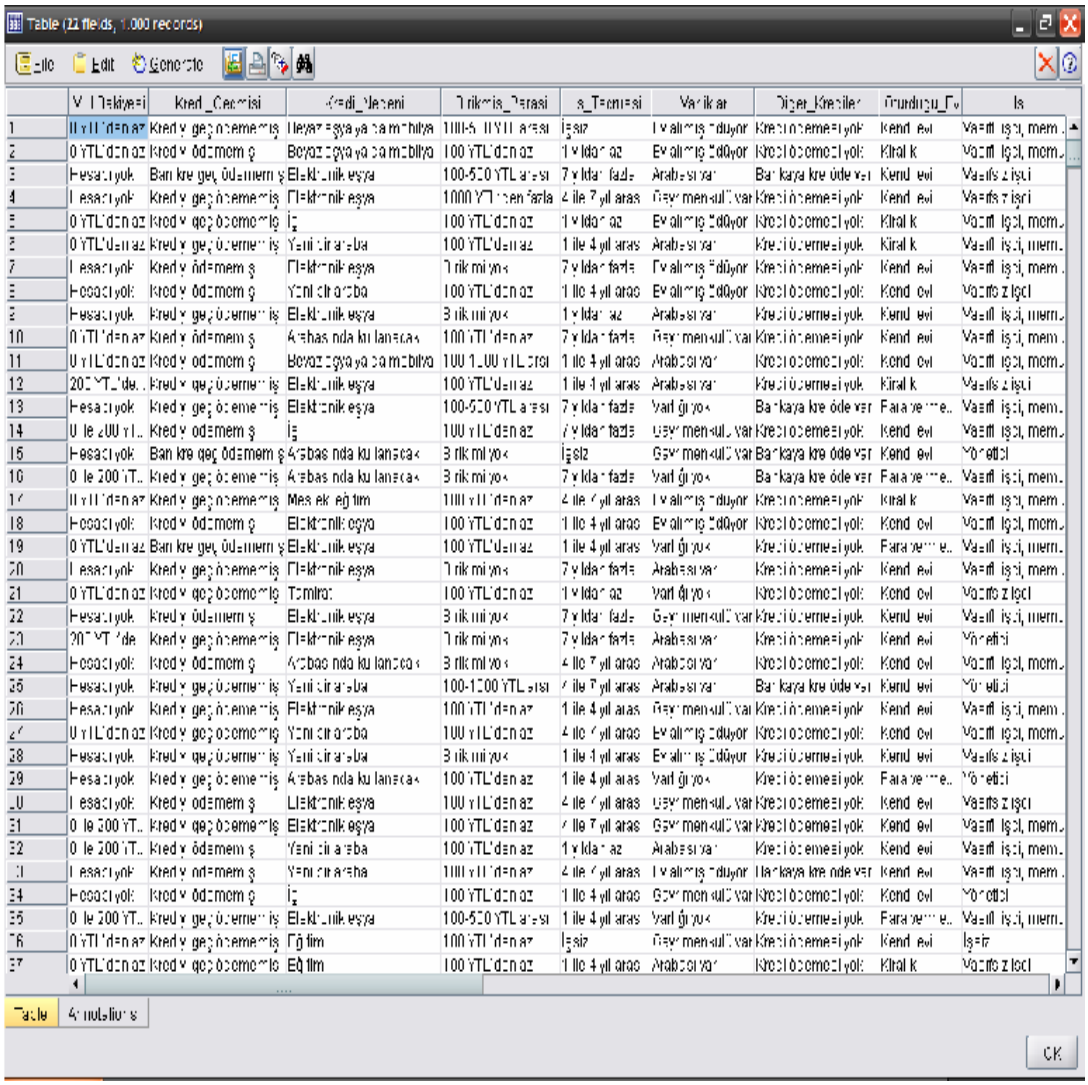
Şekil 5.17. Müşterilerin kredi geçmişi ile kredi değerlendirme sonuçları arasındaki loess regresyonu.

5.3.3. VERİ ÖNİŞLEME

Clementine programı ile verinin işlenmesi veri setinin Cementine programına aktarılması ve hedef değişkenin seçimi, normal olmayan verilerin belirlenmesi ve veri setinden çıkartılması, değerlendirme sonuçları ile ilişkisi zayıf olan değişkenlerin belirlenmesi ve veri setinden çıkartılması aşamalarından oluşmaktadır. Bu aşamaların ayrıntıları biçimde açıklamalarına izleyen alt kesimlerde değinilmiştir.

5.3.3.1. Veri Setinin Cementine Programına Aktarılması ve Hedef Değişkenin Seçimi

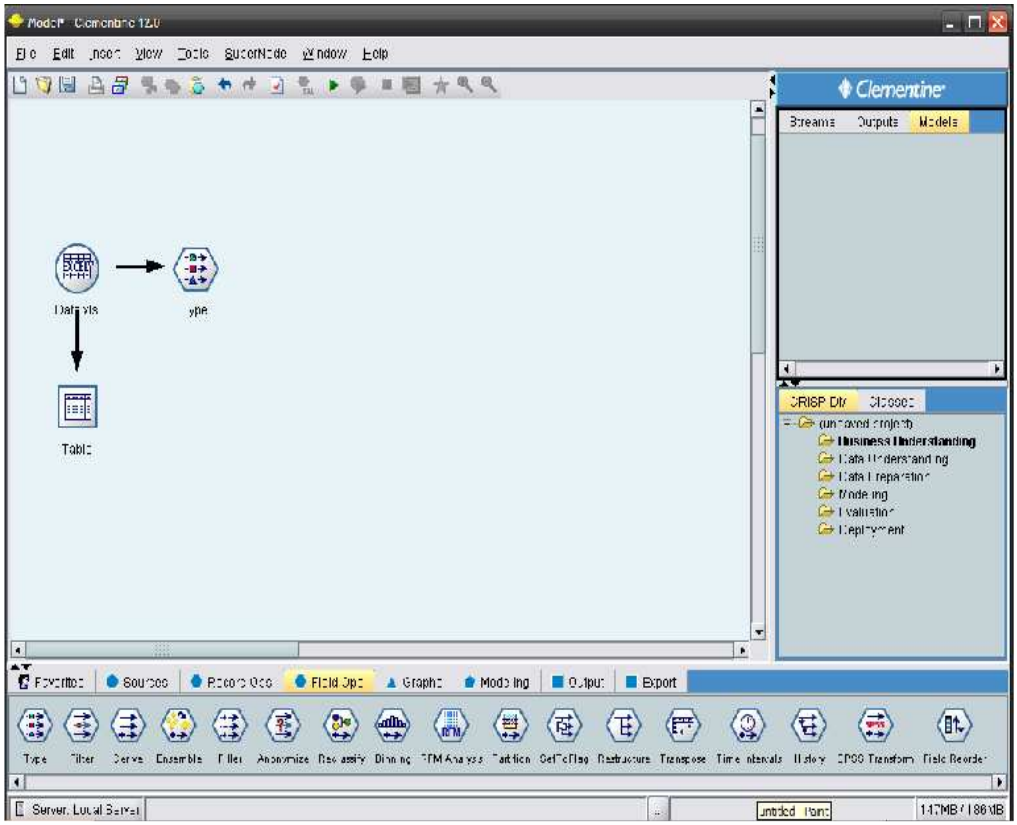
Kredi almış 1000 müşterinin verisi, Clementine programının "Source/Excel" modülünden yararlanılarak programa aktarılır. Programa aktarılan verinin tablo şeklinde gösterimini sağlamak için "Output/Table" modülünden yararlanılır. Şekil 5.18'de Clementine programına aktarılmış olan verinin tablo gösterimi yeralmaktadır.



