YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

LOJİSTİK REGRESYON İLE KREDİ SKORLAMA VE BİR UYGULAMA

İstatistikçi Handan ÖZDEMİR

FBE İstatistik Anabilim Dalında Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı : Yrd. Doç. Dr. Doğan YILDIZ

İSTANBUL, 2010

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

LOJİSTİK REGRESYON İLE KREDİ SKORLAMA VE BİR UYGULAMA

İstatistikçi Handan ÖZDEMİR

FBE İstatistik Anabilim Dalında Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı : Yrd. Doç. Dr. Doğan YILDIZ

İSTANBUL, 2010

İçindekiler

ÇİZELGE LİSTESİ

ŞEKİLLER LİSTESİ

ÖNSÖZ

Dört yıllık lisans eğitimi ve yüksek lisans eğitimim boyunca derslerini zevkle dinlediğim ve tez çalışmamda bana yol gösteren değerli hocam Yrd. Doç Dr. Doğan YILDIZ'a

YTÜ İstatistik bölümünün çok kıymetli hocalarına,

Hayatımın her döneminde koşulsuz bir biçimde destekçim olan sevgili kocaman aileme teşekkürlerimi sunarım.

ÖZET

Kredi skorlama bankacılık ve finans sektöründe istatistiksel modelleme konusunda son 40 yılın en başarılı uygulamalarından bir tanesidir. Ülkemizde de sektörde bu işle ilgilenen analistlerin sayısı artsa da bu konudaki literatür henüz çok sınırlı bir durumdadır. Bu çalışmada istatistiksel bir modelleme yöntemi olan "Lojistik Regresyon" un kredi skorlama uygulamalarında kullanımı incelenmiş ve kredi skorlama yöntemi hakkında detaylı bilgi verilmiştir. İlk bölümde regresyon teriminden başlayarak lojistik regresyon yöntemi ile modelleme ve kurulan modellerin değerlendirilmesi konusuna yer verilmiştir. İkinci bölümde ise kredi ve skorlama kavramından başlayarak kredi skorlamanın tarihçesi irdelenmiş ve bir skor kart modeli oluşturma projesinin tüm aşamaları incelenmiştir. Uygulamanın yer aldığı üçüncü bölümde ise Türkiye'de faaliyet gösteren orta ölçekli bir bankanın bireysel kredi başvurusunda bulunmuş müşterilerinin verileri kullanılarak lojistik regresyon ile bir skor kart modeli oluşturulmuştur. Modelin değerlendirilmesine ve model oluşuncaya kadar yapılan tüm çalışmalara alt bölümler halinde yer verilmiştir. Verilerin analiz edilmesi ve model kurulması için SPSS15.0, grafiklerin çiziminde ise Excel kullanılmıştır.

Anahtar Kelimeler: Lojistik Regresyon, Kredi Skorlama, ROC, GINI, Başvuru Skorlaması, Davranış Skorlaması

ABSTRACT

Credit scoring is one of the most successful applications of statistical modeling in finance and banking. In our country altough number of analysts in the industry is increasing constantly, the literature on this subject is very limited. This paper outlines "Logistic Regression" the most preferred technique in score card modeling and basic theory of developing credit scoring systems. In the first part, the key term regression and developing logistic regression models are considered. Also goodness of fit criterias are evaluated in the first part. Second part informs about score card modeling method and scoring project phases are explained step by step. Third part consist of an application about score card model development and project processes that are shown in the second part. The data, used for modeling, contains retail credit application information taken from a middle sized bank in Turkey. Logistic regression model is developed using SPSS 15.0 and graphics are drawn in Excel.

Key Words: Logistic Regression, Credit Scoring,ROC, GINI, Application Scoring, Behavioral Scoring

1. GİRİŞ

Gününüzde veri analizi, yani verilerin incelenerek anlamlı bilgiler haline dönüştürülmesi pek çok sektör için başarının anahtarı olarak görülmektedir. Kredi skorlama ülkemizde bankacılık sektöründe bu amaçla kullanılan en yeni uygulamalardan birisidir. Kredi skorlama bankacılık uygulamalarında pek çok amaçla kullanılabileceği gibi bu çalışmada yalnızca başvuru skorlaması ele alınmıştır. Kredi başvurusunda bulunan müşterilerin başvurdukları anda riskli olup olmadıklarını tahmin eden modeller kurulmasında genellikle lojistik regresyon kullanılır. Varsayımlarının az olması, sonuçlarının kolay yorumlanabilmesi gibi nedenlerle lojistik regresyon oldukça tercih edilen bir analiz tekniğidir. Bu tezde lojistik regresyon hakkında bilgi verildikten sonra kredi skorlama modeli geliştirme projesi aşamaları anlatılmıştır. Uygulama bölümünde ise lojistik regresyon ile bir skor kart modeli kurulmuştur.

2. LOJİSTİK REGRESYON

2.1. Regresyon Modelinin Tarihsel kökeni

Regresyon terimi ilk kez Francis Galton tarafından kullanılmıştır. Galton ünlü bir yazısında, uzun boylu anne babaların uzun, kısa bouylu anne babaların kısa çocukları olur eğiliminin geçerliliğine karşın, belli bir boydaki anne babaların çocuklarının ortalama boyunun genel nüfustaki ortalama boya doğru yaklaşma yani inglizce deyimiyle "regress" eğiliminde olduğunu bulmuştur. Bir başka deyişle normalden uzun ya da kısa anne babaların çocuklarının boyları nüfusun boy ortalamasına yaklaşma eğilimindedir. Galton'nun evrensel regresyon yasası, aile bireylerine ilişkin bini aşkın veri toplayan arkadaşı Karl Pearson tarafından doğrulanmıştır. Pearson bir grup uzun boylu babanın çocuklarının boy ortalamasının babalarınkinden kısa, bir grup kısa boylu babanın çocuklarının boy ortalamasının babalarınınkinden uzun olduğunu, böylece hem uzun hem kısa çocukların boylarının ortalamaya doğru çekildiğini bulmuştur.

Regresyonun günümüzdeki yorumlarından bir kaçı ise şu şekildedir:

- Regresyon çözümlemesi bir bağımlı değişkenin başka açıklayıcı değişkenlere olan bağımlılığını, birincinin ortalama değerini, ikincinin bilinen ya da değişmeyen değerleri cinsinden tahmin etme ya da kestirme amacıyla inceler.
- Regresyon, iki ya da daha çok değişken arasında doğrusal bir ilişki olup olmadığının bulunması ve bu doğrusal ilişkinin bir doğrusal denklemle nasıl ifade edildiğinin gösterilmesidir.
- Regresyon analizi yapılırken, gözlem değerlerinin ve etkilenilen olayların bir matemetiksel gösterimle yani bir fonksiyon yardımıyla ifadesi gerekmektedir. Kurulan bu modele regresyon modeli denilmektedir.
- Bir şans değişkeninin davranışının bir model kullanılarak tahminlenmesidir.

2.2. Doğrusal Regresyon Modeli

n sayıda birimin her birinden bağımlı değişken Y ve bağımsız değişken X değerleri saptanmış olsun. Bu durumda (Y_1,X_1), (Y_2,X_2),… , (Y_n,X_n) olmak üzere n sayıda gözlem değeri olacaktır. Acaba Y ve X değişkenleri arasındaki ilişkinin şekli nasıldır ve bu ilişkiyi matematiksel bir denklem olarak ifade edebilir miyiz? Bu soruların yanıtlayabilmek için (Y_i,X_i) $i = 1, \ldots, n$

gözlem çiftlerini koordinat ekseninde işaretlenirse n sayıda gözlem çiftinin her biri için serpilme diyagramında bir nokta oluşacaktır. Bu noktaların konumuna bakılarak oluşturulacak modelin doğrusal mı yoksa eğrisel mi olacağına karar verilir. Eğer noktalar bir doğru etrafında toplanıyorsa doğrusal bir model kurulur.

Doğrusal regresyon modelinin genel denklemi aşağıdaki gibi ifade edilir.

$$Y = \beta_0 + \beta_1 X + \in \quad (1.1)$$

β_0 ve β_1 modelin bilinmeyen parametreleridir. $\in$ hata terimi olup Y ve X gözlenebilen değişkenlerdir. β_0 ve β_1 parametrelerinin hesaplanması için ana kütledeki (Y_1,X_1), $(Y_2,X_2),\dots$, (Y_n,X_n) değerlerinin gözlenmiş olması gerekir. Ancak pratikte bu mümkün değildir. Bu sebeple ana kütleden rastsal olarak örneklem seçilerek (Y_1,X_1), $(Y_2,X_2),\dots$, (Y_n,X_n) değerleri tahmin edilir. Parametreleri tahmin eden modeli 1.2 denklemi ile ifade ederiz:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (1.2)$$

$\hat{\beta}_0$ ve $\hat{\beta}_1$, anakütle parametresi olan β_0 ve β_1 in tahmincileridir. $\hat{\beta}_0$ ve $\hat{\beta}_1$ in tahmin edilmesinde "En Küçük Kareler Yöntemi" kullanılır.

2.2.1. En Küçük Kareler Yöntemi

Rastsal olarak seçile n sayıda gözlem biriminde çizilen serpilme diyagramında veri noktalarının bir doğru etrafında toplandığını varsayalım. Bu doğrunun denklemindeki katsayıların tahmini için en küçük kareler yöntemi kullanılır. (Y_i,X_i) noktaları ile bu noktaların en küçük kareler yönteminde tahmin edilen doğru üzerindeki iz düşümleri arasındaki farkların toplamı sıfır olmalı ve bu farkların kareleri toplamının da minimum olması gerekmektedir.

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ doğrusunu elde ettiğimizde yukarıda bahsedilen iki koşul matematiksel olarak 1.3 ve 1.4 ile ifade edilir.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0 \quad (1.3)$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = min \qquad (1.4)$$

Yukarıda belirtilen koşulları sağlayan istatistikleri tahmin edebilmek için 1.3 de $\hat{Y}_i$ yerine değerini yazalım.

$$\sum_{1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2 = 0$$

$$S = \sum_{1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2$$

olsun.

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ doğrusunun birinci koşulu sağlaması için S nin $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistiklerine göre kısmi türevlerinin alınıp sıfıra eşitlenmesi gerekir. 1.4 de belirtilen koşulun sağlanması için de ikinci türevlerin pozitif işaretli olması gerekmektedir.

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x)$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x)\, x$$

Bu iki eşitliği sıfıra eşitleyelim.

$$-2\sum_{1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x\right) = 0$$

$$-2\sum_{1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x)\, x = 0$$

Bu iki denklem çözülürse aşağıdaki eşitlikler elde edilir.

$$\sum_{1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i \qquad (1.5)$$

$$\sum_{1}^{n} y_i x_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum {x_i}^2 \qquad (1.6)$$

Elde edilen 1.5 ve 1.6 denklemlerine "normal denklemler" denir. Bu iki denklemin birlikte çözülmesiyle $\hat{\beta}_0$ ve $\hat{\beta}_1$ istatistikleri bulunur.

Gerçekleşmesi gereken ikinci koşul için ikinci türevler alınır.

$$\frac{\partial^2 S}{\partial^2 \hat{\beta}_0} = 2n$$

$$\frac{\partial^2 S}{\partial^2 \hat{\beta}_1} = 2\sum x_i^2$$

İkinci türevler pozitif olduğundan ikinci koşul da sağlanmış olur.

İki denklemi birlikte çözerek bilinmeyen istatistikleri bulmak biraz zaman alacağından normal denklemlerden $\hat{\beta}_0$ ve $\hat{\beta}_1$ ı veren formülleri çıkaralım. Öncelikle normal denklemlerin birincisinden $\hat{\beta}_0$ ı çekelim.

$$\sum_1^n Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$$\sum_1^n Y_i - \hat{\beta}_1 \sum x_i = n\hat{\beta}_0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (1.7)$$

$\hat{\beta}_1$ i elde etmek için de ikinci normal denklemi olan 1.6 da $\hat{\beta}_0$ ı çekip birincisinde yerine yazılır, gerekli düzenlemeler yapılarak 1.8 deki sonuca ulaşılır.

$$\hat{\beta}_1 = \frac{\sum Y_i X_i - n\,\bar{X}\,\bar{Y}}{\sum x_i^2 - n\,\bar{X}^2} \qquad (1.8)$$

Bu eşitliği 1.9 denklemi ile ifade etmek mümkündür.

$$\hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y}) - (X_i - \bar{X})}{\sum(X_i - \bar{X})^2} \quad (1.9)$$

2.2.2. Doğrusal Regresyon Modelinin Varsayımları

Doğrusal regresyonda katsayı tahminlerinde kullanılan En Küçük Kareler yöntemini kullanabilmek için yukarıda da bahsettiğimiz bir takım varsayımların sağlanması gerekir. Aşağıda bu varsayımlar özetlenmiştir.

- Verilerin aralıklı ya da oransal ölçekli olması gerekir. Ayrıca bağımlı değişkenin de sürekli olması ve normal dağılıma uygunluk göstermesi gerekir.

- Bağımlı ve bağımsız değişkenin birbirlerine dönüşümünü sağlayan fonksiyon doğrusal bir fonksiyon olmalıdır.

- Gözlenen ve tahmin edilen Y değerleri arasındaki farkların beklenen değeri 0 olmalıdır. Başka bir deyişle hata teriminin beklenen değeri sıfırdır.

  $E(e_i) = 0$

- e' nin varyansı bütün bağımsız degişkenler için sabit veya aynı olmalıdır. Bu varsayım $E({e_i}^2) = {\sigma_e}^2$ ile ifade edilir. Bu varsayımın ihlali heteroskedastisite olarak da bilinmektedir.

- Bağımsız değişkenlerin hata terimleri her bir değişken için normal dağılmalıdır. (Normality of errors)

- Hata terimleri arasında ilişki olmamalıdır. Bu varsayımın ihlali baska bir deyisle hata terimlerinin serisel olarak birbiri ile bağımlı olması durumu otokorelasyon olarak bilinmektedir.

  $E(e_i, e_j) = 0$

- Hata terimleri ile bağımsız degişkenler arasında ilişki olmamalıdır. Bu varsayımın ihlali, bağımsız degişkenlerin hatalı olarak ölçülmesi, bağımlı değişkenin gecikmeli değerinin açıklayıcı değişken olarak kullanılması, gerçekte bir eşanlı denklem sistemine ait olan ve açıklayıcı değişkenleri bağımlı değişken olan bir denklemin açıklayıcı değişkenlerinin bağımsız değişken varsayılması vb. gibi problemlere neden olmaktadır.

- Çok değişkenli regresyonda bağımsız degişkenlerin hiçbiri bir diğerinin doğrusal kombinasyonu olmamalıdır. Baska bir deyişle bağımsız değişkenler arasında tam doğrusal ilişki yoktur. Bu varsayımın ihlali çoklu doğrusal bağıntı olarak bilinmektedir.

2.3. Doğrusal Olasılık Modeli

$$Y_i = \beta_0 + \beta_1 X_i + \in \quad (1.10)$$

Tek açıklayıcı değişken içeren modelde X_i i. birimin x parametre değerini, Yi ise iki kategorili olarak modelin i. birim için sonucunu göstermek üzere bu denklemin koşullu beklenen değeri doğrusal olasılık modelini verir. [1]

$$E(Yi / Xi) = \beta_0 + \beta_1 X_i \quad (1.11)$$

Bağımlı değişken Y_i'nin iki düzeyli kategorik bir değişken olduğu düşünülürse, yalnızca 0 ve 1 değerlerini alacağı anlaşılır. Buna göre Y'nin olasılık dağılımı denklem 1.12 ile belirlenir.

$$P_i = P(Y_i = 1) \ ve \ 1 - P_i = P(Y_i = 0) \quad (1.12)$$

Buna göre

$$E(Y_i) = 1(P_i) + 0(1 - P_i) = P_i$$

[1] Unvan A. Y.

$E(\mathrm{y}) = \mathrm{P}(\mathrm{y} = 1) = p_i$

Şeklinde de yazılabilir. Görüldüğü gibi bağımlı değişkenin beklenen değeri, bağımlı değişkenin 1 değerini alma olasılığına yani ilgili olayın gerçekleşme olasılığına eşittir

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik} \quad (1.13)$$

Denklemin sol tarafı olasılık olduğu çin 0 ile 1 arasında değişen değerler alabilmektedir. Bu nedenle y_i hedef değişkenin değerlerinin ikili olduğu regresyon modeline doğrusal olasılık modeli denir. Çünkü x_i i. değerini aldığında y_i nin koşullu beklenen değeri olayın gerçekleşme olasılığını verir. [2]

$$E(Y_i/X_i) = \beta_0 + \beta_1 X_i = P_i \quad (1.14)$$

$0 < \beta_0 + \beta_1 X_i < 1$ olmak üzere doğrusal olasılık modeli 1.5 ve 1.6 denklemleri ile ifade edilir.

$$P_i = 1 \qquad ise \qquad \beta_0 + \beta_1 X_i \geq 1 \quad (1.15)$$

$$P_i = 0 \qquad ise \qquad \beta_0 + \beta_1 X_i \leq 1 \quad (1.16)$$

Doğrusal olasılık modelinde Y'nin tahmin edilen değerleri olasılık gibi yorumlandığından bu modelin tahmin amacıyla kullanılması durumunda önemli bir sorunla karşılaşılmaktadır. En önemlisi tahmin edilen Y değerleri "0-1" aralığının dışına çıkabilir. Böyle bir durumda 0-1 aralığının dışındaki değerler 1'e veya 0'a eşitlenerek olasılıkların 0-1 aralığı içinde kalması sağlanabilir. Ancak böyle bir işlem sağlıklı olmamakla beraber, elde edilen tahmin değerlerinin sapmalı olmalarına neden olabilir.[3]

P_i olasılığı 0 ile 1 arasında olacağından $0 \leq E(Y_i/X_i) \leq 1$ veya $0 < \beta_0 + \beta_1 X_i < 1$ şeklinde bir sınırlama zorunlu olduğundan koşullu olasılık 0 ile 1 arasında kalmalıdır. Regresyon

[2] Ünsal A. ve Güler H.
[3] Ulupınar S. D.

modellerinde parametrelerin En Küçük Kareler (EKK) yöntemi ile bulunduğu bilinmesine rağmen bağımlı değişkenin kategorik olduğu durumda sorunlarla karşılaşılması mümkündür. Y bağımlı değişkeninin kategorik olması durumunda ei de sadece iki değer alır ve normal dağılma uygun bir dağılımı olmaz.[4]

$Y_i = \beta_0 + \beta_1 X_i + \in_i$ modelinden yola çıkarak modelden $\in_i$' yi çekersek,

$$\in_i = Y_i - \beta_0 - \beta_1 X_i \qquad (1.17)$$

(1.17) denklemini elde ederiz.

$Y_i = 1 \quad ise \quad \in_i = Y_i - \beta_0 - \beta_1 X_i$

$Y_i = 0 \quad ise \quad \in_i = -\beta_0 - \beta_1 X_i$

olarak elde edilir. Buradan $\in_i$' nin normal dağılıma değil binom dağılıma uyduğu söylenebilir.

Buna ek olarak, ei 'nin değişen varyanslı olması söz konusudur. Şöyle ki,

$Var(e_i) = E[e_i - E(e_i)]^2 = E({e_i}^2)$

$E(e_i) = 0 \; varsayımı \; ile$

$Var(e_i) = E({e_i}^2) = E(Y_i/X_i)\,[1 - E(Y_i/X_i)\,]$

$$Var(e_i) = P_i(1 - P_i) \qquad (1.18)$$

Yukarıdaki (1.18) eşitliğinden hareketle e_i' nin varyansı X'e bağlıdır, dolayısıyla sabit değildir. e_i ' nin normal dağılıma uygun olmaması durumu çok önemli değildir çünkü Merkezi Limit Teoremi gereği örneklem hacmi arttıkça e_i normal dağılmasa da katsayı tahminleri yaklaşık

[4] Topuz D. ve Çakır M.

olarak normal dağılacaklardır. Bu durumda tahminciler sapmasızlıklarını korurlar ancak değişen varyans nedeniyle artık etkin olmayacaklardır. Buna bağlı olarak standart hata ve test istatistiği yanlı olacaktır. Değişen varyansı ortadan kaldırmak için ağırlıklandırma yapmak gerekir.[5]

Ayrıca $0 < E(Y_i/X_i) < 1$ varsayımın sağlanacağı kesin olmamakla birlikte R^2 değerlerinin genellikle küçük çıkarak ilişkinin uyumunu gösteren bir ölçü olmaması da ayrı bir sorundur. Her ne kadar bu sorunlar aşılabilirse de Doğrusal Olasılık Modelinin temel sorunu Pi' nin Xi' ye doğrusal olarak arttığını varsaymasıdır. Gerçek hayatta ise Pi' nin Xi' ye dogrusal olmayan bir biçimde baglı olması istenir. Dolayısıyla Xi arttıkça Pi' nin arttığı ancak 0-1 aralığından çıkmadığı ve Pi ile Xi arasındaki ilişkinin doğrusal olmadığı durumlarda lojistik regresyon analizinin kullanımı önerilmektedir. Bağımlı değişken iki ya da daha fazla düzey içeren kesikli değişken olduğunda normallik ve sabit varyans varsayımları bozulduğundan böyle durumlarda lojistik regresyon kullanılmaktadır.

2.4. Lojistik Regresyon Analizi

Teknolojinin gelişmesiyle birlikte hemen hemen her alanda saklanan veriler ve bu verilerin boyutu da artmaktadır. Toplanan bunca veriden anlamlı sonuçlar çıkarabilmek, sonuçları yorumlayabilmek pek çok sektör açısından başarının anahtarı olarak görülmektedir. Verilerin incelenerek gizli örüntülerin ortaya çıkarılması, elde bulunan verilerin değerinin arttırılması ve verinin bilgiye dönüştürülmesi noktasında istatistiksel modellerin kullanımı yaygın bir hale gelmiştir. Elde bulunan verilerin çok boyutlu olmasından dolayı özellikle çok değişkenli istatistiksel yöntemler tercih edilmektedir.

Çok boyutlu verilerin sınıflandırılmasında kullanılan yöntemler;

- Kümeleme Analizi (Clustering)
- Diskriminant Analizi (Discriminant Analysis)
- Lojistik regresyon (Logistic Regression)
- Faktör Analizi

[5] Ürük E.

Kümeleme analizinde lojistik regresyon ve diskriminant analizinden farklı olarak gözlemlerin atanacağı küme sayısı tam bilinmemektedir. Gözlemler uzaklık veya benzerlik ölçütlerine göre kümelenmektedir. Diğer analiz yöntemlerinden farklı olarak bu analizde gruplar önceden tanımlanmaz, tersine grupların tanımlanmasına çalışılır. Kümeleme Analizinin temel hedefi ana kütle içinde dağınık halde bulunan verileri benzerliklerine göre bir araya getirip sınıflandırmaktır.

Diskriminant analizi sık kullanılan ve çok bilinen yöntem olup varsayımları daha ağırdır.

Lojistik regresyon ise diskriminant analizine alternatif olarak kullanılan son yıllarda uygulaması yaygınlaşmış bir yöntemdir. Temelde amacı regresyonun temel mantığında olduğu gibi bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi inceleyerek, bu ilişkiyi tanımlayan en iyi modeli bulmaktır. Bu yöntem çeşitli varsayımların (normallik, ortak kovaryansa sahip olma gibi) sağlanmadığı durumunda diğer yöntemlere bir alternatif olurken, hedef değişkenin 0,1 gibi ikili ya da ikiden çok düzey içeren süreksiz değişken olması durumunda normallik varsayım kısıtı olmaması nedeniyle kullanım rahatlığının yanı sıra çözümlemeden elde edilen modelin matematiksel olarak esnek olması, kolay yorumlanabilir olması yönteme olan ilgiyi artırmaktadır.

2.4.1. Lojistik Regresyonun Tarihsel Gelişimi

Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ekonomi, tarım, veterinerlik ve taşıma sahalarında yaygın olarak kullanılmaktadır.

Lojistik modelin biyolojik deneylerin analizi için kullanımı ilk olarak Berkson(1944) tarafından önerilmiş, Cox (1970) bu modeli gözden geçirerek çeşitli uygulamalarını yapmış, özet gelişmeler ise ilk Andersson (1979, 1983) tarafından verilmiştir. Ayrıca verilerin lojistik modele uyumu ile ilgili birçok çalışmalar da yapılmıştır. Bunlar arasında Aranda-Ordaz (1981) ve Johnson (1985) tarafından yapılan çalışmalar en önemlileridir. Pregibon (1981) iki grup lojistik modelde etkin (influential), aykırı (outlier) gözlemleri ve belirleme ölçütlerini (diagnostic), Lesaffre(1986), Lesaffre ve Albert (1989) ise çoklu grup lojistik modellerde etkin ve aykırı gözlemlerle belirleme ölçütlerini incelemişlerdir. Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesi, hatayı tahmin yöntemlerinin geliştirilmesi ve lojistik regresyon modellerinin daha

ayrıntılı incelenmesine sebep olmuştur. Cornfield (1962), lojistik regresyondaki katsayı tahmin işlemlerinde diskriminant fonksiyonu yaklaşımını ilk kez kullanarak popüler hale getirmiştir. Lee (1984) basit dönüşümlü (cross-over) deneme planları için doğrusal lojistik modeller üzerinde durmuştur. Bonney (1987) lojistik regresyon modelinin kullanımı ve geliştirilmesi üzerinde çalışmıştır. Robert ve ark. (1987) lojistik regresyonda standart Kikere, olabilirlik oran (G^2), "pseudo" en çok olabilirlik tahminleri, uyum mükemmelliği ve hipotez testleri üzerine çalışmalar yapmışlardır. Duffy (1990) lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşımını incelemiştir. Başarır (1990) klinik verilerde çok değişkenli lojistik regresyon analizi ve ayrımsama sorunu üzerinde çalışmıştır. Hsu ve Leonard (1995) lojistik regresyon fonksiyonlarında Bayes tahminlerinin elde edilmesi işlemleri üzerine çalışmışlar ve lojistik regresyonda Monte Carlo dönüşümünün kullanılabileceğini göstermişlerdir. Akaya ve Pazarlıoğlu (1998) lojistik regresyon modellerinin ekonomi alanında kullanımını örneklerle incelemişlerdir. Cox ve ark. (1998) kardiovasküler hastalıklar ve hipertansiyon arasındaki ilişkiyi incelemişlerdir. Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini incelemiştir. Kloiber ve ark. (1996), People ve ark. (1991), Buescher ve ark. (1993) kadınlarda düşük doğum ağırlığını etkileyen risk faktörlerini; Santos ve ark. (1998) kafein tüketimi ve düşük doğum ağırlığı arasındaki ilişkiyi, Sable ve Herman (1997) erken doğum ve düşük doğum ağırlığı arasındaki ilişkiyi incelemişlerdir. [6]

Bu çalışmanın üçüncü bölümünde lojistik regresyonun bankacılık sektöründe müşterilerin kredi başvurularının skorlanmasında lojistik regresyonun kullanımı hakkında detaylı bilgi verilmiştir.

2.4.2. Lojistik Regresyon Modelinin Doğrusal Regresyon Modeli İle İlişkisi

Lojistik regresyon ile doğrusal regresyon modelleri arasındaki farkları kısaca aşağıdaki gibi listeleyebiliriz.

---

[6] Ulupınar, S. D.

- En önemli fark, lojistik regresyonda hedef değişkenin kategorik olmasıdır.
- Doğrusal regresyon analizinde bağımsız değişkenin çoklu normal dağılım koşulu aranırken lojistik regresyonda böyle bir koşul aranmaz.
- Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilirken, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.

