İstatistiksel ve Makine Öğrenimi Yöntemleriyle Kredi Skorlama

Yunus Emre Demirbulut[1], Mehmet S. Aktaş[1], Oya Kalıpsız[1], Selçuk Bayracı[2]

[1] Bilgisayar Mühendisliği Bölümü, Elektrik-Elektronik Fakültesi
Yıldız Teknik Üniversitesi, İstanbul
[2] Ar-Ge Merkezi, Cybersoft, İstanbul
yunus.emre.demirbulut@std.yildiz.edu.tr, aktas@yildiz.edu.tr,
kalipsiz@yildiz.edu.tr, selcuk.bayraci@cybersoft.com.tr

Özet. Kredi riski bankacılık sektöründe kritik risklerden bir tanesi olması sebebiyle, finans kuruluşları kredi verme konusunda karar verme aşamasında banka personeline yardımcı olan sistemlerin geliştirilmesine önem vermektedir. Bankalar, kredi talep eden müşterilerine kredi vermeden önce çeşitli kredi değerlendirme modellerine başvurmaktadır. Kredi skorlama çalışmalarında yaygın olarak kullanılan makine öğrenme ve istatistiksel teknikler incelenmiştir. Bu çalışmada kredi skorlama sistemlerinde kullanılabilecek farklı algoritmalar incelenmiştir. Bu araştırmada, kredi talebinde bulunan müşterilerin kredi isteğinin onaylanması veya geri çevrilmesi kararının verilmesini kolaylaştıracak bir sistem geliştirilmiştir. Geliştirilen sistem K En Yakın Komşu (KNN), C4.5 Ağacı, Yapay Sinir Ağı, Destek Vektör Makinesi (SVM), Lojistik Regresyon, Probit Regresyon, Poisson Regresyon ve Genelleştirilmiş Katkı Modeli (GAM) yöntemlerinin karşılaştırılmasını sunmaktadır. Her bir algoritmadan elde edilen sonuçlar müşterinin kredi skorunun tespit edilmesine imkân sunmaktadır. Ayrıca geliştirilen sistem kuruluşların kar oranının da artmasına olanak sağlamaktadır.

Anahtar Kelimeler: Kredi Skorlama, Makine Öğrenimi, İstatistik, Regresyon, Kredi Riski

Credit Scoring With Statistical And Machine Learning Methods

Yunus Emre Demirbulut[1], Mehmet S. Aktaş[1], Oya Kalıpsız[1], Selçuk Bayracı[2]

[1] Computer Engineering Department, Electrical-Electronic Faculty
Yıldız Technical University, İstanbul
[2] Ar-Ge Center, Cybersoft, İstanbul
yunus.emre.demirbulut@std.yildiz.edu.tr, aktas@yildiz.edu.tr,
kalipsiz@yildiz.edu.tr, selcuk.bayraci@cybersoft.com.tr

Abstract. Since credit risk is one of the most crucial risks in the banking sector, financial institutions attach importance to the development of novel credit scoring techniques in loan application processes. Banks apply various credit assessment tools before granting loans to customers who demand credit. In this paper, we conduct a comparative study which evaluates the predicting accuracy of various statistical and machine learning based algorithms in credit scoring. In this research, we developed a Java based desktop application that presents the comparison of K-Nearest Neighbors (KNN), C4.5 Tree, Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression, Probit Regression, Poisson Regression and Generalized Additive Models (GAM) models. The results obtained from each algorithm allow to determine the credit score of the customer. In addition, the developed system allows organizations to increase their profit rate.

Keywords: Credit Scoring, Machine Learning, Statistics, Regression, Credit Risk

1 Giriş

Kredi endüstrisindeki genişleme müşteri sayısını ve buna bağlı olarak bankalara olan kredi talebini artırmıştır. Artan talepler tüketici kredisi piyasasındaki rekabeti ciddi bir boyuta getirmiştir. Finansal kuruluşların, başvuruda bulunan müşterilerin kredilerini zamanında ödeyip ödeyemeyeceğini öngörmesi kritik önem arz etmektedir. Çünkü, bankaların ana gelir kaynağı, müşterilerine verdiği kredilerden gelen faiz gelirleridir. Müşterilerin kredilerini ödeyememesi durumunda banka para kaybına uğramaktadır.