Şekil 5.18. Verinin Clementine programındaki görüntüsü.

Tüm verilerin Clemintine programına aktarımından sonra değişkenler içinden hedef bir değişkenin seçilmesi gerekmektedir. Çalışmada hedef değişkeni "Değerlendirme_Num" değişkeni yani kredinin "kötü" ya da "iyi" sonuçlandığını gösteren değişken olarak belirlenmiştir. Hedef değişkenin belirlenmesinde Clementine programının "Field Ops/Type" modülünden yararlanılır. Bu modülün

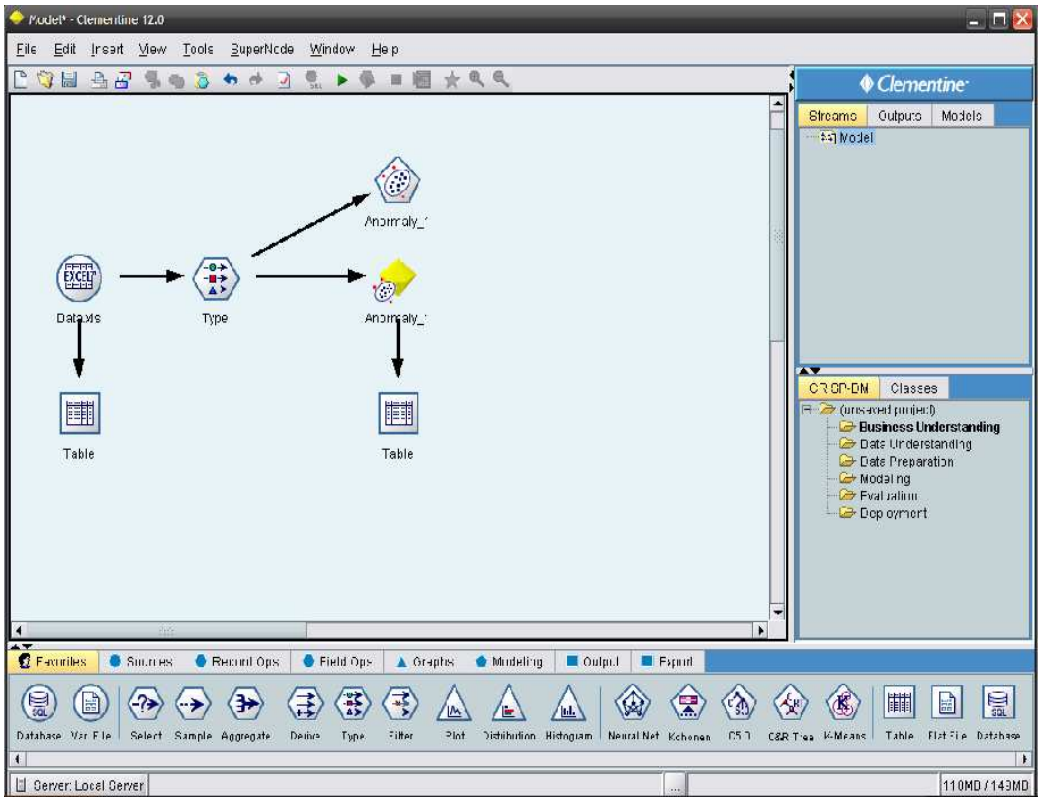
işletilmesiyle programda hedef değişken “Out” olarak işaretlenir. Clementine programında modelin görüntüsü Şekil 5.19’daki gibi oluşur.



Şekil 5.19. Veri aktarımı ve hedef değişken sonrasında Clementine sayfasında oluşan görüntü.

5.3.3.2. Anormal Verilerin Belirlenmesi ve Veri Setinden Çıkartılması

Clementine programında, veri setindeki normal olmayan verilerin belirlenmesi için “Modeling/Anolmany” modülünden yararlanılır. Bu modül “Type” modülüne bağlanarak modeli çalıştırır ve normal olmayan verileri ortaya çıkarır. Şekil 5.20’de anormallik testi sonrasında Clementine sayfasında oluşan görüntü yeralmaktadır.

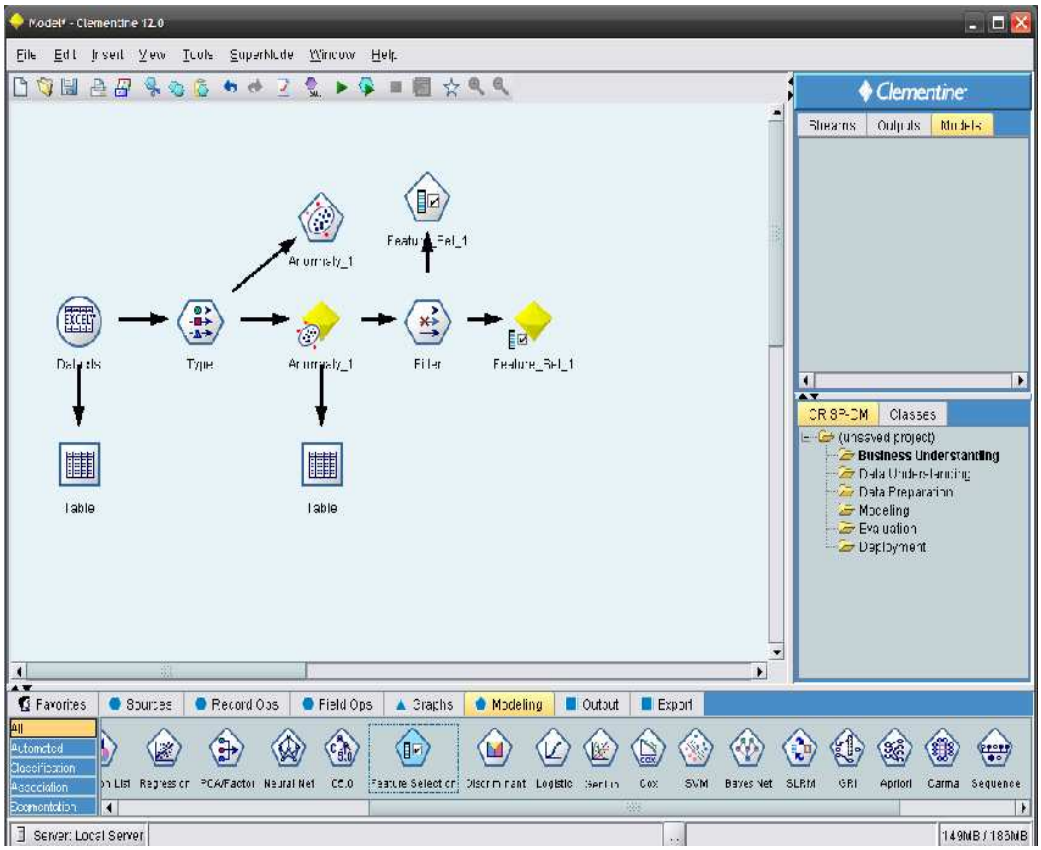


Şekil 5.20. Anormallik testi sonrasında Clementine sayfasında oluşan görüntü.