2.4.3. Lojistik Regresyonun Tercih Edilme Nedenleri

Lojistik regresyon hedef değişken kategorik olduğu durumlarda uygulanır. Hedef değişkenin kategorik olduğu durumlarda doğrusal regresyon modeli kullanılmamaktadır. Eğer hedef değişkenin kategorik olduğu durumlarda doğrusal regresyon analizi uygulanırsa aşağıda listelenen sorunlar ortaya çıkar.

i. Modelden elde edilen hata terimleri normal dağılıma uymaz.
ii. Modelden elde edilen hedef değişkenin gözlenme olasılığı olduğundan 0 ile 1 arasında değişmesi zorunluluğu sağlanmaz.
iii. Hata terimlerinin varyansı sabit olmayıp değişen varyansa sahiptir. Heteroskedastisite durumu.

Doğrusal regresyon modeli kullanabilmek için birinci ve üçüncü sorunlarla baş etmenin çeşitli yolları bulunabilir. Örneğin üçüncü problem ağırlıklı en küçük kareler yöntemi kullanılarak çözülebilir. Birinci problem için örneklem kümesi arttırılarak en küçük kareler yöntemi tahmincilerinin hata teriminin dağılımının asimptotik olarak normal olması sağlanabilir. Ancak ikinci sorunun çözümü yoktur. Doğrusal regresyon modeli ile elde edilen hedef değişkenin tahmincileri 0-1 aralığı dışında bir değer alabilmektedir. Sonuç olarak, hedef değişkenin kategorik olduğu durumlarda lojistik regresyon doğrusal regresyondan daha iyi bir çözümdür.

Varsayımlarının az olması da lojistik regresyonu tercih edilen bir yöntem olmasında etkilidir. Normallik varsayımı, varyans kovaryans matrislerinin eşit olması gibi varsayımların gerçekleşmediği durumlarda diskirminant analizi yerine tercih edilebilir. Bağımlı değişkenin kategorik olduğu durumlarda da normallik varsayımı sağlanmayacağından lojistik regresyon

tercih edilmektedir. Bunların dışında sonuçlarının kolay yorumlanabilir olması da bir başka tercih nedenidir.

2.4.4. Lojistik Regresyon Tipleri

Lojistik Regresyon analizinde bağımsız değişkenin formuna göre üç yöntem vardır.[7]

2.4.4.1. İkili Lojistik Regresyon (Binary, BLOGREG)

İkili cevap içeren hedef değişkenlerin kullanıldığı lojistik regresyon analizidir. Bir ya da daha fazla açıklayıcı değişken ile hedef değişken arasındaki ilişkiyi bir model olarak sunar.

2.4.4.2. Ordinal Lojistik Regresyon(OLOGREG)

Hedef değişkenin sıralı olduğu durumlarda uygulanan bir yöntemdir. Sıralı ölçekli cevap değişken, en az üç kategoride gözlenen değerler içermelidir. Sıralı ölçekli veriler kodlanırken ya da isimsel olarak kategorileri belirlendiğinde cevapların doğal sıralama yapısında olması gerekir. Örneğin bir yargı hakkında kişilerin görüşleri alınıyorsa "kesinlikle katılmıyorum-katılmıyorum-katılıyorum-fikrim yok" olarak kategoriler belirlenmelidir.

2.4.4.3. Nominal Lojistik Regresyon(NLOGREG)

Hedef değişkeninin isimsel olduğu durumlarda uygulanan bir yöntemdir. İsimsel ölçekli hedef değişken en az üç kategoride gözlenen değerler içermektedir. Gözlenen değerlerin kodlanması halinde bu kategorilerin bir sıra izlemesi şart değildir. Örneğin bir meslek dalları tercihlerinde sınıflar; mühendislik, bankacılık, reklamcılık vb. gibi isimsel olarak belirlenebilir.

2.4.5. Parametre Tahmin Yöntemleri

İki düzeyli bağımlı değişkenler için lojistik modelin katsayılarını tahmin etmek amacıyla kullanılan yöntemler, En Çok Olabilirlik(Maximum Likelihood, ML), Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler (Reweighted Iterative Least Square, RILS), Minimum Lojit Ki - Kare

---

[7]Özdamar K.

(Minimum Logit Chi-Square, MLCS) yöntemleridir (Tatlıdil, 1996). Bu çalısmada EÇO yöntemi üzerinde durulacaktır.

2.4.5.1. En Çok Olabilirlik Yöntemi

En çok olabilirlik yöntemi, gözlenen veri kümesini elde etmenin olasılığını maksimum yapan bilinmeyen parametrelerin değerlerini verir. Bu yöntemi uygulamak için öncelikle, en çok olabilirlik fonksiyonunun oluşturulması gerekir. Bu fonksiyon, gözlenen verilerin olasılıklarını bilinmeyen parametrelerin bir fonksiyonu olarak açıklar. Bu parametrelerin en çok olabilirlik tahmin edicileri, fonksiyonu maksimum yapan değerleri bulacak şekilde seçilir.[8]

Doğrusal regresyonda, bilinmeyen parametreleri tahmin etmek için en sık kullanılan yöntem "En küçük kareler" yöntemidir. Bu yöntemde modelimize göre y'nin gözlenen değerleri ile beklenen değerleri arasındaki sapmaların karesinin toplamını minimum yapacak β_0 ve β_1 değerlerini seçiyoruz. Doğrusal regresyondaki varsayımlarımız aynı olmak üzere en küçük kareler yöntemi bize çeşitli istenen istatistiksel özelliklere sahip tahminciler verir. Ancak iki sonuçlu bir modele en küçük kareler yöntemini uyguladıktan sonra kestiriciler aynı özelliklere sahip olamazlar. Bu sebeple Lojistik regresyon modelinde en küçük kareler yöntemi yerine "En çok olabilirlik yöntemi" kullanılır. Temel olarak, en çok olabilirlik yöntemi gözlenen veri kümesine ulaşılma olasılığını maksimum yapan bilinmeyen parametrelerin tahmin edilmesini sağlar.[9]

Bu yöntemi uygulanması için öncelikli olarak "olabilirlik fonksiyonu" fonksiyonun tanımlanması gerekir. Bu fonksiyon, gözlenen verinin olasılığını, bilinmeyen parametreler cinsinden bir model olarak verir. Bu parametrelerin "en çok olabilirlik tahmincileri" bu fonksiyonun değerini maksimum yapan değerlerden seçilir. Dolayısıyla, tahmin edilen istatistikler gözlenen veri ile en uygun olan tahminciler olurlar.

İki değeri olan bir bağımsız değişken için başarı olasılığı $p_i = p(y_i = 1/ X)$ ve başarısızlık olasılığı $1-p_i$ olmak üzere, her bir gözlemin p_i parametresi ile Bernoulli dağılımına sahip olduğu varsayılırsa i = 1, 2, ...,n için olasılık,

[8] Hosmer W. D. ve Lemeshow S.
[9] Anderson, T.W.

$P(Y_i/X_i) = P_i^{yi}(1 - P_i)^{1-yi}$ şeklinde yazılabilir. Aynı denklemi n gözlem için yazacak olursak,

$$L(Y/X) = P(Y/X) = \prod_1^n P_i^{y_i}(1 - P_i)^{1-y_i} \quad (1.19)$$

elde ettiğimiz (1.19) denklemi en çok olabilirlik fonksiyonudur.

En çok olabilirlik yönteminde, p tane bağımsız değişkene ait β tahmini, bağımlı değişkeni gözleme olasılığını yani L(Y/X) olabilirlik fonksiyonunu maksimum yapacak şekilde seçilir. Lojistik regresyon modelinin en çok olabilirlik fonksiyonu, L(Y/X) denkleminde P_i yerine açık ifadesi konularak elde edilir. Buradan, olabilirlik fonksiyonunun logaritması alınırsa (1.20) denklemi elde edilir.

$$\ln[L(Y/X,\beta)] = \sum_1^n [y_i ln P_i + (1 - y_i)\ln(1 - P_i)] \quad (1.20)$$

Denklem (1.20) de β' ya göre birinci türev alınıp sıfıra eşitlenirse (1.21) denklemi karşımıza çıkar.

$$\sum_1^n (y_i - P_i)x_{ij} = 0 \quad (1.21)$$

Bu denklemin çözülmesiyle β nın tahmin değeri olan $\hat{\beta}$ ya ulaşılır. P_i nin üstel olması nedeniyle elde ettiğimiz bu denklem doğrusal değildir. Bu sebeple $\hat{\beta}$ nın bulunabilmesi için bu denklemin iteratif olarak çözülmesi gerekmektedir. İteratif çözümlemede $\hat{\beta}$ ların başlangıç değerleri verilerek ilk tahminleri elde edilir. Her adımda ε kadar küçük miktarda ayarlamalar yapılarak türevler alınır ve en çok olabilirlik tahminleri bulunur. İteratif işlemler yakınsama sağlanıncaya dek devam eder. Yakınsama, ε düzeltme terimlerinin iterasyon değerlerini değiştirmediği noktada sağlanır.

2.4.6. Tek Değişkenli Lojistik Regresyon

2.4.6.1. Lojistik Regresyon İle Modelleme

Lojistik regresyon, hedef değişkeninin ikili, üçlü ve çok kategorili olduğu durumlarda açıklayıcı değişkenlerle neden sonuç ilişkisini belirlemede yararlanılan bir yöntemdir. Açıklayıcı değişkenlere göre hedef değişkeninin beklenen değerlerinin olasılık olarak elde edildiği bir yöntemdir.

Lojistik regresyon modelinde, hedef değişkenlerin bağımlı değişkenler üzerindeki etkileri olasılık olarak hesaplanır.

Bir regresyon probleminin tanımlanmasında açıklayıcı değişkenin verilen bir değeri için hedef değişkenin ortalama değeri kullanılır. Bu ifade "E(Y/X)" şeklinde gösterilir ve "koşullu ortalama" diye adlandırılır. Özetle E(Y /X), "Verilen bir X değeri için, Y'nin beklenen değeri" şeklinde okunur. Doğrusal regresyonda, bu ortalamayı x'e göre denklemi (1.22)'deki gibi doğrusal bir eşitlikle gösterebiliriz.

$$E(Y/X) = \beta_0+\beta_1 X \tag{1.22}$$

Bu denkleme göre x $(-\infty , +\infty)$ aralığında herhangi bir değer alabilir.
Bu denklemi lojistik regresyona modeli için yazabilmek için $\pi(x) = E(Y/X)$ yazacak olursak denklem (1.23) elde edilir.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{1.23}$$

Denklem (1.23)'e lojistik regresyon modeli denklemi denir. Elde etmiş olduğumuz bu denklem üzerinden lojistik dönüşüm uygulanırsa $\pi(x)$, denklem (1.24) ile ifade edilir.

$$g(x) = \frac{\ln[\pi(x)]}{\ln[1-\pi(x)]} = \beta_0 + \beta_1 X \tag{1.24}$$

Bu dönüşüme logit transformasyonu denir.

g(x) fonksiyonu lojistik regresyon modeli için uygun özelliklere sahiptir. Kendi parametrelerinde doğrusal olan g(x) fonksiyonu, sürekli olabilir ve x $(-\infty, +\infty)$ aralığında yer alır.

Lojistik Regresyon Modelinde hedef değişken için yaptığımız gözlem Y=E(Y/X)+ε şeklinde gösterilebilir. ε terimi"hata" olarak tanımlanır. Gözlemin koşullu ortalamadan ne kadar saptığını gösterir.

Doğrusal regresyon modelinde açıklayıcı değişken X için hedef değişkeninin koşullu dağılımı ortalaması E(Y/X) olan ve varyansı sabit olan bir normal dağılıma sahiptir. Ancak hedef değişkenin kategorik olduğunda lojistik regresyon modelinde bu durum daha farklıdır.

Lojistik regresyon modelinde verilen X açıklayıcı değişkeni için Y hedef değişkenini $Y = \pi(x)+\varepsilon$ şeklinde ifade edilir. Burada ε' nun bir veya iki olası değeri vardır. Eğer Y =1 ise π(x) olasılığıyla $\varepsilon = 1- \pi(x)$ olur. Eğer Y = 0 ise 1- π(x) olasılığıyla $\varepsilon = -\pi(x)$ olur. Dolayısıyla ε , 0 ortalamalı ve $[\pi(x)(1- \pi(x))]$ varyanslı bi dağılma sahip olur. Sonuç olarak hedef değişkeninin koşullu dağılımı aslında bir binom dağılımıdır. Hedef değişkenin ikili olduğu durumlarda ise bu dağılım Bernoulli dağılımıdır.[10]

Lojistik regresyonun varsayımlarını aşağıdaki gibi sıralayabiliriz;

- Açıklayıcı değişkenler arasında çoklu doğrusal bağlantı yoktur. Değişkenler birbirlerinin doğrusal bir fonksiyonu olarak yazılamazlar. Çoklu bağlantı olması durumunda tahmin edilen parametrelerin standart hataları artarken modelin tahmin gücü azalmaktadır.
- Hata terimleri birbirinden bağımsızdır.

2.4.6.2. Katsayıların Anlamlılık Testleri ve Güven aralıkları Tahmini

Katsayıları tahmin ettikten sonra uygun bulunan modelde ilk baktığımız şey modeldeki katsayıların anlamlılığıdır. Bunu genellikle istatistiksel hipotezleri test ederek bağımsız

[10] Hosmer W. D. ve Lemeshow S.

değişkenlerin hedef değişkeni açıklamada gerçekten önemli olup olmadığına bakılır. Bu testi yapmak için metot modelden modele değişkenlik gösterir. Tek değişkenli model için şimdi açıklayacağız.

Katsayıların Anlamlılık Testleri

Katsayıların anlamlılığının testi için bir yöntem şu soru ile başlar; Modelde yer alan açıklayıcı değişken hedef değişken ile ilgili modelde olmadığı duruma göre daha fazla şey açıklayabiliyor mu? Bu soru iki modelin kıyaslanması ile cevaplanır. Kıyaslanacak modellerden birincisi içinde ilgili açıklayıcı değişkenin bulunduğu model, ikincisi ise içinde ilgili açıklayıcı değişkenin bulunmadığı modeldir. İçinde açıklayıcı modelin olduğu modele "uygun bulunan model" diyeceğiz. Matematiksel fonksiyon hedef değişkene göre gözlenmiş ve tahmin edilmiş değerleri karşılaştırır. Eğer uygun bulunan modelde tahmin edilen değerler modelde ilgili açıklayıcı değişkenin bulunmadığı duruma göre daha güvenilir ise ilgili açıklayıcı değişken "anlamlıdır" denir.

Doğrusal regresyonda katsayıların anlamlılık testleri varyans analizi tablosu ile yapılır. Bu tablo gözlemlerin kalıntı kareleri toplamları ortalamalarına göre iki gruba ayrılır.

- Gözlemlerin regresyon doğrusundan sapmalarının kareleri toplamı; SSE(kalıntı kareleri toplamı)
- Regresyon modeline göre tahmin edilen bağımlı değişkenin kareleri toplamı, SSR (regresyon kaynaklı kalıntı kareleri toplamı)

Doğrusal regresyonda gözlenen ve tahmin edilen değerlerin karşılaştırılması ikisi arasındaki uzaklığın karesi incelenerek yapılır. i inci gözlem için y_i gözlenen değeri, $\hat{y}_i$ tahmin edilen değeri göstermek üzere karşılaştırmayı yapmak için gerekli olan istatistik SSE denklem (1.24) deki gibi yazılır.

$$SSE = \sum(y_i - \hat{y}_i)^2 \qquad (1.24)$$

Bağımsız değişken içermeyen modelde yalnızca β_0 parametresini içerecektir. Bu durumda $\bar{y}$ hedef değişkenin ortalaması olmak üzere $\beta_0 = \bar{y}$ olacaktır. Bu durumda $\hat{y}_i = \bar{y}$ olacaktır. Dolayısıyla SSE toplam varyansa eşit olacaktır. Eğer modele açıklayıcı değişken ilave edersek

SSE de bir düşüş meydana gelecektir. Bunun nedeni açıklayıcı değişkenin sıfırdan farklı bir değer almasıdır. Regresyondaki değişkenliğe bağlı olarak SSE istatistiğindeki değişim SSR olup denklem (1.25) gösterilir.

$$SSR = [\sum(y_i - \bar{y})^2] - [\sum(y_i - \hat{y}_i)^2] \quad (1.25)$$

Doğrusal regresyonda ilgi odağı SSR üzerindedir. SSR nin büyük olması bağımsız değişkenin önemli olduğunu gösterir. Küçük olması da hedef değişkeni açıklamada bağımsız değişkenin yeterli olmadığını gösterir.

Lojistik regresyonda da aynı mantıkla bağımsız değişkenin modelde olduğu durumdaki ve modelde olmadığı durumdaki tahmin edilen değerler kıyaslanır. Lojistik regresyonda bu karşılaştırma log – en çok olabilirlik fonksiyonu ile yapılır.[11]

$$\ln[\mathrm{L}(\mathrm{Y}/\mathrm{X},\beta)] = \sum_1^n [y_i ln P_i + (1 - y_i)\ln(1 - P_i)] \quad (1.26)$$

Doymuş modeli veri noktaları kadar parametre içeren bir model olarak tanımlayalım. Karşılaştırmayı daha iyi anlamak için, hedef değişkeninin gözlenen değerini aynı zamanda "Doymuş bir modelde" tahmin edilen bir değer olduğunu düşünelim.

Log-en çok olabilirlik fonksiyonu kullanılarak gözlenen değerler ile tahmin edilen değerlerin kıyaslanması denklem (1.27) ile yapılır.

$$D = -2ln\left[\frac{Uygun\ \ bulunan\ \ modelin\ \ olabilirli\ \ği}{Doymu\ş\ modelin\ \ olabilirli\ \ği}\right] \quad (1.27)$$

Denklem (1.27) de parantez içindeki ifade, benzerlik oranı olarak isimlendirilir. Parantezin başında bulunan -2ln ifadesi hipotez testinin yapılması için gerekli olan dağılımın bilinmesi varsayımını yerine getirir. Bu teste olabilirlik oranı testi denir. Denklem (1.27) denklem (1.28) ile ifade edilebilir:

[11] Menard S.

$$D = -2\sum_1^n \left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \quad (1.28)$$

D istatistiği kimi yazarlar tarafından [örneğin McCullagh, Nelder (1983)] sapma(deviance) olarak tanımlanmıştır ve uyum iyiliği testlerinde önemli rol oynamaktadır.

Sapma olarak ifade edilen şey doğrusal regresyondaki kalıntı kareleri toplamı ile aynı görevi görmektedir. Doğrusal regresyon için hesaplanan SSE lojistik regresyon için D ile gösterilir. Yani SSE ile D istatistikleri aynı rolü üstlenmektedirler.[12]

Hedef değişken 0 ve 1 değerlerini alan ikili bir değişken ise doymuş modelin olabilirliği 1 dir. Bu çıkarım $\hat{\pi}_i = y_i$ eşitliğinden yapılabilir.

$$\mathrm{L(doymuş\ model)} = \prod_1^n y_i{}^{y_i}(1 - y_i)^{(1-y_i)} = 1$$

Bu durumda D istatistiğini denklem (1.29) ile yazabilmek mümkündür:

$$D = -2\ln(uygun\ bulunan\ modelin\ olabilirliği) \quad (1.29)$$

Açıklayıcı değişkenin önemliliğinin testi için açıklayıcı değişkenin modelde yer aldığı durumdaki ve yer almadığı durumdaki D istatistikleri kıyaslanır. Açıklayıcı değişkenin modelde bulunma durumuna göre D istatistiğinde meydana gelen değişimi denklem (1.30) ile ifade edebiliriz.

$$G = D(modelde\ açıklayıcı\ değişken\ yoksa) - D(modelde\ açıklayıcı\ değişken\ varsa) \quad (1.30)$$

G istatistiği doğrusal regresyondaki F testi ile aynı görevi üstlenir. Çünkü doymuş modelin olabilirliği hesaplanan iki D istatistiği için de aynıdır. O halde G istatistiğini denklem (1.31) ile yazmak mümkündür:

[12] Hosmer W. D. ve Lemeshow S.

$$G = -2ln\left[\frac{modelde\ \ açıklay ıcı\ değişken\ \ yokken\ \ olabilirlik}{modelde\ \ açıklay ıcı\ varken\ \ olabilirlik}\right] \quad (1.31)$$

Tek açıklayıcı değişken durumunda, modelde açıklayıcı değişken yokken β_0 ın en çok olabilirlik tahmini $\ln(n_1/n_0)$ olur. Buradaki $n_1 = \sum y_i$ ve $n_0 = \sum (1 - y_i)$ olmak üzere sabit terimin tahmin edilen değeri n_1/n dir.

$$G = 2\{\sum_1^n [y_i \ln(\hat{\pi}_i) + (1 - y_i) ln(1 - \hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]\} \quad (1.32)$$

Bu durumda β_1 in sıfıra eşit olduğu H_0 hipotezi altında denklem (1.32) de verilen G istatistiği bir serbestlik dereceli bir ki kare dağılımına uyar. Ek matematiksel varsayımlar elbette ki gereklidir ancak örnek büyüklüğü n yeterince büyük olduğunda bu varsayımların pek de bir önemi kalmamaktadır.[13]

Güven Aralığı Tahmini

Parametrelerin önemlilik testleri yapıldıktan sonra parametreler için güven aralıkları tahmin edilmek istenebilir. Doğrusal regresyon modelinde sabit terim ve açıklayıcı değişkenler için güven aralıkları tahmini yapılır. Lojistik regresyonda ise aynı mantıkla parametreler için güven aralığı tahmini yapılabilir.

Doğrusal regresyonda güven aralıkları denklem (1.33) verildiği gibidir:

$$\hat{\beta}_1 \pm z_{1-\alpha \div 2} \widehat{SE}(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm z_{1-\alpha \div 2} \widehat{SE}(\hat{\beta}_0) \quad (1.33)$$

Tahmin edilen parametreler bu güven aralıkları içinde yer alırsa parametreler anlamlıdır denilebilir.

---

[13] Anderson, T.W.

Lojistik regresyonda ise logit denklemi g(x) lojistik regresyonun doğrusal kısmı olup aynı doğrusal regresyondaki gibi değerlendirilir.

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 X$$

g(x) fonksiyonunun tahmini olan $\hat{g}(x)$ fonksiyonunun varyansını alacak olursak denklem (1.34) ile gösterilen bir ifade elde ederiz;

$$\widehat{Var}[\hat{g}(x)] = \widehat{Var}(\hat{\beta}_0) + x^2\widehat{Var}(\hat{\beta}_1) + 2x\widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) \quad (1.34)$$

Genellikle toplam bir ifadenin varyansı alındığında her terimin varyansı alınır ve iki terimin kovaryansı iki ile çarpılarak toplanır. 100(1 – α/2) güven ile logit fonksiyonu için güven aralıklarının yazılımı denklem (1.35) de gösterilmiştir.

$$\hat{g}(x) \pm z_{1-\alpha\div 2}\widehat{\widehat{SE}}[\hat{g}(x)] \quad (1.35)$$

Güven aralığında yer alan $\widehat{SE}[\hat{g}(x)]$ terimi $\widehat{Var}[\hat{g}(x)]$ ifadesinin karekök alınmış halidir.

2.4.7. Çok Değişkenli Lojistik Regresyon

2.4.7.1. Modelleme

Modelde x vektörü ile gösterilen p tane açıklayıcı olduğunda bütün değişkenlerin en az aralıklı ölçekli olduğunu varsayarız. Bu durumda hedef değişkenin koşullu olasılığı P(Y=1/x)=π(x) ile gösterilir. Çok değişkenli lojistik modelin logit fonksiyonu denklem (1.36) ile ifade edilir.

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1.36)$$

Bu durumda lojistik regresyon denklemi de denklem (1.37) ile ifade edilebilir.

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (1.37)$$

Eğer modeldeki bağımsız değişkenlerin bir kısmı kesikli, cinsiyet eğitim durumu gibi nominal ölçekli değişkenler ise bu değişkenleri aralıklı ölçekli gibi modele dahil etmek mümkün olmaz. Nominal ölçekli değişkenlerin çeşitli seviyelerini gösteren sayılar aslında sayısal olarak bir anlam ifade etmezler. Böyle durumlarda u değişkenler modele gösterge değişken olarak dahil edilirler.[14]

Genellikle bir değişkenin k tane seviyesi varsa k-1 tane değişken modele dahil edilir. Bu durumlarda modelde muhakkak sabit terim bulundurulmalıdır. Notasyonu gösterebilmek için j. inci gösterge değişken olan x_j nin k tane seviyesi olsun. Bu durumda modele k-1 tane D_{jl} ile gösterilen ve katsayıları β_{jl} l = 1,2,…,k_j-1 olan gösterge değişken ilave edilir. Böylece p değişkenli ve j. inci değişkeni kesikli olan model denklem (1.38)'deki gibi yazılır.

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p \qquad (1.38)$$

(x_i,y_i) ikililerinden oluşan n birimden oluşan bağımsız bir örneklem seti olduğunu varsayalım. Çok değişkenli modellerde katsayıların tahmini bir vektör ile gösterilir. Katsayıların tahmini tek değişkenli modelde olduğu gibi maksimum olabilirlik yöntemi ile yapılır. Olabilirlik fonksiyonu da tek bir farkla aynı tek değişkenli modelde olduğu gibidir. Söz konusu tek fark ise $\pi(x)$ artık bir denklem olarak ifade edilir. Olabilirlik fonksiyonunun p+1 tane katsayıya göre diferansiyeli olacağından P+1 tane olabilirlik fonksiyonu olacaktır.[15] Olabilirlik fonksiyonları denklem (1.39) ve denklem (1.40) ile gösterilebilir.

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad (1.39)$$

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] x_{ij} = 0 \qquad (1.40)$$
$$j = 1,2,\dots,p$$

[14] Tabachnick B. G., Fidell L. S.
[15] Hosmer W. D. ve Lemeshow S.

Denklemler çözülerek katsayıların tahminleri elde edilir. Bu çözümlemeler istatistik paket programları aracılığı ile rahatlıkla yapılabilir. Katsayıların tahminleri vektörü $\hat{\beta}$ olsun. $\hat{\pi}(x)$ ise $\hat{\beta}$ ve x_i ler ile hesaplanmış lojistik regresyon modeli fonksiyonu olsun.

Varyasnların ve kovaryansların tahmin edilmesinde yine maksimum olabilirlik tahmini teorisinden faydalanılır. Bu teoriye göre varyans ve kovaryansların tahmini log-benzerlik fonksiyonun ikinci dereceden kısmi türevlerinin matrisinden elde edilir. İkinci dereceden kısmi türevlerin genel formu denklem (1.41) ve denklem (1.42) ile ifade edilir.

$$\frac{\partial^2 L(\beta)}{\partial \beta_j{}^2} = -\sum_{i=1}^{n} x_{ij}{}^2 \ \pi_i(1-\pi_i) \tag{1.41}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_i \ \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} \ x_{il} \ \pi_i(1-\pi_i) \tag{1.42}$$

$j, l = 1,2, \dots, p$

denklem (1.41) ve denklem (1.42)'deki negatif terimleri içeren (p+1)x(p+1) lik matrisi $I(\beta)$ ile gösterelim. Bu matrise *gözlenen değerler matrisi* denir. Tahmin edilen katsayıların varyans ve kovaryansları bu matrisin tersi alınarak elde edilir. Bunu $Var(\beta) = I^{-1}(\beta)$ ile gösterebiliriz. Bu matristeki i inci diyagoneldeki elemanı $Var(\beta_i)$ ile göstereceğiz. $Cov(\beta_i, \beta_l)$ ise matriste β_i ile β_l nin kesiştiği noktada yer alır ve ikisi arasındaki kovaryansı ifade eder. $\widehat{Var}(\hat{\beta})$ ile $\hat{\beta}$ nın varyansını, $\widehat{Cov}(\hat{\beta}_i, \hat{\beta}_l)$ ile $\hat{\beta}_i \ ve \ \hat{\beta}_l$ nın kovaryansını ifade edeceğiz.