Kredi skorlaması yukarıda bahsedilen sorunun yaşanmaması veya minimum seviyede yaşanması için kullanılan en yaygın çözümdür. Bankalar, kredi başvurusunda bulunan müşterinin nicel ve nitel verilerini kullanarak, çeşitli kredi skorlama modelleri vasıtasıyla müşterinin risk profilini çıkarmaktadır. Bu değerlendirmeler sonucunda kredi talebi onaylanmakta veya reddedilmektedir. Kredi skorlama modelleri aracılığıyla müşterinin temerrüt olasılığı hesaplanabilmekte veya müşteriler farklı temerrüt gruplarına ayrılabilmektedir. Bireysel krediler için kredi skorlama modellerinin kullandığı müşteri özellikleri, gelir düzeyi, sahip olunan varlıklar, yaş ve iş bilgisi bulunmaktadır. Kurumsal kredilerde ise borç-özsermaye gibi finansal oranlar önem kazanmaktadır.

Literatürde kredi skorlama için farklı yöntemler kullanılmasına rağmen, hangi yöntemin daha iyi sonuç verdiğine dair bir uzlaşı sağlanamamıştır. Bu açıdan, farklı yöntemler kullanmak ve performanslarını karşılaştırmalı olarak test etmek, sağlam ve doğru risk fiyatlaması yapmak açısından önemlidir. Bu çalışmada, bireysel tüketici kredilerinin riskini hesaplamak ve müşterileri geri ödeme durumlarına göre değerlendirmek amacıyla istatistiksel ve makine öğrenimi bazlı 8 farklı sınıflandırma yöntemi kullanılmıştır. Araştırma kapsamında kullanılan istatistiksel algoritmalar; Lojistik Regresyon, Probit Regresyon, Poisson Regresyon ve Genelleştirilmiş Katkı Modelidir. Makine öğrenimi algoritmalarından; k-En Yakın Komşular, C4.5 Karar Ağacı, Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) kullanılmıştır. Algoritmaların karartılmış bankacılık veri setleri üzerinde koşturulması sonucu elde edilen sonuçlar, Duyarlılık, Özgüllük, Doğruluk, ROC Eğrisi Altında Kalan Alan, Gini Katsayısı ve Zaman olmak üzere 6 farklı metriğe göre karşılaştırılmıştır.

Bildirinin devam eden bölümlerinde sırasıyla; Bölüm 2'de; kredi skorlama hakkında daha önceden yapılmış benzer çalışmalar anlatılacaktır. Bölüm 3'te; kullanılan istatistiksel ve makine öğrenimi algoritmalarının tanımı yapılacak ve kredi skorlama modellerinin nasıl karşılaştırıldığı açıklanacaktır. Bölüm 4'te; yapılan çalışma değerlendirilecek. Bölüm 5'te; uygulamaya ait sonuçlar incelenecek ve gelecek çalışmalar hakkında tavsiyelerde bulunulacaktır.

2 İlgili Çalışmalar

Literatür incelendiğinde, kredi skorlama konusu üzerinde yapılmış birçok çalışma bulunmaktadır.

Altman [5] tarafından gerçekleştirilen bir başka çalışmada, kurumsal iflas öngörüsü problemi için geleneksel oran analizi yöntemleri yerine diskriminant analiz yöntemi kullanılmıştır. Diskriminant analiz modeli, iflas etmiş ve iflas etmeyen grupların sınıflandırmasında %95 doğruluk oranına ulaşmıştır.

Salome Tabagari [1] tarafında yapılan çalışmada, müşterilerin kredi skorlarının nasıl hesaplanacağını göstermek için bir bankaya ait ve kredi talebinde bulunan 500 müşterinin bilgilerinden oluşan bir veri seti kullanılmıştır. Bu amaçla en sık kullanılan metotlardan bir tanesi olan lojistik regresyon yöntemi kullanılmıştır. Yapılan çalışma sonucunda %82,8 değerinde doğruluk elde edilmiştir.

Desai [2] tarafından yapılan araştırmada, YSA, doğrusal diskriminant analizi (LDA) ve Lojistik Regresyon olmak üzere 3 farklı algoritma kullanılmıştır. Çalışmada kullanılan veri seti 3 farklı kredi birliğinden toplanan ve kredi talebinde bulunan müşterilere ait bilgilerden oluşmaktadır. Çalışma sonucunda, kötü sınıfa ait kredilerin sınıflandırılmasında YSA yüksek doğruluk oranı sunmaktadır. Bununla birlikte, iyi ve kötü sınıfa ait kredilerin sınıflandırılmasında Lojistik Regresyon ve YSA yakın sonuçlar içermektedir. LDA yönteminin sonuçları ise diğer algoritmaların oldukça gerisinde kalmıştır.