"Anolmany" modülü ile veriler program tarafından dört grupta ayrı ayrı incelenmektedir. Bunun sonucunda her gruptaki farklılıklar anormallik indeksindeki değerlere göre belirlenmekte ve indeks değeri yüksek olan veriler ana veri setinden çıkartılmaktadır. Çalışmada kullanılan veri seti için Anolmany modülünün işletilmesi ile birinci grupta 2, ikinci grupta 2, üçüncü grupta 1 ve dördüncü grupta 5 olmak üzere toplam 10 anormal veri belirlenmiştir. Bu anormal veriler, veri setinden çıkartılmış böylelikle veri setinde bulunan toplam veri sayısı 990'a inmiştir. Yapılan analizler bu veri üzerinden gerçekleştirilmiştir. Veri setinden çıkartılan veriler veri setinin 47, 48, 117, 212, 284, 414, 457, 501, 707 ve 910'uncu satırlarındaki verilerdir.

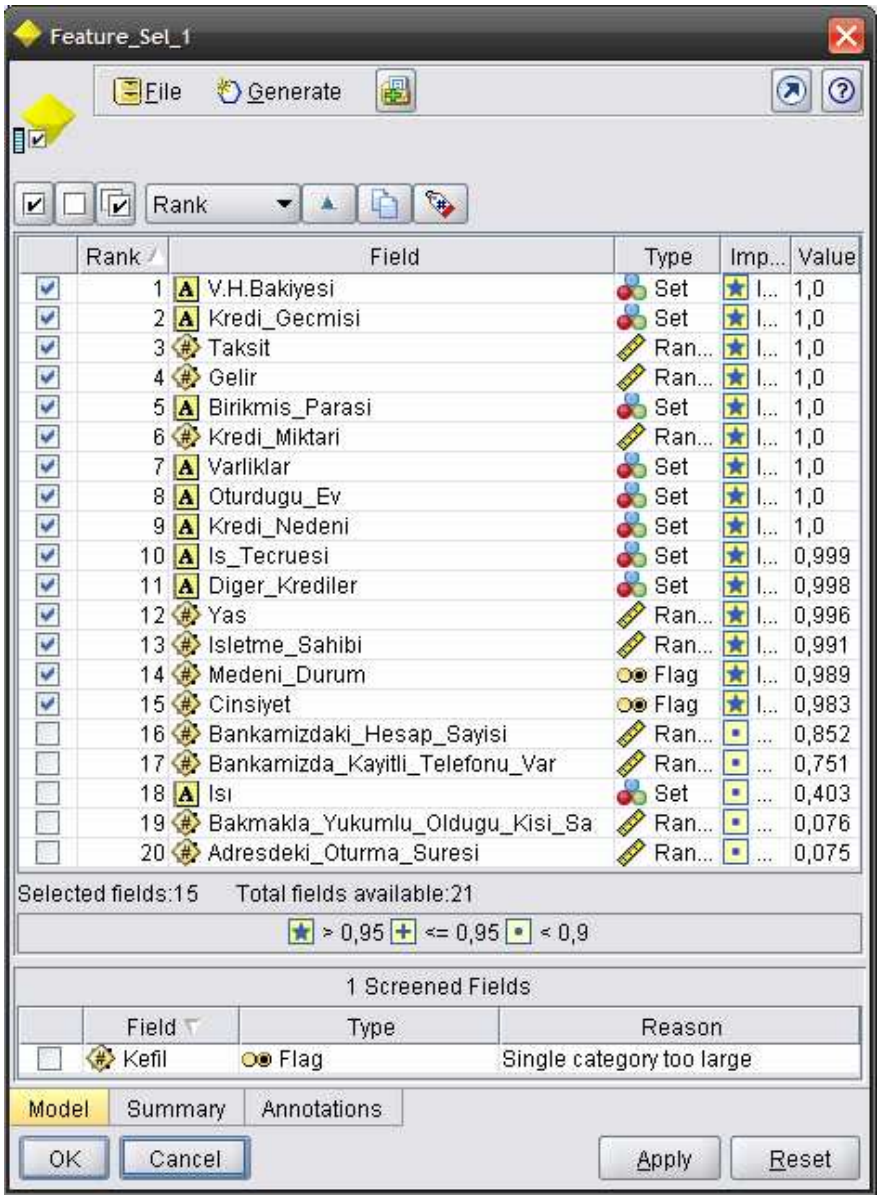
5.3.3.3. Değerlendirme Sonuçları ile İlişkisi Zayıf Olan Değişkenlerin Belirlenmesi ve Veri Setinden Çıkartılması

Anormal verilerin veri setinden çıkartılmasından sonra sıra değişkenler içinde hedef değişkenini etkileyen en zayıf yani önemsiz değişkenlerin belirlenmesi aşamasına gelinir. Bunun içinde Clementine programının "Feature Selection" modülü kullanılır. Modülün modele eklenmiş hali Şekil 5.21'de gösterildiği gibidir.



Şekil 5.21. “Feature Selection” modülünün modele eklenmesi sonrasında Clementine sayfasında oluşan görüntü.