Katsayıların anlamlılık testlerinde kullanılan standart hata ise varyansın karekökü olarak denklem (1.43)'deki gibi ifade edilir.

$$\widehat{SE}(\hat{\beta}_j) = \left[\widehat{Var}(\hat{\beta}_j)\right]^{1/2} \tag{1.43}$$

$j = 1,2, \dots, p$

Yukarıda bahsettiğimiz gözlenen değerli matrisi $\hat{I}(\hat{\beta})$ yı denklem (1.44) ile yazmak mümkündür.

$$\hat{I}(\hat{\beta}) = X'VX \tag{1.44}$$

Buradaki X matrisi gözlenen değerlerin yer aldığı n x (p+1) boyutlarındaki veri matrisidir. V matrisi ise nxn boyutlarında diyagonalinde $\hat{\pi}_i(1-\hat{\pi}_i)$ terimlerinin bulunduğu matristir.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}_3(1-\hat{\pi}_3) \end{bmatrix}$$

2.4.7.2. Katsayıların Anlamlılık Testleri ve Güven Aralıkları Tahmini

Çok değişkenli lojistik regresyon modeli kurulduktan sonra modelin başarısı değerlendirilmelidir. İki değişkenli, modelde olduğu gibi ilk aşamada katsayıların anlamlılığına bakılır.

Anlamlılık Testi

P tane bağımlı değişkenin bağımsız değişkeni ne kadar açıklayabildiğini gösterebilmek için benzerlik oranı testi aynı iki değişkenli modeldeki, yöntemle yapılır. Test daha önce bahsettiğimiz G istatistiğine göre yapılır. Aradaki tek fark $\hat{\pi}_1$ değerleri p+1 parametre içeren $\hat{\beta}$ vektörüne bağlı olarak hesaplanır.

P tane katsayının sıfıra eşit olduğu H_0 hipotezi altında G' nin dağılımı p serbestlik dereceli bir ki kare dağılımına uymaktadır.

Güven Aralığı Tahmini

$$\hat{\beta}_j \pm z_{j-\alpha \div 2}\widehat{SE}(\hat{\beta}_j) \tag{1.45}$$

$$\hat{\beta}_0 \pm z_{1-\alpha \div 2}\widehat{SE}(\hat{\beta}_0) \quad (1.46)$$

Çok değişkenli modelde logit fonksiyonu için güven aralığı denklem (1.45) ve denklem (1.46) ifade ile gösterilse de hesaplanması daha karmaşıktır. Temelde mantık aynı fakat bu sefer g(x) fonksiyonu daha fazla terimden oluşmaktadır. P tane değişken içeren logit fonksiyonunun tahminini genel olarak aşağıdaki denklem (1.47) ile gösterebiliriz.

$$g(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \quad (1.47)$$

Bu modeli göstermenin alternatif bir yolu da vektör notasyonu kullanmaktır.

$$g(x) = x'\hat{\beta}$$

$\hat{\beta}' = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \cdots]$ vektörü p+1 tane katsayının tahminlerini gösterir.

Logit fonksiyonunun tahmininin varyansı denklem (1.48) ile ifade edilir.

$$\widehat{Var}\left(\hat{g}\ (x)\right) = \sum_{j=0}^{p} {x_j}^2\, \widehat{Var}\left(\hat{\beta}_j\right) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_i x_k \widehat{cov}(\hat{\beta}_j, \hat{\beta}_k) \quad (1.48)$$

2.4.8. Lojistik Regresyon Modelinin Yorumlanması

Modelin ve katsayıların anlamlılığı test edilip uygun modele karar verildikten sonra modelin yorumlanması kısmına geçilir. Yorumlama kısmında ilk aşama bağımsız değişkenler tarafından oluşturulan doğrusal bir fonksiyonunu belirlemektir. Bu doğrusal fonksiyona bağlantı fonksiyonu denir. Doğrusal regresyon modelinde bağımlı değişken ile bağımsız değişken arasındaki ilişki doğrusaldır. Bu nedenle doğrusal regresyon modelinde bağlantı fonksiyonu modelin kendisidir. Lojistik regresyon modelinde ise logit fonksiyonu bağlantı fonksiyonu olma görevini yerine getirir.

$$g(x) = \ln\{\pi(x)/[1-\pi(x)]\} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Doğrusal regresyon modelinde eğim katsayısı β_1, bağımlı değişkenin x+1 noktasındaki değeri ile bağımlı değişkenin x noktasındaki değeri arasındaki farka eşittir. Bunu $\beta_1 = y(x+1) - y(x)$ olarak gösterebiliriz. Bu durumda β_1 in yorumu kolaydır. Bağımlı değişkendeki bir birimlik değişim bağımsız değişkende β_1 kadar değişime neden olması beklenir.

Lojistik regresyonda ise eğim katsayısı bağımsız değişkendeki değişmeyi değil logit fonksiyonundaki değişmeyi gösterir.

$$\beta_1 = g(x+1) - g(x) \tag{1.49}$$

Lojistik regresyon modelinin yorumlanması bağımsız değişkenin formuna bağlı olarak değişmektedir. Bu nedenle farklı durumlarda nasıl yorumlanacağı ayrı başlıklar halinde anlatılmıştır.

2.4.8.1. İki Düzeyli Bağımsız Değişken Olduğu Durumlar

İlk olarak regresyon modelindeki bağımsız değişkenin nominal ölçekli ve iki düzeyli olduğu durum ile başlıyoruz.

x, bağımsız değişkeninin 0 ve 1 olarak kodlandığını varsayalım. x = 0 ve x = 1 olduğunda logit fonksiyonları arasındaki fark denklem (1.50) ile gösterilmiştir.

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1 \tag{1.50}$$

Denklem (1.50)'de gösterilen basit bir matematiksel ifadedir. Bu denklemi bu aşamada detaylı olarak inceleyeceğiz. Bir modelde değişkenin etkisini yorumlayabilmek için ilk adım logit fonksiyonları arasındaki farkı model terimleri ile göstermektedir. Yukarıdaki denklem (1.50)'ye göre logit fonksiyonları arasındaki fark β_1' e eşittir. Bu sonucu yorumlayabilmek için "odds oranı" nı incelememiz gerekmektedir.

Çizelge 1.1. Mümkün Olabilecek Lojistik Olasılıkları

Hedef Değişken	Bağımsız Değişken	
	x=1	x=0
y=1	$\pi(1)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
y=0	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1 x}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_0}}$
Toplam	1	1

Mümkün olabilecek lojistik olasılıkları yukarıdaki tabloda gösterilmiştir. x=0 olduğunda hedef değişken y' nin oddsu $\pi(0)/[1-\pi(0)]$ olacaktır. Benzer şekilde x=1 olduğunda da $\pi(1)/[1-\pi(1)]$ olacaktır. x=1 in oddsu ile x=0 ın odsunun birbirlerine oranlanması ile OR olarak gösterilen odds oranı elde edilir.

$$OR=\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{1.51}$$

Lojistik regresyon modeli denklemini yukarıdaki tabloda yerine koyarsak denklem (1.52)'nin nasıl el edildiği görülecektir.

$$OR=\frac{\left[\frac{e^{\beta 0+\beta 1}}{1+e^{\beta 0+\beta 1}}\right]/\left[\frac{1}{e^{\beta 0+\beta 1}}\right]}{\left[\frac{e^{\beta 0}}{1+e^{\beta 0}}\right]/\left[\frac{1}{e^{\beta 0}}\right]}$$

$$=\frac{e^{\beta 0+\beta 1}}{e^{\beta 0}}$$

$$=e^{(\beta_0+\beta_1)-\beta_0}$$

$$=e^{\beta_1} \tag{1.52}$$

Dikotomik bağımlı değişkene sahip lojistik regresyon modelinde odds oranı ile regresyon katsayısı arasındaki ilişki denklem (1.53) ile gösterilir.

$$OR = e^{\beta_1} \tag{1.53}$$

Katsayı ile odds oranı arasındaki bu basit ilişki lojistik regresyonun güçlü ve analitik bir araştırma aracı olmasının en temel nedenidir. [16]

Odds oranı kullanım alanı geniş olan bir ölçümdür. Y nin bir bankada iyi veya kötü müşteri olma durumunu gösterdiğini varsayalım. X de evli olan müşterileri temsil etsin. Bu durumda $\widehat{OR} = 2$ olarak tahmin edilmiş olsun. Evli olan müşterilerin kötü müşteri olma olasılıkları evli olmayan müşterilerin iki katı olması beklenir şeklinde yorumlanabilir.

Bağımsız değişken 0 ve 1 ile kodlandığında $OR = e^{\beta_1}$ olduğunu göstermiştik. Bağımlı değişken o ve 1 dışında bir rakam ile kodlanmışsa odds oranı logit fonksiyonları yardımıyla bulunur.
x=a ve x=b olarak kodlandığı bir durumda odds oranının nasıl tahminleneceğini detaylı olarak denklem seti (1.53)' de gösterilmiştir.

$$\ln\left[\widehat{OR}(a,b)\right] = \hat{g}(x=a) - \hat{g}(x=b) \tag{1.53}$$

$$\ln\left[\widehat{OR}(a,b)\right] = \left(\hat{\beta}_0 + \hat{\beta}_1 a\right) - (\hat{\beta}_0 + \hat{\beta}_1 b)$$

$$\ln\left[\widehat{OR}(a,b)\right] = \hat{\beta}_1 (a-b)$$

2.4.8.2. İkiden Fazla Kategorili Bağımsız Değişken Olması Durumu

Bağımsız değişkenin ikiden fazla düzeye sahip olduğunu düşünelim. Bu düzeyler kesikli değerleri alırlar ve nominal ölçeklidirler. Nominal ölçekli değişkenler genellikle modelleme için uygun değildir. Bu nedenle bu tür değişkenleri modelleme için uygun bir forma sokmak gerekir. Bu bölümde bu tip değişkenleri nasıl dizayn edileceğini göstereceğiz.

[16] Menard S.

Bir örnek ile konuyu daha iyi anlatabiliriz. Örnek modelimizde bankadan kredi alan müşterilerin eğitim durumuna göre iyi ya da kötü müşteri olma olasılıkları tahmin edilsin. Bu durumda bağımlı değişken müşterinin iyi/kötü olması, bağımsız değişken de müşterinin eğitim düzeyi olur. Eğitim düzeyi, ilköğretim, lise, üniversite ve lisansüstü olmak üzere dört düzeyden oluşmaktadır. Değişkenlere ilişkin verileri gösteren tablo aşağıdaki gibidir.

Çizelge 1.2. Eğitim durumuna göre iyi kötü sayılarının dağılımı

	Eğitim Durumu				Toplam
Müşteri Durumu	İlköğretim	Lise	Üniversite	Lisansüstü	
İyi	5	20	15	10	50
Kötü	20	10	10	10	50
Odds Oranı	***1***	***8***	***6***	***4***	
ln($\widehat{OR}$)	***0***	***2,08***	***1,79***	***1,39***	

Bu örnek için bağımsız değişkenin dört tane düzeyi olduğu için k=4 diyebiliriz. Tablonun altında her eğitim düzeyi için ilkokul referans düzey alınarak odds oranları hesaplanmıştır. Örneğin üniversite mezunu olanlar için tahmin edilen odds oranı 15x20/5x10 = 6 olarak hesaplanmıştır. Odds oranlarının logaritması alınmış halleri de yine tablonun altında verilmiştir.

Çizelge 1.3. Eğitim Değişkeninin Kodlanması

	Dizayn Değişkenler		
Eğitim Durumu(KOD)	Eğitim(2)	Eğitim(3)	Eğitim(4)
İlköğretim(1)	0	0	0
Lise(2)	1	0	0
Üniversite(3)	0	1	0
Lisansüstü(4)	0	0	1

Tahmin edilen odds oranlarının hesaplanması için seçilen referans grup "1" ile kodlanan düzeydir ilköğretim düzeyidir. Eğitim durumu değişkeni modele üç ayrı değişken olarak dahil edilir. değerlerin kodlanması yukarıda yer alan tablodaki gibi yapılır. Eğitim(2) değişkeni Lise düzeyini,

Eğitim(3) değişkeni üniversite düzeyini, Eğitim(4) değişkeni ise lisansüstü düzeyini ifade eder. Bu üç değişkenin de "0" değer alması durumu da ilkokul düzeyini ifade eder.

Bu örnek için tahmin edilen lojistik regresyon modeli sonucu aşağıdaki tabloda yer almaktadır.

Çizelge 1.4. Eğitim değişkenine ait değerler

Değişken	Katsayılar	Standart Hata	z
Eğitim(2)	2,079	0,633	3,29
Eğitim(3)	1,792	0,647	2,78
Eğitim(4)	1,386	0,671	2,07
Sabit Terim	-1,386	0,500	-2,77

Tahmin edilen regresyon terimlerinin odds oranlarına bakalım.

$$\ln\left[\widehat{OR}\ (\text{Lise, İlköğretim})\right] = \hat{\beta}_1 = 2{,}079$$

$$\ln\left[\widehat{OR}\ (\text{Üniversite, İlköğretim})\right] = \hat{\beta}_2 = 1{,}792$$

$$\ln\left[\widehat{OR}\ (\text{Lisansüstü, İlöğretim})\right] = \hat{\beta}_3 = 1{,}386$$

Tahmin edilen katsayıların odds oranlarına eşit olması tesadüfen mi olmuştur yoksa değişkenin dizayn edilmesinin bir sonucu mudur? $\ln\left[\widehat{OR}\ (\text{Lise, İlköğretim})\right]$ için formülü açacak olursak bunun tesadüf olmadığını görebiliriz.

$$\ln\left[\widehat{OR}\ (\text{Lise, İlköğretim})\right] = \hat{g}\ (\text{Lise}) - \hat{g}\ (\text{İlköğretim})$$

$$= \left[\hat{\beta}_0 + \hat{\beta}_1 \times (\text{Eğitim}(2) = 1) + \hat{\beta}_2 \times (\text{Eğitim}(3) = 0) + \hat{\beta}_3(\text{Eğitim}(4) = 0)\right]$$

$$-\left[\hat{\beta}_0 + \hat{\beta}_1 \times (\text{Eğitim}(2) = 0) + \hat{\beta}_2 \times (\text{Eğitim}(3) = 0) + \hat{\beta}_3(\text{Eğitim}(4) = 0)\right]$$

$$= \hat{\beta}_1$$

Benzer hesaplamalar diğer lojistik regresyon ile tahmin edilen katsayıların odds oranları için yapıldığında tahmin edilen katsayıların odds oranlarına eşit olduğu görülecektir.

2.4.8.3. Sürekli Bağımsız Değişken Olması Durumu

Lojistik regresyon modeline bir sürekli değişken dahil olması durumunda sürekli değişkenin yorumu o değişkenin modele nasıl dahil olduğuna ve değişkenin belirli birimlerine bağlıdır. Lojistik regresyon modelinde sürekli bir değişkeni yorumlayabilmek için logit fonksiyonunun sürekli değişkene göre doğrusal olduğunu varsayalım.

Logit fonksiyonun x sürekli değişkeninde doğrusal olduğu varsayımı altında logit fonksiyonu $g(x) = \beta_0 + \beta_1 x$ olsun. Eğim değişkeni β_1, x'deki bir birimlik artışın oddsların logaritmasında ne kadarlık bir değişime neden olacağını gösterir. Genellikle "bir birimlik" değişimin yorumlanması yeterli görülmez. Örneğin bir bankadan kredi kullanan bir müşterinin iyi ödeme performanslı ya da kötü ödeme performanslı olduğunu açıklamaya çalışırken "yaş" değişkenini kullandığımızı varsayalım. Yaştaki bir yıllık artışın bağımlı değişken üzerindeki etkisini yorumlamak çok etkili veya anlamlı olmayabilir. Bunun yerine yaşta 10 yıllık bir artışa tekabül eden bir değişimi yorumlamak daha anlamlı olabilir. Ya da tam tersi durumlar olabilir. Ölçüm düzeyi küçük olan değişkenleri de birden küçük bir değer üzerinden yorumlamak isteyebiliriz.

Sürekli değişkenlerin olduğu lojistik regresyon modellerinde daha anlamlı yorumlara yapabilmek için c birimlik değişimleri gösteren nokta ve aralık tahminleri yapılabilir. x deki c birimlik değişim için logit fonksiyonlarının farkı $g(x+c) - g(x) = c\beta_1$ olarak gösterilir. Odds oranı ise $e^{c\beta_1}$ olarak hesaplanır. Tahmin edilen değerleri göstermek için ise β_1 yerine $\hat{\beta}_1$ yazılır. OR(c) için güven aralığı denklem (1.54)'deki gibi yazılır.

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2}\, c\, \widehat{SE}(\hat{\beta}_1)] \qquad (1.54)$$

Böylece nokta ve aralık tahminleri belirlenen c değerine göre yapılabilir. Örneğimize geri dönecek olursak yaştaki c yıllık artış kötü müşteri olma olasılığını $\widehat{OR}(c)$ kadar arttırır diyebiliriz.

2.4.9. Uyum İyiliği Testleri

Değişkenler seçilip modelleme aşaması başarılı bir şekilde sonuçlandırıldıktan sonra modelin başarısının nasıl olduğunu görmek isteriz. Modelde yer alan değişkenler hedef değişkenin açıklanmasında ne kadar etkili, bunun değerlendirmesini yapmak isteriz. Bu değerlendirmenin yapılabilmesi için uyum iyiliği testlerinin yapılması gerekir.

Modelin uyum iyiliğini ölçmek istiyorsak, iyi bir modelin nasıl olması gerektiği konusunda bilgi sahibi olmak gerekir. Hedef değişkene ait gözlenen değerlerin olduğu örneklem seti olduğunu varsayalım bunu da $y' = [y_1, y_2, y_3, \dots, y_n]$ vektörü ile gösterelim. Model ile tahmin edilen değerlere tahminci deriz ve y için bunu $\hat{y}$ ile gösteririz. Eğer tahmin edilen değer ile gözlenen değer arasındaki fark yani $(y - \hat{y})$ ne kadar küçükse model de o kadar iyidir diyebiliriz. Dolayısıyla tahmin gücü iyi bir model etmenin yolu y ile $\hat{y}$ arasındaki uzaklığın minimize edilmesinden geçmektedir. [17]

Modelde $x' = [x_1, x_2, x_3, \dots, x_p]$ ile gösterilen p tane bağımsız değişken olsun. j de x'in gözlenen değerlerinin sayısını versin. Eğer bazı birimlerde x aynı değerleri aldıysa J<n olacaktır. Aynı değerlere sahip x_j'lerin sayısını m_j ile gösterelim j = 1,2,3, . . ., j. Bu durumda $\sum m_j = n$ eşitliği sağlanır.

Doğrusal regresyon modelinde uyum iyiliği testleri kalıntı olarak isimlendirilen gözlem değeri ile tahmin edilen değer arasındaki farka $(y - \hat{y})$ bakılarak yapılır. Lojistik regresyonda ise bu farkı ölçmek için çeşitli yöntemler vardır.

2.4.9.1. Pearson Ki-Kare

Lojistik regresyonda tahminciler değişkenin gözlenme olasılığından hesaplanır.

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}} \qquad (1.55)$$

[17] Tabachnick B. G., Fidell L. S

Denklem (1.55)'deki $\hat{g}(x_j)$ ise tahmin edilen logit fonksiyonudur.

Pearson'a göre gözlemlenen değer ile tahmin edilen değer arasındaki fark olan kalıntı denklem (1.56)'daki şekilde hesaplanır.

$$r(y_i, \hat{\pi}_j) = \frac{(y_i - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1-\hat{\pi}_j)}} \tag{1.56}$$

Buradan hareketle pearson ki kare istatistiği denklem (1.57)'deki gibi ifade edilir.

$$\chi^2 = \sum_{j=1}^{J} r(y_i, \hat{\pi}_j)^2 \tag{1.57}$$

χ^2 istatistiği serbestli derecesi $J - (p + 1)$olan bir ki kare dağılımına uymaktadır.

2.4.9.2. Deviance İstatistiği

Deviance kalıntısı denklem (1.58)'deki gibi tanımlanır.

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_i \ln \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \ln \left(\frac{m_j - y_j}{m_j (1-\hat{\pi}_j)} \right) \right] \right\}^{1/2} \tag{1.58}$$

Baştaki işaret artı veya eksi ise $(y_j - m_j \hat{\pi}_j)$ nin işareti de aynıdır. $y_j = 0$ olduğu durumda deviance kalıntısı aşağıdaki denklem(1.59)'da gösterildiği gibi yazılabilir.

$$d(y_j, \hat{\pi}_j) = -\sqrt{2 m_j \left| \ln(1 - \hat{\pi}_j) \right|} \tag{1.59}$$

$y_j = 1$ olduğunda ise deviance istatistiği denklem (1.60)'da gösterilmiştir.

$$d(y_j, \hat{\pi}_j) = -\sqrt{2 m_j \left| \ln(\hat{\pi}_j) \right|} \tag{1.60}$$

Deviance kalıntıları ile heseplanan deviance istatistiğini aşağıdaki denklem(1.61) ile özet bir şekilde ifade edebiliriz.

$$D = \sum_{j=1}^{J} d\left(y_i, \hat{\pi}_j\right)^2 \quad (1.61)$$

D istatistiği de aynı pearson gibi serbestli derecesi $J - (p + 1)$ olan bir ki kare dağılımına uymaktadır.

2.4.9.3. Hosmer-Lemeshow İstatistiği

Hosmer ve Lemeshow 1980 yılında tahmin edilen olasılıkların gruplanması ile ilgili bir çalışma yaptılar. J=n olduğunu kabul edelim. Bu durumda n tane sütüna karşılık tahmin edilen olasılıkların n tane değeri olsun. ilk sütun en küçük değerli olasılıkları, n inci sütun ise en büyük olasılık değerleri göstersin. Gruplama iki farklı şekilde yapılır.

1. Tahmin edilen olasılıkların yüzdelik dilimlere göre tablo ayrıştırılır. En küçük %10 luk grup, en büyük %10 luk grup gibi.
2. Tahmin edilen olasılıkların belirlenmiş belli eşik değerlerine göre tablo ayrıştırılır.

Örneğin ilk yöntemde g=10 tane grup olduğunda ilk grup n_1=n/10 birim en küçük olasılıkları içeren değerleri içerir. İkinci yöntemde ise g=10 tane grupta k eşik değeri olmak üzere k/10 değerine göre belirlenmiş tahmin edilen olasılıklar yer alır. Örneğin ilk grup olasılık değeri 0.1 den küçük eşit olanları, onuncu grup da olasılığı 0.9 dan büyük olan değerleri içerir. y=1 satırı için beklenen değerlerin tahmini gruptaki tahmin edilen olasılıkların toplanması ile elde edilir. y=0 satırı için de bu olasılık ların toplamı birden çıkartılarak bulunur.[18] Her iki gruplama stratejisinde de Hosmer-Lemeshow istatistiği, $\hat{C}$, pearson ki kare istatistiği ve gx2 lik tabloda gözlenen değerler ile beklenen frekanslar ile hesaplanır.

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (1.62)$$

Formülde yer alan n'_k değeri k ıncı gruptaki birim sayısını gösterir. o_k ise değişkenlere karşılık gelen y değerlerinin sayısıdır. $\bar{\pi}_k$ değeri tahmin edilen ortalama olasılığı ifade eder.

[18] Hosmer W. D. ve Lemeshow S.

Hosmer ve Lemeshow göstermişlerdir ki J=n olduğu durumlarda kurulan lojistik regresyon modeli başarılı ise $\hat{C}$ istatistiği (g-2) serbestlik dereceli bir χ^2 dağılımına uymaktadır. J≠n olduğu durumlarda ise $\hat{C}$ istatistiğinin dağılımı yine (g-2) serbestlik dereceli bir χ^2 dağılımına yakınsamaktadır.

2.4.9.4. ROC Eğrisinin Altında Kalan Alan

Sınıflamanın doğru yapılıp doğru yapılmadığını gösteren bir başka gösterge de ROC (Receiver Operating Characteristic) eğrisi altında kalan alandır. Bu eğri orijinal teoriye göre gürültü olduğu durumlarda alıcı tarafından sinyal belirlenirken doğru sinyali yakalama olasılığını gösterir.

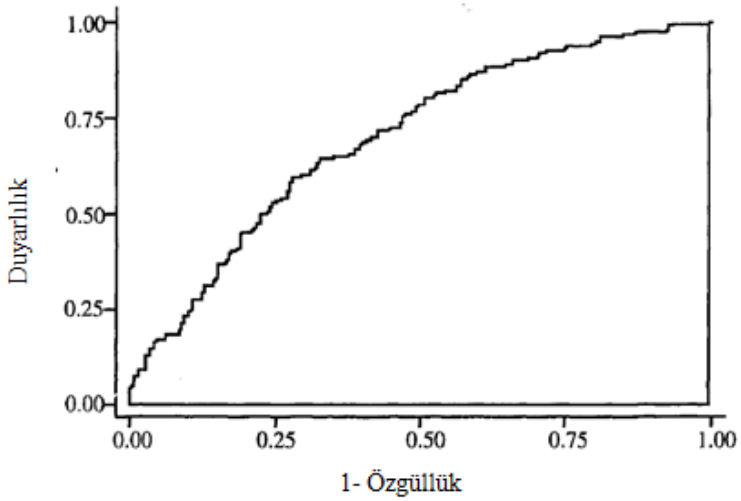
ROC eğrisi altında kalan alan 0 ile 1 arasında değerler alır. Bu alan modelin hedef değişkende belirlenen gurupları ne kadar iyi ayırabildiğinin bir göstergesidir.

Lojistik Regresyon modellemesinin sonucunda tahmin edilen hedef değişken değerleri ve gözlenen değerleri gösteren bir sınıflama tablosu oluşturulur.

Çizelge 1.5. Sınıflama Tablosu

	Gözlenen	
Tahmin Edilen	y = 0	y = 1
y = 0	a	b
y = 1	c	d
Toplam	a+c	b+d

Bu tabloya göre duyarlılık a/a+c, specificity ise d/(b+d) olarak hesaplanır. Belirlenen kesim noktasına göre bu değerler farklılık gösterebilir. Kesim noktası genellikle 0.50 olarak tercih edilir. ROC eğrisi ise duyarlılık y ekseninde, (1-özgüllük) x ekseninde olacak şekilde çizilir.



Şekil 1.1. ROC Eğrisi

Duyarlıklık, p olasılığının gerçekleşme olasılığını verir. Örneğin müşterilerin iyi ya da kötü performanslı olma durumu bağımlı değişken olarak kullanılan bir lojistik regresyon modelinde müşterilerin gecikmişe düşme olasılıkları hesaplanmak istensin. Bu durumda p müşterinin kötü olma olasılığıdır. Duyarlılık ise lojistik regresyon sonucunda tahmin edilen grup üyelikleri içinde p nin doğru tahmin edilme oranıdır diyebiliriz. Özgüllük ise lojistik regresyon modeli sonucuna göre tahmin edilen ancak gerçekte gözlenmeyen durumların oranını ifade eder.[19]

Oluşan eğri ile x ekseni arasında kalan alan da ROC eğrisi altında kalan alanı verir. Bu alan ile ilgili genel kural aşağıda verilmiştir. ROC = 0.50 ise ayırım yoktur, model başarısı düşüktür.[20]

- $0.70 \leq ROC \leq 0.80$ ise kabul edilebilir bir ayırım vardır.
- $0.80 \leq ROC \leq 0.90$ ise iyi bir ayırım söz konusudur.
- $ROC \geq 0.90$ ise mükemmel bir ayırım olduğu söylenebilir.

