T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ VERİ SETİ İLE BİR E-TİCARET ŞİRKETİ İÇİN SATIN ALMA EĞİLİMİ TAHMİNİ

Zeynep RENDECİ

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Doç. Dr. Gülder KEMALBAY

Ağustos, 2023

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ VERİ SETİ İLE BİR E-TİCARET ŞİRKETİ İÇİN SATIN ALMA EĞİLİMİ TAHMİNİ

Zeynep RENDECİ tarafından hazırlanan tez çalışması 18.08.2023 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Gülder KEMALBAY
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Gülder KEMALBAY, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Fatma NOYAN TEKELİ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Arzu BAYGÜL EDEN, Üye
Koç Üniversitesi _____

Danışmanım Doç. Dr. Gülder KEMALBAY sorumluluğunda tarafımca hazırlanan "Dengesiz Veri Seti ile Bir E-Ticaret Şirketi için Satın Alma Eğilimi Tahmini" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Zeynep RENDECİ

İmza

Bu çalışma, “Yıldız Teknik Üniversitesi Bilimsel Araştırma Proje Koordinatörlüğü’nün FYL-2022-4857” numaralı projesi ile desteklenmiştir.

TEŞEKKÜR

Destekleri, emeği ve yönlendirmesinden dolayı danışmanım

Doç. Dr. Gülder KEMALBAY'a teşekkür ederim.

Zeynep RENDECİ

İÇİNDEKİLER

TEŞEKKÜR iv
SİMGE LİSTESİ vi
KISALTMA LİSTESİ vii
ŞEKİL LİSTESİ viii
TABLO LİSTESİ ixx
ÖZET x
ABSTRACT xiii
1 GİRİŞ 1
1.1 Literatür Taraması 1
1.2 Tezin Amacı 4
1.3 Hipotez 4
2 METODOLOJİ 5
2.1 Makine Öğrenmesinde Sınıflandırma 5
2.2 Sınıflandırmada Performans Ölçümü 6
2.3 Dengesiz Veri Problemi 8
2.4 Lojistik Regresyon Analizi 11
3 UYGULAMA 13
3.1 Araştırmanın Veri Seti 13
3.2 Model Performansları 15
4 SONUÇ 25
KAYNAKÇA 27
TEZDEN ÜRETİLMİŞ YAYINLAR 30

SİMGE LİSTESİ

n^-	Gerçek Negatif Sayısı
n^+	Gerçek Pozitif Sayısı
$\hat{n}^+$	Tahmin Pozitif Sayısı
$\hat{n}^-$	Tahmin Negatif Sayısı
n	Toplam Örnek Sayısı

KISALTMA LİSTESİ

DN	Doğru Negatif
DP	Doğru Pozitif
DPO	Doğru Pozitif Oranı
YN	Yanlış Negatif
YP	Yanlış Pozitif
YPO	Yanlış Pozitif Oranı

ŞEKİL LİSTESİ

Şekil 2. 1 Roc uzayı **8**
Şekil 3. 1 Ham veri seti için karmaşıklık matrisi **15**
Şekil 3. 2 SMOTE yöntemine ait karmaşıklık matrisi **16**
Şekil 3. 3 Aşırı örnekleme yöntemine ait karmaşıklık matrisi **16**
Şekil 3. 4 Alt örnekleme yöntemine ait karmaşıklık matrisi **17**
Şekil 3. 5 Youden indeksine göre karmaşıklık matrisi **18**
Şekil 3. 6 Duyarlılığı kesinliğe eşitleyen yönteme göre karmaşıklık matrisi **18**
Şekil 3. 7 Farklı örneklem oranları için SMOTE yönteminin doğruluk oranları .. **21**
Şekil 3. 8 Farklı örneklem oranları için SMOTE yönteminin duyarlılık oranları . **22**
Şekil 3. 9 SMOTE ve Youden indeksi hibrit modeli için doğruluk oranları **22**
Şekil 3. 10 SMOTE ve Youden indeksi hibrit modeli için duyarlılık oranları **23**
Şekil 3. 11 Hibrit model için doğruluk ve duyarlılık oranları **23**

