İstatistik Araştırma Dergisi Cilt 02 No 02 Özel Sayı Sayfa 135-150 Ağustos 2003

# Veri Madenciliği Tekniklerini Kullanarak Banka Müşterileri Bölümlendirmesi Ve Kredi Skorlama Modeli

Pelin BİÇEN* S.Ümit Oktay FIRAT**


## OZET

*İşletme ve bilimsel içerikli veri tabanlarının gün geçtikçe büyümesi, veri tabanlarında bulunan verinin analiz edilmesini ve yorumlanmasını zorlaştırdı. Bu noktada, veri tabanı analiz sürecini otomatileştirecek yeni nesil tekniklere ve araçlara ihtiyaç duyulmaya başlandı. Bu anlamda, bu teknikler ve araçlar veri tabanlarında bilgi keşfi ve veri madenciliği teknikleri olarak bilinen ve çok hızlı gelişen bir alana konu oldular.*

*Bu çalışmada, ilk olarak veri tabanlarında bilgi keşfi ve veri madenciliği kavramları daha sonra da veri madenciliği modelleri açıklanmıştır. Uygulama aşamasında, günümüz işletme dünyasında çok sık karşılaşılan, müşterilerin kredi taleplerinin değerlendirilmesi ve karlılık durumlarına göre müşterilerin bölümlendirilmesi problemi, veri madenciliği sınıflandırma ve tahmin modelleri uygulanarak çözümlenmiştir. Çözüm sürecinde, SAS Enterprise Miner veri madenciliği paketi kullanılmıstır.*


***Anahtar Kelimeler***: *Veri Tabanlarında Bilgi Keşfi, Veri Madenciliği, Tahmin Modelleri, Sınıflandırma, Regresyon Analizi, Karar Ağaçları, Kümeleme Analizi, K-Ortalamalar Algoritması, Multinomial Regresyon Analizi.*

## 1. VERİ TABANLARINDA BİLGİ KEŞFİ SÜRECİ ve VERİ MADENCİLİĞİ

Karmaşık ve dinamik bir ekonomi ortamında işletmelerin karar verme süreçlerindeki etkinlikleri, çok değişkenli büyük miktarlardaki veri kümelerinde saklı bulunan bilginin elde edilmesi ve işlenmesine bağlıdır. Boyutları hızla artan veriden anlamlı bilgiler çıkarmak için bilgisayar hızlarının ve güçlerinin artmasını sağlayacak yeni teoriler ve araçlar geliştirilmektedir. Bu teoriler ve araçlar veri tabanlarında bilgi keşfi süreçlerinin konusunu oluşturmaktadır. ( Fayyad, Paitesky- Shapiro, Padhric, Uthurusamy,

* Ar. Gör., İstanbul Bilgi Üniversitesi İletişim Fakültesi Reklamcılık Bölümü (pelinb@bilgi.edu.tr)
**Prof. Dr., Marmara Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliği Bölümü (ufirat@eng.marmara.edu.tr)

Veri kendi başına değersiz olduğundan verinin amacımız doğrultusunda bilgiye çevrilmesine veri analizi (data analysis) denmektedir. Veri analizi yaparak bir mal için bir sonraki ayın satış tahminlerini çıkarabilir, müşterileri satın aldıkları mallara göre gruplayabilir, yeni çıkacak bir ürün için potansiyel müşterileri belirleyebilir, müşterilerin hareketlerini izleyerek ve inceleyerek onların davranışları ile ilgili tahminler yapabiliriz. Milyonlarca malın ve müşterinin olabileceği düşünülürse bu analizin otomatik olarak yapılmasının zorunluluğu ortaya çıkmaktadır. Bu noktada veri madenciliği devreye girmektedir. Veri madenciliği, büyük miktardaki veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır. (Alpaydın, 2000)

Veri madenciliği modelleri genel olarak tahmin edici model ve tanımlayıcı model olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici modeller veri madenciliğinin en temel görevidir. Eğitim veri kümesi (training data set) sonuçları önceden bilinen örneklerden, gözlemlerden ve kayıtlardan oluşmaktadır. Herbir gözlem girdi değişkenlerinin ve hedef değişkenin bir vektörü niteliğindedir. Eğitim veri kümesi kullanılarak girdi değişkenlerden hedef değişkenin tahmin edilmesini sağlayan bir model (kural) oluşturulur. (Potts, 1998) Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. (Akpınar, 2000)

## 2. SINIFLANDIRMA ve TAHMİN YÖNTEMLERİ KULLANARAK BANKA MÜŞTERİLERİ BÖLÜMLENDİRMESİ VE KREDİ SKORLAMA MODELİ

Çalışmanın bu kısmında gerçek bir veri kümesine veri madenciliği modellerinden sınıflandırma ve tahmin yöntemlerini baz alan veri madenciliği teknikleri (karar ağaçları, regresyon ve kümeleme analizi) uygulanarak bu tekniklerin sonuçlarının işletmeler için önemi gösterilecek ve tekniklerin performanslarının karşılaştırılması yapılacaktır.

### 2.1. Araştırmanın Amacı

Çalışmada, bir bankanın bireysel bankacılık departmanına borç konsolidasyonu ve konut yenileme nedeniyle gelen kredi taleplerinin, veri madenciliği teknikleri uygulanarak değerlendirilmesi ve kredi taleplerinin kabul veya red kararının bu doğrultuda verilmesi amaçlanmaktadır.

Çalışmada ilk olarak bankaya gelen kredi başvurularının kabul veya red kararının otomatik olarak verileceği tahmin ve denetimli (supervised) sınıflandırma modeli oluşturulacaktır. Bu amaçla tahmin modellerinden regresyon analizi ve denetimli sınıflandırma modellerinden karar ağacı analizi modele eklenecektir. Modellerin performans karşılaştırılmaları yapıldıktan sonra bankanın mevcut problemini çözmede etkin olan model değerlendirmeye alınacaktır. Bankanın kredi başvuru taleplerini değerlendirmede etkin olabilecek bir başka teknik olan kümeleme analizi de uygulamanın diğer bir aşamasını oluşturacaktır. Kümeleme analizinde müşterilere ait değişkenler detaylı şekilde incelendikten sonra kümeleme analizinin hiyerarşik olmayan yöntemlerinden K-

ortalamalar algoritması uygulanacaktır. Kümeler analitik ve işletme filtrelerinden geçirildikten sonra profillendirilecektir. Kümelerin profillerini belirleyen değişkenlerin ne oranda doğru tahmin edildiğini test edebilmek için kümeleme analizi sonunda tahmini multinomial regresyon analizi yapılacaktır.