2.4.9.5. GINI Katsayısı

GINI katsayısı gelir dağılımındaki eşitsizliği ölçmek amacıyla kullanılan bir ölçüm değeridir. İtalyan ekonomist Corrado Gini tarafından geliştirilen bu katsayı 0 ile 1 arasında çıkan bir

[19] Altman E. I.
[20] Anderson, R.

değerdir. Mutlak eşitlik doğrusu ile Lorenz eğrisi altında kalan alanın, mutlak eşitlik doğrusu altında kalan alan üçgenin alanına oranlanması ile bulunur.

Gini katsayısı skor kart modellerinde modelin iyi kötü müşterileri ayırt etme gücünü ölçmek için kullanılan oldukça yaygın ve güvenilir bir yöntemdir. Gini katsayısının formülü aşağıda denklem (1.63)'de verilmiştir.

$$D = 1 - \sum_{i=1}^{n}((cpN_i + cpN_{i-1})(cpP_i - cpP_{i-1})) \quad (1.63)$$

Formülde yer alan cpN_i iyilerin kümülatif yüzdelerini, cpP_i de kötülerin kümülatif yüzdelerini göstermektedir.

Gini değerinin kaç olması kabul edilebilir bir değerdir? Bu sorunun cevabı skor kart modelinin başvuru ya da davranış skor kart modeli olmasına göre farklılık gösterebilir. Başvuru skorlamasında %50 den büyük bir değer çıkması tatmin edici sonuçlar içerdiğini gösterir. Gini katsayısının %30 dan az olması ise sonuçların yeterince iyi olmadığını gösterir.

ROC eğrisi altında kalan alan ile benzer şekilde hesaplanır. Hatta aralarında yaklaşık bir ilişki vardır.[21] Bu ilişkiyi aşağıdaki denklem (1.64) ile gösterebiliriz.

D gini katsayısnı göstermek üzere

$$ROC \approx (D + 1)/2 \quad (1.64)$$

2.4.10. Lojistik Regresyonda Değişken Seçme Yöntemleri

Lojistik Regresyonda değişken seçme yöntemleri ile açıklayıcı değişkenlerin modele nasıl dahil edileceğine karar verilir. Farklı yöntemler kullanarak aynı değişken setinden çeşitli regresyon modelleri oluşturulabilir. Diğer çok değişkenli istatistiksel yöntemlerde olduğu gibi adımsal seçim modellerinde bir sonraki aşamada hangi değişkenin modele dahil edileceğine karar

[21] Anderson, R.

verilmektedir. İstatistiksel modellemelerde yöntem, farklı modellerin denenerek içlerinden veri seti için en uygun modeli seçilmesi şeklindedir.

Enter Yöntemi: Bütün değişkenler bir blok olarak tek aşamada modele dahil edilir.

Forward Selection (Conditional): İleriye doğru adımsal bir yöntemdir. Değişkenler modele teker teker alınarak kriterleri sağlamayanlar modelde tutulmaz. Değişkenler modele alınırken skor istatistiğinin önemine, çıkarılırken de koşullu parametre tahminlerine dayanan olabilirlik oranının olasılığına göre karar verilir.

Forward Selection (Likelihood Ratio): İleriye doğru adımsal bir yöntemdir. Değişkenler modele alınırken skor istatistiğinin önemine, çıkarılırken de maksimum kısmi olabilirlik tahminlerine dayanan olabilirlik oranının olasılığına göre karar verilir.

Forward Selection (Wald): İleriye doğru adımsal bir yöntemdir. Değişkenler modele alınırken skor istatistiğinin önemine, çıkarılırken de Wald istatistiğinin olasılığına göre karar verilir.

Backward Elimination (Conditional): Geriye dogru adımsal seçim yöntemidir. Önce tüm degiskenler modele alınır daha sonra birer birer kriterleri sağlamayan değişkenler modelden çıkartılır. Tüm geriye doğru yöntemlerde önce tüm değişkenler alınıp sonra teker teker çıkarılması yaklaşımı geçerlidir. Değişkenler modelden çıkarılırken koşullu parametre tahminlerine dayanan olabilirlik oranının olasılığına göre karar verilir.

Backward Elimination (Likelihood Ratio): Geriye doğru adımsal seçim yöntemidir. Değişkenler modelden çıkarılırken maksimum kısmi olabilirlik tahminlerine dayanan olabilirlik oranının olasılığına göre karar verilir.

Backward Elimination (Wald): Geriye doğru adımsal seçim yöntemidir. Değişkenler modelden çıkarılırken Wald istatistiğinin olasılığına göre karar verilir.

3. KREDİ SKORLAMA

3.1. Kredi ve Skorlama Kavramlarının İncelenmesi

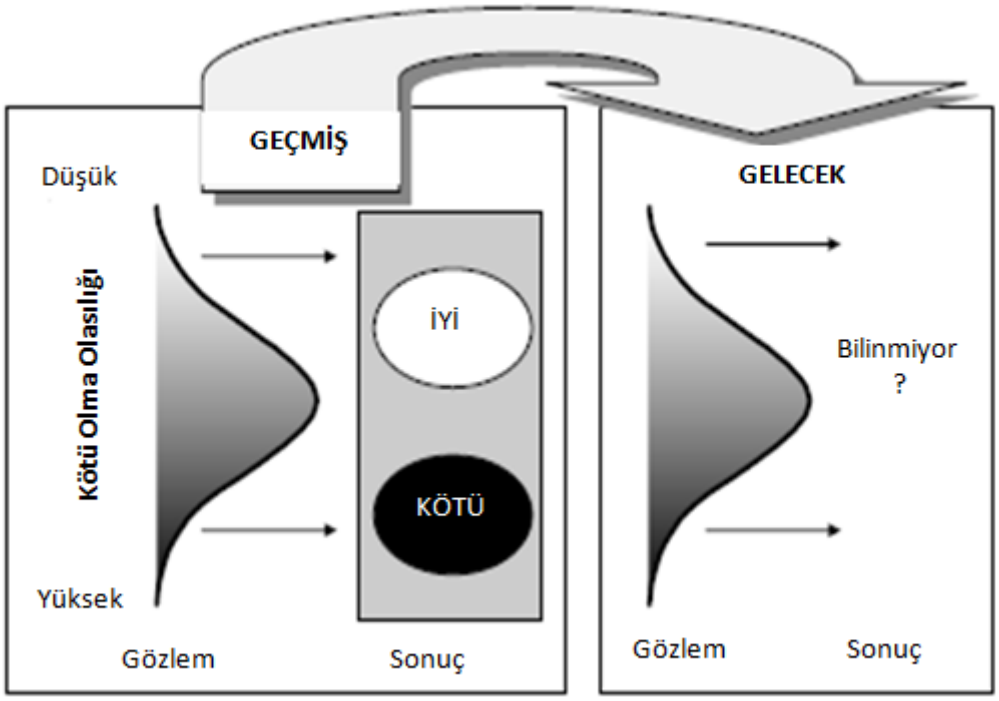
"Kredi skorlama" kavramından önce kredi ve skorlama kelimelerini incelemek gerekir.

Günümüz literatüründe **kredi** kısaca "şimdi al, sonra öde" olarak tanımlanabilir. Alışverişe konu olan ürün tüketicilerin veya işletmelerin kısa/uzun vadeli ihtiyaçları, gayrimenkul, araç vs. gibi varlıklar olabilir. Kredi kelimesi Latince kökenli olup Latincede güvenmek anlamına gelmektedir.

Skorlama bazı şeyleri kendi içlerinde (insanlar, firmalar, ülkeler, vs.) kalitesine, performansına, talep görmesine ya da satılabilir olması gibi kendine has özelliklerine göre derecelendirirken nümerik bir araç kullanmaktır. Amaç, bir bir popülasyonda bulunan birimleri birbirlerinden ayrıştırabilmek ve neticede objektif ve tutarlı kararlar alınmasını garanti etmektir.

Kredi Skorlama Kavramı

Günümüzde kredi sektörünün en önemli yönetim enstrümanlarından biri olan kredi skorlama, kredi başvuru sahibinin verilerinin istatistiksel yöntemler ile analiz ederek, müşterinin gelecekteki ödeme performansının iyi mi kötü mü olacağını bir model yardımıyla tahmin etmeyi amaçlar. Bu yöntemle elde edilen modelde kredi başvurusu esnasında müşteriden bilgi olarak alınan parametreler ağırlıklandırılarak müşterilere puan verilmektedir. Böylece kredilerin temerrüde düşme olasılığı müşteri bazında öngörülebilmektedir. Söz konusu model için geçmişte kredi başvurusunda bulunmuş müşterilerin tüm biriken bilgileri veri olarak kullanılır.



Şekil 2.1. Kredi skorlama metodolojisi[22]

Sübjektif kredi değerlendirme yöntemi, başvuru sahibinin verilerinin analizini ve kredi kararını bir kredi tahsis uzmanının yargılarına emanet ederken kredi skorlamada söz konusu analiz ve netice, çeşitli istatistiksel çalışmaların bir sonucu olarak karşımıza çıkmaktadır. Sübjektif değerlendirmede kredi kararlarındaki öznellik ve karar faktörlerinin belirsizliği skorlama modeli ile karşılanmaktadır. Skorlamanın altında yatan felsefe pragmatizm ve ampirizmdir.[23] Skorlama bir müşterinin krediyi geri ödeyip ödemeyeceğini tahmin ederken neden olarak belli bir davranışı göstermekle ilgilenmez.

Alan Greenspan'a göre skorlama sağlam istatistiksel ölçümlemeye dayanan ampirik bir türev tekniktir. Bunu destekler biçimde skorlama sistemleri benzer özelliklere sahip ve geçmişte benzer performans sergilemiş tüketicilerin ortak yanlarını istatistiksel yöntemlerle bularak benzer özelliklere sahip yeni başvurularda aynı performansı göstereceği yaklaşımı altında kararlarını oluşturmaktadır. Geçmişi incelenen verilerin istatistiksel olarak anlamlı olması gerekir.

[22] Anderson
[23] Thomas L., Edelman D. Ve Crook J

Skorlamanın ampirik ve pragmatik bir yöntem olmasının bir diğer açıklaması tüketicinin veya bulunduğu çevrenin herhangi bir özelliğinin tüketicinin talep ettiği kredinin sorunlu hale gelip gelmeyeceğini öngörmede yardımcı olmasıdır. Örneğin tüketicinin yaşı, mesleği, bakmakla yükümlü olduğu kişi sayısı gibi özelliklerin müşterilerin skorlanmasında kullanılması gibi. Müşterinin ödeme performansı etkileyen bazı değişkenlerin kullanımı yasa ile sınırlanmıştır. Örneğin ABD'de din, ırk, cinsiyet gibi değişkenlerin ayrımcılığa neden olabileceği düşüncesi ile skorlamada kullanılması yasaklanmıştır.[24]

3.2. Kredi Skorlama Tipleri

Kredi skorlama modelleri başvuru türüne göre iki farklı şekilde oluşturulmaktadır;

3.2.1. Başvuru Skorlaması

Kredi veren kuruluşların kendilerine ilk kez başvuran müşterileri değerlendirmek için geliştirilen bir model tipidir. Modelde müşterinin ödeme performansı bilgilerinden ziyade demografik parametreleri ve kredi büro bilgileri yer alabilir. Bir müşterinin başvuru skoru ile öncelikli olarak başvurusunun onaylanıp onaylanmayacağı skorlanır. Bunun yanı sıra müşterinin kaldırabileceği risk miktarı, ürün tipine göre onaylanabilecek kredi tutarı ve toplam limitinin belirlenmesi gibi konularda da karar alma noktasında başvuru skoru kullanılabilir.

3.2.2. Davranış Skorlaması

Kredi veren kuruluşların daha önce kendileriyle çalışmış olan müşterilerin yeni başvurularında veya devam eden projelerindeki temerrüde düşme olasılığını hesaplamak için kullandığı skor kart modelidir. Davranış skor kartı ile mevcut müşterilerin limit miktarlarının arttırılması veya düşürülmesi, limit yenileme, aktif limitinin pasif hale getirilmesi veya tam tersi konularda, müşteriye özel tahsilat stratejilerinin belirlenmesinde de karar alınabilir.[25]

---

[24] Thomas L., Edelman D. ve Crook J.
[25] Siddiqi N.

3.3. Kredi Skorlamanın Tarihçesi

Kredinin tarihi 5000 yıl geriye gitmesine rağmen kredi skorlamanın tarihi yalnızca 50-60 yıla dayanmaktadır. Kredi skorlama aslında bir ana kütle içinde yer alan farklı grupları, başkalarının kolaylıkla fark edemediği, grupları tanımlayan, diğerlerinden ayırıcı özellikleri bularak belirlemenin bir yoludur. Grupları belirleme probleminin ilk aşaması ana kütle içindeki grupları belirlemektir. Bu problemi ise ilk kez Fisher(1936) istatistiksel yöntemler kullanarak çözmüştür. Fisher bitkilerin fiziksel büyüklüklerine göre iki çeşit süseni ayırt edebilmiştir. 1941 yılında Durand Fisher'in kullandığı teknikler kullanılarak iyi ve kötü kredilerin birbirlerinden ayırt edilebileceğini söylemiştir. Amerikan Ulusal Ekonomik Araştırmalar bürosu için yaptığı çalışmada bunu göstermiştir.[26]

1930 larda bazı posta sipariş firmaları kredi değerlendiren kişilerin aldıkları kredi kararlarındaki tutarsızlıkları gidermek amacıyla nümarik skorlama sistemleri kullandı.(Weingartner 1966, Smalley ve Sturdirant 1973) II. Dünya Savaşı'nın başlamasıyla bütün finans kurumları ve posta sipariş firmaları kredilerin değerlendirilmesinde zorluklar yaşadılar. Kredi değerlendiren uzmanların çoğu askere alındıklarından dolayı bu işi iyi bilen uzmanların sayısında ciddi azalmalar oldu. Buna çözüm olarak da firmalarda görev yapan kredi değerlendirme uzmanları kredi verirken değerlendirdikleri kriterleri ve kuralları toparlayarak yazılı hale getirdiler (Johnson 1992). İşte bu kurallar skor kart olarak kullanılmaya başlandı. Tabi o uygulama bilgisayar aracılığı ile değil karşılaştırma anahtarları vasıtasıyla yapılıyordu. Kredi analistleri daha önce kredi verilen müşterilerin parametrelerine göre çıkarılan puan cetvellerine göre başvurulara puan vererek topluyorlardı. Parametreler arasında var olabilecek korelasyonlar göz önünde bulundurulmuyordu. Speigel firmasının yöneticisi Henry Wells istatistiksel yöntemleri kullanarak ilk kez kredi skorlamayı uygulayan kişi oldu.[27]

Savaş bittikten kısa bir süre sonra insanlar kredi kararlarının otomizasyonu ile tanıştı. İstatistikte sınıflama teknikleri geliştirildi. Kredi kararları alınırken istatistiksel tekniklerin kullanılmasının faydaları görüldü(Wonderlic 1958). Bu konuda ilk danışmanlık firması Fair Isaac, Bill Fair ve Earl Isaac tarafından 1950 li yılların başında kuruldu.

---

[26] Thomas L., Edelman D. ve Crook J.
[27] Halisdemir Ö.

1960 lı yılların sonuna doğru kredi kartları kullanılmaya başlandı. Bankalar ve diğer kredi kartı şirketleri de kredi skorlamanın önemini zamanla kavradılar. Her geçen gün kredi kartı için başvuran insanların sayısının artması ekonomik ve insan gücü bakımından kararların hızlı bir şekilde alınmasını yani otomizasyonunu gerekli kıldı. Bilgisayar teknolojisinin giderek güçlenmesi de kredi skorlamanın gelişmesine olumlu yönde katkı sağladı. Kredi skorlamanın yaygın hale gelmesiyle temerrüt oranının %50 den fazla azaldığı görüldü(Myers ve Forgy 1963). Böylece kredi skorlamanın iyi bir tahmin metodu olduğu kanıtlanmış oldu.[28]

1980 li yıllarda kredi kartlarında kredi skorlamanın başarılı olmasıyla bankalar ev, araç ve tüketici kredilerindede kredi skorlamayı kullanabileceklerini fark ettiler. 1990 larda doğrudan pazarlama sektöründeki büyüme skor kartların reklam kampanyalarına verilen tepkinin ölçümlenmesi için kullanılmasına katkı sağladı. Bu aslında skor kartların ilk kullanım amaçlarından biriydi. Sears 1950 li yıllarda hangi müşterilere katalog gönderip göndermeyeceğini skor kartlar yardımıyla belirliyordu(Lewis 1992). Bilgisayar teknolojisinin gelişmesi skor kart modellemesinde farklı tekniklerin kullanılmasına imkan sağladı. 1980 li yıllarda bugün en çok kullanılan iki temel teknik olan lojistik regresyon ve doğrusal programlama tanıtıldı. Daha sonraları yapay zeka ve sinir ağları gibi farklı teknikler de uygulandı.

Günümüzde ise kredi skorlama ile amaç müşterinin temerrüde düşme olasılığının minimize edilmesinin yanı sıra bir müşteriden elde edilebilecek karın maksimize edilmesini de amaçlıyor. Bunun yanı sıra "Müşteri hangi ürünü kullanacak?", "Müşteri ne zaman ter edecek?", "Müşterinin kredi geri ödemelerinde gecikmeler olursa temerrüt değişik yöntemlerle nasıl önlenebilir?", "Bir başvuru hileli olabilir mi?" gibi soruların yanıtları da kredi skorlama ile cevaplanmaktadır.

Çizelge 2.1. Kredi ve kredi Skorlamanın tarihçesi

Tarih	**Olay**
M.Ö. 2000	Asya, Babil ve Mısır'da kredi ilk kez kullanıldı.
1100 ler	Avrupa'da ilk emanetçiler hayır kurumları tarafından açıldı,

[28] Anderson R.

	1350 li yıllara gelince bu dükkanlar ticari amaçlar için kullanılmaya başlandı.
1536	Faiz ödenmesi Protestan kilisesi tarafından kabul edildi.
1730	Kredi ile ilgili ilk reklam yapıldı. Londra'da Christopher Thornton mobilyaların haftalık ödemelerle alınabileceğini söyledi.
1780 ler	İngiltere'de ilk çekler kullanılmaya başlandı.
1803	İlk müşteri raporları Londra'daki Mutual Communications Society tarafından hazırlandı.
1841	İlk Amerikan kredi raporlama ajansı, Mercantile Agency açıldı.
1851	John M. Bradstreet ilk kez ticari kredi verenler için kredi notunu kullandı.
1856	Singer Dikiş Makineleri ilk kez tüketici kredisini önerdi.
1869	İlk Amerikan tüketici bürosu Retailers Commercial Agency(RCA) Brooklyn'de kuruldu.
1906	Amerika'da Ulusal Bireysel Kredi Ajansları Birliği kuruldu.
1909	John M. Moody ilk kez ticari bonolar için kredi derecelendirme notlarını açıkladı.
1927	Almanya'da ilk kredi bürosu kuruldu.
1934	Almanya'da ilk kredi başvuru kurumu PCR kuruldu.
1936	R.A. Fisher istatistiksel teknikleri kullanarak farklı iris türlerini birbirinden ayırdı.
1941	David Durand bir rapor yazarak, istatistiğin kredi kararlarında kullanılabileceğini savundu.
1947	Henry Wells Spiegel Inc. Firmasında kredi skorlamayı kullandı.
1950	Diners Club ve American Express ilk ödeme kartlarını tanıttı.
1950 ler	Sears eğilim skor kartlarını katolog siparişleri için kullandı.
1956	Fair-Isaac danışmanlık firması California'da kuruldu.
1958	Amerikan yatırımcıları tarafından ilk kez başvuru skorlaması kullanıldı.

1960 lar	Kredi kart şirketlerinde kredi skorlamanın kullanımı yaygınlaştı.
1966	Credit Data Corp. ilk otomatik kararlar alan kredi bürosu oldu.
1975	Fair-Isaac ilk kez davranış skor kartını Wells Fargo'nun sistemine entegre etti.
1978	Güney Afrika'da Stannic ilk kez araç finansmanı için skor kart kullandı.
1982	CCN, ilk kez kendi tüketici kredi bürosu için, CAIS(Credit Account Information Sharing) yani kredi bilgilerinin paylaşılmasını önerdi.
1984	Fair-Isaac büro skorlarını ilk kez kredi risk izleme için kullandı.
1987	MDS, iflas tahmini yapan büro skorlarını geliştirdi.
1995	Freddy ipotek emanetçileri Mac and Fannie Mae kredi skorlamayı kullandı.
2000	Moody's KMV finansal oran skorlamasını(FRS) tanıttı.
2000 ler	Basel II pek çok banka tarafından uygulanmaya başlandı.

3.4. Kredi Skorlamanın Yöntemi

3.4.1. Skorkartlar

Pek çok insan skor kart deyince bir yarışmada katılımcıların puanlarının yazılı olduğu bir kağıt parçası veya pano anlar. Kredi skorlamada da benzer bir durum mevcuttur ancak sonuçları ve uygulama yöntemi farklıdır. Kredi skorlama durumların gelecekte iyi ya da kötü olma olasılıklarına göre derecelendirmek için tahminleyici modeller kullanılmasıdır. Mantık basittir, iyi müşteri sizin dört gözle beklediğiniz müşteri-düşük risk, kötü müşteri ise bir daha gelmesini istemediğiniz müşteri-yüksek risktir. Skoring modellerinin geliştirilmesinde genellikle lojistik regresyon kullanılır.

Çizelge 2.2. Skor kart

Kriter	Skor
Yaş < 29 ise	-20
Yaş ≥ 45 ise	+50
Ev Telefonu yok ise	-30
Eski müşteri ise	+20

Bir skor kartın son hali yukarıdaki gibi bir tabloda alt alta birkaç koşul sıralanması olarak gösterilebilir. Aşağıdaki tabloda ise biraz daha gelişmiş bir skor kart modeli görülebilir. Bu tabloda ise karakteristiklerin kendi içlerindeki özelliklerine göre kıyaslanarak gösterildiğini görüyoruz. Tabloda örnek bir müşteri skorlanarak nihai skoru hesaplanmıştır. Bu skor kart güney Afrikalı bir motorlu araç firması olan Stannic'in 1978 yılından 1983 yılına kadar kullanmış olduğu bir skor karttır.

Çizelge 2.3. Skor kart

Karakteristik	**Cevaplar**					**Alınan Puan**
Adresinde bulunma süresi	< 3 yıl 30	3-6 yıl 36	> 6 yıl 38		Boş 35	38
İşinde çalışma süresi	< 2 yıl 30	2-8 yıl 39	9-20 yıl 43	> 20 yıl 64	Boş	43
Ev telefonu	Kayıtlı 47				Kayıtlı değil 30	30
Ev mülkiyet durumu	Kendinin 41	Kira 30	Ailesinin 39		Diğer 36	41
Kredi büro sonucu	Temiz 20	1 -16	2 -30	3 -54		20
Geçmiş kayıtlar	Kayıt yok 3	Eski müşteri 36			Temerrüd / Red -50	3
					Toplam skor	175

3.4.2. Skorlama Modeli Geliştirilmesinde Kullanılan Parametrik Yöntemler

İlk olarak geleneksel skor kartlar diskriminant analizi(DA), doğrusal olasılık modellemesi(LPM), ve lojistik regresyon gibi parameterik yöntemler kullanılarak geliştirildi. Bu yöntemler güçlü olmalarıyla birlikte sağlanması gereken pek çok varsayımları vardır ve çoğu zaman bu varsayımlar gerçekleşmez. Bütün uygulamalarda DA ve LPM birlikte kullanıldığı için ikisini tek bir yöntem olarak görmek mümkündür. Doğrusal olasılık modellemesi diskiriminant analizinin bir parçasıdır. LPM kolay ve hızlıdır, yıllarca çözüm olarak tercih edilmiş bir yöntemdir. LPM yönteminin denklemi aşağıdaki denklem (2.1) ile gösterilir.

$$y \cong G/(G+B) \tag{2.1}$$

Denklemdeki G ve B ler örneklemdeki iyi ve kötü sayılarıdır. LPM ikili değişkenlerin modellemesine uygun olmadığı gerekçesi ile pek çok eleştiri almıştır. Şimdilerde de yaygın olarak kullanılmaktadır çünkü geçmişte yapılan eleştiriler modelin nasıl uygulandığına yönelmiştir.

Diğer yöntemlere göre daha yavaş olan lojistik regresyon ise ikili değişkenlerin modellenmesinde çok daha başarılı olup, en yaygın olarak kullanılan tekniktir.

3.4.2.1. Lojistik Regresyon

Birinci bölümde detaylı olarak anlatılmıştır.

3.4.2.2. Doğrusal Olasılık Modeli

Birinci bölümde detaylı olarak anlatılmıştır.

3.2.1.3.Diskriminant Analizi(DA)

Diskriminant analizi iki grup arasındaki ayrımı yapan ve kategorik bir bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi doğrusal bir fonksiyon olarak ortaya koyan istatistiksel bir

tekniktir. Bu fonksiyon yardımıyla mevcut birimlerin ve yeni gelen birimlerin hangi grupta olacağına karar verilmektedir. Temel mantık grup içindeki dağılımın minimum, gruplar arası dağılımın da maksimum olmasını sağlamaktır.[29] Ayırıcı fonksiyon sayısı her zaman grup sayısından bir tane az olacak şekilde sonuçlanır. Örneğin 2 gruplu bir çalışmada bir tane diskriminant fonksiyonu olacaktır. Diskriminant analizinin skor kart modellemesinde kullanıldığı durumlar daha basit modellerin söz konusu olduğu durumlardır. Eğer hedef değişken yalnızca iki kategorili ise yani belirsizler veya diğer gruplar dahil edilmeyip yalnızca iyi ve kötüler dikkate alındıysa kullanılabilir.

Diskriminant analizinin temel varsayımları aşağıdaki gibi sıralanabilir.

- Bağımlı değişken veri seti çok değişkenli normal dağılıma uygundur.
- Grupların varyans kovaryans matrisleri homojendir.
- Bağımsız değişkenlerin kendi aralarında anlamlı bir korelasyon bulunmamaktadır.
- Değişkenler arasında çoklu doğrusal bağlantı yoktur.
- Bağımsız değişken veri seti grupların birbirinden ayrılmasında rol oynamayacak gereksiz değişkenler içermemelidir.

Diskriminant analizinin etkin olarak kullanılabilmesi için yukarıda listelenen varsayımların yerine getirilme zorunluluğu vardır. Bu nedenle uygulaması esnek olmayan ve skor kart modellemesinde çok tercih edilmeyen bir tekniktir.

3.4.3. Skorlama Modeli Geliştirilmesinde Kullanılan Parametrik Olmayan Yöntemler

Parametrik yöntemlerde pek varsayımın sağlanması gerektiği için pek çok insan onların yerine varsayım gerektirmeyen non-parametrik yöntemleri kullanmayı denemişlerdir. Bu yöntemler yapay sinir ağları, K-yakın komşu algoritması, doğrusal programlama, karar ağaçları olarak sayılabilir. Şeffaf olmamaları en büyük eleştiri kaynağıdır.

3.4.3.1. Karar Ağaçları

[29] Özdamar K.

