Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(2) (2022) 137-144
DOI: 10.38016/jista.1036047

Makine Öğrenmesi Yöntemleri ile Banka Müşterilerinin Kredi Alma Eğiliminin Karşılaştırmalı Analizi

Ali Tezcan Sarızeybek[1*], Onur Sevli[1*]

[1] Burdur Mehmet Akif Ersoy Üniversitesi, Yazılım Mühendisliği Bölümü, Bucak Teknoloji Fakültesi, Burdur, Türkiye

[2] Burdur Mehmet Akif Ersoy Üniversitesi, Bilgisayar Mühendisliği Bölümü, Mühendislik Mimarlık Fakültesi, Burdur, Türkiye

atsarizeybek@mehmetakif.edu.tr, onursevli@mehmetakif.edu.tr

Öz

Bankacılık, müşterilerle sık sık iletişime girilmesi gereken bir sektördür. Bankalar müşterilerine, onların durumlarına uygun bir kredi vermek istediğinde müşteriyi telefonla ararlar. Çoğu zaman müşteri, teklif edilen krediyi reddeder, bu da müşteriyle iletişime geçen personelin zamanından büyük bir kayıptır. Bu çalışmada, banka müşterilerinin verilerinin bulunduğu veri seti ele alınarak ve çeşitli makine öğrenmesi sınıflama modelleri kullanılarak müşterinin kredi alıp almayacağı tahmin edilmiştir. Elde edilen çalışma sonuçlarına göre, makine öğrenmesi yöntemleri ile müşterinin kredi alma eğilim tahmini başarılı bir şekilde gerçekleşmiştir. Çalışma sonucunda K-Best uygulanan modellerin arasında doğruluk değeri en yüksek olan sınıflandırıcı modelinin %98,86 ile Rastgele Orman algoritması olduğu, özellik seçimi yapılmadan eğitilen modellerin arasında en yüksek olan modelin %93,66 ile Rastgele Orman algoritması olduğu, cross-validation ve grid search ile eğitilen modellerin arasında ise en yüksek değerin %98,6 ile Rastgele Orman algoritmasında olduğu görülmüştür.

Anahtar kelimeler: Makine Öğrenmesi, Bankacılık, Kredi Tahminleme, Sınıflandırma Algoritmaları.

A Comparative Analysis of Bank Customers' Loan Propensity Using Machine Learning Methods

Abstract

Banking is a sector that requires frequent communication with customers. When banks want to give their customers a loan that suits their situation, they call the customer by phone. The often time customer rejects the loan offer, which is a huge waste of time from the staff contacting the customer. In this study, the customer's tendency to take credit was estimated using the data set of bank customers' data and various machine learning classification models. According to the results of the study, the prediction of the customer's tendency to take credit with machine learning methods has been successfully realized. As a result of the study, among the models applied K-Best, the classifier model with the highest accuracy value was found to be the Random Forest algorithm with 98.86%. Among the models trained without feature selection, the highest model was found to be the Random Forest algorithm with 93.66%. Among the models trained with cross-validation and grid search, the highest value was found in the Random Forest algorithm with 98.6%.

Keywords: Machine Learning, Banking, Credit Estimation, Classification Algorithms.

1. Giriş (Introduction)

Bu çalışma, Kranti Walke'nin hazırlamış olduğu "Bank Personal Loan Modelling" veri setinden yola çıkarak banka müşterilerinin teklif edilen krediyi alma eğilimlerinin farklı makine öğrenme yöntemleriyle tahmin edilmesini ve performans analizlerini konu almaktadır. Bankacılık sektöründe bankalar kendi müşterilerine onları telefonla arayarak ilgili müşteriye kredi teklifi yaparlar. Müşteri teklifi değerlendirir ve onay verip vermediğini belirtir; fakat bu belirtilen teklifi müşteriye ilettikten sonra alınan olumsuz cevap hem zaman hem de iş gücü almaktadır. Bunun önüne geçebilmek için müşterilerin verileri analiz yapılmalı elde edilen bilgilerden yola çıkarak kredi alma

* Sorumlu yazar.
E-posta adresi: atsarizeybek@mehmetakif.edu.tr

Alındı : 13 Aralık 2021
Revizyon : 8 Şubat 2022
Kabul : 22 Şubat 2022

eğilimlerini tahmin etmek zaman ve iş gücü tasarrufu açısından büyük yarar sağlayacaktır. Müşteri sayısının az olduğu durumlarda insan eli ile analiz yapmak basittir ancak müşteri sayısının çok fazla olduğu bir bankada analiz yapmak zor, hatta imkânsız hale gelir. Bu yüzden bilgisayarda incelemelerin yapılması, müşterinin kredi alma profilinin çıkarılması gerekir, bunun için de devreye makine öğrenmesi girmelidir. Sonuç sağlayabilmek için veri setinden müşterilerin maaş, aylık kredi kartı kullanımı, limit bilgileri gibi verilerin incelenerek bir makine öğrenmesi modeli oluşturulmalıdır. Müşteri verileri analiz edilerek müşterinin krediyi alıp almayacağı tahmin edildiği için diğer özellikler hedef alınarak farklı çıkarımlarda da bulunulabilir, örnek olarak kredi kart harcamalarından aylık gelir tahmini yapılabilir.