## 2.2. Anakütle ve Değişkenlerin Tanımı

Uygulamada kulllanılan veri kümesi tüm ana kütle olarak belirlenmiştir. Veri kümesinde 5960 müşteriye ait 13 adet değişken bulunmaktadır. Bu değişkenlerin isimleri, değişken tipleri (ikili, nominal, sırasal, aralık ölçekli, oran ölçekli), veri madenciliği modelindeki rolleri ve kod açılımları Tablo 1'de ayrıntılı olarak verilmektedir.

**Tablo 1**. Değişkenlerin tanımlanması

| Değişken İsmi | Değişkenin Modeldeki Rolü | Değişken Tipi | Değikenin Kod Açılımı |
|---|---|---|---|
| BAD | girdi | ikili | 1=Borcunu ödememiş<br>0=Borcunu ödemiş |
| REASON | girdi | ikili | HomeImp:ev yenileme<br>Debcon: Borç konsolidasyonu |
| JOB | girdi | Nominal | 6 meslek kategorisi |
| LOAN | girdi | ralık ölçekli | Talep edilen kredi miktarı |
| MORTDUE | girdi | ralık ölçekli | Konut ipotek değeri |
| VALUE | girdi | ralık ölçekli | Mevcut mal varlığının bugünkü değeri |
| DEBTINC | girdi | ralık ölçekli | Borç gelir oranı |
| YOJ | girdi | ralık ölçekli | Müşterinin mevcut mesleğinde geçirdiği toplam sene |
| DEROG | girdi | ralık ölçekli | Borç ihbar belgesi sayısı |
| CLNO | girdi | ralık ölçekli | Kredi başvuru sayısı |
| DELINQ | girdi | ralık ölçekli | Ödenmeyen kredi sayısı |
| CLAGE | girdi | alık ölçekli I | İlk yapılan kredi başvuru süresinden itibaren ay bazında geçen toplam sure |
| NINQ | girdi | ralık ölçekli | Kredi soruşturma sayısı |

Kaynak: Wielenga Doug , Lucas Bob & George Jim, "Applying Data Mining Techniques- Course Notes", USA, SAS Ins., 1999, s:106

Veri kümesindeki tüm değişkenler model oluşturma sürecinde aktif rol oynayacağından bütün değişkenler kullanılır statüdedir. Verinin SAS Enterprise Miner'a

yerleştirilmesinden sonra veri madenciliği modelinde eğitim ve değerlendirme kümesi olarak kullanacağımız veri kümelerinin oranlarının belirlenmesi gerekmektedir. Eğitim ve değerlendirme kümelerinin oranları sırasıyla %70 ve %30 olarak belirlenmiştir. Ana kitlede "borcunu ödemiş" ve "borcunu ödememiş" müşteri sayısının tüm ana kitleye oranı sırasıyla %80 ve %20 olduğundan bu oranların oluşturulan eğitim ve değerlendirme kümesinde korunması sınıflandırma ve tahmin modelinin performansını olumlu yönde etkileyecektir. Bu amaçla eğitim ve değerlendirme kümelerindeki müşteriler katmanlı örneklem yöntemine (stratified sampling method) göre seçilmiştir.

### 2.3. Betimsel İstatistikler

Veri kümesimizde bulunan 10 aralık ölçekli (interval) değişkenin betimsel istatistiği incelenerek normal dağılıma uymayan ve aykırı değerlere sahip değişkenler tablo 2'de ayrıntılı biçimde verilmiştir. Veride betimsel istatikler incelenerek normal dağılıma uymayan değişkenler için uygun transformasyon yöntemleri, eksik değerler için ise uygun tamamlama yöntemleri belirlenecektir.

**Tablo 2.** Aralık Ölçekli Değişkenlerin Betimsel İstatistiği

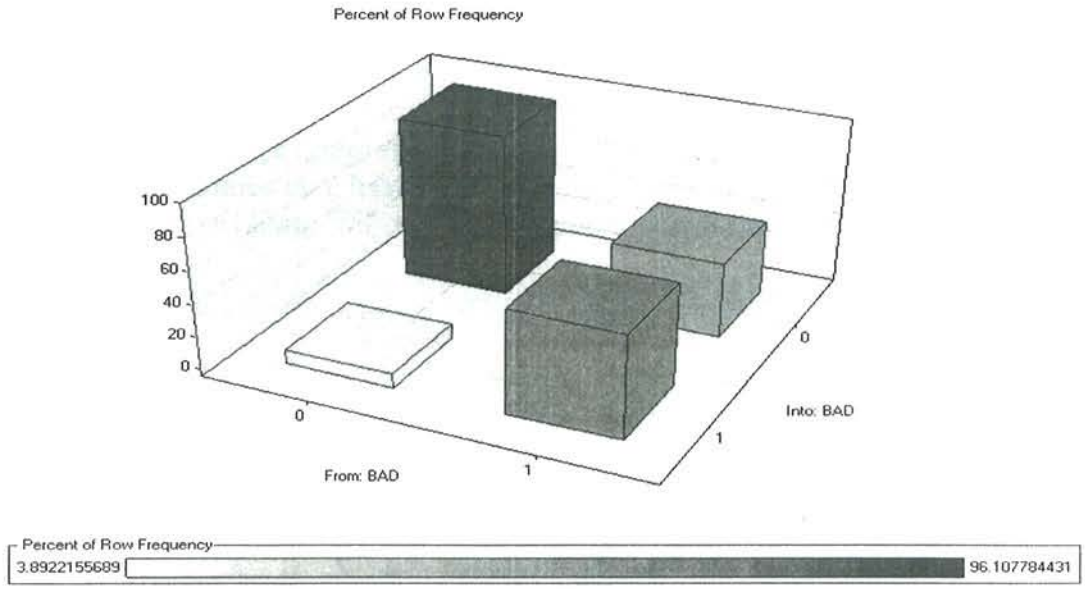
| Değişken ismi | Min. değer | Maks. değer | Aritmetik ort. | Standart sapma | Eksik değer(%) | Skewness | kurtosis |
|---|---|---|---|---|---|---|---|
| CLAGE | 0 | 1168.2 | 179.77 | 85.81 | %5 | 1.34 | 7.60 |
| CLNO | 0 | 71 | 21.296 | 10.14 | %4 | 0.78 | 1.16 |
| DEBTINC | 0.52 | 203.31 | 33.78 | 8.60 | %21 | 2.85 | 50.5 |
| DELINQ | 0 | 15 | 0.45 | 1.13 | %10 | 4.02 | 23.36 |
| DEROG | 0 | 10 | 0.25 | 0.85 | %12 | 5.32 | 36.87 |
| LOAN | 1100 | 89900 | 18608 | 11207 | %0 | 2.02 | 6.93 |
| MORTDUE | 2063 | 399550 | 73761 | 44458 | %9 | 1.81 | 6.48 |
| NINQ | 0 | 17 | 1.1861 | 1.73 | %9 | 2.62 | 9.78 |
| VALUE | 8000 | 855909 | 101776 | 57386 | %2 | 3.05 | 24.36 |
| YOJ | 0 | 41 | 8.92 | 7.574 | %9 | 0.98 | 0.37 |