Muhtemelen pek çoğumuz karar ağacı şekline aşinayızdır. Karar ağaçları grafiksel bir yöntem olup olayların olası seçeneklerini olasılıklarıyla sunar. Karar vericiye tüm seçenekleri olasılıklarıyla görüp karar verme imkânı sağlar. Ayrıca kümeleme tahmin etme problemlerinde verinin görselleştirilmesinde de kullanılır. Daha çok kuralların karar verici tarafından belirlendiği durumlarda kullanılır. Skor kart modellemesinde kullanılmak için skor kart kurallarının kişilerin iş tecrübeleri ile kuralların oluşturulması gerekir ki bu da çok elverişli bir yöntem değildir.

3.4.3.2. Yapay Sinir Ağları

İnsanoğlu yüzyıllardır bulunduğu durumdan daha iyiye ulaşabilmek için çaba sarfetmiştir, ancak son birkaç yüzyılda insanoğlunun bu çabası işlerini makinelere yaptırma yönünde gelişmiştir. Son birkaç yılda ise otomatik karar verme süreçleri üzerinde yoğunlaşılmıştır. Temel amaç hayatı kolaylaştırmak iken pek çok insan bilgisayarların artık insan gibi düşünebiliyor olacağından korkuyor. Aslında bu bir bilim kurgu konusu iken günümüzde soyut zeka ile bir parça ilerleme sağlanmıştır. Karar vermede tahminsel yöntemlerin kullanılması soyut zekanın bir parçası olarak görülse de soyut zekanın özünde başka bir teknik yatmaktadır. Bu teknik "Yapay Sinir Ağları" (YSA) dır.

Yapay sinir ağları çevreden öğrenen, çevreye uyum sağlayan ve etkilere tepki veren bilgi işlemleri ağıdır. Bu yöntem insanın kendi kendine öğrenme kabiliyetini taklit etmektedir. İstatistiksel teknikler gibi formüller kullanmaksızın tekrarlı örnekler yardımıyla modelleme yapar. Karar ağacına benzer bir sonuç çıkar. Karar kuralları da aynı şekilde karar ağaçlarına benzer ancak buradaki kurallar çok daha karmaşıktır.

Pek çok yapay sinir ağı tekniği mevcuttur. Skor kart modellemesinde ise genellikle çok katmanlı yapay sinir ağları(multilayer perceptron-MLP) tekniği kullanılır. Bu teknik eğrisel ilişkilerde ve değişkenler arasında etkileşim olduğu durumlarda avantajlıdır. Radial Basis Function, Self-Organising Maps ve Kohonen ağları gibi başka teknikler de mevcuttur.

Özellikle firmaların karar sebeplerini müşteriye tavsiye etmek zorunda olduğu, karar mantığının çok iyi anlaşıldığı kurumlarda kullanılabilir. Bu nedenle yapay sinir ağları skor kart

modellemesinde nadiren kullanılır. Çok az veri olduğu durumlarda tercih edilebilen bir tekniktir. Fraud skorlamasında yaygın olarak kullanılmaktadır.

3.4.3.3. k-En yakın komşu Tekniği

İlk kez Fiz ve Hudges tarafından geliştirilen k-en yakın komşu tekniği parametrik olmayan istatistiksel bir yöntemdir. Kredi skorlama tarihinde ilk olarak 1970 ylında Chatterjee ve Barcun bu tekniği kullanarak skor kart geliştirmiştir. Daha sonra 1996 yılında Henley ve Hand da bu yöntemi kullanmışlardır. Bu yöntemin temelindeki düşünce başvurular arasında belli bir uzaklık saptamaktır. Başka bir deyişle bir başvurunun yakın olduğu diğer iki başvuru tespit edilir. Skorlama modeli oluşturulurken geçmiş veri kullanılır, yeni bir müşteri geldiğinde ise k tane en yakın komşuların iyi kötü oranlarına göre iyi ya da kötü sınıfına alınır. Yani yeni gelen başvuru geçmiş veriye göre incelenerek, yeni gelen başvurunun en yakın komşusu bulunur.

3.4.3.4. Doğrusal Programlama

Doğrusal programlama yöneylem araştırmaları alanından gelen bir yöntemdir. Genelde yöneylem araştırmaları araçlarının amacı karar vericilere kaynakları en uygun biçimde kullanarak karar vermelerinde yardımcı olmaktır. Bu konudaki ilk araştırmalar 1930 lu yıllarda taşıma problemleri(Kantorovich), oyunlar teorisi(Morgenstern ve von Neumann) çalışmaları ile başlamıştır. Kantorovich and von Neumann doğrusal programlamanın öncüsü olarak görülseler de onların öncesinde 1947 yılında George Dantzig US Air Foce'da çalışırken simplex yöntemiyle lojistik kapasitelerini arttırmak için kullanmıştır. 1950 li yıllarda Rand Corporation ve US National Bureau of Standards firmaları tarafından kullanılmıştır. Bilgisayarlar ile birlikte doğrusal programlama türevleri metotlar daha da gelişmişleridr. Kredi skorlamada aşağıda yer alan denklemdeki β ların çözülmesi ile kullanılır.

$\sum e_i^2$ değeri minimum olmak üzere

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \qquad (2.2)$$

$$\beta_1 < \beta_3 , \quad \beta_2 > 0$$

Diğer bir deyişle doğrusal programlama esasında regresyon denklemini hata terimlerini minimize ederek çözmeyi amaçlıyor. Kredi skorlama açısından doğrusal programlamanın en büyük avantajı

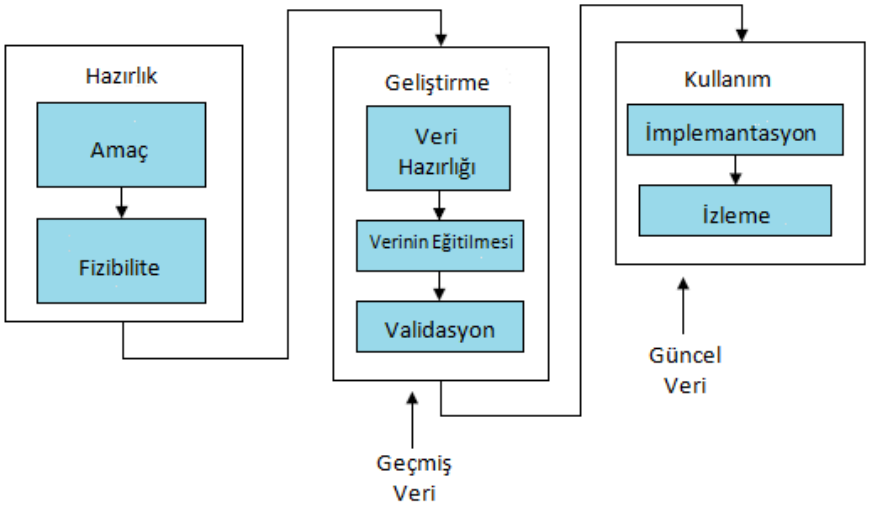
skor kartı geliştiren kişinin sonuç skorları üstünde kontrol etme yetkisi vardır. Örneğin yaş > 62 olanlar, yaşa uygulanan diğer skorların maksimum değerine eşit olmalıdır. Kredi skorlamada çok nadir olarak kullanılan bir tekniktir. Sonuçları istatistiksel olarak değerlendirilemez böylece kurulan modelin başarısı hakkında her hangi bir ölçüm yapmak mümkün olmaz.

3.4.4. Parametrik ve Parametrik Olmayan Yöntemlerin Kıyaslanması

Çeşitli teknikler için pek çok istatistiksel tartışma mevcuttur, hala da tartışma bitmiş değildir. Her tekniğin derecelendirme yeteneği yani gücü ile ilgili karşılaştırmalar henüz bitmemiştir. LPM bilgisayarlar büyük ve yavaş olduğu zamanlarda hızlı olmasından dolayı en çok tercih edilen yöntemdi. Bugünlerde ise lojistik regresyon en yaygın kullanılan tekniktir. İkili değişkenler için uygun modelleme yapılabilmesi ve skorların kolaylıkla olasılık tahminlerine dönüştürülebilmesi en büyük iki avantajı olarak sayılabilir. Lojistik regresyon elbette her durum için uygun olmayabilir. LPM nin bazı koşullarda lojistik regresyona göre üstün yanları vardır.

3.5. Skor kart Geliştirme Aşamaları

3.5.1. Proje Hazırlığı



Şekil 2.2. Kredi skorlama projesi aşamaları[30]

Temel olarak bir skor kart projesini basit haliyle yukarıdaki şemadaki gibi özetleyebiliriz.

[30] Anderson R.

3.5.1.1. Amaç Tanımlama

Hazırlık aşaması olan ilk aşamada temel amacın ne olduğuna karar verilir. Buna karar verirken

- Müşteri ihtiyaçları nelerdir
- Piyasadaki rakipler bu konuda neler yapıyorlar, skor kartları nasıl ve ne amaçla kullanıyorlar
- Kanunlar neyi gerektiriyor

gibi konuların gözden geçirilmesi faydalı olabilmektedir.

3.5.1.2. Fizibilite Çalışması

Fizibilite çalışmasında temel olarak üç soruya yanıt aranır.

- Modelleme yapılabilecek yeterli veri mevcut mudur?
- Projeyi yürütebilmek için yeterli para ve insan kaynağı mevcut mudur?
- Projeyi destekleyecek yeterli teknoloji gücü mevcut mudur?

Proje ekibinin belirlenmesi ve görevle dağılımın yapılması da hazırlık aşamasında yapılması gereken bir adımdır.

3.5.2. Veri Hazırlama

Analiz ve karar verme aşamasına geçmeden evvel veri üzerinde sıkı çalışmak gerekmektedir. Modellerin başarısı çok büyük ölçüde modellemede kullanılan veriye bağlıdır. Eğer veri standart değilse modelin başarısını düşürecektir. Veri kalitesindeki temel sorunlar kayıp veriler, yanıltmak amacıyla verilmiş yanlış olağan dışı bilgi, kullanıcı kaynaklı yanlış kaydedilen bilgiler olarak sıralanabilir.

Skor kart projesinde modelleme aşamasına gelmeden önce veri hazırlığının ayrıntılı bir şekilde yapılması gerekir. Yapılacak istatistiksel modellerin anlamlılığı bakımından verinin belirli aşamalardan geçtikten sonra modellenmesi başarı sağlayacaktır. Bu aşamada aşağıda listelenen soruların cevapları aranmalıdır.

- Proje hangi durumları kapsayacak?
- İyi müşteri kötü müşteri tanımı nasıl yapılacak?
- Örneklem periyodu olarak hangi tarih aralığı seçilecek?
- Örneklem büyüklüğü ne olacak?
- Karakteristikler ne olacak?
- Veri nasıl bir araya getirilecek?

3.5.2.1. Projenin Kapsamının Belirlenmesi

Veri seti oluşturulmadan önce belirlenen hedefe göre skor kart modellemesinde kullanılacak veride hangi durumların veya hangi tarih aralıklarının yer alacağı kararlaştırılmalıdır. Burada proje ekibinin iş bilgisine dayanarak karar verilmelidir. Örneğin olağanüstü durumların yaşandığı dönemler, ekonomik kriz dönemleri, ekonomik dalgalanmaların olduğu dönemler veya verinin temiz olmadığı düşünülen kısımlar veri kapsamına dâhil edilmeyebilir. VIP hesapları ya da kullanımdan kaldırılan ürünler gibi sıra dışı durumlar varsa bunların örneklem dışı tutulması modelin gücünü olumlu yönde etkileyecektir. Verilerin kirli olması, kullanıcı hatası nedeniyle girilen yanlış, mantıksız veriler modelin tahmin gücünü olumsuz yönde etkileyecektir. Bu nedenle modelleme aşamasından önce veride mantıksal kontrollerin yapılması işi kolaylaştıracaktır.

Aynı şekilde kurum içindeki yeni uygulamaların başladığı dönemler, kampanya dönemleri de dikkatle gözden geçirilmelidir. Çünkü bu tip dönemlerde oluşan veriler normal durumu yansıtmayacağı için model sonuçlarının olması gerektiğinden farklı çıkmasına neden olabilir.

3.5.2.2. Verinin Bir Araya Getirilmesi

Skor kart modellemesinde kullanılacak veriler sistem üzerinde farklı tablolarda veya faklı veri tabanlarında tutuluyor olabilir. Verilerin doğru ve tutarlı bir şekilde toplanması gerekmektedir. Son olarak verinin nasıl birleştirileceğine proje ekibi tarafından karar verilir.

3.5.2.3. **İyi Müşteri Kötü Müşteri Tanımının Yapılması**

Modelleme aşamasında geçmişte kredi kullanan müşterilerin verileri kullanılarak gelecekte başvuracak müşterilerin iyi mi yoksa kötü mü performans sergileyeceklerini tahmin edebilen en iyi modeli bulmak amaçtır. Bir başka deyişle skor kart modellemesindeki temel amaç iyi veya kötüyü tahmin etmektir. Dolayısıyla skor kart modelinde kullanılacak hedef değişken müşterinin iyi veya kötü olma durumudur. Hedef değişken olarak kullanacağımız değişkenin nasıl tanımlanacağı proje ekibinin iş bilgisi, kurum kuralları, kanunlar çerçevesinde belirlenebilir.

İyi kötü tanımlaması genellikle kişilerin ödeme performanslarındaki kaçan ödeme sayılarına göre yapılmaktadır. Örneğin; "Son yirmi dört aylık ödeme performansında üç taksit gecikmesi olan müşteri kötü müşteridir ."veya "Son altı aylık performansında hiç gecikmesi olmayan müşteri iyi müşteridir." gibi tanımlamalar yapılarak bir kurallar bütünü ile iyi kötü tanımı oluşturulur.

İyi kötü tanımlamasında "belirsiz" olarak isimlendirilebilecek üçüncü bir grubun oluşturulması skor kartın gücünü arttıran bir uygulama olacaktır. Belirsizler grubunda ödeme performansı hakkında yeteri kadar bilgi sahibi olamadığımız yani ne iyi ne de kötü müşteri diyebildiğimiz kayıtların yer alması uygun olacaktır. Bu tür bir grup iyi müşteriler ve kötü müşterilerin arasında bir grup olacağından ikisinin daha iyi ayrıştırılması noktasında olumlu etki yaratacaktır.

3.5.2.4. Örneklem Periyodunun Seçilmesi

Örneklem periyodunun seçiminde yapılacak bir hata da modelin tahmin gücünü azaltacaktır. Geleceği tahmin etmek amacıyla bir model kuruluyorsa ana kütleyi en iyi temsil eden ve güncel zamana en yakın veriler örneklem olarak seçilmelidir.
Örneklem periyodunun hangi tarih aralıklarında olacağına karar verilirken veride bulunan iyi kötü kayıtların oranlarına bakılır. Etkin bir örneklem periyodu seçilmesi için veri içinde bulunan kayıtların ödeme performanslarının olgunlaşma periyodu seçilmelidir. Ayrıca mevsimsellik etkisi, veri toplama sisteminde olmuş olabilecek değişiklikleri içeren süreler gibi özel durumlar proje ekibinin iş bilgisi de kullanılarak örneklem periyodu seçiminde dikkate alınmalıdır.

3.5.2.5. Örneklem Büyüklüğünün Belirlenmesi

İstatistiksel modeller kullanılarak oluşturulan skor kartlarda örneklem büyüklüğü de modelin başarısını etkileyen bir faktördür. Modellemede kullanılacak örneklem veri setinde az 1500 adet iyi, 1500 adet kötü ve 1500 tane ret başvurunun olması yeterli olabilir. Baktığımızda bu sayılar çok büyük değil ancak bazen kötü sayısı beklendiğinden çok daha az olabilmektedir.

3.5.2.6. Veriye Dâhil Olacak Karakteristiklerin Belirlenmesi

Geçmişte müşteri ile alakalı olarak toplanan tüm veriler değerlendirilerek güvenilirliği ve doluluk oranı yüksek olan, skor kart modelinde yer alması iyi olabilecek değişkenler proje ekibi tarafından seçilir. Ayrıca mevcut verilerden oran, süre gibi yeni karakteristikler de üretilebilir.

3.5.3. Skor Kart Modellemesi

Skor kart modelleme kısmında aşağıda listelenen süreçlerden geçilir.

- Veri dönüştürülmesi
- Değişken Seçimi
- Retlerin Anlamlandırılması ve Performans manipülasyon Yöntemleri
- Segmentasyon
- Modelin Değerlendirilmesi

3.5.3.1. Veri Dönüştürmesi

Modelleme aşamasında ilk adım verinin kullanılabilir bir şekle dönüştürülmesidir. Veri pek çok değişken ile beraber büyük boyutta olabilir ancak kullanılabilir durumda olmadığı sürece pek bir anlam ifade etmez. Değişkenleri dönüştürmek için grup oluşturma, gösterge değişkenlere dönüştürme gibi teknikler kullanılabilir.

3.5.3.2. Değişken Seçimi

Değişkenlerin iyi kötü oranı analizleri ve anlamlılıkları yapılarak hangilerinin model içinde yer alması gerektiğine karar verilir. Bu yapılırken özellikle regresyon analizi kullanılıyorsa bağımsız

değişkenlerin kendi aralarındaki korelâsyonun düşük bağımlı değişken ile aralarındaki korelasyonun yüksek olmasına dikkat edilir.

Tahmin gücü olan değişkenlerin seçilmesi modelin tahmin gücü açısından önemlidir. Bir değişkenin tahmin gücü olup olmadığı χ^2, F istatistiği gibi çeşitli istatistiksel testler yardımıyla tespit edilmelidir.
Pek çok durumda model için kullanılan değişkenler birbirleriyle ilişkili olur. Özellikle de finansal oranlar gibi benzer şekilde hesaplanan değişkenler için korelasyon yüksek olur. Bu da çoklu doğrusal bağlantı olmasına neden olur. Bu durumlarda değişkenler incelenerek bir kısmı modele dahil eldir.

Değişkenlerin sürekli olarak var olmasından da emin olunmalıdır. Yani modellemede kullanılan değişkenlerin özellikle de anlamlı değişkenlerse veri tabanında gelecekte de sürekli olarak olacağı bilinmeli, bunların kontrolü yapılmalıdır. Örnek olarak kullanımdan kaldırılan ürünlere ait bilgiler ya da tam tersi daha yeni kullanılmaya başlanmış ürünlere ait bilgiler modellemeye dahil edilmeyebilir. Bu noktada proje ekibinin iş bilgisi önemlidir.

Kısacası mantıklı, tahmin gücü olan, diğer değişenler ile korelasyonu düşük, müşteri ile ilişkili, herhangi bir yasal engeli bulunmayan değişkenler modelleme için seçilmelidir.

3.5.3.3. Retlerin Anlamlandırılması

Başvuru skor kartı geçmiş veriye dayanarak modellenir. Bu veri kredi kullanan ve belli bir ödeme performansı oluşmuş müşterilerden oluşmaktadır. Fakat reddedilen veya kredi limit açılan ancak kullanmayan bazı müşterilerin istenen performansı oluşmamıştır. Ortalama olarak retlerin performansının genelden daha kötü olacağı düşünülür. Fakat eğer retler kabul olsaydı ödeme performansları iyi mi ya da kötü mü olurdu konusu hiçbir zaman kesin olarak bilinemez. Kabul edilen başvuruların içinden ödeme performansı kötü olan müşterilerin çıktığı gibi reddedilen başvuruların içinden de iyi ödeme performansına sahip müşteriler de olacaktır. Bir örneklem setinde normal durumda %25-30 civarında ret başvuruların olduğu düşünülürse veri setinin bu kadarlık kısmı eğer retler anlamlandırılmazsa hedef değişkenin tahminlenmesinde kullanılmayacaktır. Bu durumun temel olarak iki dezavantajı olduğu söylenebilir. Birincisi, ret

verileri hiç modellemeye dâhil edilmediğinden analiz için veri seçimi yanlı olacaktır. İkincisi, ret başvurular çıkartılınca modelleme için kullanılacak veri sayısı azalacaktır. Kısmi olarak gözlemlenen ret başvuruların da modellemeye dâhil edilmesi modelin tahmin gücünü olumlu yönde etkileyecektir.

Retlerin anlamlandırılması için çeşitli teknikler bulunmaktadır. Bu teknikler aşağıdaki gibi listelenebilir.

- Rassal Belirleme(Random Supplementation) : Veri seçimi tamamen rassaldır. Hand ve Hanley' e göre(1993) bu en ideal yöntemdir. Seçimin rassal olması yanlılık ihtimalini ortadan kaldırmaktadır.
- Birleştirme(Augmentation)
- Ekstrapolasyon
- İki Değişkenli Yöntem

3.5.3.4. Segmentasyon

Segmentasyon denince ilk olarak daha çok pazarlama ile ilgili konular akla geliyor. Kredi skorlamada segmentasyon, farklı ürün gruplarının, farklı müşteri grupları için farklı skor kart geliştirilmesi gerekip gerekmediği noktasında kullanılır. Modelleme için kullanılacak değişkenlerin analizleri yapıldıktan sonra diğer değişkenlerden bağımlı değişkeni açıklama bakımından farklı olduğu gözlemlenebilir. Örneğin ürün tipi müşterinin iyi ya da kötü olma durumu iyi açıklayabilen bir değişken olabilir, ürün tipinin içinde de konut kredileri diğerlerine göre farklı özellikler gösteriyor olabilir. Bu durumda konut kredileri için ayrı bir skor kart geliştirilebilir. Ya da pazarlama için kullanılan müşteri tiplerine göre de skor kart oluşturulabilir. VİP müşteriler için ayrı, banka ile ilişkileri daha seviyeli olan müşteriler için de ayrı skor kartlar modellenebilir.

3.5.4. Modelin Değerlendirilmesi

Yukarıda bahsedilen aşamalardan geçilip uygun yöntem tercih edilerek model kurulduktan sonra modelin başarılı olup olmadığı merak konusudur. İyi bir modelin tahmin gücü yüksek, iyi ve kötü müşterileri birbirinden iyi ayırt edebilir. Modelin başarısını ölçmek için çeşitli teknikler

mevcuttur. Bu yöntemlerin bir kısmı birinci bölümde açıklanmış olup bir kısmı da burada kısaca açıklanacaktır.

Şüphesiz ki kesin doğruyu tahmin edebilecek bir model geliştirmek mümkün değildir. Her zaman gözden kaçırılan, gözlenemeyen, elde olmayan dış faktörler olacaktır. Zamanla veri geliştikçe modelin tahmin etme gücü maksimum düzeye ulaşacaktır.

3.5.4.1. Sınıflama Matrisi

Sınıflama matrisi model kurulduktan sonra belirlenen bir eşik değere göre tahmin edilen iyi kötü frekansları ve mevcut iyi kötü sayıları kullanılarak oluşturulur. Birimlerinin yüzde kaçının doğru veya yanlış tahmin edildiği görülebilir.

Öncelikle altında kalanların reddedileceği bir eşik skor belirlenir. Bu skorun altında puan alan başvurular kötü başvurular, bu skorun üstünde puan alan başvurular ise iyi başvurular olacaktır. Tahmin edilen iyi-kötü frekansları ve mevcut iyi-kötü frekansları kullanılarak bir çapraz tablo oluşturulur. Oluşturulan tablo üzerinde doğru tahmin etme ve yanlış tahmin etme yüzdeleri hesaplanır. Örnek bir tablo aşağıda verilmiştir.

Çizelge 2.4. Sınıflama matrisi

Mevcut	Tahmin Edilen		
	İyi	Kötü	Toplam
İyi	83.275 56%	17.850 12%	100.125 67%
Kötü	16.700 11%	32.175 22%	48.875 33%
Toplam	99.975 67%	50.025 33%	149.000 100%

Bu tabloya göre iyi müşterilerin %56 sı doğru olarak tahmin edilmiştir. İyi müşterilerin %12 si iyi olmasına rağmen kötü olarak tahmin edilmiştir. Kötü müşterilerin %22 si doğru olarak tahmin

edilmiş, %11 i ise yanlış tahmin edilmiştir. Toplamda ise iyi müşteriler %67 oranında doğru tahmin edilmiş olduğunu söyleyebiliriz.

3.5.4.2. WOE(Weight of Evidence)

$$W_i = \ln\left(\frac{N_i}{P_i}\right) - \ln\left(\frac{\sum N}{\sum P}\right) \quad (2.3)$$

Yukarıdaki denklem (2.3) ile hesaplanır. P = pozitif oluş/kötü, N = P'nin tersi yani oluşmama/iyi. Formülün ilk kısmı ilgili grup için odds oranını gösterir. İkinci kısım ise örneklem ile ana kütle arasındaki sabit oranı gösterir.

3.5.4.3. Bilgi İstatistiği, F testi

Bilgi istatistiği iki dağılım arasındaki farkı ölçer. Formülü denklem (2.4) ile verilmiştir.

$$F = \sum_{i=1}^{n}\left[\left(\frac{N_i}{\sum N} - \frac{P_i}{\sum P}\right).WoE_i\right] \quad (2.4)$$

N: negatif olaylar/iyiler, P: pozitif olaylar/kötüler

F değeri her zaman pozitif değerler alır. F değeri 0.10 dan küçük olan değişkenler zayıf olarak değerlendirilebilir. F değeri 0.30 dan büyük olan değişkenler ise model için başarılı bir şekilde yer alabilir. F değeri düşük olan zayıf değişkenler diğer değişkenlerle kombine edildiğinde modelin tahmin gücünü arttıran bir değişken olabilir, bunu da unutmamak gerekir[31].

3.5.4.4. Kolmogorov Smirnov

Kolmogorov Simirnov(KS) istatistiği deneysel kümülatif dağılım fonksiyonu analizi üzerine geliştirilmiştir. Uyum iyiliğinin ölçülmesi için parametrik olmayan iyi bir ölçüdür. Bu yöntemde bir eğri kullanılarak uyum iyiliği ölçülür.

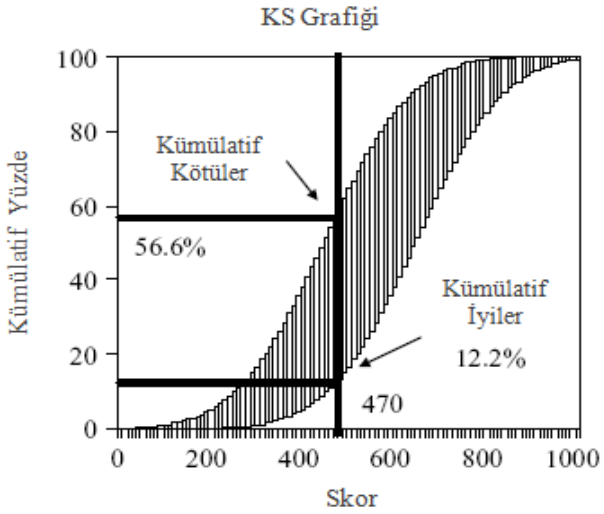
---

[31] Anderson

Mays'e göre KS Amerika'da derecelendirme modellerinin tahmin gücünü ölçmek için en çok kullanılan istatistiktir. KS eğrisi skor kart verisini görselleştirerek skor kartın tahmin gücünü ölçer. Grafik deneysel kümülatif dağılım fonksiyonu yüzdeliklerini iyi ve kötüleri skorlara göre gösterir. Aşağıda sol taraftaki grafikte kötülerin %56.6 sı 470 skorunun altında yer almaktadır. Buna karşılık iyilerin yalnızca %12.2 si 470 skorunun altında kalmaktadır. KS istatistiği iki eğrinin farkının mutlak değerinin maksimumuna eşittir.

$0 < D_{KS} < 1$

$$D_{KS} = maks\left\{ \left| cpY - cpx \right| \right\} \tag{2.5}$$



Şekil 2.3. KS istatistiğinin grafiksel gösterimi[32]

KS anlaşılması çok basit bir istatistiktir. Yukarıda grafikte gösterilen örnek için 470 skorunda $D_{KS} = maks\left\{ \left| cpY - cpx \right| \right\} = 44{,}4$ olarak hesaplanır.