Bu çalışmada yapılan işlem sonucunda müşterinin kredi alma eğilimi doğru tahmin edilirse müşteriye yapılacak teklif tekrardan değerlendirilebilir, teklif paketi kapsamı bu çıkarıma göre genişletilebilir ya da daraltılabilir. Bu sayede banka müşteriye yüksek veya düşük bir teklifte bulunmaz. Müşterinin ihtiyacı doğrultusunda teklif yapılacağı için de müşteri servisten memnun kalır ve bankanın sadakat puanı yükselir. Teknolojinin hızla ilerlediği bu yıllarda bankacılık sektörünün de ilerlemesi gerekmektedir. Yapay zekanın bankacılık sistemlerinde uygulanması müşteri-banka arasındaki köprüyü sağlamlaştırmakla kalmayıp yapacağı çıkarımlarla müşteri ve bankanın detaylı analizlerinin yapılmasını sağlar. Bu sayede veriler ve müşteri ilişkileri sağlamlaşır. Bu çalışmada, bir bankanın müşteri verilerinden kredi alma eğilimi tahmin edilmeye çalışılmış, bunun için makine öğrenme modeli eğitilmiş, test verilerinin sonuçları ve gerçek verilerin kredi alma eğilimlerinin sonuçları karşılaştırılarak performans analizi yapılmıştır. Bu çalışmada elde edilmesi amaçlanan başarım makine öğrenmesi yöntemlerinde çapraz doğrulama ve ızgara arama yöntemlerinin sonuçlara nasıl katkı sağladığını gözlemlemek ve benzer veri seti kullanılmış çalışmalarla karşılaştırmaktır. Literatürde bu veri setini kullanan başka bir çalışma bulunmamış, bu yüzden müşterilerin davranışlarının tahminlemesi üzerine yapılan bu çalışmada diğer modellerin performans sonuçları ile karşılaştırması yapılmamıştır, literatürde bu veri seti ilk kez kullanıldığı için literatürde ilk olacaktır. Çalışmada sınıflandırma yöntemleri olarak k-En Yakın Komşu, Lojistik Regresyon, Karar Ağaçları, Rastgele Orman ve Destek Vektör Makinesi yöntemleri kullanılmış, sonuç olarak müşterinin kredi alıp almayacağının tahmin edilmesi amaçlanmıştır.

1.1. Benzer çalışmalar (Similar studies)

Yapılan literatür taramasında, çalışmalarda genel olarak kredi risk tahminleri yapılmış, bu çalışmadaki gibi müşterinin krediyi kabul etme tahminlemesi değil, bankanın müşterinin başvurusunu kabul etme tahminlemesi yapılmıştır.

Arun, Ishaan ve Sanmeet (2016) tarafından yapılan çalışmalarda müşterilerin başvuruları değerlendirilerek krediye en uygun müşterinin tahminlemesi amaçlanmış ve çalışmalar yapılmıştır. Cinsiyet, medeni durum, vasilik, eğitim durumu, gelir, ek gelir, başvurulan kredi miktarı, kredi süresi, sahip olunan taşınmaz mallar gibi veriler üzerinden kredinin kabul edilip edilmeyeceğinin tahmini amaçlanmıştır. Metot olarak Karar Ağaçları, Rastgele Orman, Lineer Modeller, Neural Net ve AdaBoost kullanılmış, sınıflandırma gerçekleştirilmiştir. Performans test sonuçları makaleye verilmemiştir.

Alan (2020) tarafından yapılan çalışmada, en iyi sınıflandırma modelinin belirlenmesi için çapraz doğrulama yöntemlerinden hold-out ve k-katlı çapraz doğrulama kullanılmış, bu işlemler 32 ayrı veri setine uygulanmıştır. Veri setlerinin arasında Portekiz Bankası Pazarlama veri setinde çapraz doğrulama ile en iyi modelin %88,75 doğruluk oranına sahip olan DVM sınıflandırıcısı olduğu tespit edilmiş, AUC oranı %53,32 ve F1 skor değeri ise %61,07 olarak elde edilmiştir. Veri seti üzerindeki çalışma sonucunda doğruluk değeri %86,35, AUC değeri %60,23 ve F1 skoru ise %61,83 olarak elde edilmiştir.