Uygulamada, tamamlama yöntemlerinden aralık ölçekli, nominal ve ikili değişkenler için ağaç tamamlanması (tree imputation) yöntemi ve model kurma

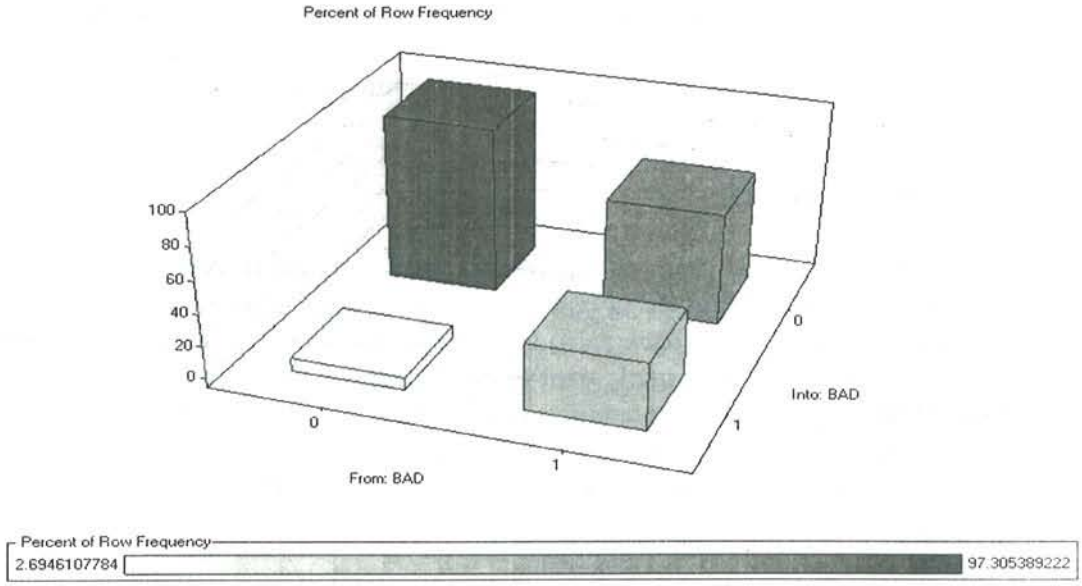
aşamasında SAS Enterprise Miner'da değişken dağılımlarının normalizasyonunu hedefleyen transformasyon yöntemlerinden girdi değişkenlerinin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu (optimal binning for relationships to target tranformation) yöntemi en uygun yöntem olarak belirlenmiştir.

## 2.4. Regresyon Analizi

Tahmin edilecek bağımlı değişkenin model rolü ikili değişken (1/0) olduğundan uygulamada lojistik regresyon uygun regresyon tekniği olarak belirlenmiştir. Regresyon methodlarından geri adım (backward) ve adım adım (stepwise) methodları alternatif olarak analize eklenmiştir. Uygulamada yapılan analiz sonucu da bu doğrultuda gerçekleşmiştir. Adım adım ve geri adım yöntemine göre kurulan lojistik regresyon modelinin sonuçları sırasıyla şekil 1 ve 2'de karşılaştırma matrisi (confusion matrix) olarak verilmiştir. Bu matrislere göre adım adım metodunu kullanan lojistik regresyon modeli daha başarılı olmuştur.

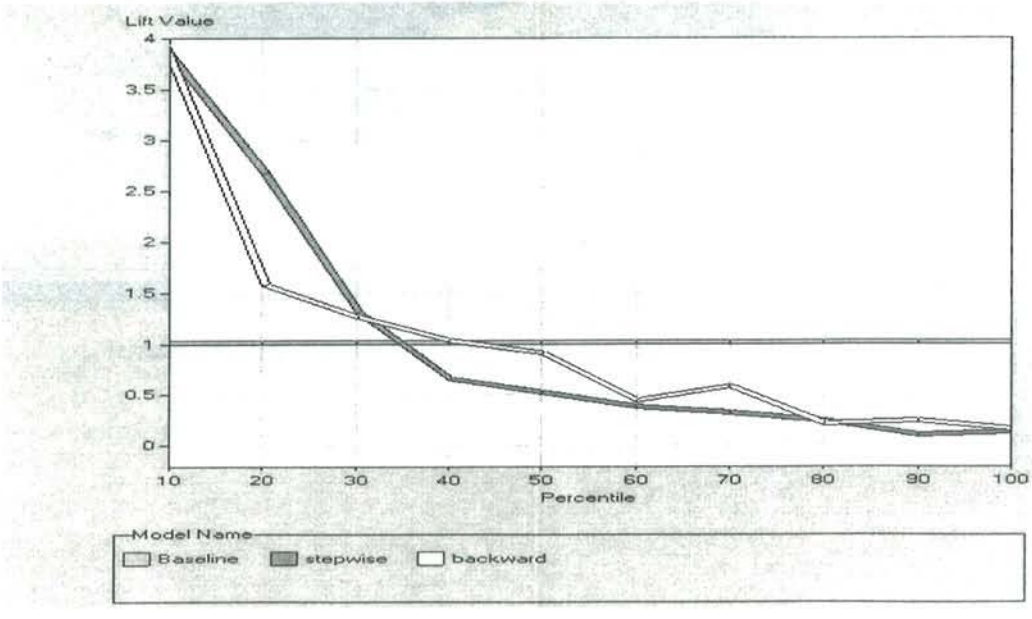



**Şekil 1.** Adım adım lojistik regresyon modeline göre karşılaştırma matrisi




**Şekil 2.** Geri adım lojistik regresyon modeline göre karşılaştırma matrisi

Oluşturulan regresyon modelinin değerlendirilmesi aşamasına gelindiğinde modelin asansör grafiğinin yorumlanması gerekmektedir. Şekil 3'de adım adım ve geri adım lojistik regresyon modellerinin kümülatif olmayan asansör (lift) grafikleri incelenmektedir.