KS'nin temel kullanım alanı tahmin gücünü ölçmek ve iki dağılımın birbirlerinden farklı olup olmadığını test etmektir. KS_{kritik} değerleri kıyaslanarak hipotez testleri yapılır. Bulunan değer KS_{kritik}'den küçükse o zaman dağılımların farklı olduğu söylenebilir. $D_{KSkritik}$, $c/n^{1/2}$ formülü ile hesaplanır. c, anlamlılık seviyesine ve dağılın cinsine göre farklılık gösterir. n, örneklem

[32] Anderson

büyüklüğüdür. Çoğu durumda dağılımın normal olduğu ve anlamlılık seviyesinin 0.05 olduğu kabul edilir. Bu durumda c=1.36 olarak kullanılır.

3.5.4.5. ROC

İkinci bölümde bu konu detaylı olarak anlatılmıştır.

3.5.4.6. GINI

İkinci bölümde bu konu detaylı olarak anlatılmıştır.

3.5.4.7. Ki-Kare Testleri

Birinci bölümde anlatılan Hosmer Lemeshow.

3.5.5. Skor Kartın Kullanıma Alınması

Skor kart modelle aşaması sonuçlanıp uygun modele karar verildikten sonra modelin gerçek ortamda nasıl çalıştırılacağı söz konusu olur. Bu aşama skor kart projesinin son aşamasıdır diyebiliriz. Bu aşamanın tamamlanabilmesi için de aşağıda listelenen adımların izlenmesi gerekir.

- Eşik puanın belirlenmesi
- İzlenecek stratejinin belirlenmesi
- Sisteme entegrasyon ve test
- İzleme
- Validasyon

3.5.5.1. Kesim Puanın ve İzlenecek stratejinin Belirlenmesi

Skor kart modelleme aşaması tamamlandıktan sonra el edilen skor dağılımları incelenerek bir kesim değer(cut-off) belirlenmesi gerekir. Kesim puan seçilirken belli stratejiler izlenebilir. Mevcut portföyün iyi kötü oranları, red oranı veya kabul oranı göz önünde bulundurularak bir değer saptanabilir. Bulunan değerin yüksek olması aslında iyi olabilecek müşteri kaybının artmasına, düşük olması da skor kartın amacına ulaşmamasına sebep olur. Bu nedenle skorların dağılımı dikkatle analiz edilmelidir.

Kesim puanın altında skor alan başvurular reddedilip, sütünde puan alanlar da kabul edilebilir. Ancak genellikle bu kullanılan bir yöntem değildir. Böyle bir uygulama olabilmesi için skor kart modelinin çok güçlü tahmin gücüne sahip ve güvenilir olması gerekir. Buna karşılık bir başka alternatif olarak iki adet kesim puan belirlenebilir. Düşük olan değerin altında kalanlar reddedilirken, yüksek değerin üstünde puan alanlar da kabul edilebilir. Büyük değer ile küçük değer arasında puan alan başvurular da manüel olarak değerlendirilebilir. Bu daha uygulanabilir bir yöntemdir. Nasıl bir yöntem izleneceği proje ekibinin iş bilgisi, tecrübeleri ve skor kartın kullanım amacına bağlıdır.

3.5.5.2. Sisteme Entegrasyon ve Test

Skor kart kullanıma alınmadan önce kurum içindeki süreçlerin içinde nasıl ve nerede çalışılacağına karar verilmelidir. Kullanım alanına göre skor kartın entegre olacağı sistem farklılık gösterebilir. Örneğin başvuru skor kartı kredi başvuru sürecine entegre olurken, davranış skor kartı kredi izleme sürecine entegre edilmek istenebilir. Başvuru skor kartlarında genel uygulama kredi başvuru sürecine entegre edilmesi ve skorların eş zamanlı olarak hesaplanmasıdır.

Sisteme entegrasyon tamamlandıktan sonra sistemin doğru çalışıp çalışmadığı test edilir. Parametrelerin aldığı puanlar kontrol edilip belirlenen stratejinin düzgün çalışıp çalışılmadığına bakılır. Eğer testlerde bir sorun çıkmazsa skor kart gerçek ortamda çalıştırılmaya başlanabilir.

3.5.5.3. Validasyon

Validasyon aşaması skor kart gerçek ortama atıldıktan 6 ay ya da bir yıl gibi belli bir süre sonra yapılır. Başvuru skor kartları için çalışma tarihinden itibaren 6 ay sonra validasyon çalışmasına başlanabilir. Bu aşama için yeniden bir veri hazırlanarak benzer skor kart proje aşamalarından geçilir. Mevcut modelin tahmin gücündeki değişime bakılır. Modelin tahmin gücünde olumsuz yönde bir değişim mevcutsa, bunun nedenleri araştırılır ve model yeniden düzenlenir.

4. UYGULAMA

Uygulamaya ait veriler Türkiye'de faaliyet gösteren, büyümekte olan bir bankaya aittir. Veriler ilgili bankanın bireysel kredi başvurusunda bulunmuş müşterilerin verilerinden derlenmiştir.

4.1. Proje Hazırlığı

4.1.1. Amaç Tanımlama

Bireysel krediler portföyünde kredi başvurusunda bulunan tüm müşteriler manüel olarak klasik değerleme yöntemlerine göre değerlendirilmektedir. Rakip bankaların kredi kartlarının yanı sıra bireysel kredilerde de skor kart modelleri kullanarak başvuruların performanslarını ileriye dönük olarak tahminleyerek otomatik ret veya otomatik kabul sistemlerini devreye almak suretiyle iş yüklerini azalttıkları gözlenmiştir. Bankacılık Denetleme ve Düzenleme Kurumu (BDDK) bireysel kredi başvurularında skor kart kullanımını denetlemektedir. BASEL II'ye geçilmesi durumunda skor kart kullanılması zorunlu hale gelecektir. Bu nedenle bireysel kredi başvurularında skor kart kullanılması zaruri hale gelmiştir.

Projenin amacı bireysel krediler portföyü için başvuru skor kartı modellemek, geliştirilen modeli başvuru sistemine entegre etmektir. Amacın gerçekleştirilmesiyle iş yükünü azaltmak, başvuruda bulunan müşterilerin risklerini ölçümlemek, yasal zorunlulukların yerine getirilmesi hedeflenmiştir.

4.1.2. Fizibilite Çalışması

Proje başlangıç aşamasında fizibilite çalışması yapılmıştır. Modelleme yapılabilecek yeteri veri mevcuttur. Projeyi yürütebilmek için insan kaynağı toplanmış, gerekli eğitim verilmiştir. Verileri analiz edebilmek için gerekli teknoloji desteği sağlanmıştır. Bu çalışmaya konu olan uygulama SPSS.15 ile yapılmıştır. Proje ekibinin belirlenmiş ve görev dağılımının yapıldığı kabul edilmiştir.

4.2. Veri Hazırlama

4.2.1. Projenin Kapsamının Belirlenmesi

Proje yalnızca bireysel kredi başvurusunda bulunmuş müşterilerin başvuru bilgilerini kapsayacaktır. 2003 yılından sonra yapılan sistem değişikliği nedeniyle 2003 yılı öncesi ve sonrası arasında veri bakımından farklılıklar mevcut olduğundan 2003 yılı öncesi başvurular dikkate alınmayacaktır. Bireysel kredi başvuru formu incelenerek ilk bakışta bütün karakteristikler incelenmiş ve alanların veri kalitesine bakılmıştır.

Veride yer alan karakteristikler çizelge (3.1)'de verilmiştir.

Çizelge 3.1. Değişkenler

Hesap Bilgileri	Hesap No
	Başvuru Numarası
Kimlik Bilgileri	Kimlik Belge Türü
	T.C. Kimlik No
	Cinsiyet
	Uyruk
	Yaş
	Medeni Hal
	Askerlik Durumu
Kişisel Bilgiler	Eğitim Durumu
	Çocuk Sayısı
	Eş Çalışma Durumu
	Sosyal Güvenlik Kurumu
	Sosyal Güvenlik Numarası
	E-mail Adres Durumu
	Ev Sahipliği
	Ev Telefonu
	İş Telefonu
	Cep Telefonu
	Banka Bilgisi
	Ev İkamet Süresi
İş ve Gelir	Çalışma Şekli
	Meslek

Bilgileri	İş Yeri Faaliyet Konusu
	İş Yerindeki Ünvanı
	Sektör
	İş Yeri Çalışma Süresi
	Maaş Geliri
	Serbest Meslek Geliri
	Menkul Değer Geliri
	Gayrimenkul Değer Geliri
	Aylık Net Gelir
	Araba Durumu
Kredi Bilgileri	Finansman Türü
	Ürün Peşinat Oranı
	Ürün Tipi
	Araç Durumu
	Vade
	Araç Yılı
	Araç Marka
	Skor Puan
	Kredi Kayıt Bürosu Sorgu Sonucu

Veri kalitesinin düşük olması nedeniyle modellemeye dahil edilmeyecek değişkenler aşağıda listelenmiştir.

- **İş Yeri Faaliyet Konusu :** Bu alan başvuru sisteminde serbest metin olarak kaydedilmektedir. Yani kullanıcıya çoktan seçmeli seçenekler sunulmayıp faaliyet konusunun tanımının yapılması istenmiştir. Bu nedenle bu alanda girilen bilgiler içinde sistematik bir ilişki bulunamamıştır. Yani veriler çok çeşitlilik gösterdiği için tasnif edilememiştir. Bu nedenle analize dahil edilmemiştir.
- **İş Yerindeki Ünvanı :** Bu alan da iş yeri faaliyet konusu isimli alanla benzer özellikleri taşımaktadır. Analize uygun olmadığı için analiz dışında tutulmuştur.

Veriye bağımsız değişken olarak kullanılması amacıyla müşterinin kredi ödeme performansını gösteren yeni bir alan hesaplatılmıştır. Bu alanda müşterinin her ay geciktirdiği taksit sayısı her rakam bir ayı gösterecek şekilde 0,1…,9 ile kodlanmıştır.

4.2.2. Verinin Bir Araya Getirilmesi

Veride yer alması istenen başvuru ve performans bilgileri veri tabanında ayrı tablolarda tutulmaktadır. Başvuru ile performans bilgisini birebir eşleştirecek bir anahtar alan belirlenmiştir. Anahtar alan başvuru numarası olarak seçilmiştir, bu alan başvuruya özel tekrarlı olmayan bir değerdir. Başvuru ve performans bilgileri belirlenen anahtar alana göre birleştirilecektir.

4.2.3. İyi-Kötü Müşteri Tanımının Yapılması

Proje ekibi iyi-kötü tanımını kurum politikaları ve iş tecrübelerine dayanarak çeşitli alternatiflerle yapmıştır. Birkaç tane alternatif iyi kötü oluşturulmasının nedeni her bir tanım sonucu ortaya çıkan dağılımı gözlemlemek ve en uygun olanını seçmektir.

Çizelge 3.2. İlk iyi-kötü tanımı

Güncel Ay	Son 6 Ay	İ/K
>=3		Kötü
0	>=3	Belirsiz
1	>=3	Belirsiz
2	>=3	Kötü
0	<=2	İyi
1	<=2	İyi
2	=2	Belirsiz
2	<=1	Belirsiz

Çizelge 3.3. İkinci iyi-kötü tanımı

Güncel Ay	Son 6 Ay	İ/K
0	>=2	Kötü
0	<=1	İyi
1	<=1	Belirsiz
1	>=2	Kötü
2+		Kötü

Çizelge 3.2'de yer alan iyi kötü tanımına göre veri çekildiği anda 3 taksit ve daha fazlası gecikmiş ise o müşteri kötü olacaktır. Veri çekildiği anda hiç taksit gecikmemişse ve son altı ayda 3ve daha fazla gecikmişi olmuşsa belirsiz olacaktır. Veri çekildiği anda bir veya iki taksit

gecikmişse ve son altı ayında 3 ve daha fazla taksit gecikmişi varsa belirsiz olacaktır. Güncel durumu sıfır veya bir ise ve son altı ayında 2 ve daha az gecikmişi varsa iyi olacaktır. Güncel durumda iki taksit gecikmişi varsa ve son altı ayında en az bir taksit gecikmiş taksiti varsa belirsiz olacaktır. Güncel ayında bir taksit gecikmişte ise ve son altı ayında iki ve daha fazla gecikmişi varsa kötü olacaktır. Veri çekildiği anda iki ve daha fazla gecikmişi varsa kötü olarak tanımlanacaktır.

Çizelge 3.3'de yer alan iyi kötü tanımına göre veri çekildiği anda hiç gecikmişi yoksa ve son altı ayında 2 ve daha fazla taksitini gecikmeli ödemişse kötü olacaktır. Veri çekildiği anda hiç gecikmişi yoksa ve son altı ayında en çok bir taksitini gecikmeli ödemişse iyi olacaktır. Veri çekildiği anda bir taksit gecikmişte ise ve son altı ayında en çok bir taksitini gecikmeli ödemişse belirsiz olacaktır.

Yukarıda listelenen iyi kötü tanımlarına göre iyi kötü sayılarına bakılmıştır. Birinci tanıma göre bakıldığında müşterilerin %86.3'ü iyi, yalnızca %6.3 ü kötü olduğu görülmüştür. Bu veriler yapılan ilk tanımın iyimser bir yaklaşım olduğunu göstermektedir. Portföyün genel yapısına bakıldığında müşterilerin büyük bir bölümü 3 döneme varan gecikmeler yaşamamaktadır. Bu tanımdan elde edilen sonuçtan hareketle ikinci tanım elde edilmiştir. İkinci tanımın sonuçları aşağıdaki tabloda verilmiştir.

Çizelge 3.4. İlk iyi-kötü tanımı sonuçları

İyi/Kötü Tanımı_v1 Sonuçları	
İyi	86.3%
Kötü	6.1%
Belirsiz	5.3%
İyi/Kötü Odds	14.13

Çizelge 3.5. İkinci iyi-kötü tanımı sonuçları

İyi/Kötü Tanımı_v2 Sonuçları	
İyi	61.5%
Kötü	18.5%
Belirsiz	19.2%
İyi/Kötü Odds	3.32

4.2.4. Örneklem Periyodunun Belirlenmesi

Modelleme için alınan veri için zamana göre iyi kötü oranlarını gösteren aşağıdaki grafik çizilmiştir. Grafik incelendiğinde 01.05.2004 ile 30.04.2006 tarihleri arasındaki kredilerin iyi-kötü odds değerlerinin benzer bir dağılım gösterdiğini söyleyebiliriz. Bu da bu tarih aralığındaki performansların olgunlaşmış olduğunu gösterir. Bu tarih aralığındaki başvurular modelleme için kullanılacaktır.

Şekil 3.1. İyi-Kötü odds değerinin zamana göre grafiği ve örneklem periyodu

4.2.5. Örneklem Büyüklüğünün Belirlenmesi

Tane 8,603 iyi müşteri, 2,595 tane kötü müşteri ve 2,799 tane belirsiz ödeme performansına sahip müşteri alınacaktır.

Çizelge 3.6. Örneklem büyüklüğü

Performans	Hesap Sayısı
Kötü	2,595
İyi	8,603
Belirsiz	2,799
Toplam	**13,997**

4.2.6. Modellemeye Dahil olacak Değişkenlerin Belirlenmesi

Değişkenler tek tek hedef değişken olan iyi kötü durumu(İKD) değişkenine göre incelenmiştir.

4.2.6.1. Medeni Hal

Çizelge 3.7. Medeni hal değişkenine göre dağılım

			Medeni Hal		Toplam
			Evli	Bekar	
İKD	Kötü	Adet	2050	545	2595
		% Kötü	79,0%	21,0%	23,2%
	İyi	Adet	7368	1235	8603
		% İyi	85,6%	14,40%	76,8%
Toplam		Adet	9418	1780	11198
		% Toplam	84,1%	15,9%	100,0%

Modelleme için kullanılan müşterilerin %84,1'i evli, %15,9'u ise bekârdır. Ödeme performansı kötü olan müşterilerin %79'u evli, %21'i ise bekârdır. İyi müşterilerin %85,6'sı evli, kötü müşterilerin %21'i ise bekârdır.

Medeni hal ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için medeni hal değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Çizelge 3.8. Medeni hal değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	62,524	1	,000

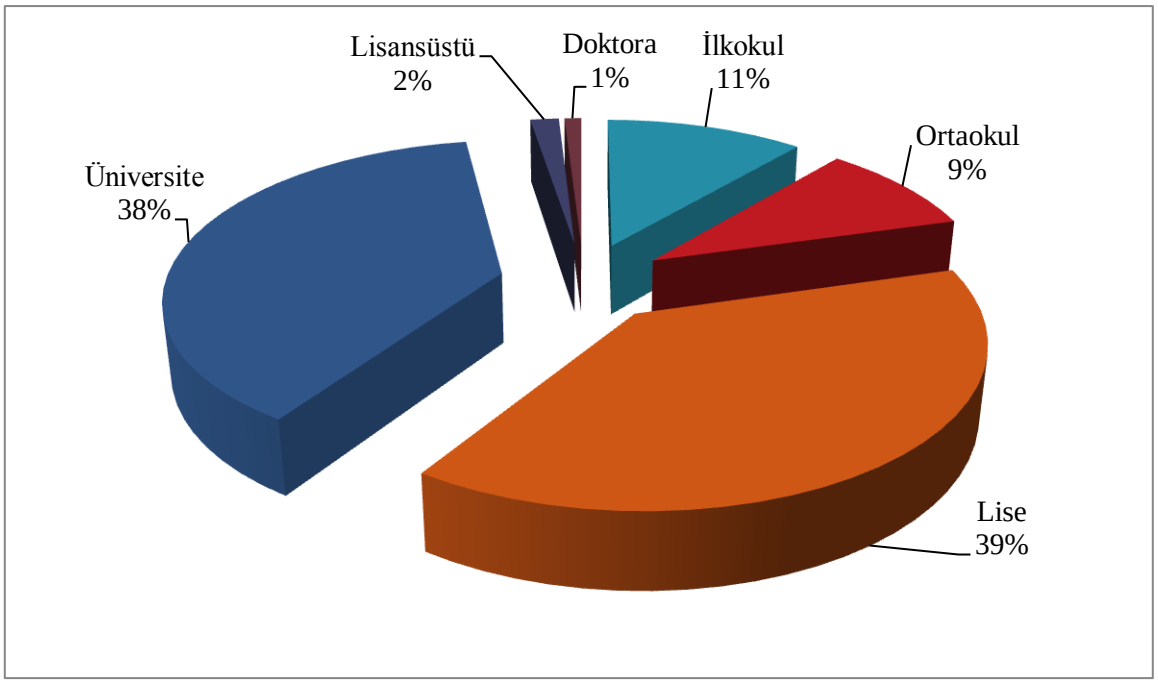
4.2.6.2. Eğitim Durumu

Çizelge 3.9. Eğitim Durumu değişkenine göre dağılım

			Eğitim_Durumu						Total
			İlkokul	Ortaokul	Lise	Üniversite	Lisansüstü	Doktora	
İKD	Kötü	Adet	351	291	1076	822	33	22	2595
		% Kötü	13,5%	11,2%	41,5%	31,7%	1,3%	,8%	100,0%
	İyi	Adet	905	757	3262	3446	148	85	8603
		% İyi	10,5%	8,8%	37,9%	40,1%	1,7%	1,0%	100,0%
Toplam		Adet	1256	1048	4338	4268	181	107	11198
		% Toplam	11,2%	9,4%	38,7%	38,1%	1,6%	1,0%	100,0%

Eğitim durumu ilkokul, ortaokul, lise, üniversite, lisansüstü ve doktora olmak üzere 5 seviyeden oluşmaktadır.

Modelleme için kullanılan veriye göre müşteriler içinde en yüksek paya sahip olan %38,8 lik pay ile lise mezunlarıdır. Onu izleyen grup ise %38,1 lik pay ile üniversite mezunlarıdır.



Şekil 3.2. Eğitim durumuna göre dağılım

Eğitim durumu ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için eğitim durumu değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

H_0 = Değişken anlamsızdır.
H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Çizelge 3.10. Eğitim durumu değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	60,966	1	,000

4.2.6.3. Eş Çalışma Durumu

Çizelge 3.11. Eş Çalışma Durumu değişkenine göre dağılım

Medeni Hal				Eş Çalışma Durumu		Toplam
				Çalışmıyor	Çalışıyor	
Evli	İKD	Kötü	Adet	1924	126	2050
			% Kötü	93,9%	6,1%	100,0%
		İyi	Adet	6833	535	7368
			% İyi	92,7%	7,3%	100,0%
	Toplam		Adet	8757	661	9418
			% Toplam	93,0%	7,0%	100,0%

Bu alanda kredi başvurusunda bulunan kişilerin eşlerinin çalışıp çalışmadığı durumu sorulmuştur. Yalnızca evli müşterilerin bu soruya cevap vermeleri istenmiştir.

Evli müşterilerin %93'ünün eşi çalışmamaktadır. İyi müşterilerin %92,7'sinin eşi çalışmamaktadır. Kötü müşterilerde ise bu oran %93,9 dur. Bu betimsel istatistiklere bakarak bu değişkenin iyi ve kötü performanslı müşterileri yeterince iyi ayırt edemediği tahmin edilebilir.

Çizelge 3.12. Eş Çalışma Durumu değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	3,640	1	,058

Eş çalışma durumu ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık > 0,05 olduğu için eş çalışma durumu değişkeninin İKD değişkenini açıklama bakımından yeterli olmadığı söylenebilir.

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında 0.05 anlamlılık seviyesinde H_0 hipotezini reddetmekle yapacağımız hata daha büyük olduğu için H_0 reddedilemez. Değişkenin anlamsız olduğunu söyleyebiliriz.

4.2.6.4. Çocuk Sayısı

Çocuk sayısı isimli değişkene ait ayrıntılı frekans bilgileri aşağıdaki tabloda verilmiştir.

Çizelge 3.13. Çocuk Sayısı değişkenine göre dağılım

Medeni Hal				Çocuk Sayısı								Toplam
				0	**1**	**2**	**3**	**4**	**5**	**6**	**8**	
Evli	İKD	Kötü	Adet	1771	106	106	46	17	3	1	0	2050
			% Kötü	86,4%	5,2%	5,2%	2,2%	,8%	,1%	,0%	,0%	100,0%
		İyi	Adet	6168	441	509	178	54	12	3	3	7368
			% İyi	83,7%	6,0%	6,9%	2,4%	,7%	,2%	,0%	,0%	100,0%
	Toplam		Adet	7939	547	615	224	71	15	4	3	9418
			% İKD	84,3%	5,8%	6,5%	2,4%	,8%	,2%	,0%	,0%	100,0%
Bekar	İKD	Kötü	Adet	538	1	3	3	0				545
			% Kötü	98,7%	,2%	,6%	,6%	,0%				100,0%
		İyi	Adet	1215	11	6	1	2				1235
			% İyi	98,4%	,9%	,5%	,1%	,2%				100,0%
	Toplam		Adet	1753	12	9	4	2				1780
			% İKD	98,5%	,7%	,5%	,2%	,1%				100,0%

Çocuk Sayısı ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için Çocuk Sayısı değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

H_0 = Değişken anlamsızdır.
H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Çizelge 3.14. Çocuk Sayısı değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	11,980	1	,001

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.5. Sosyal Güvenlik Kurumu

Kredi başvurusunda bulunan müşterinin hangi sosyal güvenlik kurumuna bağlı olduğu sorulur. Toplamda SSK, Emekli Sandığı, Bağkur, Diğer ve yok olmak üzere beş adet seviye mevcuttur.

Çizelge 3.15. Sosyal Güvenlik Kurumu değişkenine göre dağılım

			Sosyal Güvenlik Kurumu					Toplam
			Yok	SSK	Emekli Sandığı	Bağkur	Diğer	
İKD	Kötü	Adet	284	1275	353	605	57	2574
		% Kötü	11,0%	49,5%	13,7%	23,5%	2,2%	100,0%
	İyi	Adet	614	4212	2208	1383	116	8533
		% İyi	7,2%	49,4%	25,9%	16,2%	1,4%	100,0%
Toplam		Adet	898	5487	2561	1988	173	11107
		% İKD	8,1%	49,4%	23,1%	17,9%	1,6%	100,0%

Çizelge 3.16. Sosyal Güvenlik Kurumu değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	,308	1	,579

Sosyal güvenlik kurumu değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık > 0,05 olduğu için Sosyal güvenlik kurumu değişkeninin İKD değişkenini açıklama bakımından yeterli olmadığı söylenebilir.

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha büyük olduğu için H_0 hipotezi reddedilemez. Değişken anlamsızdır diyebiliriz.

4.2.6.6. Sosyal Güvenlik Numarası

Çizelge 3.17. Sosyal Güvenlik Numarası değişkenine göre dağılım

			Sosyal Güvenlik Numarası		Toplam
			Yok	Var	
İKD	Kötü	Adet	866	1729	2595
		% Kötü	33,4%	66,6%	100,0%
	İyi	Adet	2394	6209	8603
		% İyi	27,8%	72,2%	100,0%
Toplam		Adet	3260	7938	11198
		% İKD	29,1%	70,9%	100,0%

İyi müşterilerin %72,2'sinin sistemde kayıtlı sosyal güvenlik numarası bulunmaktadır. Kötü müşterilerin %33,4'ünün sistemde kayıtlı sosyal güvenlik numarası bulunmamaktadır.

Sosyal güvenlik numarası değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için Sosyal güvenlik numarası değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

Çizelge 3.18. Sosyal Güvenlik Numarası değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	29,159	1	,000

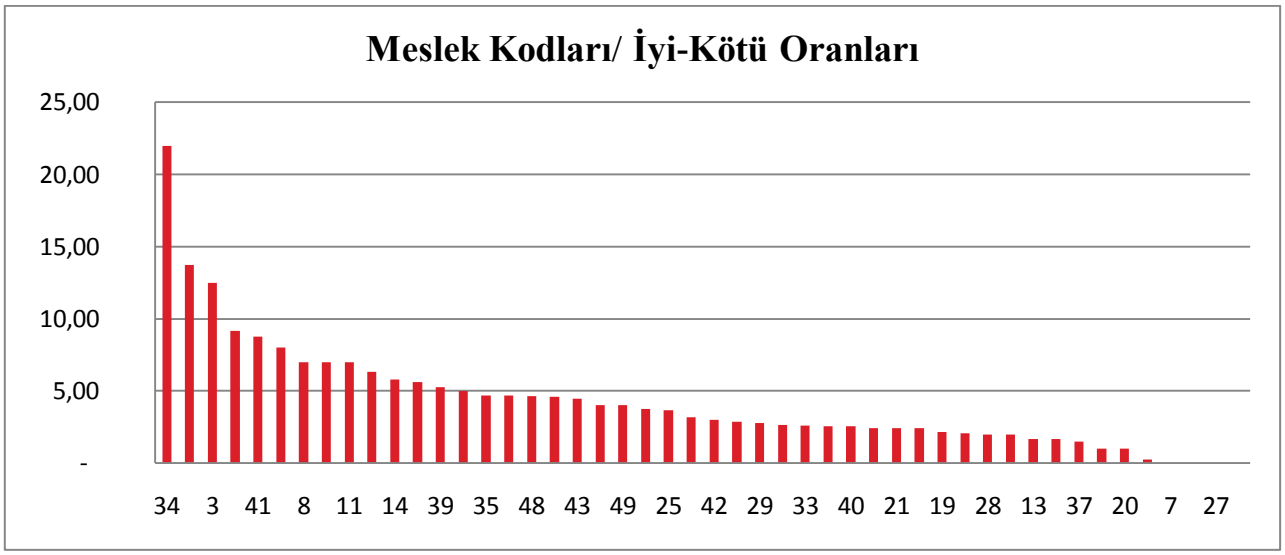
H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.7. Meslek

Meslek değişkeni 49 seviyeden oluşmaktadır. Bu 49 seviye hem analizini kolaylaştırmak hem de daha kolay yorumlanabilmesi için aza indirildi. Bu işlemin yapılabilmesi için her bir meslek seviyesindeki (iyi_sayısı/kötü_sayısı) oranlarına bakıldı. Benzer dağılım gösteren meslekler bir grupta toplandı.