Serengil, İmece, Tosun, Büyükbaş ve Köroğlu (2021) tarafından yapılan çalışmada tahsili geciken kredinin farklı sınıflandırma algoritmaları ile tahmin edilmesi ve karşılaştırılması amaçlanmıştır. Özel bir bankanın 181.276 adet müşteri verisinden oluşan bir veri seti kullanılmıştır. Veri setinde müşteri ödeme geçmişi, bilançolar, önceki kredi kartı ödemeleri, diğer bankalardan gelen risk ve limit verileri gibi özellikler bulunmaktadır. Sonuç olarak elde edilecek değerler sağlıklı, gecikmeli ve tahsili gecikmiş kredi değerleridir. Veri seti farklı modellerde test edilmiş ve karşılaştırmalar sonucunda problem için en uygun modelin 0.90 özgüllük, 0.87 kesinlik, 0.77 duyarlılık ve 0.82 F1 skoru olarak elde edilmiştir.

Yapılan literatür taramasına göre bu probleme en uygun modeli bulmak için farklı sınıflandırıcılar kullanmalı ve aşırı uyma problemini ortadan kaldırmak için grid search ve modelin en iyi durumunu görmek için çapraz doğrulama yapılmalıdır.

1.2. Sınıflandırma yöntemleri (Classification methods)

Yapılan bu çalışmada makine öğrenmesi uygulamaları için çeşitli sınıflandırma algoritmaları kullanılmıştır. Çalışma kapsamında kullanılan veri setleri üzerinde k-En Yakın Komşu (kNN), Lojistik Regresyon, Karar Ağacı ve Rastgele Orman sınıflandırıcı algoritmaları uygulanmış ve performans değerlendirme analizleri yapılmıştır. Alt başlıklarda çalışmada kullanılan algoritmalar basitçe kısaca açıklanmıştır.

1.2.1. k_En Yakın Komşu Algoritması (k-Nearest Neighbor Algorithm)

k-En Yakın Komşu Algoritması, regresyon ve sınıflandırma için kullanılan popüler algoritmalardan biridir. Adından da anlaşılacağı üzere algoritmaya verilen değerler üzerinde, verilen k sayısı kadar en yakın noktalara bakılır, verilen örneğe en yakın olan k kadar değer seçilir ve çoğunluğa sahip olan değer örneğe atanır. K sayısında çakışma olmaması için genel olarak tek sayı verilmelidir (Rasjid, Setiawan,2017). Bu mesafelerin ölçülmesinde genellikle 1. eşitlikte verilen Öklid mesafe formülünü kullanılır. (Kılınç v.d.,2016)

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (1)$$

1.2.2. Lojistik Regresyon Analizi (Logistic Regression Analysis)

Kesikli olarak iki değerin içinde, yani 1 veya 0 gibi, bir sonuç veren, kesikli veya sürekli verilerin incelenebildiği bir algoritmadır. Anlamlılık tespiti ile modelin ve parametrelerin olabilirlikleri tespit edilir ve tahminleme yapılır (Field,2009). 2. eşitlikte değerlerin gerçekleşme olasılığının formülü verilmiştir. (Coşkun v.d.,2004)

$$\Pi\,(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (2)$$

1.2.3. Karar Ağaçları (Decision Trees)

Düğüm, dal ve yaprak olmak üzere 3 adet kısımdan oluşur. Her özellik bir düğüme atanmıştır, veriler kök düğümde saklanır. Kök tarafından itibaren yukarı doğru dalı olmayan düğümlere ve yapraklara gelene kadar bu sorular sorulur. Performans iyileştirmesi yapılması için birbirini bağlayan düğümlerin bağlantısını kesip yerine yaprak koyarak budama yapılmalı, veri ile alakasız dalların gereksiz yere işlenmesinden kaçınılmalıdır. Sınıflara ait entropi 3. eşitlikte verildiği gibi hesaplanır.

$$Entropi\,(T) = \sum_{1}^{n} p_i log_2\,(p_i) \quad (3)$$

Entropi hesaplandıktan sonra ise bölünme sonucunda elde edilen kazanç da 4. eşitlik gibi hesaplanır. (Akar, Güngör, 2012)

$$Kazanç\,(B,T) = Entropi\,(T) - \sum_{i=1}^{n}\frac{|T_i|}{|T|}Entropi\,(T) \quad (4)$$

1.2.4. Rastgele Orman (Random Forest)

Regresyon ve sınıflandırma için de kullanılabilen Rastgele Orman algoritması karar ağaçlarındaki overfitting problemini giderir (Ali, vd., 2012). Rastgele karar ağaçlarından bir orman oluşturur ve bunları birleştirir. İlk aşamada oluşan sınıflandırıcıya göre tahminleme yapılır. N adet özellik içinden K kadar özellik seçilir ve seçilen özellikler arasından en iyi noktaya göre sonraki düğüm hesaplanır. Düğüm, çocuk düğümlere ayrılır ve hedeflenen düğüm sayısına ulaşana kadar özellik seçme ve çocuk düğümlere ayırma tekrarlanır. Tahminleme aşamasına gelindiğinde ise rastgele seçilen bir karar ağacının oyları hesaplanır ve en yüksek oy alan tahmin seçilir. (Biau, Scornet, 2016)