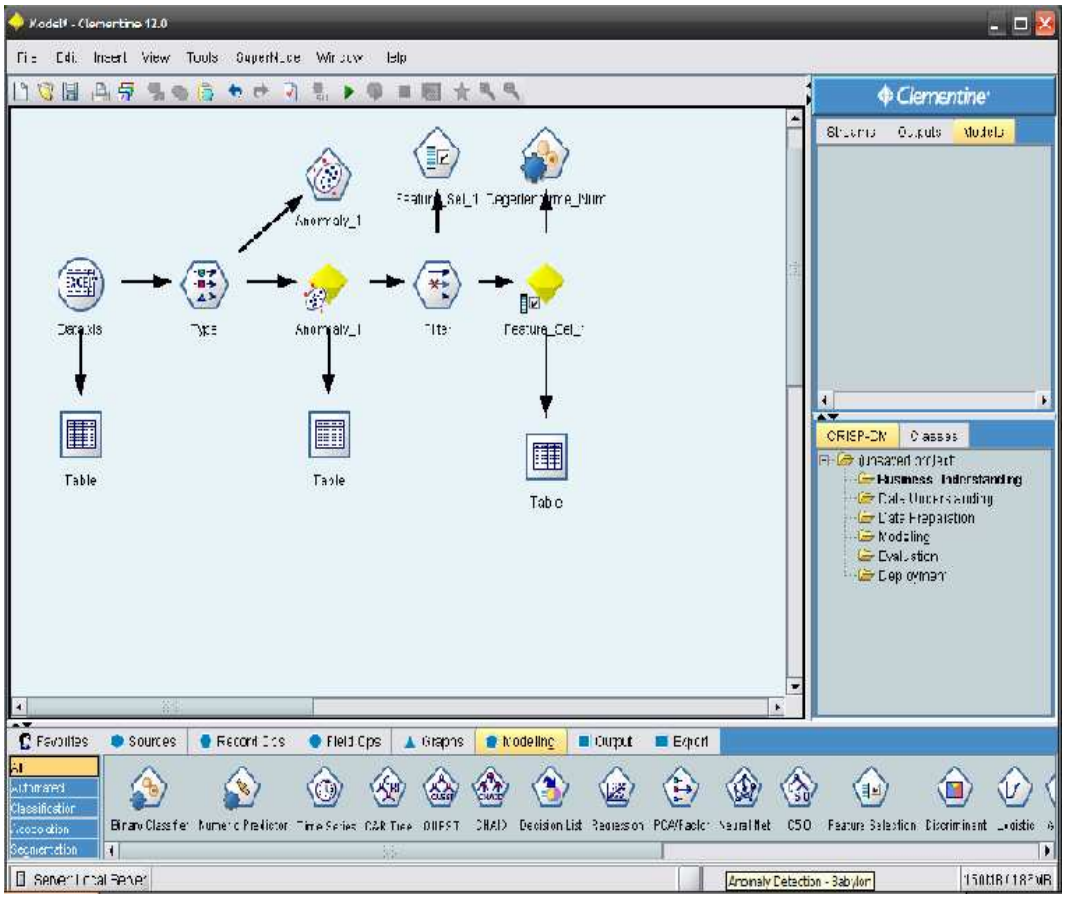
Modele “Feature Selection” modülünün de eklenip çalıştırılması sonucunda bankadaki hesap sayısı, bankada kayıtlı telefonu var, işi, bakmakla yükümlü olduğu kişi sayısı, adresteki oturma süresi ve kefil değişkenlerinin hedef değişken üzerinde önemli bir etkiye sahip olmadıkları belirlenmiştir. Daha sonra bu değişkenler veri setinden çıkartılıp bunların yerine hedef değişken olan “Değerlendirme_ Num”a etkisi en fazla olan değişkenler modele eklenmiştir. “Feature Selection” modülünün işletilmesi sonucunda elde edilen görüntü Şekil 5.22’deki gibidir.



Şekil 5.22. “Feature Selection” test sonucu.

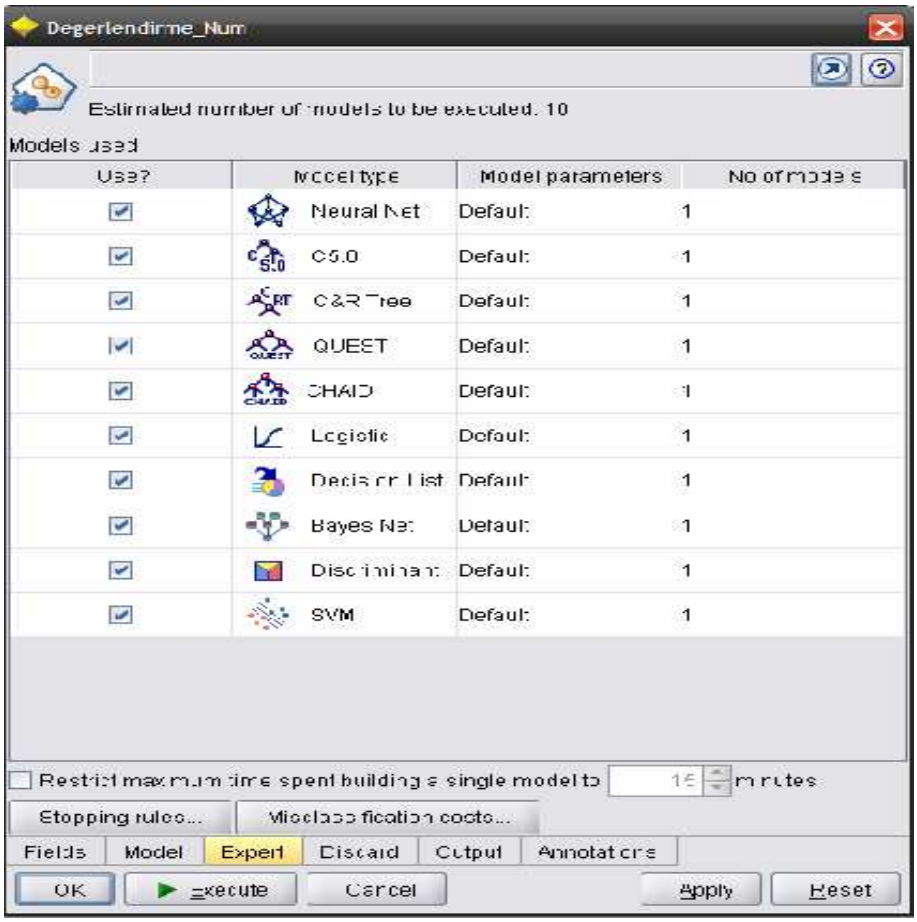
5.3.4. MODELLEME

Bu aşama gerçek sonuçlara en uygun olan scorecard modelinin belirlenmesini içermektedir. Veri setinin hazırlığı bittikten sonra sıra modelin kurulması aşamasına gelinir. Clementine programında en uygun modelin seçilmesi için kullanılan modül “Binary Classifier” modülüdür. Şekil 5.23’de “Binary Classifier” modülünün modele eklenmesinden sonra elde edilen görüntü yeralmaktadır.



Şekil 5.23. “Binary Classifier” modülü sonrasında Clementine sayfasında oluşan görüntü.

“Binary Classifier” modülü kullanıma hazır olan veri setindeki değişkenler için en uygun modelin seçilmesini sağlamaktadır. Modül çalıştırılmadan önce model türlerinin seçiminin gerçekleştirilmesi gerekmektedir. Şekil 5.24’te kullanılacak modellerin belirlenmesini sağlayan görüntü yeralmaktadır.