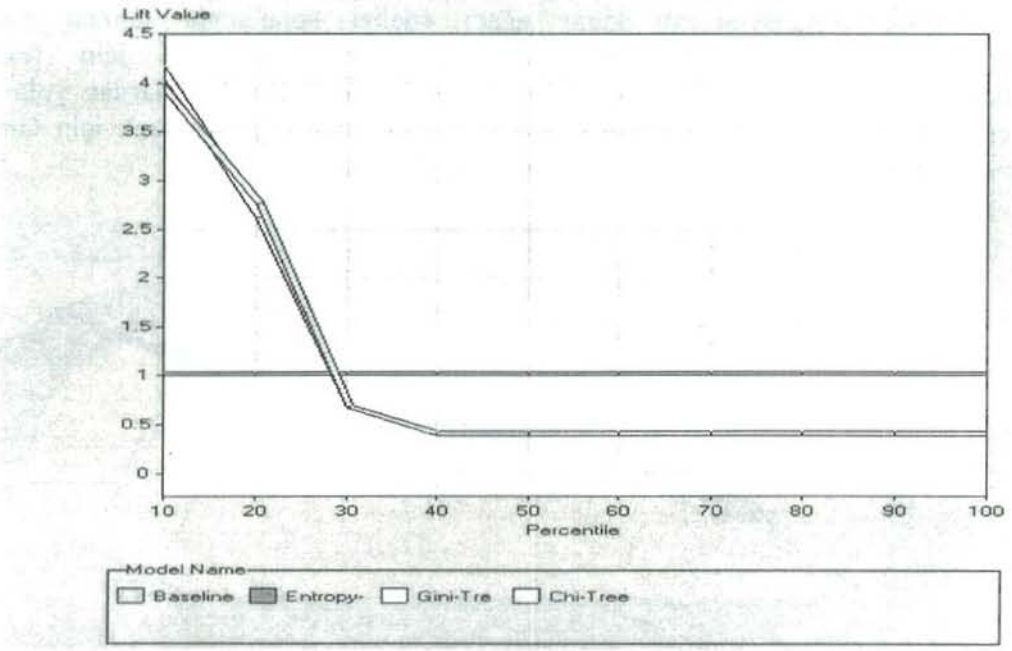

**Şekil 3.** Adım adım ve geri adım regresyon modellerinin kümülatif olmayan asansör grafiklerinin karşılaştırılması

Şekil 3.'deki kümülatif olmayan asansör grafiği incelendiğinde daha önce tespit ettğimiz adım adım regresyon modelinin başarı grafiği daha net gözlenmektedir. Ayrıca geri adım regresyon modelinin kümülatif olmayan asansör grafiğinin iniş çıkışlar göstermesi başarısız bir model olduğunun başka bir göstergesidir.

## 2.5. Karar Ağaçları Analizi

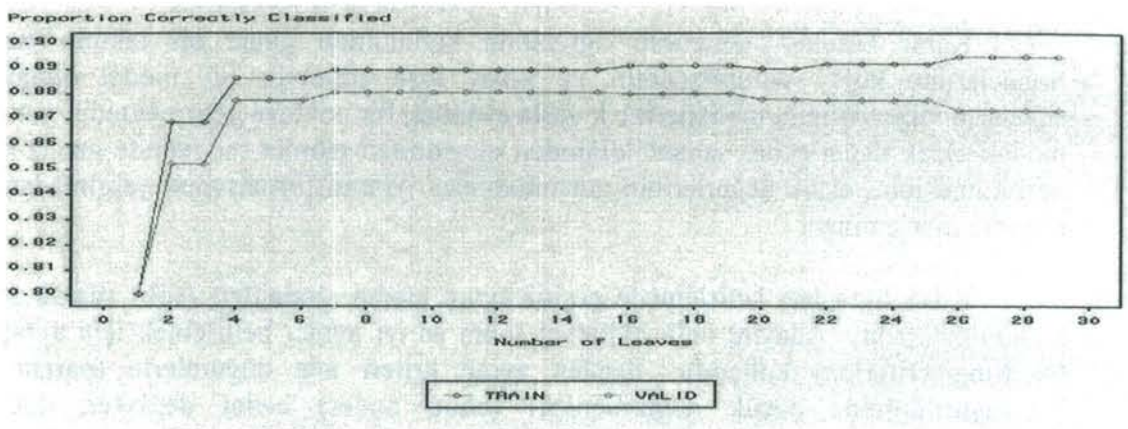
Karar ağaçları, denetimli öğrenimin kullanıldığı güçlü bir tahmin modelidir. Sonuçlarının kolay yorumlanabilir ve kolay inşa edilebilir bir model olması karar ağaçlarını diğer tahmin modellerine kıyasla avantajlı bir noktaya getirmektedir. Karar ağacı modeli eksik değerler ile çalışabildiğinden oluşturulan tahmin modelinde analiz edilecek veri kümesinin eksik değerlerinin tamamlanması ve transformasyonu yapılmadan önceki durumu incelenmiştir.

Aday ayraçları belirlemede çeşitli ayraç arama stratejileri (split search strategy) kullanılmaktadır. Adaylar belli olduktan sonra en iyi ayracı belirlemek için ayraç kriteri (splitting criterion) kullanılır. Seçilen ayraç kriteri ana düğümlerle (parent nodes) karşılaştırıldığında çocuk düğümlerdeki (child nodes) hedef değişken dağılımının değişkenliğinin azalırlığını ölçmektedir. Sınıflandırma ağaçlarında üç tür ayraç kriteri kullanılmaktadır: Entropi, Gini ve Ki-kare testi. Analiz sürecimizde üç ayraç kriterini de performans karşılaştırması yapmak amacıyla inceleyeceğiz. Şekil 4'de Gini, Entropi ve Ki-kare testi ayraç kriterlerine göre oluşturulmuş karar ağacı modellerinin kümülatif olmayan asansör grafikleri karşılaştırılmalı olarak verilmektedir.