1.2.5. Destek Vektör Makinesi (Support Vector Machine)

Destek Vektör Makinesi(DVM), veri setinde değişkenler arasındaki ilişkilerin bilinmediği, sınıflara ait verileri ayırabilmek için sınıflandırma problemlerinde kullanılır. Lineer verilerde doğrular üzerinde marjini en yüksek olan doğrunun seçilmesi amaçlanmış, lineer olmayan verilerde ise veriyi daha yüksek boyutta bir uzaya taşınır ve en iyi hiper düzlem bulunur. (Akşehirli, v.d., 2013) (Burges, 1998)

1.3. Algoritmaların karşılaştırılmasında kullanılan kriterler (Criteria used in comparing algorithms)

Algoritma sınıflandırma işlemi sonucunda elde edilen hata matrisinde 4 adet değer vardır. True Positive tahminin doğru, özelliğin de doğru olması, True Negative tahminin yanlış fakat özelliğin doğru yerde olması, False Positive tahminin doğru fakat özelliğin yanlış olması, False Negative ise tahminin ve özelliğin yanlış olması durumlarıdır. (Gök, 2017) Tablo 1'de karmaşıklık matrisinin değerleri vardır ve TP, TN, FP ve FN değerleri burada yerlerine yazıldıkları değerleri alırlar.

Tablo 1. Karmaşıklık Matrisi (Confusion Matrix)

		Tahmin Edilen	
		Pozitif	Negatif
Asıl Değer	Pozitif	TP	FN
	Negatif	FP	TN

1.3.1. Doğruluk Oranı (Accuracy)

Doğruluk oranı, modelde doğru tahmin edilen verilerin tüm verilere oranı ile bulunur. Örnek olarak 100 adet veri setinde doğru tahmin edilen veri sayısı 60 ise doğruluk oranı %60 olacaktır. Doğruluk oranı hesaplamak için 5. eşitlikteki formül kullanılır.

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

1.3.2. Kesinlik (Precision)

Kesinlik oranı, modelde pozitif olarak tahmin edilen verilerin kaç adetinin gerçekte kaç tanesinin pozitif olduğunun oranıdır. Kesinlik oranı hesaplamak için 6. eşitlikteki formül kullanılır.

$$Kesinlik = \frac{TP}{TP+FP} \quad (6)$$

1.3.3. Duyarlılık (Recall)

Duyarlılık oranı, pozitif olması gereken değerlerin ne kadarının pozitif olarak tahmin edildiğini gösteren bir orandır. Duyarlılık hesaplaması için 7. eşitlikteki formül kullanılır.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (7)$$

1.3.4. F-Değeri (F-Score)

F-Skor oranı, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Uç durumların da hesaba dahil edilmesi için harmonik ortalama alınır. F-Skor hesabı için 8. eşitlikteki formül kullanılır.

$$F1 = 2 * (\frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik}) \quad (8)$$

1.3.5. Alıcı Çalıştırma Karakteristik Eğrisi (Receiver Operating Characteristic Curve)

Alıcı Çalıştırma Karakteristik Eğrisi (ROC), tüm sınıflandırma eşiklerindeki sınıflandırıcının performansını gösterir. Sınıflandırma eşiğini düşürmek daha fazla değerin sınıflandırılması demektir. ROC eğrinin True Positive Rate (TPR) ve False Positive Rate (FPR) olmak üzere iki adet parametresi vardır. TPR, duyarlılığa benzerdir ve 9. eşitlik gibi hesaplanır.

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

FPR ise 10. eşitlik gibi hesaplanır.

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

1.3.6. AUC Oranı (Area Under the Curve Rate)

Area Under the Curve (AUC), ROC eğrisinin altında kalan alanı hesaplar, tüm sınıflandırma eşikleri arasında performans hesaplarının toplamlarını gösterir. AUC oranı 0-1 arasındadır, %100 hatalı tahmin eden bir modelin AUC oranı 0.0 iken %100 doğru tahmin eden bir modelin AUC oranı 1.0'dır. (Zhang, 2016)

2. YÖNTEM (Method)

2.1. Veri seti (Data set)

Bu yazıda Kranti Walke'nin hazırlamış olduğu "Bank Personal Loan Modelling" adlı, Thera Bank adında bir bankanın müşterilerinin verilerini içeren bir veri seti kullanılmıştır. Veri setinde yaş, deneyim, gelir, posta kodu, ailedeki kişi sayısı, aylık kredi kart harcama ortalaması, eğitim durumu, eğer varsa evinin mortgage değeri, menkul kıymet hesap varlığı, mevduat hesap varlığı, online işlem kullanıcısı olup olmadığı, kredi kart varlığı ve son teklif edilen kredi teklifini kabul edip etmediği özelliklerini içerir. Veri seti 14 öznitelik ve 5000 adet örnekten oluşmaktadır. Veri setine ait öznitelikler ve açıklamaları Tablo 2'de verilmiştir.