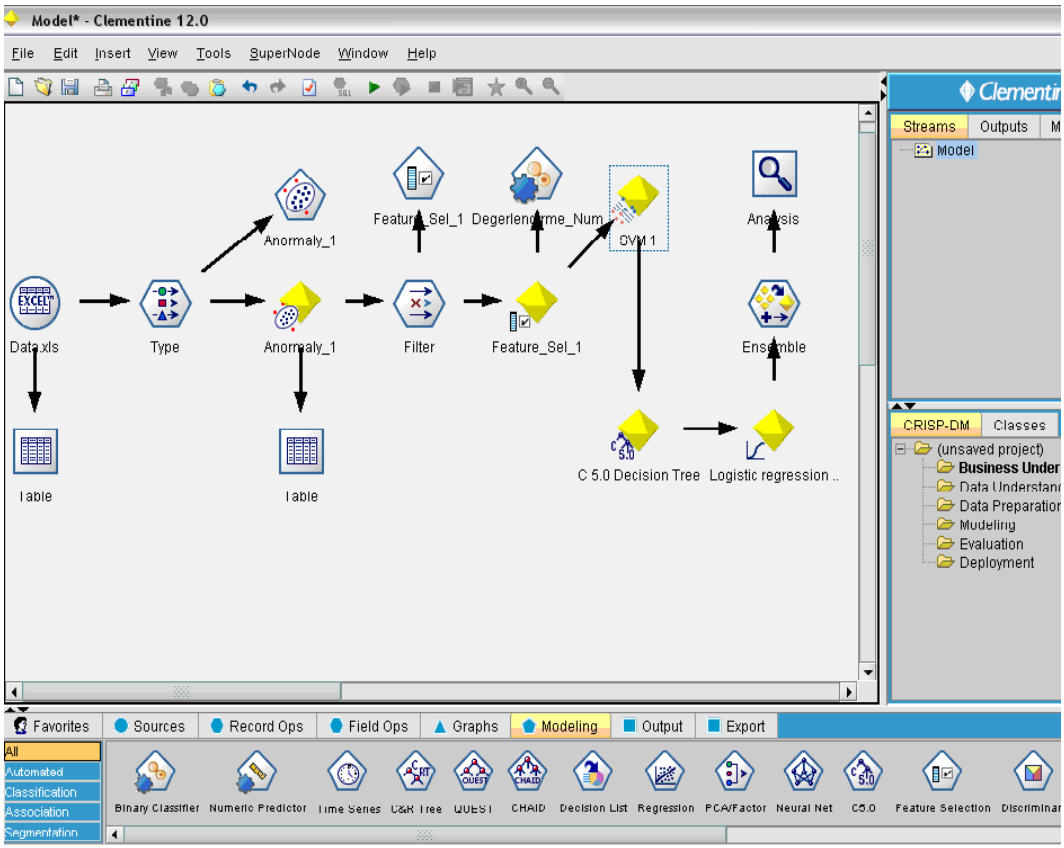


Şekil 5.24. Kullanılacak modellerin belirlenmesi.

Veri seti üzerinde işletilecek olan modellerin belirlenmesinin ardından “Binary Classifier” modülü çalıştırılarak elde edilen bulgular doğrultusunda scorecard için en iyi modelin seçim aşamasına gelinmiş olunur.

5.3.5. DEĞERLENDİRME

Scorecard modelini oluşturacak en iyi modelin bulunması için tüm modeller ana modele eklenip program çalıştırılmıştır. Tüm modellerin ana modele eklendiği son görüntü Şekil 5.25’de gösterildiği gibidir.



Şekil 5.25. Scorecard modelinin Clementine sayfasında oluşan görüntüsü.

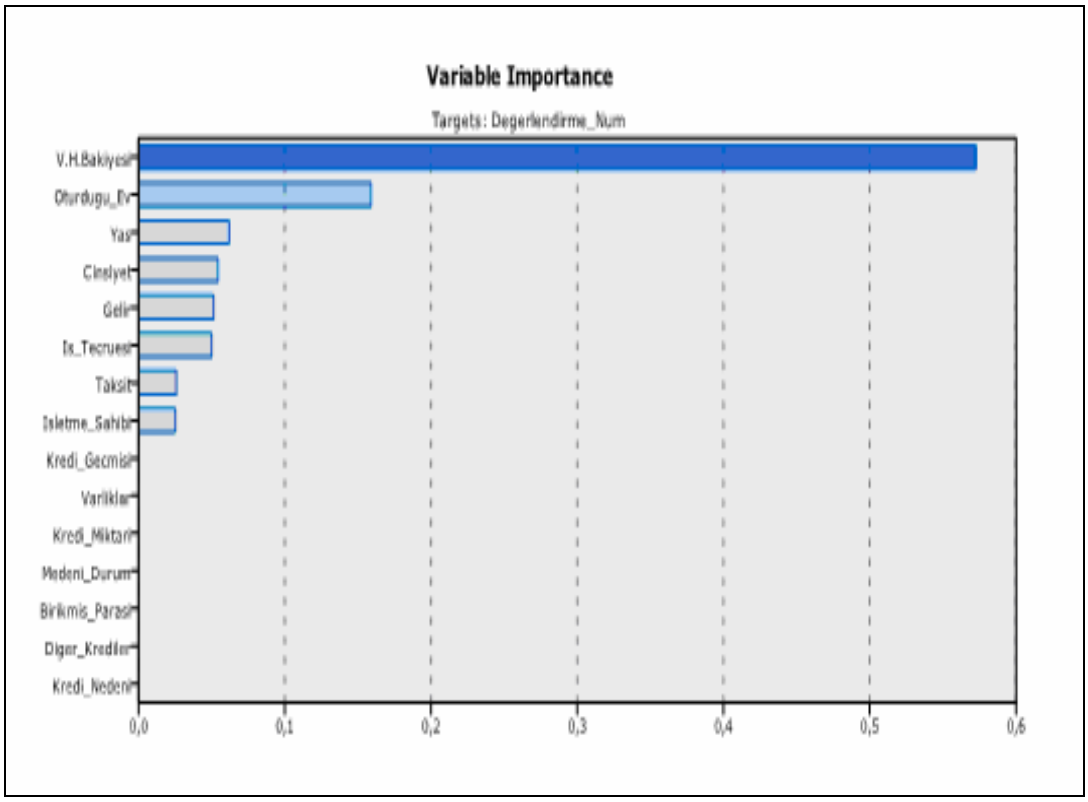
990 müşteriye ait verinin kullanıldığı uygulamada, analiz için uygulanan modellerden her müşteri için, aldıkları kredileri batırıp batırmayacağının tahmini yapılmıştır. Daha sonra elde edilen değerler ile gerçek değerler karşılaştırılıp müşterilerin aldıkları krediyi batırıp batırmayacağını en doğru şekilde tahmin eden model belirlenip bu model en iyi scorecard modeli olarak seçilmiştir.

Uygulamada scorecard modelini belirlemek amacıyla SVM (Support Vector Machine), karar ağacı ve lojistik regresyon modellerinden yararlanılmıştır.

5.3.5.1. SVM Modeli

SVM modeli, uygulamada kullanılan modeller içinde tahmin doğruluğu en yüksek olan modeldir. Modelin doğru tahminde bulunma oranı %95,25'tir.