**Şekil 4.** Gini, Entropi ve Ki-kare testi ayraç kriterlerine göre oluşturulmuş karar ağacı modellerinin karşılaştırmalı kümülatif olmayan asansör grafikleri

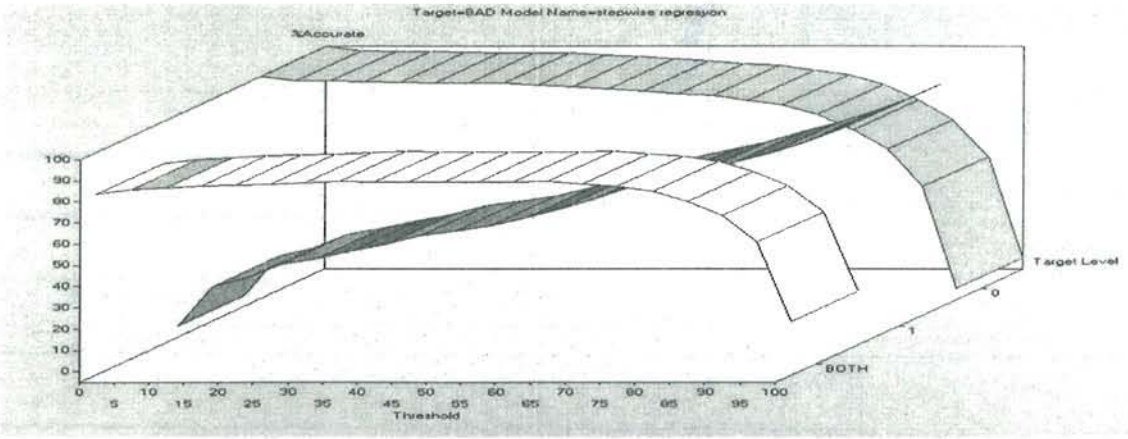
Şekil 4.'deki asansör grafiklerinde Entropi ve Ki-kare testi kriterlerine göre oluşturulan karar ağacı modelleri aynı sonucu vermektedir. Gini kriterine göre oluşturulan karar ağacı modeli ise diğer iki modele göre daha iyi bir sonuç üretmektedir. Şekil 5'de Gini kriterine göre oluşturulan karar ağacı modelinde yaprak oluşumunu eğitim ve değerlendirme kümeleri çerçevesinde ifade eden bir grafik bulunmaktadır. Bu grafiğe göre Gini karar ağacı 7 yapraktan oluşmaktadır.



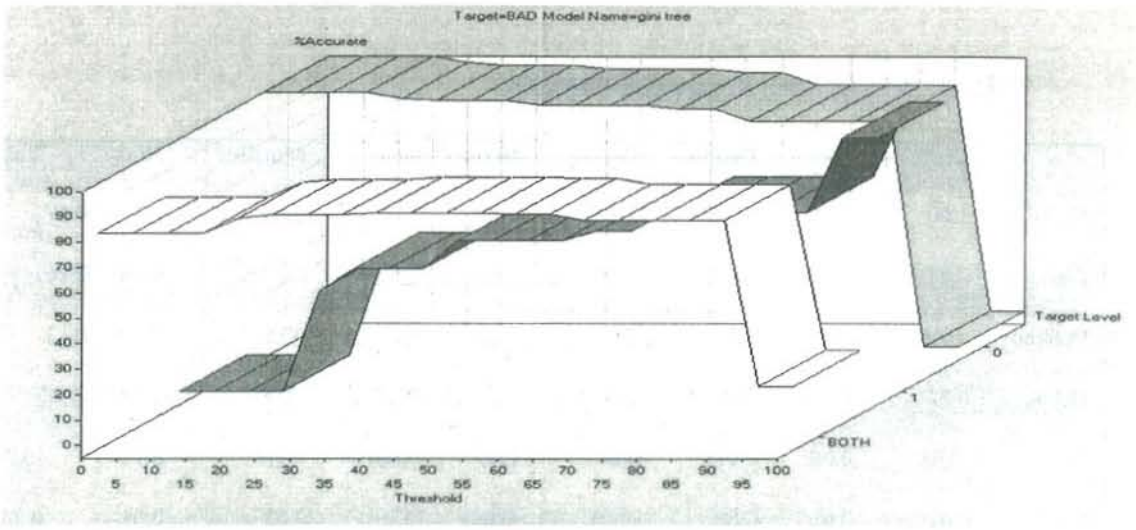

**Şekil 5.** Yaprak sayısının belirlenmesi

## 2.6. Uygulanan Tahmin ve Denetimli Sınıflandırma Modeli Yorumu

Yapılan regresyon ve karar ağacı analizi sonucunda tahmin modelinin performansını arttıran veri madenciliği tekniğini bulabilmek için tekniklerin karşılaştırılmalı asansör grafiklerine bakmak gerekecektir. Bu sonuçlardan yola çıkarak mevcut probleme dair en iyi tahmin modelini sunan analizi bulabilmek için Gini karar ağacı analizi ve adım adım regresyon analizi karşılaştırılacaktır.




**Şekil 6.** Adım adım regresyon analizinde eşik değere bağlı doğru sınıflandırma oranı




**Şekil 7.** Gini karar ağacı analizinde eşik değere bağlı doğru sınıflandırma oranı

Şekil 6 ve 7'deki grafikleri incelediğimizde, adım adım regresyon modelinin kötü ve iyi müşteri sınıflandırma oranlarının Gini karar ağacı analizine göre daha iyi sonuç verdiğini görmekteyiz.