Veri seti içindeki eşsiz veri sayısı, ortalama, medyan ve standart sapma değerleri Tablo 3'te verilmiştir

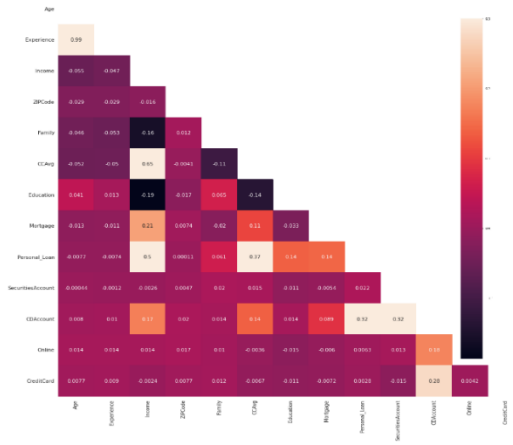
Tablo 2. Banka müşterileri veri seti öznitelikleri ve açıklamaları (Bank customers dataset attributes and descriptions)

Öznitelik	Açıklama
Age	Tamamlanan yıl olarak yaşı
Experience	Yıl olarak iş deneyimi
Income	Kişinin yıllık geliri (Gelir = Veri * 1000 $)
ZIPCode	Bulunduğu yerin posta kodu
Family	Yaşadığı yerdeki veya ailesinde bulunan kişi sayısı
CCAvg	Aylık ortalama kredi kartı harcama (Harcama = Veri * 1000 $)
Education	Eğitim Durumu 1 = Lisans 2 = Lisansüstü 3 = Doktora
Mortgage	Eğer varsa evinin mortgage değeri (Değer = Veri * 1000 $)
Personal_Loan	Son kampanyada müşterinin kredi teklifini kabul edip etmemesi 1 = Evet 0 = Hayır
SecuritiesAccount	Müşterinin bankada mevduat hesabının varlığı 1 = Var 0 = Yok
CDAccount	Müşterinin bankada menkul kıymet hesabının varlığı 1 = Var 0 = Yok
Online	Müşteri bankanın internet bankacılığını kullanıyor mu? 1 = Evet 0 = Hayır
CreditCard	Müşteri kredi kartı kullanıyor mu? 1 = Evet 0 = Hayır

Tablo 3. Veri setindeki eşsiz veri, standart sapma, ortalama ve medyan değerleri (Unique data, standard deviation, mean and median values in the data set)

Özellik	Eşsiz Veri Sayısı	Standart Sapma	Ortalama	Medyan	Min	Max
Age	45	11.463166	45.3384	45	23	67
Experience	47	11.467954	20.1046	20	0	43
Income	162	46.033729	73.7742	64	8	224
ZIPCode	467	2121.852197	93152.503	93437	9307	96651
Family	4	1.147663	2.3964	2	1	4
CCAvg	108	1.747659	1.937938	1.5	0	10
Education	3	0.839869	1.881	2	1	3
Mortgage	347	101.713802	56.498	0	0	635
Personal_Loan	2	0.294621	0.096	0	0	1
SecuritiesAccount	2	0.305809	0.1044	0	0	1
CDAccount	2	0.238250	0.0604	0	0	1
Online	2	0.490589	0.5968	1	0	1
CreditCard	2	0.455637	0.294	0	0	1

Veri setinin korelasyon matrisi ve sıcaklık haritası Şekil 1'de verilmiştir.



Şekil 1. Veri setinin sıcaklık haritası (The temperature map of the dataset)

2.2. Nitelik seçme (Feature selection)

Bu veri setinde özellikler, tek değişkenli analiz yöntemi olan K-Best yöntemi ile çıkarılmıştır. K-Best yöntemi, veri seti üzerinde hedefe yönelik tek değişkenli istatistik testler yapar ve en yüksek skor yapan özellikleri belirtilen k sayısı kadar alır. Bu çalışma için k sayısı, çalışmalara göre doğruluk ve kesinlik oranlarının en yüksek olduğu, 10 olarak belirlenmiştir, 9 olduğunda en yüksek doğruluk değeri %94, 8 olduğunda ise %88 çıkmaktadır. 11 Olduğunda da aynı şekilde doğruluk değeri düşmektedir. Özellik seçme aşamasının sonucunda özellik olarak kullanılarak sütunlar: Yaş (Age), gelir (Income), ailedeki kişi sayısı (Family), kredi kartı aylık ortalama harcama (CCAvg), eğitim durumu (Education), mortgage durumu (Mortgage), menkul kıymet hesap varlığı (SecuritiesAccount), mevduat hesap varlığı (CDAccount), online işlem kullanıcısı olup olmadığı (Online) ve kredi kart kullanıcısı olup olmadığı (CreditCard) sütunlarıdır, hedef sütun ise kredi alıp almadığı (Personal Loan) sütunudur, yani tahmin edilmek istenen veri bu sütundadır.