Şekil 3.3. Meslek değişkenine göre iyi kötü odds değerinin dağılımı

Buna göre akademisyen, astsubay, bankacı, bilgisayar programcısı, doktor, hemşire, İşçi, memur, denetçi, mühendis, öğretmen, polis, şöfor ve teknisyen isimli meslekler birinci meslek grubunu, kalan diğer meslekler de ikinci meslek grubunu oluşturmaktadır. Model bu yeni değişkene göre oluşturulacaktır.

Meslek değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için meslek değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

Çizelge 3.19. Meslek değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	141,142	1	,000

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.8. E-mail Adresi

Çizelge 3.20. E-mail Adresi değişkenine göre dağılım

			E-mail		Toplam
			Yok	Var	
İKD	Kötü	Adet	2557	38	2595
		% Kötü	98,5%	1,5%	100,0%
	İyi	Adet	8381	222	8603
		% İyi	97,4%	2,6%	100,0%
Toplam		Adet	10938	260	11198
		% İKD	97,7%	2,3%	100,0%

İyi ödeme performansına sahip müşterilerin %98,5'i, %97,7'sinin sistemde kayıtlı e-mail adresi bulunmamaktadır.

E-mail Adresi değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için E-mail adresi değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

Çizelge 3.21. E-mail Adresi değişkeni için ki-kare

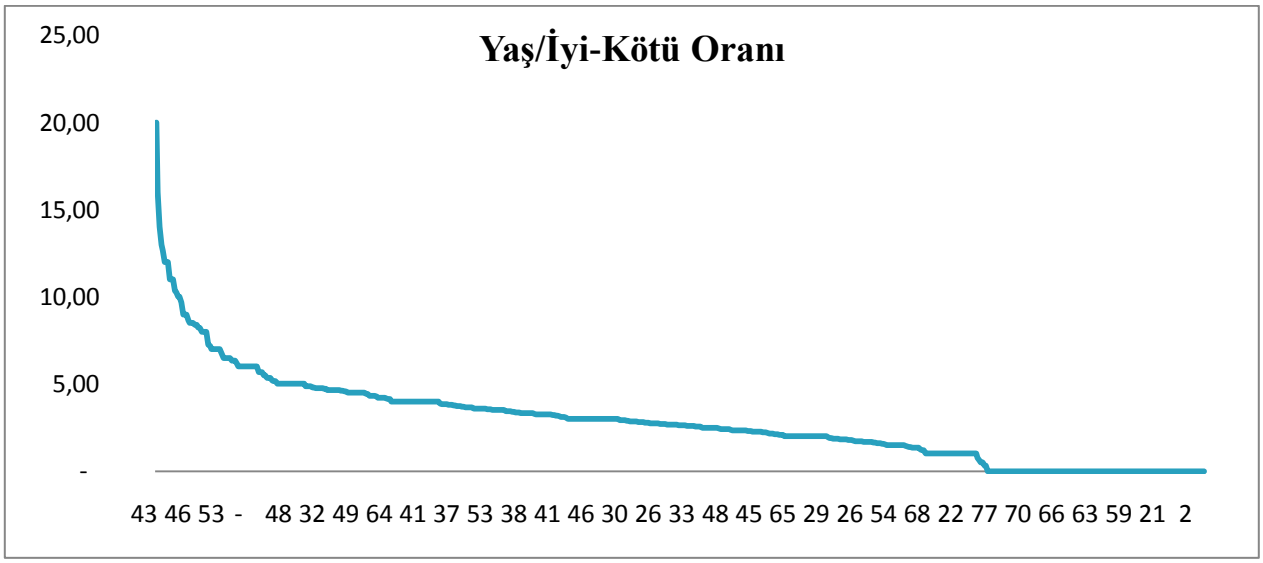
	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	12,117	1	,000

H_0 = Değişken anlamsızdır.
H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.9. Yaş

Yaş değişkeni için de her bir yaş seviyesi için (iyi_sayısı/kötü_sayısı) oranları hesaplandı. Aşağıdaki grafikten de görüleceği gibi riskli yaş grubu olarak tanımlayabileceğimiz 60 yaş üstü ve 25 yaş altı kişilerde bu oranlar daha düşük görünüyor.



Şekil 3.4. Yaş değişkenine göre iyi kötü odds değerinin dağılımı

Müşteriler yaşlarına göre 25 yaş altı ve 60 yaş üstü birinci grup, 26 ile 40 yaş arası ikinci grup, 41 ile 59 yaş arası ise üçüncü grup olmak üzere üç seviyeye indirgendi.

Çizelge 3.22. Yaş değişkenine göre dağılım

			Yaş			Toplam
			1,00	2,00	3,00	
İKD	Kötü	Adet	415	1355	825	2595
		% Kötü	16,0%	52,2%	31,8%	100,0%
	İyi	Adet	1164	4741	2698	8603
		% İyi	13,5%	55,1%	31,4%	100,0%
Toplam		Adet	1579	6096	3523	11198
		% İKD	14,1%	54,4%	31,5%	100,0%

Kötü müşterilerin %52,2'si 26-40 yaş arasında yer almaktadır. Riskli yaş grubundaki iyi müşteri oranı %13 iken kötü müşteri oranı %16 dır.

Yaş değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için yaş değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

H_0 = Değişken anlamsızdır.
H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Çizelge 3.23. Yaş değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	11,534	2	,003

4.2.6.10. Ev Durumu

Ev durumu alanında müşterinin ikamet ettiği evin mülkiyet durumu sorulmaktadır. İyi müşterilerin %45,6sı kendi evinde otururken, kötü ödeme performansına sahip müşterilerin %54,3'ü kendi evinde oturmamaktadır.

Çizelge 3.24. Ev durumu değişkenine göre dağılım

			Ev_Durumu					Toplam
			Kendisinin	Ailesinin	Kira	Lojman	Diğer	
İKD	Kötü	Adet	1168	836	511	61	2	2578
		% Kötü	45,3%	32,4%	19,8%	2,4%	,1%	100,0%
	İyi	Adet	3885	2374	1898	365	1	8523
		% İyi	45,6%	27,9%	22,3%	4,3%	,0%	100,0%
Toplam		Adet	5053	3210	2409	426	3	11101
		% İKD	45,5%	28,9%	21,7%	3,8%	,0%	100,0%

Ev durumu değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için ev durumu değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

H_0 = Değişken anlamsızdır.
H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Çizelge 3.25. Ev durumu değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	43,213	4	,000

4.2.6.11. Banka Bilgisi

Banka bilgisi alanında kredi başvurusunda bulunan müşterinin başka bankalarla çalışmaları olup olmadığı sorulmuştur. Modelleme için alınan veri setinde müşterilerin yalnızca %3'ünün banka bilgisi olduğu görülmektedir.

Çizelge 3.26. Banka Bilgisi değişkenine göre dağılım

			Banka_Bilgisi		Toplam
			yok	var	yok
İKD	Kötü	Adet	2527	68	2595
		% Kötü	97,4%	2,6%	100,0%
	İyi	Adet	8332	271	8603
		% İyi	96,8%	3,2%	100,0%
Toplam		Adet	10859	339	11198
		% İKD	97,0%	3,0%	100,0%

Banka bilgisi değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık > 0,05 olduğu için banka bilgisi değişkeninin İKD değişkenini açıklama bakımından yeterli olmadığı söylenebilir.

Çizelge 3.27. Banka Bilgisi değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	1,968	1	,161

H_0 = Değişken anlamsızdır.

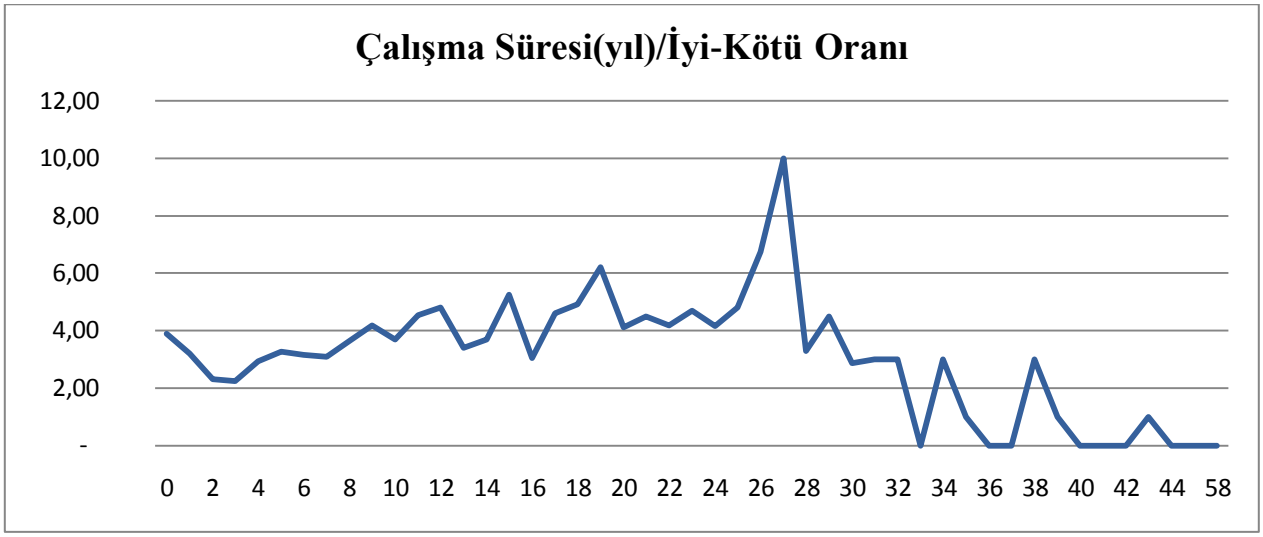
H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha büyük olduğu için H_0 reddedilemez. Değişkenin anlamsız olduğunu söyleyebiliriz.

4.2.6.12. İş Yeri Çalışma Süresi

İş yeri çalışma süresi kredi başvurusunda bulunan müşterinin çalıştığı iş yerinde ne kadar zamandır çalıştığı bilgisini ay bazında içermektedir. Bu değişken sürekli bir değişken olup ay bilgisi 12'ye bölünerek yıl olarak incelenmiştir.

Sürekli bir değişken olan iş yeri çalışma süresi iyi/kötü oranı bakımından incelenmiş ve 5 yılın kritik bir nokta olduğuna karar verilmiştir. Bu nedenle 0-5 yıl birinci grup, 5 ve daha fazlası ikinci grup olmak üzere iş yeri çalışma süresi iki seviyeye indirgenerek kategorik bir hale getirilmiştir.



Şekil 3.5. İş yeri çalışma süresi değişkenine göre iyi kötü odds değerinin dağılımı

İyi müşterilerin %61'i 5 yıldan fazla süredir başvuru esnasında çalıştığı iş yerinde çalışmaktadır. Kötü ödeme performansına sahip müşterilerin %45'i başvuru esnasında çalıştığı iş yerinde 5 yıldan az bir süredir çalışmaktadır.

Çizelge 3.28. İş Yeri Çalışma Süresi değişkenine göre dağılım

			İş Yeri Çalışma Süresi		Total
			1,00	2,00	1,00
İKD	Kötü	Adet	1164	1431	2595
		% Kötü	44,9%	55,1%	100,0%
	İyi	Adet	3388	5215	8603
		% İyi	39,4%	60,6%	100,0%
Total		Adet	4552	6646	11198
		% İKD	40,7%	59,3%	100,0%

İş Yeri Çalışma Süresi değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için İş Yeri Çalışma Süresi değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Çizelge 3.29. İş Yeri Çalışma Süresi değişkeni için ki-kare

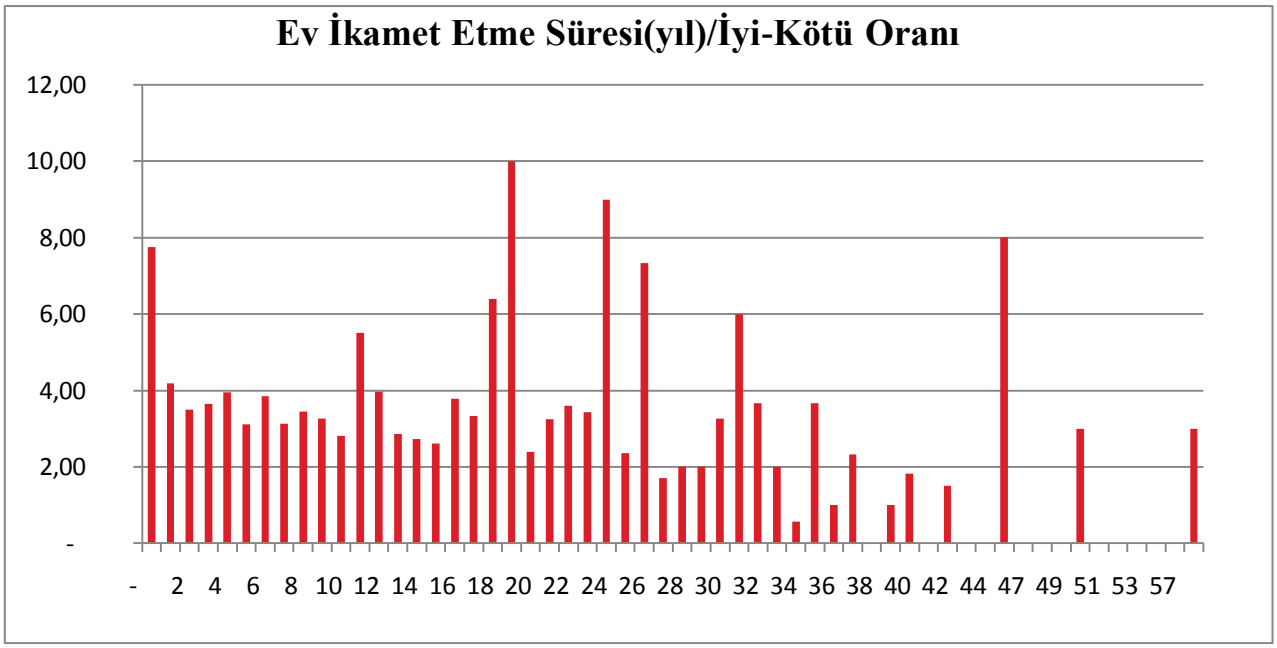
	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	24,599	1	,000

4.2.6.13. Ev İkamet Etme Süresi

Ev ikamet etme süresi kredi başvurusunda bulunan müşterinin başvurduğu esnada oturduğu evde toplam ne kadar süredir ikamet ettiği bilgisini içerir. Bu değişken de aynı çalışma süresi gibi sürekli bir değişkendir. Bu değişkenin de iyi/kötü oranına göre dağılımı incelenmiş ve temelde 3 grupta kümelenme olduğu gözlenmiştir. Birinci grup 0 ile 5 yıl arası, ikinci grup 6 ile 10 yıl, üçüncü grup ise 10 yıldan fazla süredir ikamet edenleri içerir.

Çizelge 3.30. Ev İkamet Etme Süresi değişkenine göre dağılım

			Ev İkamet Etme Süresi			Toplam
			1,00	2,00	3,00	
İKD	Kötü	Adet	794	1048	753	2595
		% İKD	30,6%	40,4%	29,0%	100,0%
	İyi	Adet	3054	3362	2186	8602
		% İKD	35,5%	39,1%	25,4%	100,0%
Toplam		Adet	3848	4410	2939	11197
		% İKD	34,4%	39,4%	26,2%	100,0%



Şekil 3.6. Ev ikamet etme süresi değişkenine göre iyi kötü odds değerinin dağılımı

Ev ikamet etme süresi değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için Ev ikamet etme süresi değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir

.

Çizelge 3.31. Ev İkamet Etme Süresi değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	24,835	2	,000

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.14. Çalışma Şekli

Çalışma şekli ücretli, serbest, emekli, emekli çalışan, çalışmayan, ev hanımı ve öğrenci olarak 7 seviyeden oluşmaktadır. Emekli, emekli çalışan, çalışmayan, ev hanımı ve öğrenci kategorilerinin frekansları oldukça düşük olduğu için bu gruplar serbest kategorisiyle birleştirilmiştir.

Çizelge 3.32. Çalışma Şekli değişkenine göre dağılım

			Çalışma Şekli							Toplam
			Ücretli	Serbest	Emekli	Emekli Çalışan	Çalışmayan	Ev Hanımı	Öğrenci	
İKD	Kötü	Adet	1438	808	183	70	33	39	3	2574
		% Kötü	55,9%	31,4%	7,1%	2,7%	1,3%	1,5%	,1%	100,0%
	İyi	Adet	5544	1731	783	250	109	109	7	8533
		% İyi	65,0%	20,3%	9,2%	2,9%	1,3%	1,3%	,1%	100,0%
Toplam		Adet	6982	2539	966	320	142	148	10	11107
		% İKD	62,9%	22,9%	8,7%	2,9%	1,3%	1,3%	,1%	100,0%

Çalışma Şekli değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için Çalışma Şekli değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

Çizelge 3.33. Çalışma Şekli değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	70,256	2	,000

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.15. Peşinat

Verileri kullanılan bankada tüketici kredileri için %0- %12, %13-%24 ay, %25-%36 ay, %36-%48 ay, %48-%60 ay ve %61 ve üzeri olmak üzere 6 seviyede incelenmiştir.

Çizelge 3.34. Peşinat değişkenine göre dağılım

			Peşinat						Toplam
			1,00	2,00	3,00	4,00	5,00	6,00	1,00
İKD	Kötü	Adet	198	216	1339	422	260	160	2595
		% İKD	7,6%	8,3%	51,6%	16,3%	10,0%	6,2%	100,0%
	İyi	Adet	391	491	3609	1699	1318	1095	8603
		% İKD	4,5%	5,7%	42,0%	19,7%	15,3%	12,7%	100,0%
Toplam		Adet	589	707	4948	2121	1578	1255	11198
		% İKD	5,3%	6,3%	44,2%	18,9%	14,1%	11,2%	100,0%

Peşinat değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için Peşinat değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

Çizelge 3.35. Peşinat değişkeni için ki-kare

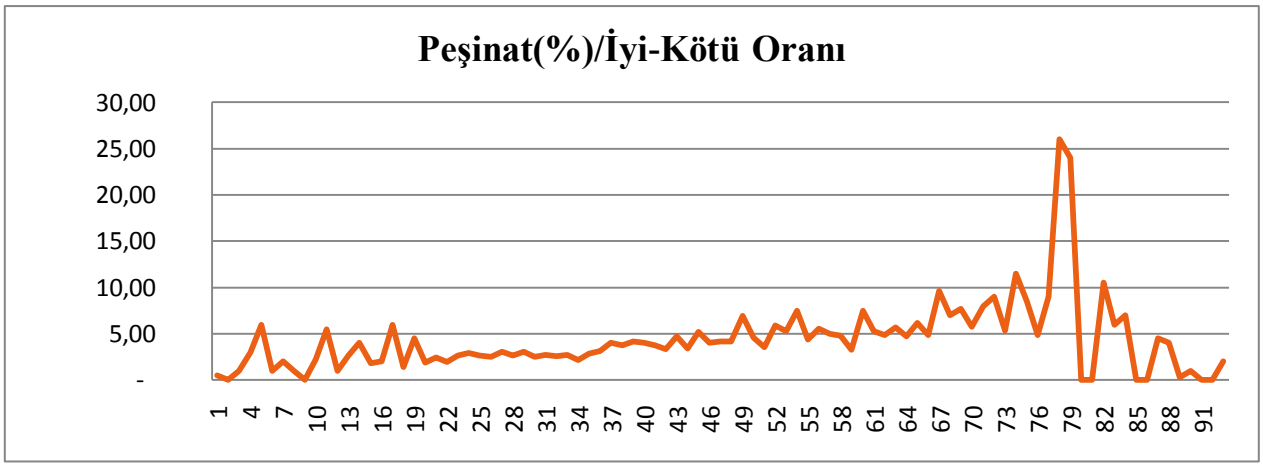
	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	237,480	5	,000

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

Peşinat değişkenine göre iyi/kötü oranının dağılımı aşağıdaki grafikte verilmiştir.



Şekil 3.7. Peşinat değişkenine göre iyi kötü odds değerinin dağılımı

4.2.6.16. Ürün Tipi

Çizelge 3.36. Ürün tipi değişkenine göre dağılım

			Ürün Tipi				Toplam
			Diğer	Konut-İşyeri	Taşıt	İhtiyaç	
İKD	Kötü	Adet	74	772	1630	118	2594
		% Kötü	2,9%	29,8%	62,8%	4,5%	100,0%
	İyi	Adet	152	3702	4545	198	8597
		% İyi	1,8%	43,1%	52,9%	2,3%	100,0%
Toplam		Adet	226	4474	6175	316	11191
		% İKD	2,0%	40,0%	55,2%	2,8%	100,0%

Müşterinin başvurduğu kredi tipini ürün tipi değişkeni göstermektedir. Bu alanda iyi ve kötü ödeme performanslı müşteriler için ürüne göre iyi bir ayrışma olduğunu söylemek mümkündür. İyi ödeme performansına sahip müşterilerin %43'ü konut kredisi alırken, kötü müşterilerin %63'ü araç kredisi almıştır.

Ürün Tipi değişkeni ile İKD değişkeni arasındaki tek değişkenli lojistik regresyon sonucuna göre 0,05 anlamlılık seviyesinde olasılık < 0,05 olduğu için Ürün Tipi değişkeninin İKD değişkenini açıklama bakımından yeterli olduğu söylenebilir.

Çizelge 3.37. Ürün Tipi değişkeni için ki-kare

	Ki-Kare	Serbestlik Derecesi	Olasılık
Model	170,463	2	,000

H_0 = Değişken anlamsızdır.

H_1 = Değişken anlamlıdır.

Hipotezleri altında H_0 hipotezini reddetmekle yapacağımız hata daha küçük olduğu için H_0 reddedilebilir. Değişkenin anlamlı olduğunu söyleyebiliriz.

4.2.6.17. Diğer Değişkenler

Araba durumu, kredi döviz türü, ev telefonu, iş telefonu, cep telefonu, sektör değişkenleri müşterinin iyi kötü performanslı olma durumunu yeterince açıklayamadığı için modelleme dışında tutulmasına karar verilmiştir.

Finansman türü değişkeni talep edilen kredinin nakdi mi gayri nakdi mi olduğunu belirten bir alandır. Bankanın bireysel kredi portföyünde çoğunlukla nakdi krediler kullandırıldığı için gayri nakdi krediler müşterinin iyi kötü performanslı olma durumunu yeterince açıklayamadığı için modelleme için kullanılmayacaktır. Benzer şekilde uyruk, kimlik belge türü değişkenleri de ayırıcı olarak tespit edilmemiştir.

4.3. Skor Kart Modellemesi

4.3.1. Değişkenlerin Dönüştürülmesi

Değişkenlerin tamamı iyi kötü dağılımlarına göre incelenmiştir. Bazı değişkenler oldukları halleriyle anlamlı çıkmayıp farklı formlara dönüştürülmesi gerekmiştir.

Yaş değişkeni sürekli bir değişken olmasına karşın kategorik olarak incelenmesi daha anlamlı olmuştur. Müşteriler yaşlarına göre 25 yaş altı ve 60 yaş üstü birinci grup, 26 ile 40 yaş arası ikinci grup, 41 ile 59 yaş arası ise üçüncü grup olmak üzere üç seviyeye indirgendi.

İş yeri çalışma süresi değişkeni de aynı şekilde sürekli bir değişken iken iyi/kötü oranlarının dağılımındaki benzerlikten faydalanılarak 2 düzeyli bir kategorik değişken haline getirilmiştir.

Ev ikamet etme süresi değişkeni de sürekli bir değişken iken üç deüzeyli kategorik bir değişken haline getirilmiştir.

4.3.2. Değişken Seçimi

Aşağıdaki değişkenlerin modelleme için kullanılmasına karar verilmiştir.

- Yaş
- Medeni Hal
- Eğitim Durumu
- Meslek
- İş Yerinde Çalışma Süresi
- Sosyal Güvenlik Numarası
- Talep Edilen Ürün Tipi
- Peşinat Oranı
- İstenilen Vade
- Çalışma Şekli
- Ev Durumu
- E-mail Bilgisi
- Aylık Net Gelir
- Kredi Büro Sorgu Sonucu-Son 6 Aydaki En Kötü Ödeme Durumu
- Kredi Büro Sorgu Sonucu-Başvuru Anına Kadar En Kötü Ödeme Durumu

4.3.3. Retlerin Anlamlandırılması

Ret başvuruların modele dahil edilmesi çalışması bu uygulama için yapılmamıştır.

4.4. Modelin Değerlendirilmesi

4.4.1. Modelin Uyum İyiliği Testleri

Değişkenlerin tamamı İyi-Kötü durumu isimli değişken ile birlikte lojistik regresyon ile modellenmiştir.

Çizelge 3.38. Modelin anlamlılık testleri

Model	-2 Log Benzerlik	Ki-Kare	Serbestlik Derecesi	Olasılık
Değişkenler Modelde Değilken	11686,108			
Final Model	8673,926	3012,182	2192	,000

Modelde yalnızca sabit terim varken ve değişkenler dahil edildiği durumdaki modeli kıyaslayalım. Yukarıdaki tabloya göre 0.05 anlamlılık düzeyinde final modelin anlamlı olduğu söylenebilir.

Çizelge 3.39. Modelin uyum iyiliği testleri

Uyum İyiliği Testleri	Ki Kare	Serbestlik Derecesi	Olasılık
Pearson	8115,545	8493	,998
Deviance	8595,247	8493	,216
Hosmer&Lemeshow	9,287	8	,319

Tabloda modelin uyum iyiliği testleri yer almaktadır.

H_0: Model verilere uygundur.

H_1: Model verilere uygun değildir.

Hipotezleri altında 0.05 anlamlılık seviyesinde H_0 hipotezini reddetmekle yapacağımız hata daha büyük olduğu için H_0 hipotezi kabul edilir. Modelin verilere uygun olduğunu söyleyebiliriz.

Çizelge 3.40. Modelde yer alan değişkenlerin uyum iyiliği testleri

	-2 Log Benzerlik (Azaltılmış Model)	Benzerlik Oranı Testi - Ki Kare	Serbestlik Derecesi	Olasılık
Sabit	8673,926	,000	0	.
Çalışma Şekli	8684,590	10,664	2	,005
Son 6 Ay içindeki En Kötü Ödeme Durumu	8695,638	21,712	4	,000
En Kötü Ödeme Durumu	8779,111	105,185	4	,000
Peşinat	8741,321	67,395	5	,000
Vade	8689,138	15,212	3	,002
Çalışma Süresi	8685,183	11,257	1	,001
Eğitim	8713,430	39,504	2	,000
Meslek	8684,100	10,174	1	,001
Ürün Tipi	8725,636	51,710	2	,000
Ev Durumu	8682,780	8,854	4	,025
Medeni Hal	8699,078	25,152	1	,000
E-mail	8696,304	22,378	1	,000
Yaş	8674,066	,140	2	,932
Sosyal Güvenlik Numarası	8680,014	6,088	1	,014

H_0:Değişken anlamsızdır.

H_1: Değişken anlamsız değildir.