TABLO LİSTESİ

Tablo 2. 1 İkili sınıflandırma için karmaşıklık matrisi **7**
Tablo 3. 1 Veri seti için betimsel istatistikler-1 ..... **14**
Tablo 3. 2 Veri seti için betimsel istatistikler-2 ..... **14**
Tablo 3. 3 Tüm modeller için performans metrikleri..... **19**
Tablo 3. 4 Örneklem oranı 0.1-0.5 için Smote-Youden indeks model metrikleri **.20**
Tablo 3. 5 Örneklem oranı 0.6-1 için Smote-Youden indeks model metrikleri ... **21**

ÖZET

Dengesiz Veri Seti İle Bir E-Ticaret Şirketi İçin Satın Alma Eğilimi Tahmini

Zeynep RENDECİ

İstatistik Anabilim Dalı

İstatistik Programı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Gülder KEMALBAY

Günümüzde çevrimiçi alışveriş alışkanlıklarının artması ile müşteriler, birçok farklı firmanın ürününe ulaşarak fiyat karşılaştırması yapmaları sonucunda satın almaya karar vermektedirler. Bu durum e-ticaret şirketleri arasında rekabeti arttırırken, geçmiş alışveriş verilerinden hareketle müşterilerin gelecekteki satın alma eğilimlerini tahmin etmenin önemini ortaya çıkarmıştır. Müşteriler, tüm alternatifleri inceledikten sonra bile fiyatta veya kalitede beklenti sebebi ile alışveriş yapmadıkları bilinmektedir. Bu nedenle, müşteri davranışlarını, isteklerini tahmin etmek, müşteriye uygun iletişim ve pazarlama faaliyetleri yapmak firmaların pazardaki satış performansını artırmaktadır. Satış performansını artıracak analitik çözümlerden biri de müşterilerin satın alma eğiliminin tahmin edilmesidir. Satın alma eğilimi verisinde genellikle sınıf dengesizliği problemi ile karşılaşılmaktadır.

Sınıf dengesizliği problemi, sınıflandırma analizinin performansını düşürürken yanlı tahmin yapılmasına sebep olmaktadır. Bu çalışmada, sınıf dengesizliğine sahip bir veri seti üzerinden sınıflama analizi yapılacaktır. Çalışmada, dengesiz veri

problemini gidermek için iki farklı yöntem kullanılacaktır: İlki, dengesiz veri seti için en uygun kesim noktası ROC analizi ile bulunacaktır. İkincisi, yeniden örnekleme teknikleri kullanılarak veri seti dengeli hale getirilecektir. ROC analizi için Youden indeksi, duyarlılık=kesinlik kriteri; yeniden örnekleme teknikleri için aşırı örnekleme alt örnekleme ve sentetik veri üretilmesi kullanılacaktır. Dengesiz veri setini sınıflandırmak için lojistik regresyon analizi yöntemi kullanılacaktır. Araştırmanın amacı, dengesiz veri seti ile sınıflandırma analizi yaparken ROC analizi ve yeniden örnekleme tekniklerinin model performansına etkisini incelemek ve en başarılı performans gösteren modeli belirlemektir. Çalışmanın yöntemi, bir e-ticaret şirketine ait veri seti üzerinde uygulanacaktır. Çalışmadan elde edilecek sonuçlara göre, müşterilerin alışveriş yapma ya da yapmama eğilimlerini sınıflandırmanın e-ticaret şirketlerinin pazarlama faaliyetlerine yön vererek karlarını arttırmalarına yardımcı olması umulmaktadır.

Anahtar Kelimeler: Dengesiz veri, satın alma eğilimi, sınıflandırma analizi, makine öğrenimi.