Moares [3] tarafından yapılan çalışmada, müşterilerin kredi profillerini iyi veya kötü olarak sınıflandırmak için C4.5 karar ağacı ve YSA kullanılmıştır. Yapılan çalışma sonucunda, C4.5 karar ağacı ile %90.07 doğruluk oranı elde edilmiştir. YSA ise %95,58'lik bir doğruluk oranı sunmuştur.

Ceren [4] tarafından yapılan çalışmada, şirketlerin başarısızlığını finansal oranlara dayanarak öngörmek için DVM algoritmasını uygulamış ve sonuçları Lojistik Regresyon ile karşılaştırmıştır. DVM algoritmasının doğruluğunun artırılması için çalışmada ızgara arama yöntemi kullanılmıştır. Çalışma sonucunda DVM algoritması ile %75 oranında doğruluk değeri elde etmiştir. Buna karşın Lojistik Regresyon ile %71,8 oranında doğruluk değerine ulaşmıştır.

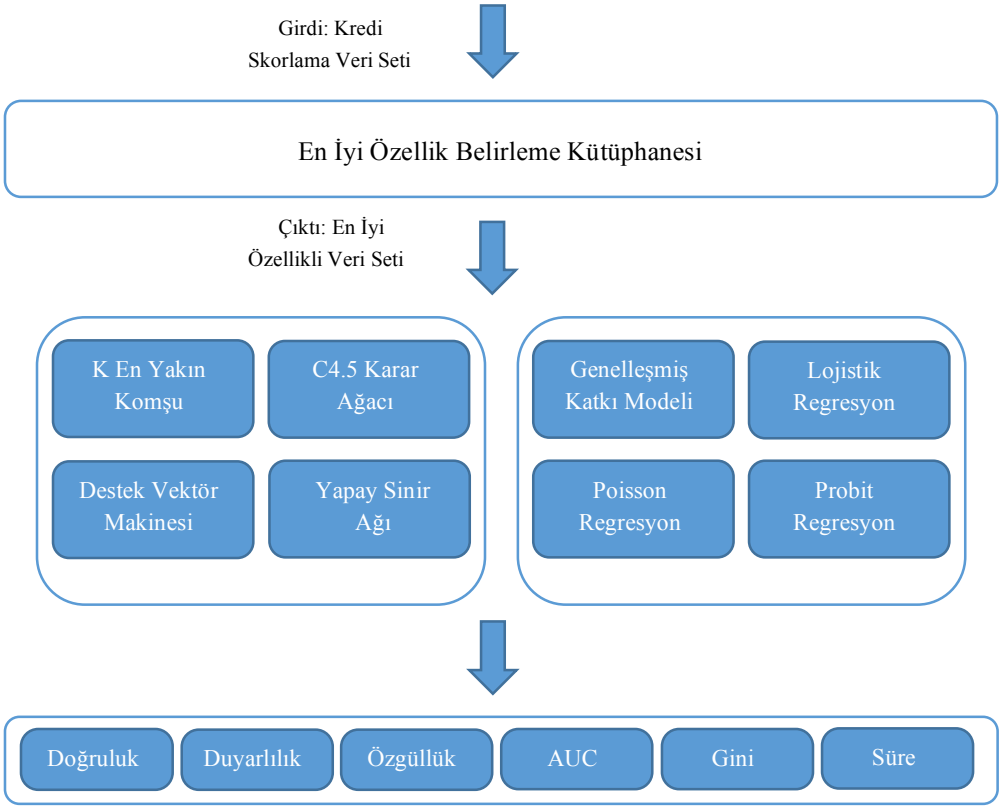
Bart [6] tarafından yapılan araştırmada, kredi skorlaması için en gelişmiş sınıflandırma algoritmaları karşılaştırılmıştır. Çalışma sonucunda, 41 farklı sınıflandırıcının sınıflandırma performansı, gerçek dünya kredi skorlama veri setleri ile karşılaştırılmıştır. Sonuçlara göre YSA algoritmalarının en güçlü sınıflandırıcı olduğu tespit edilmiştir.