3. Çalışma Sonuçları (Study Results)

Bu çalışmada makine öğrenmesi sınıflandırma algoritmalarından kNN, Lojistik Regresyon, Karar Ağacı ve Rastgele Orman algoritmaları kullanılmıştır. Veri setinde model oluşturmak için veriler %70 eğitim, %30 test olarak ayrılmıştır. Bütün algoritmalarda rastgele durum 0 olarak belirlenmiştir. Rastgele Orman algoritmasında ağaç sayısı 10 olarak belirlenmiştir. kNN algoritmasında komşu sayısı 3 olarak belirlenmiştir. Rastgele Orman algoritmasında ağaç sayısı ve kNN algoritmasında komşu sayısı, sayılar denenip sonuç olarak doğruluk oranı en yüksek elde edilen sayılar alınmıştır. Veri seti üzerindeki null değerler silinmiş, bazı verilerdeki deneyim değeri 0'dan küçük olan veriler saptanmış, bu veriler silinmiştir.

Müşterilerden kredi teklifini reddedenlerin kredi kart harcama ortalamalarının orta değeri 1400 olarak elde edilmiş, kredi teklifini reddedenlerin kredi kart harcama ortalamaları ise 3800 olarak elde edilmiştir.

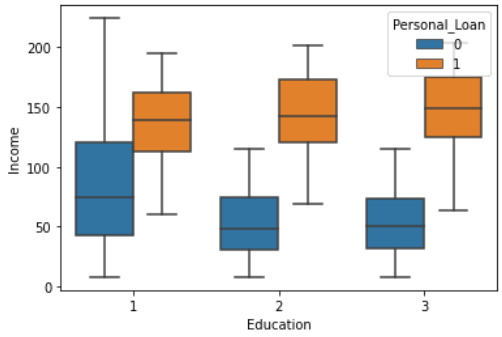
Tüm özelliklerin ortalaması Tablo 4'te verilmiştir.

Ortalama değerlere göre kredi teklifini kabul edenler müşterilerin %9,6'lık bir kısmını oluşturmaktadır. Bunun yanında internet bankacılığı kullananların oranı %59,68, kredi karta sahip olan müşteri oranı %29,4, aylık ortalama kredi kartı harcaması ise 1937,938$'dır. Yıllık gelir ortalaması ise 73.774,20$'dır.

Eğitim durumu ve yıllık gelir açısından kredi teklifini kabul etme durumu Şekil 2.'de verilmiştir.

Tablo 4. Özelliklerin ortalama değerleri (Average values of features)

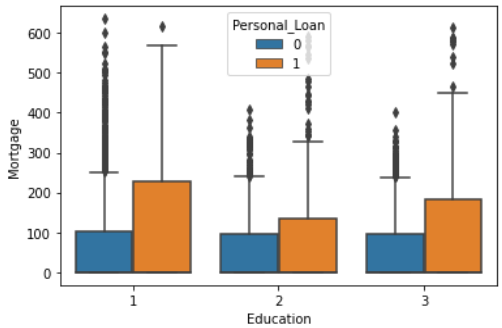
Özellik	Ortalama
Age	45.338400
Experience	20.104600
Income	73.774200
ZIPCode	93152.503000
CCAvg	1.937938
Education	1.881000
Mortgage	56.498800
Personal_Loan	0.096000
SecuritiesAccount	0.1044000
CDAccount	0.060400
Online	0.596800
CreditCard	0.294000



Şekil 2. Eğitim durumu ve yıllık gelire göre kredi teklifini kabul etme durumu kutu grafiği (Box chart of accepting loan offer by education and annual income)

Şekil 2'deki grafiğe göre geliri yüksek olan müşteriler daha çok kredi alma eğilimine sahip iken, lisans mezunu olan müşterilerde teklifi reddedenlerin gelir aralığı biraz daha fazladır.

Eğitim durumu ve mortgage değerine göre kredi teklifini kabul etme durumu Şekil 3'te verilmiştir.



Şekil 3. Eğitim durumu ve mortgage değerine göre kredi teklifini kabul etme durumu kuru grafiği (Graph of accepting a loan offer by education level and mortgage value)

Şekil 3'teki sıcaklık haritasına göre kredi teklifini kabul etme durumunun en çok yıllık gelir ve kredi kartı harcaması ile ilişkisi vardır.

Veri setinden ID, ZIPCode ve Experience özellikleri çıkarılarak %70 eğitim ve %30 test olarak ayrılmışlardır. Eğitim ve test için ayrılan veriler ölçeklenmiştir. ID ve ZIPCode özelliklerinin çıkarılma sebebi sonuca ulaşmak için gerekli olan özellikler olmadıkları içindir. Experience özelliğinin çıkarılma sebebi ise Age özelliği ile neredeyse aynı olduğu ve bu yüzden tekrarlı bir veri gibi görüleceği için çıkarılmıştır.