ABSTRACT

Prediction of Propensity To Buy For An E-Commerce Company With An Unbalanced Dataset

Zeynep RENDECİ

Department of Statistics

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Gülder KEMALBAY

Today, with the increase in online shopping habits, customers decide to purchase as a result of price comparisons by reaching the products of many different companies. While this situation increases the competition among e-commerce companies, it has revealed the importance of predicting the future purchasing tendencies of customers based on past shopping data.

Class imbalance problem is usually encountered in purchasing propensity data. This problem reduces the performance of classification analysis and causes biased estimation. In this study, classification analysis will be done on a data set with class imbalance. Two different methods will be used to solve the unbalanced data problem: First, the most suitable cutoff point for the unbalanced data set will be found by ROC analysis. Second, the data set will be balanced using resampling techniques. Youden Index, and recall=precision criterion will be used for ROC analysis. Over sampling, under sampling and synthetic data generation will be used for resampling techniques. The unbalanceed data set will be classified by logistic regression analysis method. The aim of the research is to examine the effect of ROC analysis and resampling techniques on the model performance while performing classification analysis with an unbalanced data set and to determine the model with

the most successful performance. The method of the study will be applied on the data set of an e-commerce company. According to the results of the study, it is hoped that predicting customers' propensity to buy will help e-commerce companies to increase their profits by directing their marketing activities.

Keywords: Imbalanced data, propensity to buy, classification, machine learning.

1
GİRİŞ

Günümüzde, çevrimiçi alışveriş alışkanlıklarının artması ile müşterilerin ürünlere ulaşımı, ürünleri değerlendirmesi, fiyat karşılaştırması yapması kolaylaşmıştır. Müşteriler birçok farklı firmanın ürünlerine, fiyatlarına ulaşıp değerlendirme yapıp, bu değerlendirmelerine göre alışveriş yapmaya başlamışlardır. Bu durum firmalar arası rekabeti de arttırmıştır. Bu rekabetçi ortamda müşterilere daha iyi bir alışveriş deneyimi sunmak, müşteriyi tanımak, müşteri için kişiselleştirilmiş deneyimler yaratmak değerli hale gelmiştir. Firmalar müşterilerin geçmiş alışkanlık ve hareketlerini veriden inceleyerek, müşterinin gelecek hareketlerini tahmin etmeye başlamıştır. Böylece bu tahminlere göre yol haritası çizerek piyasada etkin olmaya, varlıklarını sürdürmeye ve müşteriyi elde tutmaya çalışmaktadırlar.

Müşterilerin, tüm seçenekleri inceledikten ve değerlendirdikten sonra bile alışveriş işlemini tamamlamadığı, ertelediği bilinmektedir. Bu ertelemenin başlıca sebepleri, ürünle ilgili daha fazla bilgiye ihtiyacı olması, daha iyi bir alternatif bulma isteği, fiyatta düşüş ya da kalitede artış isteğidir. Burada firmaların veri ile bakış açısı elde edip harekete geçmesi, müşteri özelinde aksiyon alması, pazarlama faaliyetleri kurgulaması gerekir. Müşterinin alışveriş yapıp-yapmama durumunu sınıflandırması, buna göre hedefli müşteri kitlesine ilgili ürün veya indirim teklifi etmesiyle rakipleri arasından sıyrılarak, müşteri dönüş oranlarını arttıracaktır.

Tahmine dayalı analitik, sınıflandırma çalışmaları, eğilim modelleri firmalar için büyük önem arz etmektedir. Tüm pazarlama ve iletişim faaliyetlerini bu çıktılara göre kurgulamaktadırlar. Eğilim modelleri sıklıkla müşterinin terk etme, teklife cevap verme eğilimi, alışveriş yapma, satın alma gibi eğilimlerini tahmin etmede kullanılmaktadır [1].