Clementine programında kernel özelliğinin simple olduğu durum için SVM modelinin işletilmesiyle elde edilen değişkenlerin önem durumları Şekil 5.26'da gösterildiği gibidir. Buna göre "V. H. Bakiyesi" değişkeni en yüksek öneme sahip değişken olup önem derecesi %57,3'tür. Diğer değişkenlerin önem derecelerinin yüzdeleri ise Tablo 5.10'da gösterildiği gibidir.



Şekil 5.26. SVM modeli değişkenlerin önem durumu.

Tablo 5.10. SVM modelinde değişkenlerin önem derecelerinin yüzdesel dağılımı.

Değişken	**Önem Yüzdesi (%)**
V.H.Bakiyesi	57
Oturduğu_Ev	16
Yaş	6
Cinsiyet	5
Gelir	5
Is_Tecrübesi	5
Taksit	3
İşletme_Sahibi	3
Kredi_Nedeni	0
Diğer_Krediler	0
Birikmiş_Parası	0
Medeni_Durum	0
Kredi_Miktarı	0
Varlıklar	0
Kredi_Geçmişi	0

Tablo 5.11'de kernel özelliğinin olmadığı SVM modelinden elde edilen sonuçlar yeralmaktadır. Bu modelden elde edilen sonuç ile gerçekleşen sonuçlar karşılaştırıldığında doğruluk oranının %95,25 olduğu görülmüştür.

Tablo 5.11. Kernel çeşidi kullanılmadan çalıştırılan SVM modeli sonuçları.

Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlışlık Oranı (%)
Doğru	943	95,25
Yanlış	47	4,75
Toplam	990	100

SVM modelinde kernel çeşidinin lineer olarak alınmasıyla elde edilen sonuçlar Tablo 5.12'deki gibidir. Tablodan, lineer modelin kullanılmasıyla elde edilen modelin doğru tahmin etme oranının %78,18 olduğu görülmektedir. Buna göre doğruluk oranında kernel olmadan oluşturulan SVM modeline göre önemli ölçüde bir düşün olduğu görülmektedir.

Tablo 5.12. Kernel çeşidi lineer olan SVM modeli sonuçları.

Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlışlık Oranı (%)
Doğru	774	78,18
Yanlış	216	21,82
Toplam	990	100

SVM modelinde kernel çeşidinin polynomial olarak alınmasıyla elde edilen sonuçlar Tablo 5.13'deki gibidir. Sonuçlardan pratikte elde edilme olasılığı düşük fakat teoride yüksek olan en iyi modelin elde edildiği görülmektedir. Bu model teoride %100 sonuç veriyor gibi görünse de %100'e yakın bir sonuç beklenmelidir.

Tablo 5.13. Kernel çeşidi polynomial olan SVM modeli sonucu.

Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlışlık Oranı (%)
Doğru	990	100
Yanlış	0	0
Toplam	990	100

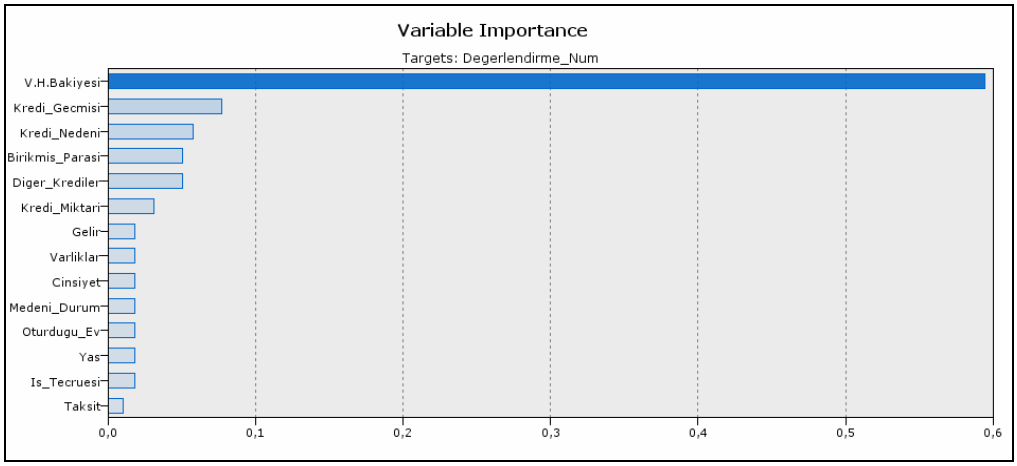
5.3.5.2. Karar Ağacı Modeli

Uygulamada SVM modelinden sonra C5 algoritmalı karar ağacı modeli kullanılan modeller içinde tahmin doğruluğu en yüksek olan ikinci modeldir. Modelin doğru tahmin etme oranı %82,83'tür.

Tablo 5.14. C5 algoritmalı karar ağacı modeli sonuçları.

Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlışlık Oranı (%)
Doğru	820	82,83
Yanlış	170	17,17
Toplam	990	100

Clementine programından çıkan sonuçlara göre karar ağacı modeli için değişkenlerin önem durumlarını gösteren grafik Şekil 5.27'deki gibidir. Grafiğe göre "V. H. Bakiyesi" değişkeni en yüksek öneme sahip değişken olup %59,5 değerinde öneme sahiptir. Geri kalan diğer değişkenlerin önem durumları ise Tablo 5.15.'de gösterilmiştir.



Şekil 5.27. Karar ağacı modeli değişkenlerinin önem durumu.

Tablo 5.15. Karar ağacı modeline göre değişkenlerin önem derecelerinin yüzdesel dağılımı.

Değişken	Önem Yüzdesi (%)
V.H.Bakiyesi	59
Kredi_Geçmişi	8
Kredi_Nedeni	6
Diğer_Krediler	5
Birikmiş_Parası	5
Kredi_Miktarı	3
Is_Tecrübesi	2
Yas	2
Oturduğu_Ev	2
Medeni_Durum	2
Cinsiyet	2
Varlıklar	2
Gelir	2
Taksit	1

Bu model sınıflandırma modülüne uygundur. Bu model karar ağaçlarının ya da kural setlerinin yapılması için C5.0 algoritmasını kullanır. Bir C5.0 modeli en fazla bilgiyi sağlayacak şekilde bölünmesiyle çalışmaktadır. Her bir örneklem ilk bölünmeyle tanımlanır, sonra yeniden bölünür. Genellikle bu örneklem daha fazla bölünemeyecek bir örneklem haline gelinceye kadar bu süreç devam eder. Sonuç olarak en düşük seviyedeki bölünme tekrarlanır ve modelin değerine katkıda bulunacak anlam taşımayan değerler model dışında bırakılır.

5.3.5.3. Lojistik Regresyon

Clementin programından elde edilen lojistik regresyon sonuçları Tablo 5.16'daki gibidir. tabloya göre lojistik regresyon modelinin doğru tahmin etme oranı %76,57'dir.

Tablo 5.16. Lojistik regresyon modeli sonucu.

Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlışlık Oranı (%)
Doğru	758	76,57
Yanlış	232	23,43
Toplam	990	100

Uygulamada lojistik regresyon model denkleminin elde edilebilmesi için Clementin programında lojistik regresyonun "Forwards" yöntemindan yararlanılmıştır. Tablo 5.17'de modelin adım adım işletilmesiyle elde edilen sonuçlar yeralmaktadır.