Yang [7] tarafında yapılan bir çalışmada ise, bir çekirdek öğrenme algoritması temelinde yeni bir uyarlanabilir kredi skorlama tekniği sunulmuştur. Bu yöntem, gerçek hayatta doğrusal olmayan kredi skorlama görevlerini kolaylaştırmaktadır. Ayrıca veri ön işleme ve değişken analiz için zaman maliyetinin azaltılmasına yardımcı olmaktadır.

3 Metodoloji

Geliştirilen yazılımda, bankacılık alanındaki veri setleri aracılığıyla kredi skorlama modelleri oluşturmak ve bu modellerin uygunluğunu karşılaştırmak için sekiz farklı algoritma kullanılmıştır. Sekiz algoritmanın dört tanesi istatistiksel algoritmalardan, geri kalan dört tanesi ise makine öğrenimi algoritmalarından seçilmiştir. İstatistiksel algo-

ritmalar için Lojistik Regresyon, Probit Regresyon, Poisson Regresyon ve genelleştirilmiş katkı modeli yöntemleri tercih edilmiştir. Makine öğrenimi algoritmalarından ise K En Yakın Komşu, C4.5 Karar Ağacı, DVM ve YSA kullanılmıştır. Kullanılan algoritmalar ile ilgili bilgiler Bölüm 3.1 ve Bölüm 3.2'de yer almaktadır. Sistemin genel mimarisi Şekil 1'de gösterilmektedir.



Şekil. 1. Sistemin Genel Mimarisi

3.1 Çalışmada Kullanılan İstatistiksel Algoritmalar

Lojistik Regresyon [8] (Bknz. Şekil 2), bağımlı değişkeni ikili (binary) yapıda olan veri setleri üzerinde uygulanacak bir regresyon analizidir. Diğer tüm regresyon analizlerinde olduğu gibi, lojistik regresyon da bir tahmin analizidir. Lojistik regresyonda amaç, ikili yapıdaki bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi en

uygun şekilde açıklayan modelin bulunmasıdır. Lojistik regresyon ile sınıflandırma yapılırken modele ait her bir beta katsayısı Formül 1'deki formülde yerine yazılarak elde edilir.

Algoritma 1: Lojistik Regresyon

1. Başla
2. Her bir özellik için beta katsayısını sıfıra eşitle
3. Repeat
    a. Güncel beta katsayılarını kullanarak Logit fonksiyonunu (Formül 1) veri setindeki tüm örnekler için sırasıyla çalıştır ve elde edilen sınıf bilgilerini gerçek sınıf bilgisinden çıkararak hata miktarını hesapla.
    b. Veri setindeki her bir örneğe ait hata miktarını topla.
    c. Toplam hata miktarını öğrenme katsayısı ile çarp ve sonucu tüm beta katsayılarından çıkar.
    d. Adım C'de elde edilen sonuç ile beta katsayılarını güncelle.
4. Until Beta katsayıları yakınsayana kadar.
5. Bitir

Şekil. 2. Lojistik regresyona ait sözde kod

$$\ln\left(\frac{p}{1+p}\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_n \mathrm{X}_n \quad (1)$$

Poisson Regresyon [9] (Bknz. Şekil 3), düzenli çoklu regresyonlara benzemektedir. Bağımlı değişken ise Poisson dağılımını (Bknz. Formül 2) izleyen gözlenen bir sayıdır. Böylece bağımlı değişkenin olası değerleri negatif olmayan tamsayılardır. Dolayısıyla, Poisson regresyon, lojistik regresyona benzerlik göstermektedir. Fakat bağımlı değişkenler lojistik regresyondaki gibi sınırlı değildir.

Algoritma 2: Poisson Regresyon

1. Başla
2. Her bir özellik için beta katsayısını sıfıra eşitle
3. Repeat
    a. Güncel beta katsayılarını kullanarak Poisson Dağılımı (Formül 2) fonksiyonuna ait $\boldsymbol{\lambda}$ değerini 0 ve 1 çıktı değeri için hesapla.
    b. En yüksek olasılığa sahip sonucu seç.
    c. Adım B işlemini veri setindeki tüm örnekler için sırasıyla çalıştır ve elde edilen sonuçları gerçek sınıf bilgisinden çıkararak hata miktarını hesapla.
    d. Veri setindeki her bir örneğe ait hata miktarını topla.
    e. Toplam hata miktarını öğrenme katsayısı ile çarp ve sonucu tüm beta katsayılarından çıkar.
    f. Adım E'de elde edilen sonuç ile beta katsayılarını güncelle.
4. Until Beta katsayıları yakınsayana kadar.