1.1 Literatür Taraması

Müşterilerin satın alma eğilimleri tahmin etmek için literatürde yapılan çalışmalar incelendiğinde genellikle dengesiz veri seti problemini çözmek için yeniden

örnekleme tekniklerine başvurdukları gözlenmiştir. Bazı çalışmalarda ise farklı sınıflandırma algoritmaları ile karmaşıklık matrisinden doğruluk yanı sıra elde edilen çeşitli performans metrikleri kullanılmıştır. Literatürde yapılan çalışmalar izlendiğinde, dengesiz veri seti üzerinde ROC eğrisi analizi ile en uygun kesim noktası bulan çalışmaların genellikle medikal sektöründe yoğunlukla kullanıldığı gözlenmiştir. Satın alma eğilim modelleri ve e-ticaret alanında dengesiz veri seti üzerinden yapılan çalışmalar izleyen şekilde taranmıştır:

Maloof tarafından 2003 yılında yapılan çalışmada, dengesiz veri seti kullanılarak yanlış sınıflandırma maliyetlerinin eşit olamadığı veya bilinmediği durumda sınıflandırma performansını arttırmaya yönelik analizler yapılmıştır [2]. Dengesiz veri probleminden kurtulmak için ROC eğrisi altında kalan alan değerleri farklı algoritmalarda karşılaştırılmıştır. C5.0, K-nn, çok katmanlı algılayıcılar, Naive Bayes sınıfılandırıcı algoritmalarının ROC sonuçları karşılaştırılmıştır. Aynı zamanda, veri dengeleştirme yöntemleri olan alt örnekleme ve aşırı örnekleme yöntemleri denenmiştir. Bu yöntemlerin de etkili olduğu görülmüştür.

Duman ve arkadaşları, veri setinde yüksek derecede dengesizliğin olduğu ve azınlık sınıfının belirlenmesinin daha önem kazandığı uygulama alanlarından biri olan bankacılık sektöründe bir araştırma yapmıştır [3]. Uygulamada, bir bankaya ait müşterilerin bir yatırım ürününü satın alma eğilimlerini tahmin etmek için CHAID analizi, sinir ağları ve lojistik regresyon analizi yöntemleri uygulanmış ve sınıflama analizi performanslarını karşılaştırmak için ROC eğrisi altında kalan alan kullanılmıştır. Sonuç olarak, CHAID analizi başarılı bulunmuştur.

Hajian-Tilaki tarafından yayınlanan makalede, hastalık sınıflandırması için ROC analizine başvurulmuştur [4]. Farklı kesim değerlerinde doğruluk ve duyarlılık değerleri incelenmiştir. ROC eğrisinde optimal kesim noktasını belirlemedeki avantajlar ve dezavantajlar, yanlılık tartışılmıştır.

Song ve arkadaşları tarafından yapılan araştırmada dengesiz veri seti üzerinde sınıflandırma analizi yapılmıştır [5]. Dengesiz veri setinden kaynaklı düşük çıkan sınıflandırma performansı çeşitli yöntemler ile yükseltilmeye çalışılmıştır. Sınıflandırma yöntemi olarak, Destek Vektör Makineleri ve Rassal Orman algoritmaları kullanılmıştır. Dengesiz veri problemini gidermek için olağan kesim

değeri 0.5'in yanı sıra maliyet duyarlılık öğrenme ve alt örnekleme yöntemleri denenmiştir. Doğruluk performansında kritik bir değişiklik gözlenmese bile, kesinlik metriği 0.5 değerinden 0.88 değerine yükselmiştir.

Santos, yapmış olduğu doktora tez çalışmasında ferdi kaza sigortası ürünlerini satın alma eğilim tahmini için bir sınıflandırma modeli geliştirmiştir [6]. Çalışmada, karar ağaçları, lojistik regresyon analizi, yapay sinir ağları ve topluluk öğrenmesi yöntemleri denenmiştir. Performans metriklerinin çoğunda, topluluk öğrenmesi yöntemi daha başarılı bulunmuştur.