Tablo 5.17. Forward yöntemi ile uygulanarak adım adım oluşturulan lojistik regresyon modeli sonuçları.

Adım Sayısı	Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlış Oranı (%)
1. Adım	Doğru	694	70,10
	Yanlış	296	29,90
2. Adım	Doğru	726	73,33
	Yanlış	264	26,67
3. Adım	Doğru	743	75,05
	Yanlış	247	24,95
4. Adım	Doğru	751	75,86
	Yanlış	239	24,14
5. Adım	Doğru	750	75,76
	Yanlış	240	24,24
6. Adım	Doğru	747	75,45
	Yanlış	243	24,55
7. Adım	Doğru	751	75,86
	Yanlış	239	24,14

Tablo 5.17. Forward yöntemi ile uygulanarak adım adım oluşturulan lojistik regresyon modeli sonuçları. (Devam).

8. Adım	Doğru	752	75,96
	Yanlış	238	24,04
9. Adım	Doğru	758	76,57
	Yanlış	232	23,43

Tablodan da görüldüğü gibi ilk adımda tahmin doğruluk oranı %70,2 iken 9. adımda bu oran % 76,8'e çıkmaktadır. Lojistik regresyon denkleminin katsayıları 9. adım sonrasında elde edilen katsayılar oluşturmaktadır. Tablo 5.18'de lojistik regresyon modeli değişkenlerinin katsayıları yeralmaktadır.

Tablo 5.18. Lojistik regresyon modeli değişkenlerinin katsayıları.

Değişken	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Vadeli_Hesap_Bakiyesi	-0,584	0,069	71,193	1	0	0,557
Aylık_Taksit_Sayısı	0,034	0,007	25,092	1	0	1,035
Kredi_Geçmişi	-0,332	0,078	18,19	1	0	0,718
Birikmiş_Parası	-0,209	0,057	13,56	1	0	0,811
İş_Tecrübesi	-0,136	0,068	4,037	1	0,045	0,872
Medeni_Durum(1)	0,36	0,166	4,693	1	0,03	1,434
Varlıkları	0,175	0,081	4,661	1	0,031	1,192
Diğer_Krediler	-0,267	0,108	6,128	1	0,013	0,766
İşletme_Sahibi	-1,48	0,76	3,792	1	0,051	0,228
Constant	3,061	0,936	10,698	1	0,001	21,355

Yukarıdaki tablodan 9. adımdaki katsayılar (1.1.) denkleminde yerine yazılırsa:

$$Z = a + a_1*X_1 + a_2*X_2 + a_3*X_3 + a_4*X_4 + a_5*X_5 + a_6*X_6 + a_7*X_7 + ... \quad (1.1.)$$

Z = 3,061 - 0,584 * (Vadeli_Hesap_Bakiyesi) + 0,034 * (Aylık_Taksit_Sayısı) - 0,332 * (Kredi_Geçmişi) - 0,209 * (Birikmiş_Parası) - 0,136 * (Medeni_Durum(1)) + 0,175 * (Varlıkları) - 0,267 * (İşletme_Sahibi) - 1,48 * (İşletme_Sahibi)

$$P(Y) = 1/(1+Exp(-Z)) \quad (1.2.)$$

(1.2.) denklemine müşteri bilgileri girildiğinde kredisini batırıp batırmayacağı yani P(Y) olasılığı hesaplanır. P(Y)>0.5 olması iyi sonuçlanabilecek kredi anlamına gelip bu durumda Y=0 olarak alınır. P(Y)<0.5 olması ise kötü sonuçlanabilecek kredi demek olup Y =1 olması anlamına gelir.

6. SONUÇLAR VE TARTIŞMA

VM, istatistiğin teknoloji ile kaynaşması sonucu oluşmuş bir yöntemler serisidir. Bilgi teknolojilerinin gelişmesi ve konu ile ilgili SPSS Clementine gibi yeni programların üretilmesi, bu konuda yapılan çalışmaları kolaylaştırmıştır. Ancak VM sadece program kullanmak değildir. Büyük veri yığınları içinde karar vericilere yön gösterecek bilgileri keşfetmektir.

VM için iş deneyimine, sorunları tanımlayabilme ve istatistik bilgisine ihtiyaç vardır.VM veriden bilgi üreterek ortalama kararlar yerine veriye dayalı kararlar verilmesini destekleyen satışları, karlılığı, yenilikçiliği ve kaynak kullanımında etkinliği artıran önemli bir yönetim aracıdır. Veriye dayalı kararların kalitesi ve güvenilirliğini artırır.

Bireysel müşterilerin scorecard analizlerini gerçekleştirmek üzere yapılan bu çalışmada SVM, karar ağacı ve lojistik regresyon modelleri kullanılmıştır. Modellerde kullanılan veri, bankanın geçmişte kredi verdiği 1000 adet bireysel müşteriye ait bilgileri kapsamaktadır. Bu müşteriler aldıkları kredileri ya ödememiş ya da zamanında düzgün ödemiştir. Müşterilere ait bu bilgiler kulanılarak SPSS Clementine programında yukarıda sıralanan tekniklerin kullanılmasıyla geliştirilen kredi değerlendirme modelleri ile bundan sonra kredi başvurusu yapacak bireysel müşterilerin krediye uygun olup olmadıkları, krediyi geri ödeyip ödeyemeyecekleri tahmin edilebilecek ve müşterilere kredilerin verilip verilmeyeceği bu model sonucuna göre belirlenecektir. Tablo 6.1'de en iyi modeller ve başarı yüzdeleri yeralmaktadır. Tablodan da görüleceği gibi bireysel müşterilerin scorecard analizlerinin gerçekleştirildiği bu çalışmada en iyi model SVM olarak bulunmuştur. Kredi almak isteyen müşterilere ait bilgilerin bu modele girmesiyle kredilerini batırıp batırmayacakları büyük olasılıkla tahmin edilebilmektedir. %95,25 doğruluk oranına sahip SVM modelinden sonra sırayla %82,83 oranıyla karar ağaçları ve %76,57 oranıyla lojistik regresyon modelleri iyi tahminde bulunan modeller arasında yeralmıştır.

Tablo 6.1. En iyi modeller ve başarı tahmin yüzdeleri.

Model	Karşılaştırma Sonucu	Müşteri Sayısı	Doğru/Yanlış Oranı (%)
SVM	Doğru	943	95,25
	Yanlış	47	4,75
	Toplam	990	100
Karar Ağacı	Doğru	820	82,83
	Yanlış	170	17,17
	Toplam	990	100
Lojistik Regresyon	Doğru	758	76,57
	Yanlış	232	23,43
	Toplam	990	100

Değişen büyük ekonomik göstergeler ve ortaya çıkan krizlerde aynı scorecard'ların kullanımı müşterilerin davranış ve tutumlarını değiştireceği için çok da doğru bir yaklaşım olmayacaktır. Bu gibi durumlarda bankaların risklerini minimize edecek modeller geliştirmeleri gerekmektedir. Bu nedenle toplanan verilerden yeni bir scorecard modeli oluşturulup kredi almak isteyen müşterilerin bu model üzerinden değerlendirimesi gerekmektedir.