Hipotezleri ile modele dahil olan değişkenlerin model içindeki anlamlılığı en çok benzerlik oranı testi ile test edilir. 0.05 anlamlılık seviyesinde olasılık < 0.05 olduğu için Yaş dışındaki bütün değişkenlerin anlamlı olduğu söylenebilir. Yaş değişkeninin 0.05 anlamlılık seviyesinde anlamsız olduğu söylenebilir.

4.4.2. Modelde Yer Alan Parametreler

Çizelge 3.41. Modelde yer alan değişkenlerin uyum iyiliği testleri

İKD	B	Standart Hata	Wald	S.D.	Olasılık	Exp(B)
Sabit	1,624	338,678	,002	1	,041	
Çalışma Şekli = Ücretli	,547	,182	9,056	1	,003	1,728
Çalışma Şekli = Serbest, emekli, emekli çalışan	,001	,077	,000	1	,985	1,001
ÇalşmaŞekli= Çalışmayan, Ev hanımı, Öğrenci	0	.	.	0	.	.
KBSS-Son 6 Aydaki En Kötü Ödeme Durumu =0	,177	,076	5,486	1	,019	1,194
KBSS-Son 6 Aydaki En Kötü Ödeme Durumu = 1,00	-,737	,187	15,526	1	,000	0,479
KBSS-Son 6 Aydaki En Kötü Ödeme Durumu =2,00	-,750	,376	3,978	1	,046	0,472
KBSS-Son 6 Aydaki En Kötü Ödeme Durumu =3,00 ve daha fazla	-1,110	1,188	,873	1	,350	0,329
KBSS-Başvuru Anına Kadar En Kötü Ödeme Durumu=0	,673	,075	81,018	1	,000	1,960
KBSS-Başvuru Anına Kadar En Kötü Ödeme Durumu =1,00	,230	,089	6,685	1	,010	1,258

KBSS-Başvuru Anına Kadar En Kötü Ödeme Durumu =2,00	-,053	,123	,188	1	,665	,589
KBSS-Başvuru Anına Kadar En Kötü Ödeme Durumu= 3,00 ve daha fazla	-,335	,183	3,367	1	,067	,715
Peşinat=48%-60%	,634	,199	10,185	1	,001	1,885
Peşinat=13%-24%	,607	,147	16,942	1	,000	1,829
Peşinat=25%-36%	,681	,112	37,140	1	,000	1,976
Peşinat=37%-48%	,311	,122	6,536	1	,011	1,365
Peşinat=0%-12%	,193	,129	2,235	1	,135	1,212
Vade=0-12 ay	,361	,172	4,399	1	,036	1,434
Vade=13-24 ay	,185	,114	2,646	1	,104	1,203
Vade=25 ay ve fazlası	-,037	,100	,138	1	,710	0,963
Çalışma Süresi=0-60 ay	,197	,059	11,288	1	,001	1,219
Çalışma Süresi=61 ay ve fazlası	0	.	.	0	.	.
Eğitim=İlköğretim	-,490	,078	39,349	1	,000	,613
Eğitim=Lise	-,233	,066	12,537	1	,000	,792
Eğitim=Üniversite ve üstü	0	.	.	0	.	.
Meslek Kodu=1, 2, 6, 9, 14, 24, 25, 30, 33, 35, 39, 41, 47, 48	,206	,065	10,142	1	,001	1,229
Meslek Kodu=Diğer	,126	.	.	0	.	1,134
Ürün Tipi=Diğer	-,427	,180	5,619	1	,018	,652
Ürün Tipi=Konut	,605	,094	41,655	1	,000	1,831
Ürün Tipi=Taşıt	0	.	.	0	.	.
Ev Durumu=Kendisinin	1,186	1,400	,718	1	,397	3,273
Ev Durumu=Ailesinin	1,089	1,400	,606	1	,436	2,971
Ev Durumu=Kira, Lojman, Diğer	0	.	.	0	.	.
Medeni Hal=Evli	,377	,075	25,527	1	,000	1,457

Medeni Hal=Bekar	0	.	.	0	.	.
E-mail=Yok	-,947	,216	19,205	1	,000	0,387
E-mail=Var	0	.	.	0	.	.
Yaş=Riskli Yaş Grubu	-,003	,090	,001	1	,972	,997
Yaş=26-40 yaş	,020	,067	,092	1	,761	1,021
Yaş=40-59 yaş	0	.	.	0	.	.
Sosyal Güvenlik No=Yok	-,163	,066	6,111	1	,013	0,849
Sosyal Güvenlik No=Var	0	.	.	0	.	.

Yukarıdaki tabloda İKD değişkenini seçilen bağımsız değişkenler ile birlikte lojistik regresyon modeli uygulandığında çıkan sonuç görülmektedir.

B isimli sütunda bağımsız değişkenlerin tahmin edilen katsayıları yer almaktadır. Standart Hata isimli sütunda ise bu katsayıların standart hataları hesaplanmıştır. Wald isimli sütunda modele giren her bir bağımsız değişkenin Wald değerleri yer almaktadır. Olasılık sütununda ise Wald istatistiğine karşılık gelen olasılık değerleri verilmiştir. Exp(B) sütununda bağımsız değişkenlerin katsayılarının ekponansiyeli hesaplanmıştır. Bu sütun bize her bir bağımsız değişkenin odds değerini vermektedir.

Tabloya göre çalışma şekli ücretli olan müşterilerin iyi olma olasılığının diğerlerine göre 1,728 kat daha fazla olduğu söylenebilir.

%48 ile %60 arasında peşinat veren müşterilerin iyi olma olasılığı diğer müşterilere göre 1.885 daha fazladır. Bu değer %13 ile %24 arasında peşinat verenler için 1.829 dur.

0 ile 12 ay arasında vade talep eden müşterilerin iyi olma olasılığı diğer müşterilere göre 1,434 kat daha fazladır. 25 ay ve daha fazla vade talep eden müşterilerde ise iyi olma olasılığı diğer müşterilere göre 0,963 kat daha fazladır.

İş yerinde 5 yıldır çalışan müşterilerin iyi olma olasılıkları diğer müşterilere göre 1,219 kat daha fazladır.

Kredi alma amacı konut olan müşterilerin diğer müşterilere göre iyi olma olasılıkları 1,831 daha fazladır.

Oturduğu ev kendisine ait olan müşterilerin iyi ödeme performansına sahip olma olasılığı diğer müşterilere göre 3,273 kat daha fazladır. Oturduğu ev ailesine ait olan müşterilerde bu oran 2.971 olarak hesaplanmıştır.

Tabloya göre lojistik regresyon denklemi aşağıda verilmiştir.

$$\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = 1{,}624 + 0{,}547(\text{Çalışma Şekli} = \text{Ücretli}) + 0{,}001(\text{Çalışma Şekli} = \text{Serbest, emekli, emekli çalışan}) + 0{,}177(\text{KBSS} - \text{Son 6 Aydaki En Kötü Ödeme Durumu} = 0) + \cdots + 0{,}126(\text{Aile Net Geliri}) \quad (3.1)$$

Modele göre denklem (3.1)'de verilen ifadenin sonucu *y* olsun. Buna göre müşterinin kötü olma olasılığı denklem (3.2)'de verilmiştir.

$$p = \frac{1}{1+e^{-y}} \quad (3.2)$$

4.4.3. Skor Kart

Modeli daha görsel hale getirmek amacıyla aşağıdaki Çizelge (3.42) oluşturulmuştur. Tabloda yer alan parametrelerin alacağı puanlar yani lojistik regresyon modelindeki katsayılar 1000 ile çarpılarak daha görsel bir hale getirilmiştir.

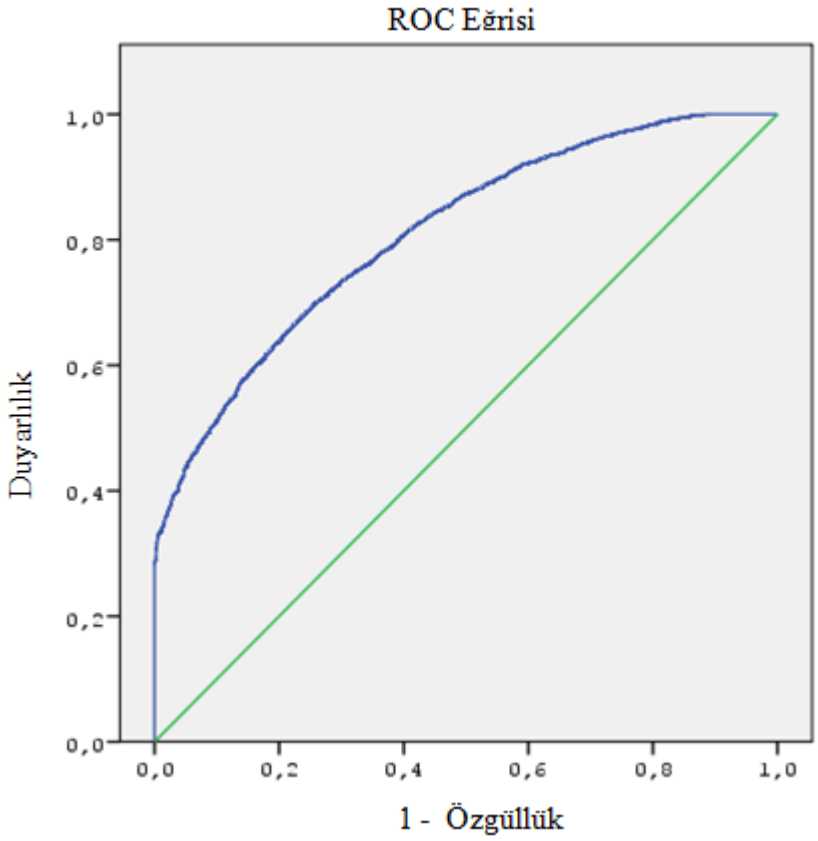
Çizelge 3.42. Elde edilen skor kart

Parametre		Skor
Sabit		1624
Çalışma Şekli		
	Ücretli	547
	Serbest, emekli, emekli çalışan	1

	Çalışmayan, Ev hanımı, Öğrenci	0
KBSS-Son 6 Aydaki En Kötü Ödeme Durumu		
	0	177
	1	-737
	2	-750
	3	-1110
KBSS-Başvuru Anına Kadar En Kötü Ödeme Durumu		
	0	,673
	1,00	,230
	2,00	-,053
	3,00 ve daha fazla	-,335
Peşinat		
	48%-60%	634
	13%-24%	607
	25%-36%	681
	37%-48%	311
	0%-12%	193
Vade		
	0-12 ay	361
	13-24 ay	185
	25 ay ve fazlası	-37
Çalışma Süresi		
	0-60 ay	197
	61 ay ve fazlası	0
Eğitim		
	İlköğretim	-490
	Lise	-233
	Üniversite ve üstü	0
Meslek Kodu		
	1, 2, 6, 9, 14, 24, 25, 30, 33, 35, 39, 41, 47, 48	206

	Diğer	126
Ürün Tipi		
	Diğer	-427
	Konut	605
	Taşıt	0
Ev Durumu		
	Kendisinin	1186
	Ailesinin	1089
	Kira, Lojman, Diğer	0
Medeni Hal		
	Evli	377
	Bekar	0
E-mail		
	Yok	-947
	Var	0
Yaş		
	Riskli Yaş Grubu	-3
	26-40 yaş	20
	40-59 yaş	0
Sosyal Güvenlik No		
	Yok	-163
	Var	0

4.4.4. ROC, GINI ve Sınıflama Tablosu



Şekil 3.8. Modele ait ROC eğrisi

Çizelge 3.43. ROC eğirsi altında kalan alanın anlamlılık testi sonucu

ROC Eğrisi Altında Kalan Alan	Standart Hata	Olasılık	95% Güven aralığı	
			Üst sınır	Alt Sınır
,811	,004	,000	,802	,819

Tahmin edilen modelin tahmin ettiği olasılıklara göre ROC eğrisi yukarıdaki grafikte görülmektedir. ROC eğrisi ile x ekseni altında kalan alan 0,811 olarak hesaplanmıştır. Bu değer modelin oldukça iyi bir tahmin gücü olduğunu göstermektedir.

ROC eğrisi altında kalan alandan yola çıkarak yaklaşık olarak GINI katsayısını da hesaplayabiliriz. GINI ile ROC arasındaki ilişkiyi daha önce belirtmiştik. Buna göre GINI katsayısının yaklaşık değerini [(0,811x2)-1] = 0,622 olarak bulabiliriz. GINI katsayısı için 0,62 değeri de modelin tahmin gücünün başarılı olduğunu gösteren yeterli büyüklüktedir.

Çizelge 3.44. Lojistik regresyon modeli soncuna göre sınıflama tablosu

Gözlenen	Tahmin edilen		
	Kötü	İyi	Doğru yüzdesi
Kötü	666	1856	26,4%
İyi	265	8066	96,8%
Toplam Yüzde	8,6%	91,4%	80,5%

Çizelge (3.44) deki sınıflama tablosuna göre kötü müşterilerin 666 tanesi gözlenen veride kötü iken lojistik regresyon sonucuna göre de kötü olarak tahmin edilmiştir. İyi müşterilerin 8066 tanesi gözlenen veride iyi iken lojistik regresyon sonucuna göre de iyi olarak tahmin edilmiştir. Toplamda lojistik regresyon modeli tüm verilerin %80,5'ini doğru olarak tahmin edebilmiştir.

4.5. Modelin Kullanıma Alınması

Kurulan skor kart modeli tahmin gücü bakımından yeterli bulunmuştur. Bu nedenler tüketici kredileri başvuru sisteminde başvuran müşterilerin iyi veya kötü performanslı olma durumunu tahmin etmek için kullanılmasına karar verilmiştir.

4.5.1. Kesim Puanının Belirlenmesi

Modelden elde edilen parametrelere göre elde bulunan veride yer alan müşterilerin alacağı skorlar hesaplanmıştır. Hesaplanan skorlar büyükten küçüğe doğru sıralanarak %10 luk gruplara ayrılmıştır. Çizelge (3.44)'de bu gruplar gösterilmektedir.

Çizelge 3.45. Skor bantları

Skor Aralığı	İyi Sayısı	Kötü Sayısı	İyi/Kötü
894 ve fazlası	1017	68	14.96
893-842	931	154	6.05
843-793	893	193	4.63
794-746	837	248	3.38
747-693	770	315	2.44
693-625	741	344	2.15
624-522	630	456	1.38
521 den az	342	744	0.46

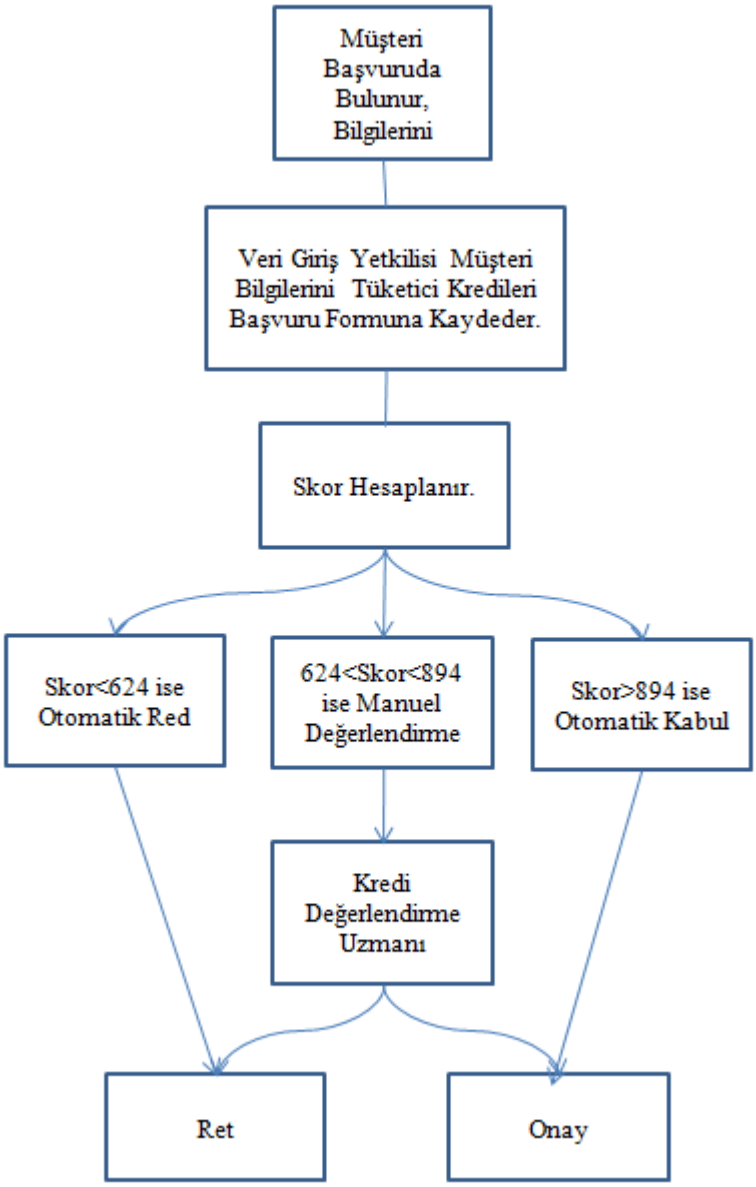
Çizelge (3.44)'de %10 luk gruba tekabül eden her skor bandındaki (iyi müşteri sayısı/kötü müşteri sayısı) oranları yer almaktadır. 894 skorundan yüksek puan alan ve en yüksek skor bandında yaklaşık olarak 15 iyi müşteriye bir kötü müşteri düşmektedir. Düşük skor bantlarına inildikçe (iyi müşteri sayısı/kötü müşteri sayısı) oranlarının da azaldığı görülmektedir.

İyi kötü oranı en yüksek ve en düşük olan skor bantları kesim noktasının belirlenmesinde yol gösterici olacaktır. En yüksek skor bandı otomatik kabul, en düşük skor bantları da otomatik ret için kullanılabilir.

Çizelge (3.44.)'e göre en sonda yer alan en düşük iki %10 luk grubun otomatik red için uygun olduğu görülmektedir. Buna göre otomatik ret için kesim noktası 624 olarak belirlenmiştir. En yüksek ilk skor bandı ise otomatik kabul için kullanılacaktır. 894 skorundan yüksek alan müşteriler otomatik olarak kabul edilecektir.

4.5.2. İzlenecek Strateji ve Entegrasyon

Kullanıma hazır olan skor kartın başvuru sistemine entegre edilmesinden önce izlenilecek strateji belirlenmiştir. Strateji Şekil (3.9.)' da gösterilmiştir.



Şekil 3.9. İzlenecek Strateji

Belirlenen stratejiye göre skor kart tüketici kredileri başvuru sistemine entegre edilmiştir. Skorların doğru hesaplanıp hesaplanmadığı ve belirlenen stratejinin düzgün çalışıp çalışmadığı kontrol edilmiştir. Bütün kontroller olumlu bir şekilde neticelendikten sonra skor kart uygulaması kullanılmaya başlanmıştır.

SONUÇLARIN DEĞERLENDİRİLMESİ

Günümüzde veri analizi ve ileriye yönelik tahmin, iş dünyası için rakiplerinden bir adım öne geçmek için en önemli araçlardan birisidir. Bu noktada kredi skorlama bankacılık ve finans sektöründe dünyada yaklaşık son 40 yıldır kullanılmaya başlanmış ve gittikçe yaygın bir hale gelmiştir. Geçmişe dönük verilerin kullanılarak ileriye yönelik tahmin yapılarak müşterilerin iyi veya kötü performanslı olma olasılıkları hesaplanabilmektedir. Bu da öncelikle riski ölçümlemeyi mümkün kılarak temerrüt yüzdelerini azaltmaktadır. Ayrıca skor kartları kurumlar için zaman, insan kaynağı ve masraf açısından tasarruf edilmesini sağlamaktadır.

Bu çalışmada lojistik regresyon ile kredi başvurularında kullanılan skor kart modellerinin kurulması anlatılmıştır. Lojistik regresyon skor kart modellemelerinde en yaygın kullanılan istatistiksel tekniktir. Varsayımlarının az olması, bağımlı değişkenin sürekli değişken olmadığı durumlarda rahatlıkla kullanılabilmesi ve kolay yorumlanabilir olması önemli tercih nedenleridir.

İlk bölümde detaylı olarak lojistik regresyon modelinin kurulması, tahmin yöntemi, modelin anlamlılık testleri, güven aralıkları, değişkenlerin anlamlılık testleri ve lojistik regresyon modelinin yorumlanması anlatılmıştır.

İkinci bölümde kredi kavramından başlanarak kredi skorlama ile ilgili kavramlar anlatılmıştır. Bir skor kart modelleme projesinin aşamaları adım adım gösterilmiştir. Verinin hazırlanması skor kart modeli oluşturma projesinin en önemli aşamasıdır diyebiliriz. Bu aşamada yapılacak doğru tespitler modelleme aşamasına olumlu katkılar sunacaktır. Bu nedenle bu aşama titizlikle yürütülmelidir. Modelleme aşamasında hangi değişkenlerin modele ne kadar katkıda bulunduğu gözlenmeli, bir çok denemeden sonra modelin son hali elde edilmelidir. Bu aşamada verilerin nasıl yani hangi formda kullanılacağına karar verilir ki bu da modelin performansını etkileyen önemli bir faktördür. Modelleme aşamasının son adımı olan modelin değerlendirilmesi projenin en zevkli kısmıdır diyebiliriz. Kurulan modellerin performansları yani tahmin güçleri değerlendirilir ve en iyisine karar verilir.

Üçüncü bölümde diğer 2 bölümde yer alan bilgiler ışığında örnek bir başvuru skor kart modeli kurulmuştur. İlk aşamada projenin amacı tüketici kredileri skor kart modeli kurulması olarak belirlenmiştir. Veri hazırlama aşamasında tüketici kredileri başvuru formunda yer alan tüm

alanları içerecek şekilde geçmiş müşterilere ait bir veri seti oluşturulmuştur. Veri alanları incelenerek kullanılabilecek alanlar seçilerek analize tabi tutulmuştur. Kurum prensiplerine göre iyi-kötü tanımı yapılarak veri setinde yer alan kredi ödeme performansı oluşmuş müşterilere performans ataması yapılmıştır. İyi-kötü oranlarına göre modellemenin yapılacağı örneklem periyodu seçilmiş ve büyüklüğü belirlenmiştir. Modellemeye dahil olacak değişkenler bağımlı değişken ile tek değişkenli lojistik regresyon ile modellenerek seçilmiştir. Çalışma Şekli, Son 6 Ay içindeki En Kötü Ödeme Durumu, En Kötü Ödeme Durumu, Peşinat, Vade, Çalışma Süresi, Eğitim, Meslek, Ürün Tipi, Ev Durumu, Medeni Hal, E-mail, Yaş ve Sosyal Güvenlik Numarası değişkenleri modelleme için kullanılmıştır. Bu değişkenlerden Çalışma Süresi, yaş gibi bazı sürekli değişkenler iyi kötü dağılımlarına göre gruplara ayrılarak kategorik hale getirilmiştir. Modelin uyum iyiliği testleri incelenerek modelin veriler ile uyumlu olduğuna karar verilmiştir. Model parametreleri de anlamlı bulunmuştur. Modelde yer lan değişkenlerin katsayılarına ait ekponansiyel değerleri yorumlanmıştır. Skor kart modelini daha görsel bir hale getirmek için de modelde yer alan değişkenlerin katsayıları 1000 ile çarpılarak nihai skor kart modeli oluşturulmuştur. Modele ait tahminler kullanılarak ROC eğrisi altında kalan alan hesaplanmış ve 0,81 gibi yüksek bir rakam elde edilmiştir. ROC değerinden yola çıkarak GINI katsayısı hesaplanmıştır. GINI katsayısı da 0,62 olarak bulunmuştur. Bulunan bu değerler kurulan başvuru skor kart modelinin yeterli olduğunu göstermiştir. Kesim noktası otomatik retler için 624, otomatik kabuller için de 894 belirlenmiş ve uygulamada izlenecek strateji de tanımlanmıştır.

KAYNAKÇA

Altman E. I. (2002), Revisiting Credit Scoring Models in a BASEL 2 Environment, London Risk Books, London.

Anderson, R. (2007), The Credit Scoring Toolkit, Oxford University Press, New York.

Anderson, T.W. (2003), An Introduction to Multivariate Statistical Analysis, New Jersey.

Halisdemir Ö. (2004), Bireysel Kredilerin Değerlendirilmesinde Skorlama Yönteminin Kullanımı, Doktora Tezi, Marmara Üniversitesi Bankacılık ve sigortacılık Enstitüsü.

Hosmer W. D. ve Lemeshow S. (2000), Applied Logistic Regression, John Wiley & Sons, New York

İşyar Y. (1999), Ekonometrik Modeller, Vipaş A.Ş., Bursa

Kalaycı Şeref v.d. (2008), SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, Asil Yayın Dağıtım, Ankara.

Menard S. (2002), Applied Logistic Regression Analysis, Sage Publications, London.

Özdamar K. (2004), Paket Programlar ile İstatistiksel Veri Analizi 1, Kaan Kitapevi, Eskişehir.

Özdamar K. (2004), Paket Programlar ile İstatistiksel Veri Analizi 2, Kaan Kitapevi, Eskişehir.

Tabachnick B. G., Fidell L. S.(2007), Using Multivariate Statistics, Pearson Education,California.

Tinsley H. E., Brown S. D. (2000), Handbook of Applied Multivariate Statistics and Mathematical Modeling, Academic Press, California

Thomas L., Edelman D. ve Crook J. (2002), Credit Scoring and Its Applications, Siam, Philadelphia.

Thomas L., Edelman D. ve Crook J. (2004), Readings in Credit Scoring, Oxford University Press, New York.

Topuz D. ve Çakır M. (2003), "Lojistik Regresyon Analiz Tekniğinin Eğitim Bilimleri Araştırmalarında Uygulanabilirliği ile ilgili bir Araştırma", A.İ.B.Ü.Eğitim Fakültesi Dergisi.

Siddiqi N. (2006), Credit Risk Scorecards, John Wiley & Sons, New Jersey.

Shirley D., Stanley W. ve Daniel C. (2004), Statistics for Research, John Wiley & Sons, New York.

Subhash S. (1996), Applied Multivariate Techniques, John Wiley & Sons, Inc., Toronto.

Ulupınar, S. D. (2007), 2001 Kriz Dönemi, Öncesi ve Sonrasında Türk Ticari Bankalarının Karlılıklarının Lojistik Regresyon Analizi ile İncelenmesi, Yüksek Lisans Tezi, Marmara Üniversitesi.

Unvan A. Y. (2006), Koşullu Lojistik Regresyon Çözümlemesi ve Avrupa Birliği Verisi Üzerinde Bir Uygulama, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü.

Ünsal A. ve Güler H., "Türk Bankacılık Sektörünün Lojistik Regresyon ve Diskiriminant Analizi İle İncelenmesi",

Ünver Ö. ve Gamgam H. (1999), Uygulamalı İstatistik Yöntemler, Siyasal Kitabevi, Ankara.

Ürük E. (2007), İstatistiksel Uygulamalarda Lojistik Regresyon Analizi, Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü.

ÖZGEÇMİŞ

Doğum Tarihi	13.07.1985	
Doğum Yeri	Çorum	
Lise	2000-2003	Üsküdar Çağrı Bey Anadolu Lisesi
Lisans	2003-2007	Yıldız Teknik Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü
Yüksek Lisans	2007-2010	Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Bölümü
Çalıştığı Kurumlar		
	2008-…	Kuveyt Türk Katılım Bankası Bireysel ve İşletme Krediler/Bilgi ve Karar Sistemleri Yönetimi