Stubseid ve Arandjelovic tarafından 2018 yılında yayımlanan bu makalede müşterilerin satın alma eğilimleri sınıflandırma modelleri ile tahmin edilmiştir [7]. Dengeli veri üzerinde çalışılmıştır. Rassal Orman ve Naive Bayes algoritmaları uygulanıp, karmaşıklık matrisi metrikleri Doğruluk, F1 skoru, AUC değerleri karşılaştırılmıştır.

2019 yılında Munoz ve Hedström tarafından yapılan çalışmada, lojistik regresyon kullanılarak müşterinin alışveriş yapma eğilimi bulunması amaçlanmıştır. %88 oranında doğrulukla alışveriş yapma eğiliminde olan müşterilerin doğru hesaplandığı görülmüştür [8]. Makalede dengesiz veri probleminden bahsedilmemiştir fakat karmaşıklık matrisi incelendiğinde, doğruluk yüksek olmasına rağmen kesinlik değerinin %34 seviyesinde kaldığı gözükmektedir. Bu da çoğunluk sınıfa yönelik yanlı bir sınıflandırma olduğuna işaretlerden biridir.

Gramegna ve Giudici, sigortacılık sektöründe yaptıkları çalışmalarında müşterilerin sigorta poliçesi satın alma eğilimini ve mevcut olan bir müşterinin terk etme riskini tahmin eden yapay zeka temelli bir model geliştirmişlerdir [9]. Geliştirilen model, oldukça yüksek doğruluk oranına sahip olan XGBoost sınıflandırma algoritmasından elde edilmiş Shapley değerlerine benzerlik kümelemesinin uygulanmasından oluşmaktadır. Elde edilen sonuçlara göre; geliştirilen modelin müşterilerin satın alma veya terk etme davranışlarını iyi tahmin edebildiği ve benzer özelliklere göre etkili ve hızlı bir şekilde müşterileri gruplayabildiği gösterilmiştir.

2021 yılında Zhang tarafından yapılan çalışmada, e-ticaret verisi kullanılmış ve müşterilerin satın alma eğilimleri tahmin edilmiştir [10]. Sınıflandırma algoritması olarak lojistik regresyon ve rassal orman yöntemleri kullanılmıştır. Karmaşıklık

matrisi sonuçlarına göre metrikler incelenmiştir. Doğruluk oranı, rassal orman yönteminde daha yüksek çıkarken, kesinlik ve F1 skorları ise lojistik regresyonda daha yüksek çıkmıştır. Kullanılan veride, dengesiz veri problemi vardır. Bunun için doğruluk metriği yanında diğer metrikler de göz önünde bulundurulmuştur.

2023 yılında Dani ve Ginting tarafından yapılan çalışmada, müşterilerin reklam tıklamalarını tahmin etmek için lojistik regresyon analizi ve k-En Yakın Komşular yöntemleri kullanılmıştır [11]. Algoritmaların performansları; doğruluk, duyarlılık, F1 skoru ve AUC-ROC gibi bazı ölçütler kullanılarak değerlendirilmiştir. Elde edilen sonuçlara göre, lojistik regresyon analizi daha başarılı olmuştur.

Bu araştırmada, literatürde yapılan çalışmalardan farklı olarak bir e-ticaret sitesine ait dengesiz veri seti üzerinde satın alma eğilimini sınıflandırmak için yeniden örnekleme ve ROC analizi teknikleri ayrı ayrı ve hibrit model olarak ele alınarak makine öğrenimi algoritmalarından biri olan lojistik regresyon analizinin sınıflama performansı üzerindeki etkileri incelenecektir.

1.2 Tezin Amacı

Bu tezin amacı, sınıflandırma analizinde ortaya çıkan dengesiz veri seti problemini çözmek için yeniden örnekleme yöntemleri ve ROC analizi yöntemini uygulamak ve bu yöntemlerin makine öğrenimi model performansı üzerindeki etkilerini araştırmaktır. Tezin amacına ulaşmak için bir e-ticaret şirketine ait engesiz veri seti kullanılarak müşterilerin gelecek 1 ay içinde alışveriş yapma olasılıklarını tahmin ederken ROC analizi ve yeniden örnekleme tekniklerinin lojistik regresyon analizi performansı üzerindeki etkileri karşılaştırmalı olarak incelenecektir. Böylece, dengesiz veri problemi ortadan kalkarken, lojistik regresyon model performansı üzerindeki iyileşme gösterilecektir.