## 2.7. Kümeleme Analizi Uygulaması

Segmentasyon analizine başlamadan önce bu analizin sürekli bir süreç olduğunu anlamak gerekmektedir. Bir başka önemli nokta ise müşteri segmentasyonunun doğru ve yanlış bir yolu olmadığının, asıl amacın işletme değeri olan segmentler yaratmak olduğunun anlaşılmasıdır. Kümeleme analizine başlamadan önce ilk yapılması gereken, değişkenlerin iyi tanımlanmasıdır. Değişkenlerden bazıları segmentlerin oluşturulmasında önemli rol oynayacağından bu değişkenler aktif (active) olarak tanımlanmıştır. Diğer değişkenler ise segmentler oluşturulduktan sonra, segmentlerin işletme değeri olduğunun anlaşılmasında rol oynayacağından betimsel (descriptive) olarak adlandırılmıştır.Aktif ve betimsel değişkenlerin belirlenmesi sürecinde değişkenlerin birbirleriyle ne kadar ilişkili olduğunu anlayabilmek için korelasyon analizi yapmak gerekmektedir. Bu analiz sonucu yüksek korelasyona sahip değişkenlerin birarada analize alınmaması kümeleme analiz sürecini çok farklı etkileyecektir. Korelasyon analizinde korelasyon değeri 0.5'ten fazla çıkan değişkenler analiz sürecinde beraber değerlendirilmeyecektir. Tablo 3'de değişkenlerin korelasyon değerleri bulunmaktadır.

**Tablo 3.** Değişkenlerin Korelasyon Değerleri

| | Clage | Clno | Debtinc | Delinq | Derog | Loan | Mortdue | Ninq | Value | Yoj |
|---|---|---|---|---|---|---|---|---|---|---|
| Clage | 1.00 | 0.24 | -0.05 | 0.22 | -0.08 | 0.09 | 0.14 | 0.11 | 0.17 | 0.20 |
| Clno | 0.24 | 1.00 | 0.18 | 0.16 | 0.06 | 0.07 | 0.32 | 0.08 | 0.26 | 0.02 |
| Debtinc | -0.05 | 0.18 | 1.00 | 0.05 | 0.02 | 0.08 | 0.15 | 0.14 | 0.13 | -0.05 |
| Delinq | 0.22 | 0.16 | 0.05 | 1.00 | 0.21 | -0.03 | -0.001 | 0.07 | -0.01 | 0.04 |
| Derog | -0.08 | 0.06 | 0.01 | 0.21 | 1.00 | -0.001 | -0.04 | 0.17 | -0.04 | -0.06 |
| Loan | 0.09 | 0.07 | 0.08 | -0.03 | -0.001 | 1.00 | 0.23 | 0.04 | 0.33 | 0.10 |
| Mortdue | 0.14 | 0.32 | 0.15 | -0.001 | -0.04 | 0.23 | 1.00 | 0.03 | **0.87** | -0.08 |
| Ninq | 0.11 | 0.08 | 0.14 | 0.07 | 0.17 | 0.04 | 0.03 | 1.00 | -0.004 | -0.007 |
| Value | 0.17 | 0.26 | 0.13 | -0.01 | -0.04 | 0.33 | **0.87** | -0.004 | 1.00 | 0.007 |
| Yoj | 0.20 | 0.02 | -0.05 | 0.04 | -0.06 | 0.10 | -0.08 | -0.07 | 0.007 | 1.00 |

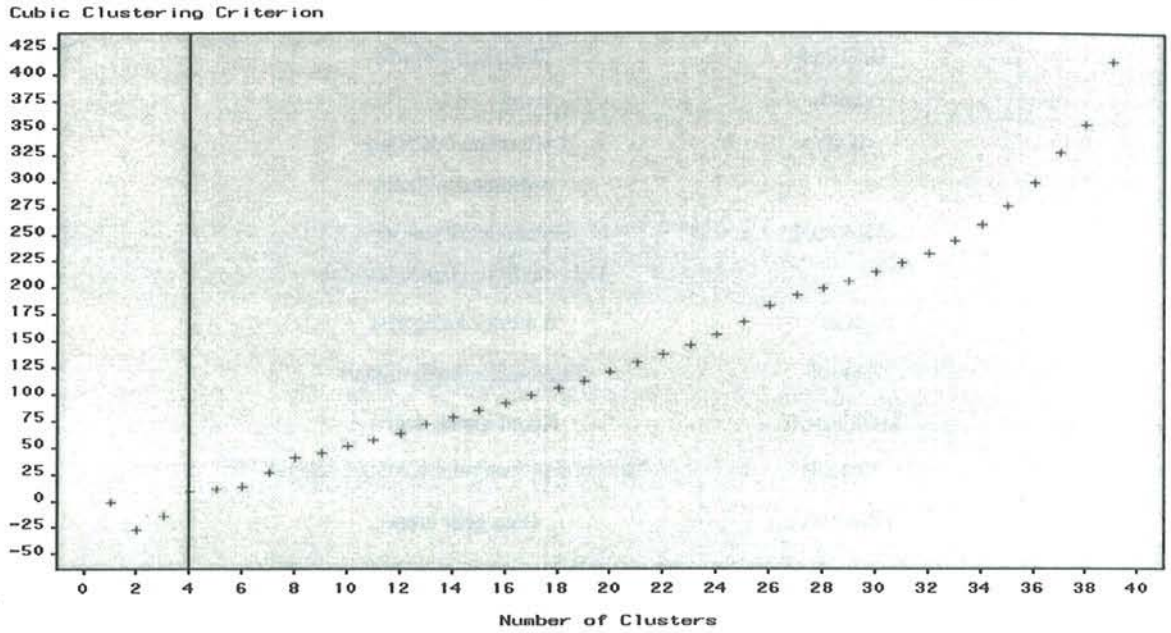
Değişkenlerin bankacılık sektöründeki tanımları gereği ve korelasyon analizi sonucu kümeleme analizinde değişkenlerin rolü tablo 4'de ayrıntılı biçimde verilmiştir.