Tüm sınıflandırma algoritmalarına göre veriler modellenmiş ve doğruluk, kesinlik, duyarlılık ve F-Skor değerleri Tablo 5'te verilmiş ve doğruluk oranlarına göre bir sıraya konulmuştur. Elde edilen sonuçlara göre bu çalışma için uygulanabilecek en kötü algoritmanın Lojistik Regresyon yöntemi, en iyi algoritmanın ise Rastgele Orman algoritması olduğu gösterilmektedir.

Tablo 5. K-Best ile Model performans değerleri (Model performance values with K-Best)

İsim	Kesinlik	Duyarlılık	F-Skor	Doğruluk	Sıra
Lojistik Regresyon	0.93	0.95	0.98	0.95	5
Karar Ağacı	0.93	0.95	0.94	0.9793	3
Rastgele orman	0.99	0.94	0.96	0.9886	1
DVM	0.96	0.92	0.93	0.9832	2
kNN	0.96	0.83	0.88	0.966	4

Tablo 6. Özellik seçimi yapılmadan ölçülen model performans değerleri (Model performance values measured without feature selection)

İsim	Kesinlik	Duyarlılık	F-Skor	Doğruluk	Sıra
Lojistik Regresyon	0.88	0.94	0.91	0.93666	2
Karar Ağacı	0.90	0.89	0.89	0.88733	5
Rastgele orman	0.91	0.94	0.91	0.93666	1
DVM	0.88	0.91	0.90	0.91544	4
kNN	0.90	0.93	0.91	0.93133	3

Tablo 6'daki özellik seçimi yapılmadan uygulandığında elde edilen performans sonuçlarına göre 0.91 kesinlik ve 0.93666 doğruluk oranı ile Rastgele Orman algoritmasının en iyi sonuç verdiği elde edilmiştir. Tablo 6'ya ve Tablo 5'e göre özellik seçimi yapıldıktan sonra performans iyileşmesi gözlemlenmiştir.

Çalışma sonrasında aşırı uymayı (overfitting) engellemek için Grid Search (ızgara arama) ve bağımsız veri setinde en iyi modeli seçmek için çapraz doğrulama (cross-validation) uygulanıp testler uygulanmış, testin performans sonuçları Tablo 5'te verilmiştir. ROC eğrisi ise Şekil 4'te verilmiştir.

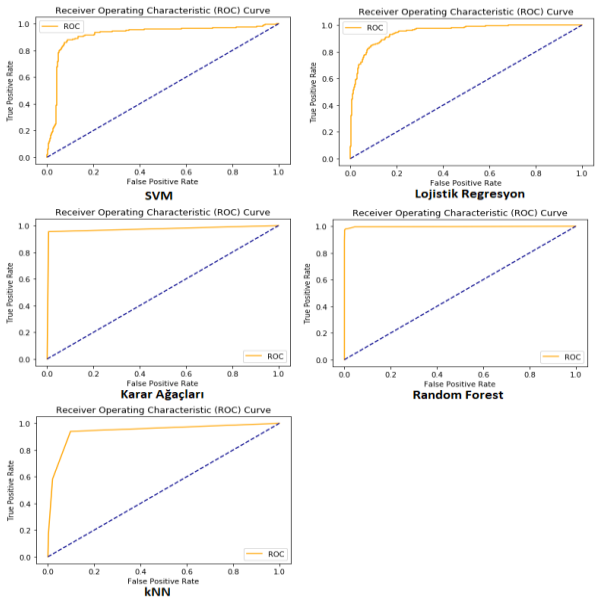
Çapraz doğrulama Stratified K-Fold yöntemi ile gerçekleştirilmiş, katlama(fold) oranı 5, diğer parametreler varsayılan olarak belirlenmiştir. Izgara aramada ise karar ağaçları algoritmasında en fazla derinlik(max depth) 2, 3, 4 olarak belirlenmiş; en az

örnek sayısı (min samples leaf) 0.12, 0.14, 0.16 ve 0.18 olarak belirlenmiştir, diğer algoritmalarda ise varsayılan değerler atanmış ve ızgara arama işlemi gerçekleştirilmiştir.

Tablo 7'ye göre bu problem için en uygun olan modelin çapraz doğrulama oranı 0.997, doğruluk oranı 0.986, AUC oranı 0.998 olan Rastgele Orman algoritması olduğu elde edilmiştir.