1.3 Hipotez

Araştırmanın hipotezi, bir e-ticaret sitesine ait dengesiz veri seti üzerinde yeniden örnekleme ve ROC analizi tekniklerine dayanarak oluşturulan lojistik regresyon sınıflandırma modeli ile müşterilerinin önümüzdeki 1 ay içinde alışveriş yapma eğilimlerini en az %70 duyarlılık oranı ile tahmin edilebileceğini savunmaktadır.

2
METODOLOJİ

2.1 Makine Öğrenmesinde Sınıflandırma

Makine öğrenmesi, bilgisayar sistemlerinin verilerden öğrenme ve çıkarım yapma yeteneğini sağlayan bir yapay zeka disiplinidir. Sınıflandırma, makine öğrenmesinin önemli bir alanıdır ve veri kümesindeki örnekleri belirli sınıflara atama amacını taşır. Sınıflandırma algoritmaları, verileri analiz ederek desenleri ve ilişkileri belirleyerek bu sınıflandırmayı gerçekleştirir. Sınıflandırma algoritmaları, genellikle denetimli öğrenme yöntemleri kullanır. Denetimli öğrenme, giriş verileri ile ilişkili hedef çıktıları içeren bir eğitim veri kümesi üzerinden modelin öğrenilmesini sağlar. Sınıflandırma algoritmaları, eğitim veri kümesini kullanarak giriş özelliklerini analiz eder ve yeni örneklerin hangi sınıflara ait olduğunu tahmin etmek için bir model oluşturur. Sınıflandırma analizinin aşamaları izleyen şekilde özetlenebilir:

Veri Toplama ve Hazırlama: Sınıflandırma algoritmasının kullanılacağı bir veri kümesi toplanır ve gerektiğinde önişleme adımları uygulanır. Veriler genellikle sayısal formatta olmalıdır ve eksik değerler, aykırı değerler veya gereksiz özellikler işlenmelidir.

Özellik Seçimi ve Mühendisliği: Veri kümesindeki özelliklerin seçimi veya oluşturulması önemlidir. İyi seçilmiş özellikler, sınıflandırma performansını artırabilir. Ayrıca, verilerin normalleştirilmesi veya özellik vektörlerinin dönüştürülmesi gibi işlemler de yapılabilir.

Eğitim ve Model Oluşturma: Önceden belirlenmiş bir sınıflandırma algoritması seçilir ve eğitim veri kümesine uygulanır. Bu adımda, algoritma, giriş özelliklerini hedef sınıflara en iyi şekilde eşleştiren bir model oluşturur. Model, eğitim veri kümesini doğru bir şekilde sınıflandırmayı hedefler.

Test ve Doğrulama: Eğitilen model, test veri kümesi üzerinde doğrulama işlemine tabi tutulur. Modelin doğruluğu, kesinlik, hassasiyet ve geri çağırma gibi

performans metrikleriyle değerlendirilir. Eğer model başarılı bir şekilde performans gösteriyorsa, gerçek dünya verilerinde kullanılmaya hazır hale gelir.

Tahmin ve Sınıflandırma: Eğitilen model, yeni giriş örneklerini analiz ederek sınıflandırma tahminleri yapabilir. Model, özelliklerin belirli sınıflarla en uyumlu olduğunu belirleyerek yeni verileri sınıflandırır.

Makine öğrenmesinde sıklıkla kullanılan sınıflandırma algoritmaları; lojistik regresyon, karar ağaçları, destek vektör makineleri, k-en yakın komşu, yapay sinir ağlarıdır [12].