5. Bitir

Şekil. 3. Poisson regresyona ait sözde kod

$$P(X) = \lambda^x e^{-\lambda}/x! \quad (2)$$

Probit Regresyon [10] (Bknz. Şekil 4), ikili yapıdaki sonuç değişkenine sahip veri setleri üzerinde regresyon yapılmasını sağlar. Probit regresyon, bir değerin olası ikili sonucundan birine düşme ihtimalini hesaplar. Regresyon yöntemlerinden bir tanesi olan lojistik regresyona oldukça benzemektedir. Fakat probit regresyonda sınıflandırma standart normal dağılım (Bknz. Formül 3) ile hesaplanır.

Algoritma 3: Probit Regresyon

1. Başla
2. Her bir özellik için beta katsayısını sıfıra eşitle
3. Repeat
    a. Güncel beta katsayılarını kullanarak Standart Normal Dağılım fonksiyonunu (Formül 3) veri setindeki tüm örnekler için sırasıyla çalıştır ve elde edilen sınıf bilgilerini gerçek sınıf bilgisinden çıkararak hata miktarını hesapla.
    b. Veri setindeki her bir örneğe ait hata miktarını topla.
    c. Toplam hata miktarını öğrenme katsayısı ile çarp ve sonucu tüm beta katsayılarından çıkar.
    d. Adım C'de elde edilen sonuç ile beta katsayılarını güncelle.
4. Until Beta katsayıları yakınsayana kadar.
5. Bitir

Şekil. 4. Probit regresyona ait sözde kod

$$\phi^{-1}(p_i) = \sum_{k=0}^{k=n} \beta_k X_{ik} \quad (3)$$

Genelleştirilmiş Katkı Modeli [11] (Bknz. Şekil 5), lineer olmayan değişkenlerin smooth fonksiyonları ile hesaplandığı bir modelleme tekniğidir. Bu yöntemde lineer değişkenler lojistik regresyonda olduğu gibi logit fonksiyonu ile modellenir. Genelleştirilmiş katkı modeli ile sınıflandırma Formül 4'deki formül ile yapılmaktadır.

Algoritma 4: Genelleştirilmiş Katkı Modeli

1. Başla
2. Nümerik özellikler ile kategorik özellikleri ayrı veri setlerine böl.
3. Nümerik özellikler için spline fonksiyonu hesapla.
4. Her bir kategorik özellik için beta katsayısını sıfıra eşitle
5. Repeat
    a. Güncel beta katsayılarını kullanarak Logit fonksiyonunu veri setindeki tüm örnekler için sırasıyla çalıştır ve elde edilen sınıf bilgilerini gerçek sınıf bilgisinden çıkararak hata miktarını hesapla.

b. Veri setindeki her bir örneğe ait hata miktarını topla.
c. Toplam hata miktarını öğrenme katsayısı ile çarp ve sonucu tüm beta katsayılarından çıkar.
d. Adım C'de elde edilen sonuç ile beta katsayılarını güncelle.

6. Until Beta katsayıları yakınsayana kadar.
7. Bitir

Şekil. 5. Genelleştirilmiş katkı modeline ait sözde kod

$$g\big(E(y)\big) = \alpha + s_1(X_1) + \beta_1 X_2 + \cdots s_n(X_n) + \beta_m X_k \quad (4)$$

3.2 Çalışmada Kullanılan Makine Öğrenimi Algoritmaları

K En Yakın Komşu [13] algoritması (Bknz. Şekil 6), sınıflandırma işlemi esnasında veri setine ait özelliklerden, sınıflandırılacak olan yeni örneğin daha önceki örneklerden k tanesine olan yakınlığına bakılmasıdır.