Tablo 7. Çapraz Doğrulama ve Grid Search uygulanan modellerin performans sonuçları (Performance results of models with Cross Validation and Grid Search applied)

Model	Çapraz Doğrulama Doğruluk Oranı	Model Doğruluk Oranı	AUC Oranı
Karar Ağaçları	0.99634	0.97933	0.97721
Lojistik Regresyon	0.88495	0.92933	0.95029
Rastgele orman	0.99734	0.986	0.99882
DVM	0.9279	0.942	0.95356
kNN	0.95840	0.91333	0.92927



Şekil 4. ROC Eğrisi (Roc Curve)

4. Sonuçlar (Conclusions)

Bankalarda yapay zekâ ve makine öğrenmesi kullanımı her geçen gün artmakta, örnek olarak müşterilerin ilgi alanları ve para çekme eğilimleri gibi işlemlerin tahminlemesi yapılmaktadır. Bu tahminlemelerin yapılabilmesinin bir insan tarafından yapılması imkansıza yakındır, çünkü bankaların binlerce, hatta milyonlarca müşterisi bulunmaktadır. Bu yüzden yapay zekâ ve makine öğrenmesinin kullanılması gerekmektedir.

Bu çalışmada makine öğrenmesi sınıflandırma algoritmalarından kNN, Rastgele Orman, Lojistik Regresyon, Karar Ağacı kullanılmış, K-Best ile yapılan özellik seçiminde yapılan çalışma sonucunda elde edilen verilere göre bu çalışma için en uygun olan sınıflandırma algoritması doğruluk değeri 0,9886 ile Rastgele Orman algoritmasıdır. İkinci en iyi algoritma ise 0,9832 ile Destek Vektör Makineleri algoritmasıdır. En kötü algoritma ise 0,95 ile Lojistik Regresyon algoritmasıdır. Özelliklerin fazla olması ve korelasyon matrislerinde çoğu özelliklerin değerlerinin birbirlerine yakın olmaması yüzünden kötü bir performans beklenirken, iki özellik arasında değil, çok sayıdaki özelliğin birbirlerine bağlantılı olduklarını gösterir. Özellik seçimi yapmadan yapılan çalışma sonucunda ise 0,91 Kesinlik, 0,94 Duyarlılık, 0,91 F-Skor ve 0,93666 Doğruluk ile Rastgele Orman en yüksek doğruluğa sahip model olmuş, özellik seçimi yapmanın önemi tüm sonuçların düşmesinden anlaşılmıştır. Çapraz doğrulama ve grid search uygulanan modelde en iyi sonucu çapraz doğrulama oranı 0,997, doğruluk oranı 0,986 ve AUC oranı 0,998 ile Rastgele Orman algoritması elde etmiştir, bu da veri setinde aşırı uyma problemini ve en iyi sonuç elde edecek olan modeli elde etmeye yardımcı olmuştur.

Bu çalışmanın sonucunda müşterilere teklif gönderirken müşterilerin kabul edebilme durumlarını tahmin edebilen bir uygulamanın yapılabileceği, bu sayede müşteri temsilcilerinin iş gücünden ve zamandan tasarruf edilebileceği düşünülmektedir. Daha önce bu veri seti ile bir çalışma yapılmadığı için gelecekte bu veri seti ile yapılacak olan çalışmalara katkı sağlaması düşünülmektedir.

Kaynaklar (References)

Akar, Ö., & Güngör, O., 2012. Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, 1 (2), 139-146.

Akşehirli, Ö. Y., Ankaralı, H., Aydın, D., Saraçlı, Ö., 2013. Tıbbi Tahminde Alternatif Bir Yaklaşım: Destek Vektör Makineleri. Turkiye Klinikleri Journal of Biostatistics, 5(1).

Alan, A., 2020. Makine öğrenmesi sınıflandırma yöntemlerinde performans metrikleri ile test tekniklerinin farklı veri setleri üzerinde değerlendirilmesi (Master's thesis, Fen Bilimleri Enstitüsü)

Ali, J., Khan, R., Ahmad, N., & Maqsood, I., 2012. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272.

Arun, K., Ishan, G., & Sanmeet, K., 2016. Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18 (3), 18-21.

Biau, G., & Scornet, E., 2016. A random forest guided tour. Test, 25 (2), 197-227.

Burges, C. J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

Coşkun, S., Kartal, M., Coşkun, A., & Bircan, H., 2004. Lojistik regresyon analizinin incelenmesi ve diş hekimliğinde bir uygulaması. Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi, 7 (1), 42-50.

Field, A., 2013. Discovering statistics using IBM SPSS statistics. sage.

Gök, M., 2017. Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 5 (3), 139-148.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M. & Özçift, A., 2016. KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi. Marmara Fen Bilimleri Dergisi, 28 (3), 89-94. DOI: 10.7240/mufbed.69674

Rasjid, Z. E., & Setiawan, R., 2017. Performance comparison and optimization of text document classification using k-NN and naïve bayes classification techniques. Procedia computer science, 116, 107-112.

Serengil, S. I., Imece, S., Tosun, U. G., Buyukbas, E. B., & Koroglu, B., 2021. A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction. In 2021 6th International Conference on Computer Science and Engineering (UBMK) (pp. 326-331). IEEE.

Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4 (11) .