KAYNAKLAR

[1] **Argüden, Y. ve Erşahin B.,** 2008, VM, ARGE Danışmanlık Yayınları, İstanbul.

[2] **Berger, J.O. and Selke, T.,** 1987, Testing a Point Null Hypothesis: The Irreconcilability of P-Values And Evidence, *J. Am. Stat. Assoc., Vol.* 82, 397, pp. 112-122.

[3] **Cabena, P., Hadjinian P., Stadler, R.,Verhess, J. and Kamber, M.,** 1998, Discovering Data Mining: From Concept to Implementation, Prentice, Hall, New Jersey.

[4] **Cios J.K., Pedrycz W., Swiniarski W. R.,Kurgan A.L.,** 2007, Data Mining A Knowledge Discovery Approach. Springer, New York.

[5] **Cleveland, W.S.**, 1979, Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 74:829-836.

[6] **Çomak E.,** 2004, Destek Vektör Makineleri çoklu sınıf problemleri için çözüm önerileri , *Yüksek Lisans Tezi,* Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

[7] **Dunham M.H.,** 2003, Data Mining Introductory and Advanced Topics, Pearson Education Inc. New Jersey.

[8] **Elif C. C.,** 2007, VM ve VA'larnın Perakendecilik Sektöründe Uygulamaları, *Tezsiz Yüksek Lisans Projesi,* Dokuz Eylül Sosyal Bilimler Enstitüsü, İzmir.

[9] **Fabris, P.** 1998, Advanced Navigation, CIO, May 15.

[10] **Fung, G. ve Mangasarian, O. L.,** 2002, Incremental Support Vector Machine Classification Second SIAM International Conference on Data Mining.

[11] **Hamparsum Bozdogan,** 2004, Statistical Data Mining and Knowledge Discovery. CRC Press LLC , New York.

[12] **Han, J. and Kamber, M.,** 2001, Data Mining Concepts and Techniques, Academic Pres. New York.

[13] **Hui, S.C. and Jha, G.,** 2000, Mastering Data Mining for Customer Service Support, *Information & Management*, 38(1), 1-13.

[14] **Perner P.,** 2002, Data Mining on Multimedia Data, Springer-Verlag Berlin Heidelberg.

[15] **P. Ravi Kumar, V. Ravi,** 2007, "Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques-A Review", *European Journal of Operational Resaerch*, Elseiver, Vol. 180, 2007.

[16] **Roiger, R.J. and Geatz M.W,** 2003, Data Mining : A Tutorial-Based Primer, Pearson Education Inc. USA.

[17] **Sumathi S. and Sivanandam S.N.,** 2006, Introdution to Data Mining and Its Applications, Springer-Verlag Berlin Heidelberg.

[18] **Thomas L.C.,** 2000, A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16, 149-172.

[19] **Tolun S.,** 2008, Destek Vektör Makineleri banka başarısızlığının tahmini üzerine bir uygulama, *Doktora Tezi*, İstanbul Üniversitesi Sosyal Bilimleri Enstitüsü, İstanbul.

[20] **Vecihe E. G.,** 2007, Veri Madenciliği'nde Duyarlılık, *Yüksek Lisans Tezi*, İ.T.Ü. Fen Bilimleri Enstitüsü, İstanbul.

[21] **Yurtsever U.,** 2002, Veri Madenciliği ve Uygulaması, *Yüksek Lisans Tezi*, Sakarya Üniversitesi Sosyal Bilimleri Enstitüsü, Sakarya.

ÖZGEÇMİŞ

Adı Soyadı: Akın SÖYLEMEZ

Doğum Tarihi : 02.01.1976

Doğum Yeri : Sivas

Medeni Durum: Evli

Mezun Olduğu Üniversite: İstanbul Teknik Üniversitesi, İşletme Fakültesi, İşletme Mühendisliği Bölümü (1994-1999).

İş Tecrübeleri:

2006- (halen devam ediyor): Finansbank, Süreç Yönetimi Yönetmeni, İstanbul.

Birim Süreç Projeleri: Genel Müdürlük birimleri ve şubelerinin süreçlerinin çıkartılması ve süreçler üzerinden öneriler geliştirilmesi, işyükü hesaplamalarının yapılması. Projeler; Şubeler Projesi, Hazine Birimi Projesi, Operasyon Merkezi ile Şube Bağlantısı Projesi, Özel Bankacılık Birimi Projesi.

Ürün Süreç Projeleri: Bankanın ürünlerinin süreç bazlı analizi, çevrim süreleri, işyükü hesabı ve önerilerin geliştirilmesi. Projeler; Mortgage Projesi, POS Projesi.

Norm Kadro Projesi: Şubede çalışan tüm personelin ünvan bazında norm kadrolarının aylık olarak sistemden alınması için microstrategy üzerinden model kurulması ve uygulamaya geçirilmesi, takibinin yapılması.

2004-2006: Türk Ekonomi Bankası, İş Geliştirme Yetkilisi, İstanbul

Üst yönetime sunulmak üzere Genel Müdürlük birimleri ve şubeler ile ilgili aylık raporlar ve sunumların hazırlanması ve her ay bu raporların geliştirilmesi.

Şubelerin ve Genel Müdürlük birimlerinin norm kadrolarının hesaplanması ve sonuçların üst yönetim ve ilgili departmanlar ile paylaşılması.

Süreç çizimlerinin ve süreç geliştirme çalışmalarının yapılması.

2002-2004: UPS Müşteri Temsilcisi, İstanbul.

Bölgedeki müşterilere UPS hizmetlerinin tanıtılması ve satış anlaşmalarının yapılması, müşteri memnuniyetinin sağlanması.

1998-2000: MNG Bank (T-Bank), İş Geliştirme Analisti, İstanbul

Genel Müdürlük birimlerinin ve Şubelerin iş süreçlerinin oluşturulması ve süreçlerin dokümantasyonunun yapılması.

Yabancı Dili: İyi derecede İngilizce.

Bilgisayar Bilgileri: MS Office (Excel - İleri Düzeyde), MS Project, Control ES, Visio, Minitab, SPSS Clementine, Microstrategy, Microsoft System Engineering Certificate kursunu tamamladım.

Aldığı Eğitimler: Six Sigma Yeşil Kuşak Eğitimi, Matris Danışmanlık 04.02.2008-07.05.2008

Sertifika Bilgileri: ABC (Activity Base Costing)-Sistema, Balanced Scorecard-Ironman Consulting, Problem Çözme Teknikleri-Bankalar Biriliği, Süreç Yönetimi - İstanbul Kurumsal, Microsoft Project - İstanbul Kurumsal, Yaratıcı Düşünme Teknikleri - Bankalar Birliği, Değişimlerin Oluş Nedenleri - Bankalar Birliği