Algoritma 5: K En Yakın Komşu

1. Başla
2. K değerini belirle.
3. Sınıflandırılmak istenen örneğin veri setindeki tüm örnekler ile olan Öklid, Manhattan veya Chebyshev uzaklığını hesapla.
4. Uzaklıkları küçükten büyüğe doğru sırala.
5. Sıralanmış uzaklıklardan ilk k tane komşuyu seç.
6. K tane komşunun kategorilerini topla.
7. En uygun kategoriyi seç.
8. Bitir

Şekil. 6. K En yakın komşu algoritmasına ait sözde kod

C4.5 Karar Ağacı [14] (Bknz. Şekil 7), bilgi entropisi kavramını temel alarak eğitim veri setini kullanarak bir karar ağacı üretir. Karar ağacında bölüm kriteri bilgi kazanımı değeridir. En yüksek bilgi kazanımına sahip özellik ağaca eklenir. Bu işlem veri setindeki tüm özellikler için uygulanarak ağaç oluşturulur.

Algoritma 6: C4.5 Karar Ağacı

1. Başla
2. MaxBilgi değişkenini sıfıra eşitle.
3. Repeat
4. Güncel veri setinin entropisini hesapla.
5. FOR veri setindeki her bir özelik
    a. Güncel veri setindeki özelliklerden bir tanesini seç.
    b. Seçilen özellik için bilgi değerini hesapla.
    c. Seçilen özellik için bilgi kazanımını hesapla.
    d. IF bilgi kazanımı > MaxBilgi then
        i. MaxBilgi değişkenine bilgi kazanımı değerini ata.
6. ENDFOR

7. MaxBilgi değerine sahip özelliği ağaca ekle.
8. MaxBilgi değerine sahip özeliği veri setinden sil.
9. Until veri setindeki özellik sayısı = 0
10. Bitir

Şekil. 7. C4.5 Karar Ağacı sözde kod

Destek Vektör Makineleri [12], girdi olarak verilen veri setindeki veriler arasındaki ilişkilerin bilinmediği durumlarda kullanılmak üzere geliştirilmiş bir sınıflandırma algoritmasıdır. Destek vektör makinesinin amacı, veri setinde bulunan tüm sınıflara en uzak hyperplane bulmaktır. Test işlemi hyperplane kullanılarak gerçekleştirilir.

Yapay sinir ağları [15], yapay nöronlardan oluşur ve insan beynin basit bir modelini gerçeklemektedirler. Fakat bu ağlar gerçek beyin yapısı işle karşılaştırıldığında çok basit kalmaktadır. Optimum ağırlık değerlerinin bulunması doğruluk oranını artırır. Yapay sinir ağları, eğitim veri setinden faydalanarak yapısında bulunan nöronlar arsındaki bağlantıların ağırlıklarını bulmaya çalışır.

3.3 İstatistiksel ve Makine Öğrenimi Modellerinin Karşılaştırılması

Modellerin karşılaştırma işlemi için aşağıdaki adımlar takip edilmiştir.

- Veri setine ait özelliklerin tipi belirlenir. İki tip özellik kullanılmaktadır; nümerik ve kategorik.
- Veri seti özellik seçimi işleminden geçirilir.
- Veri seti dört tanesi istatistiksel, dört tanesi makine öğrenimi olmak üzere toplam sekiz algoritma üzerinde koşturulur.
- Algoritmaların çalıştırılması sonucu elde edilen modeller test edilerek her bir algoritmaya ait karmaşıklık matrisi elde edilir.
- Karmaşıklık matrisinden faydalanılarak modelin doğruluk, özgüllük, duyarlılık, ROC eğrisi altında kalan alan ve gini katsayısı elde edilir.
- ROC Eğrisi altından kalan alan ve her bir modelin oluşturulma süresi bir grafik ile sunulur.

Karmaşıklık matrisi, bir sınıflandırma işleminin tahmini sonuçlarının özetidir. Her sınıfa ait doğru ve yanlış tahminlerin özetini sunar. Sınıflandırma modelinin test edilmesi sonucu dört farklı çıktı elde edilir. Bu çıktılar;

- True Positive (TP): Doğru pozitif tahminlerin sayısı.
- False Positive (FP): Yanlış pozitif tahminlerin sayısı.
- True Negative (TN): Doğru negatif tahminlerin sayısı.
- False Negative (FN): Yanlış negatif tahminlerin sayısı.