**Tablo 4.** Değişkenlerin Kümeleme Analizindeki Rolleri

| Değişken İsmi | Değişken Tanımı | Değişken Rolü |
|---|---|---|
| BAD | 1=Borcunu ödememiş<br>0=Borcunu ödemiş | Betimsel |
| REASON | HomeImp:ev yenileme<br>Debcon: Borç konsolidasyonu | Betimsel |
| JOB | 6 meslek kategorisi | Betimsel |
| LOAN | Talep edilen kredi miktarı | Aktif |
| MORTDUE | Konut ipotek değeri | Betimsel |
| VALUE | Mevcut mal varlığının bugünkü değeri | Aktif |
| DEBTINC | Borç gelir oranı | Aktif |
| YOJ | Müşterinin mevcut mesleğinde geçirdiği toplam sene | Betimsel |
| DEROG | Borç ihbar belgesi sayısı | Aktif |
| CLNO | Kredi başvuru sayısı | Aktif |
| DELINQ | Ödenmeyen kredi sayısı | Aktif |
| CLAGE | İlk yapılan kredi başvuru süresinden itibaren ay bazında geçen toplam sure | Betimsel |
| NINQ | Kredi soruşturma sayısı | Betimsel |

## 2.8. Kümeleme Analizinin Analitik ve İşletme Değeri Kontrolü

Verideki değişkenlerin aktif ve betimsel olarak tanımlanması, aykırı değerlerin elenmesi ve eksik değerlerin tamamlanmasından sonra kümeleme analizi sürecine başlayabiliriz. Kümeleme analizinde hiyerarşik olmayan yöntemlerden K-ortalamalar yöntemi kullanılmıştır. SAS Enterprise Miner'da kümeleme analizi yapan CCC (Cubic Clustering Criterion) aracına göre ilk olarak küme sayısı otomatik olarak belirlenir. Şekil 8'de CCC grafiği görülmektedir.




**Şekil 8.** Cubic Clustering Criterion grafiği

CCC grafiğine göre otomatik olarak 4 küme belirlenmiştir. Oluşturulan kümelerin istatistikleri tablo 5'de incelenmektedir.

**Tablo 5.** K-ortalamalar algoritması sonucu oluşan kümelerin istatistik değerleri

| Kümeler | Küme Frekansları | Küme merkezinden maksimum uzaklık | En yakın küme | En yakın kümeye uzaklık |
|---|---|---|---|---|
| 1 | 2340 | 1.147 | 2 | 0.781 |
| 2 | 2387 | 1.079 | 1 | 0.781 |
| 3 | 584 | 1.315 | 2 | 1.096 |
| 4 | 649 | 1.385 | 1 | 0.997 |

Sonuç olarak, yapılan kümeleme analizi analitik olarak değerlendirildiğinde K-ortalamalar algoritmasının başarılı olduğu gözlenmektedir. Analitik kontrolden geçen kümeleme analizinin sonuçlarının genelleştirilebilmesi için oluşturulan kümelerin işletme değeri taşıması gerekmektedir.Analitik olarak değerlendirilen 4 kümenin işletme değeri açısından anlamlı sonuçlarının kontolü aşamasına kümeleme analizinde profilleme ve geçerleme süreci denmektedir. Bu aşamada aktif değişkenlerin yanında betimsel değişkenlerde rol oynamaktadır. Kümelerin işletme değerini tam olarak anlayabilmek için sınıflayıcı ve aralık ölçekli değişkenlerin ayrı ayrı küme-içi istatistik değerlerinin incelenmesi gerekmektedir. Sınıflayıcı, aktif ve betimsel aralık ölçekli değişkenlerin

genele göre dağılımlarını incelediğimizde oluşan 4 kümenin işletme değerlerinin olduğunu görmekteyiz. Analitik olarak geçerlenen kümeler işletme değeri açısından da geçerlenmektedir.

## 2.9. Kümelerin Profilleri

Kümeleme analizinin analitik ve işletme değeri açısından başarılı olarak geçerlenmesi ve doğrulanmasından sonra ortaya çıkarılan kümelerin profillendirilmesi kümelerin anlamlı olduğu sonucunu vermektedir. Bu anlamda, banka için en değerli müşterilerin birinci kümede olduğu ve bankanın bu müşteri segmenti için her türlü yatırımı yapması gerektiği ortaya çıkmıştır. İkinci kümedeki müşterilerin banka için bir zarar nedeni olmadığı bunun yanında uzun vadede kar getirebilecek müşterilerden oluştuğu görülmektedir. Üçüncü ve dördüncü kümedeki müşteriler banka için bir zarar kaynağıdır. Bu amaçla üçüncü kümedeki müşterilerin ve dördüncü kümedeki müşterilerin banka müşteri portföyünden çıkarılması bankanın boş yere para harcamasını engelleyecektir. Küme profilleri tablo 6'da özetlenmektedir.

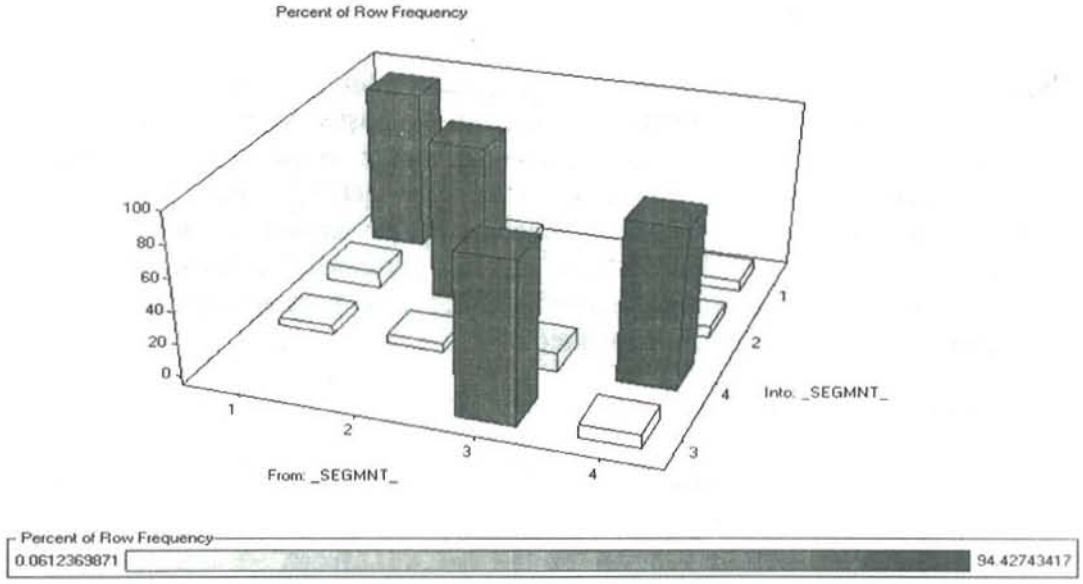
**Tablo 6.** Küme Profilleri

| | 1.küme | 2.küme | 3.küme | 4.küme |
|---|---|---|---|---|
| **Borç ödeme durumu** | EN SADIK | SADIK | SADIK DEĞİL | SADAKAT DERECESİ EN DÜŞÜK |
| **Kredi talep nedeni** | Borç konsolidasyonu | Borç konsolidasyonu | Borç konsolidasyonu | Borç konsolidasyonu ve ev yenileme |
| **Borç gelir oranı** | EN YÜKSEK | EN DÜŞÜK | YÜKSEK | YÜKSEK |
| **Talep edilen kredi tutarı** | EN YÜKSEK | EN DÜŞÜK | YÜKSEK | YÜKSEK |
| **Mevcut mal valık değeri** | EN YÜKSEK | EN DÜŞÜK | DÜŞÜK | ÇOK DÜŞÜK |
| **Kredi soruşturma sayısı** | YOK | YOK | EN YÜKSEK | YÜKSEK |