Doğruluk (Bknz. Formül 5), doğru tahminlerin sayısının, test veri setindeki tüm örneklerin sayısına bölümünden elde edilir.

$$Doğruluk = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (5)$$

Duyarlılık (Bknz. Formül 6), doğru pozitif tahminlerin sayısının, test veri setindeki tüm pozitif örneklerin sayısına bölümünden elde edilir.

$$Duyarlılık = \frac{TP}{(TP+FN)} \tag{6}$$

Özgüllük (Bknz. Formül 7), doğru negatif tahminlerin sayısının, test veri setindeki tüm negatif örneklerin sayısına bölümünden elde edilir.

$$Özgüllük = \frac{TN}{(TN+FP)} \tag{7}$$

4 Değerlendirme

Araştırma kapsamında geliştirilen uygulama üzerinde gerçekleştirilen testler, Intel Core i5 2.4 GHz işlemcili ve 8GB RAM sahip bilgisayar ile yapıldı. Testler uygulama içerisindeki tüm algoritmalar için gerçekleştirilmiştir. Algoritmalar 5KB ve 10KB olmak üzere iki farklı veri seti üzerinde 100 defa çalıştırılmıştır ve algoritmalara ait sonuçlar gösterilene kadar geçen süre ortalama ve standart sapma olarak Tablo 1'de gösterilmiştir. Çalışma süreleri hesaplanırken Java programlama dili içerisinde bulunan Timestamp sınıfı kullanılmıştır ve süreler nanosaniye olarak elde edilmiştir. Uygulamanın çalışma süresi açısından maliyeti veri setinin boyutuna göre değişkenlik göstermektedir.

Tablo 1. Algoritmalara Ait Zaman Maliyetleri

Algoritmalar	5 KB		10 KB	
	Ort. (sn)	**Std. (sn)**	**Ort. (sn)**	**Std. (sn)**
KNN	0.0013	0.0019	0.0047	0.003
C4.5	0.0458	0.028	0.1737	0.0899
SVM	0.0571	0.0339	0.0743	0.0372
ANN	0.2876	0.0381	0.5483	0.0481
GAM	0.8544	0.0277	4.0316	0.1881
Lojistik Regresyon	3.3540	0.0339	6.9897	0.352
Poisson Regresyon	3.7459	0.0370	3.5936	0.2026
Probit Regresyon	3.7104	0.0522	7.7404	0.3915

Sonuçlar incelendiğinde, makine öğrenimi algoritmalarının çalışma süreleri ile istatistiksel algoritmaların çalışma süreleri arasında belirgin bir fark olduğu görülmektedir. İstatistiksel algoritmalar, kategorik verileri kukla verilere dönüştürdüğü için veri setinin boyu enine büyümektedir dolayısıyla model oluşturma süresi artmaktadır. Makine öğrenimi algoritmalarının istatistiksel algoritmalara oranla zaman maliyeti açısından daha verimli olduğu gözlemlenmiştir.

Tablo 2. Algoritmalara Ait Sonuçlar

Algoritmalar	Duyarlılık	Özgüllük	Doğruluk	AUC	Gini
KNN	0.9588	0.2173	0.8614	0.622	0.244
C4.5	0.9161	0.1978	0.8209	0.5831	0.1662
SVM	1.0	0.0	0.8685	0.5	0.5
ANN	0.9638	0.2826	0.8742	0.8148	0.6296
GAM	0.9983	0.0869	0.8785	0.7319	0.4838
Lojistik Regresyon	0.9851	0.2826	0.8928	0.8027	0.6054
Poisson Regresyon	1.0	0.0	0.8685	0.5085	0.0171
Probit Regresyon	0.9819	0.3478	0.8985	0.8068	0.6837

Tablo 2, istatistiksel ve makine öğrenimi algoritmalarının karşılaştırılabilmesi için gerekli bilgileri içermektedir. Sekiz farklı sınıflandırma algoritmasının veri seti üzerinde koşturulması sonucu elde edilen karmaşıklık matrisleri sistem tarafından değerlendirilip tablodaki bilgiler elde edilmiştir. Bu sonuçlar incelendiğinde, istatistiksel yöntemler ile makine öğrenimi yöntemlerinin yakın sonuçlar elde ettiği görülmektedir. Sınıflama modellerinin doğruluğu açısından istatistiksel yöntemlerin makine öğrenimi yöntemlerinden daha başarılı olduğu gözlemlenmiştir. Tablodaki AUC değerleri dikkate alındığında, en yüksek skora sahip algoritmanın yapay sinir ağı olduğu görülmektedir. AUC değeri, hangi modelin sınıflandırma için daha başarılı olduğunu göstermektedir. Yapay sinir ağı algoritmasının AUC değeri en yüksek olmasına rağmen, istatistiksel algoritmaların AUC ortalamalarının, makine öğrenimi algoritmalarının AUC ortalamalarından daha yüksek olduğu gözlemlenmiştir.