## 2.10.Kümeleme Analizinin Doğrulanması Süreci

Kümeleme analizi, çoğu veri madenciliği projesinde ilk yapılan modellemedir. Veri kümesinde benzer verileri gruplandıran kümeleme analizinin ardından kurulan tahmin modelleri oluşan kümelerin etkinliğini ölçmektedir. Uygulamamızda, kümeleme analizi sonucu oluşan 4 adet kümeyi birbirinden ayırt eden değişkenleri belirlemek ve kümelerin

etkinliğini ölçmek için, tahmin modellerinden adım adım regresyon tekniği kullanılmıştır. Kümeleme analizi sonucu oluşan küme sayısı ikiden fazla olduğundan dolayı en uygun regresyon tekniği multinomial regresyon olarak belirlenmiştir. Şekil 9' da kümelerin etkinliğini belirlemede kullanılan multinomial adım adım regresyon analizinin karşılaştırma matrisi bulunmaktadır.




**Şekil 9.** Multinomial Adım adım Regresyon Tekniğinin Karşılaştırma Matrisi

## 3. SONUÇ

Günümüz dünyasında işletmelerin günlük işlemlerinde transfer edilen verilerin sayısı her geçen gün artmaktadır. Veri madenciliği ve bilgi keşfi büyük miktarlardaki veri içerisinden anlamlı sonuçlar çıkarmada kullanılmaktadırlar. Çalışmamızda veri madenciliği modelleri iki ana başlıkta incelenmiştir: tahmin modeli ve tanımlayıcı model. Çalışmamızın uygulama aşamasında, bir bankanın bireysel bankacılık departmanına gelen konut kredisi başvurularının kabul red kararının otomatik olarak verilmesi ve başvuruda bulunan müşterilerin ortak özelliklerinin belirlenerek anlamlı müşteri segmentleri yaratma süreci incelenmiştir. Kredi talebinin otomatik olarak incelenmesi ve cevaplanması süreci, veri madenciliği tahmin ve sınıflandırma modeli oluşturularak değerlendirilmiştir. Performans karşılaştırması yapmak ve en iyi modeli seçmek amacıyla çalışmamızda tahmin modellerinden lojistik regresyon ve sınıflandırma modellerinden karar ağacı uygulanmıştır. Tahmin modeli olarak lojistik regresyon ve denetimli sınıflandırma tekniği olarak karar ağacı analizlerinin kendi içlerinde değerlendirme süreçlerinden sonra adım adım lojistik regresyon modelinin performansı daha başarılı bulunmuştur.

Tahmin ve denetimli sınıflandırma modellerinden sonra hedef değişkenin bulunmadığı, müşterilerin 13 değişkene göre ortak özelliklerinin incelenerek segmentlere

ayrıldığı denetimsiz sınıflandırma tekniklerinden kümeleme analizi yapılmıştır. Çalışmamızda kümeleme analizinin hiyerarşik olmayan yöntemlerinden K-ortalamalar algoritması incelenmiştir. Kümeleme analizinin analitik ve işletme değeri açısından başarılı olarak geçerlenmesi ve doğrulanmasından sonra ortaya çıkarılan kümelerin profillendirilmesi kümelerin anlamlı olduğu sonucunu vermektedir. Kümelerin analitik ve işletme değeri açısından değerlendirilmesi ve küme profillerinin oluşturulması aşamalarından sonra kümeleme analizinin geçerlenmesi ve doğrulanması süreci bulunmaktadır. Bu aşamada denetimli sınıflandırma tekniklerinden multinomial regresyon kullanılmıştır.

## KAYNAKLAR

AKPINAR, Haldun. (2000). "*Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği*". www.isletme.edu.tr/dergi/nisan2000 (24.01.2002)

ALPAYDIN, Ethem. (2000*)*, *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*, Bilişim 2000 Veri Madenciliği Eğitim Semineri.

BİÇEN, Pelin (2002), *Veri Madenciliği:Sınıflandırma ve Tahmin Yöntemlerini Kullanarak Bir Uygulama*, Basılmamış Yüksek Lisans Tezi.

FAYYAD, Usama, PIATESKY-SHAPIRO Gregory, PADHRIC, Symtih ve UTHURUSAMY, Ramasomy (1996). *Advances in Knowledge Discovery and Data Mining*. USA: MIT Press

FIRAT, Ümit Oktay ve BİÇEN, Pelin. (2002), *Veri Madenciliği Teknikleri*, XXIII. Yöneylem Araştırması ve Endüstri Mühendisliği Ulusal Kongresi Bildiri Özetleri, s: 62.

POOTS, William J. E. (2000) *Data Mining Primer: Overview of Applications and Methods*. USA: SAS Inst.

WIELENGA, Doug, LUCAS, Bob ve GEORGES Jim (1999), *Enterprise Miner: Applying Data Mining Techniques Course* Notes. USA: SAS Inst

# Customer Segmentation and Credit Scoring Model in Banking Sector by Using Data Mining Techniques

## ABSTRACT

*The explosive growth of many business and scientific databases has far exceeded the ability to interpret the data. At this point, there was a creating need for a new generation of tools and techniques for automated databases analysis. The tools and techniques are the subject of the rapidly emerging field of knowledge discovery in databases and data mining*

*techniques. In this research, first of all data mining and knowledge discovery in databases concepts and then data mining models were explained. In the second part, credit scoring and customer segmentation problem that was steadily encountered in current business world was solved with predictive and classification data mining modeling techniques. In the solution period, SAS Enterprise Miner data mining package was used.*

**Key Words:** *Knowledge Discovery in Databases, Data Mining, Predictive Modeling, Classification, Regression Analysis, Decision Trees, Clustering, K-Means Algorithm, Multinomial Regression Analysis.*