5 Sonuçlar ve Gelecekteki Çalışmalar

Bu araştırma kapsamında, bankacılık alanındaki veri setleri üzerinde koşturulan ve en uygun kredi skorlama modelinin tespitini sağlayan bir çözüm geliştirilmiştir. Araştırma kapsamında güncel makine öğrenimi ve istatistiksel sınıflandırma algoritmaları kullanılmıştır. Algoritmaların çalıştırılması sonucu elde edilen modellerden faydalanılarak bankaların müşterilerinin kredi taleplerini değerlendirme aşamasında karar verme süreçleri hızlandırabilmektedir.

Üçüncü bölümde detaylı bir şekilde incelenen algoritmalar makine öğrenimi algoritmalarının süre maliyeti açısından daha verimli olduğu gözlemlenmiştir. Regresyon algoritmaları kategorik verilerden ziyade nümerik verilerden oluşan veri setlerinde daha verimlidir. Regresyon algoritmaları kategorik verileri kukla verilere dönüştürerek veri setinin boyutunu enine artırmaktadır fakat makine öğrenimi algoritmaları veri setinin boyutunu sabit tutamaktadır. Sistemin doğru çalışabilmesi için, sisteme yüklenen veri setinin kredi skorlama işlemine uygun olması ve eksik veri analizi vb. ön inceleme işlemlerine tabii tutulmuş olması gerekmektedir.

Gelecek çalışmalarda, Random Forest ve Naive Bayes gibi güncel ve sıkça kullanılan makine öğrenimi algoritmaları sisteme dahil edilebilir. Bunlara ek olarak, kümeleme analizi ve bulanık mantık (fuzyy logic) gibi metodolojilerin kullanılması sistemin başarısını artırabilir.

Teşekkür

Bu araştırma kapsamında veri ve çalışma ortamı sağlayan Cybersoft Ar-Ge birimine teşekkürlerimi sunarım.

Kaynaklar

1. Tabagari, Salome. *Credit scoring by logistic regression*. Diss. Tartu Ülikool, 2015.
2. Desai, Vijay S., Jonathan N. Crook, and George A. Overstreet. "A comparison of neural networks and linear scoring models in the credit union environment." *European Journal of Operational Research* 95.1 (1996): 24-37.
3. Sousa, Marcos de Moraes, and Reginaldo Santana Figueiredo. "Credit analysis using data mining: application in the case of a credit union." *JISTEM-Journal of Information Systems and Technology Management* 11.2 (2014): 379-396.
4. Önder, Ceren. "Bankruptcy prediction with support vector machines." (2010).
5. E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The journal of finance*, vol. 23, no. 4, pp. 589-609, 1968.
6. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring", *Journal of the operational research society*, vol. 54, no. 6, pp. 627-635, 2003.
7. Y. Yang, "Adaptive credit scoring with kernel learning methods", European Journal of Operational Research, vol. 183, no. 3, pp. 1521-1536, 2007.
8. Logistic Regression, http://www.statisticssolutions.com/what-is-logistic-regression/, son erişim 2017/06/16
9. Poisson Regression, https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf, son erişim 2017/06/19
10. An Introduction to Logistic and Probit Regression Models, https://liberalarts.utexas.edu/prc/_files/cs/Fall2013_Moore_Logistic_Probit_Regression.pdf, son erişim 2017/06/18
11. GAM, http://multithreaded.stitchfix.com/blog/2015/07/30/gam/, son erişim 2017/06/19
12. Destek Vektör Makineleri, https://www.slideshare.net/ozgur_dolgun/destek-vektr-makineleri, son erişim 2017/06/16
13. KNN, K-En Yakın Komşu, http://bilgisayarkavramlari.sadievrenseker.com/2008/11/17/knn-k-nearest-neighborhood-en-yakin-k-komsu/, son erişim 2017/06/16
14. C4.5 Karar Ağaçları, http://bilgisayarkavramlari.sadievrenseker.com/2012/11/13/c4-5-agaci-c4-5-tree/, son erişim 2017/06/18
15. ANN, Artificial Neural Network, https://en.wikipedia.org/wiki/Artificial_neural_network, son erişim, 2017/06/18