2.2 Sınıflandırmada Performans Ölçümü

Sınıflandırma algoritmalarında model performans sonuçlarını değerlendirmek için en çok kullanılan yöntemlerden biri, karmaşıklık matrisi oluşturmaktır. Karmaşıklık matrisi, örnek verisindeki gerçek sınıf ve tahmin sınıf sayılarını göstermektedir. İki sınıf içeren bir veride sınıflar genellikle 1 ve 0 olarak kodlanır ve 1 sınıfı pozitif, 0 sınıfı negatif sınıf olarak adlandırılır.

Doğru Pozitif (DP): Doğru pozitif değeri, gerçekte pozitif yani 1 olan gözlemlerin model sonucunda kaçının 1 olarak sınıflandırıldığını gösterir.

Doğru Negatif (DN): Doğru Negatif değeri, gerçekte negatif yani 0 olan gözlemlerin model sonucunda kaçının 0 olarak sınıflandırıldığını gösterir.

Yanlış Pozitif (YP): Yanlış pozitif değeri, gerçekte negatif yani 0 olan gözlemlerin model sonucunda kaçının 1 olarak sınıflandırıldığını gösterir.

Yanlış Negatif (YN): Yanlış pozitif değeri, gerçekte pozitif yani 1 olan gözlemlerin model sonucunda kaçının 0 olarak sınıflandırıldığını gösterir.

İkili sınıflandırma problemi için karmaşıklık matrisi Tablo 2.1'de verilmiştir [13]. Karmaşıklık matrisi ile model sonucunda tahmin edilen sınıflar ve verideki gercek değerler karşılaştırılarak performans metrikleri oluşturulur.

Tablo 2.1 İkili sınıflandırma için karmaşıklık matrisi [13]

		Gerçek		
		Pozitif	Negatif	Toplam
Tahmin	Pozitif	DP	YP	Tahmin pozitif sayısı $(\hat{n}^+)$
	Negatif	YN	DN	Tahmin negatif sayısı $(\hat{n}^-)$
	Toplam	Gerçek pozitif sayısı (n^+)	Gerçek negatif sayısı (n^-)	Toplam örnek sayısı (n)

Sınıflandırma modellerinde en çok kullanılan performans metrikleri ve formulasyonları Denklem (2.1)-(2.8)'de verilmiştir.

Doğruluk Oranı: Sınıfları doğru tahmin edilen gözlemlerin tüm gözlemlere oranıdır.

$$Doğruluk\ Oranı = (DP + DN)/n \tag{2.1}$$

Hata Oranı: Sınıfları yanlış tahmin edilen gözlemlerin tüm gözlemlere oranıdır. 1-Doğruluk Oranı ile de hesaplanır.

$$Hata\ Oranı = (YP + YN)/n \tag{2.2}$$

Doğru Pozitif Oranı (DPO): Bu orana, Duyarlılık da denir. Doğru tahmin edilen pozitif gözlemlerin tüm pozitif gözlemlere oranıdır.

$$DPO = DP/(DP + YN) = DP/n^+ \tag{2.3}$$

Doğru Negatif Oranı (DNO): Bu orana, Belirleyicilik de denir. Doğru tahmin edilen negatif gözlemlerin tüm negatif gözlemlere oranıdır.

$$DNO = DN/(YP + DN) = DN/n^- \tag{2.4}$$

Yanlış Pozitif Oranı (YPO): Yanlış tahmin edilen negatif gözlemlerin tüm negatif gözlemlere oranıdır. 1-Belirleyicilik ile de hesaplanır.

$$YPO = YP/(YP + DN) = YP/n^- \tag{2.5}$$

Yanlış Negatif Oranı (YNO): Yanlış tahmin edilen negative gözlemlerin tüm pozitif gözlemlere oranıdır. 1-Duyarlılık ile de hesaplanır.

$$YNO = YN/(YN + DP) = DN/n^+ \tag{2.6}$$

Kesinlik: Pozitif olarak doğru tahmin edilen gözlemlerin tüm pozitif tahmin edilen gözlemlere oranıdır.

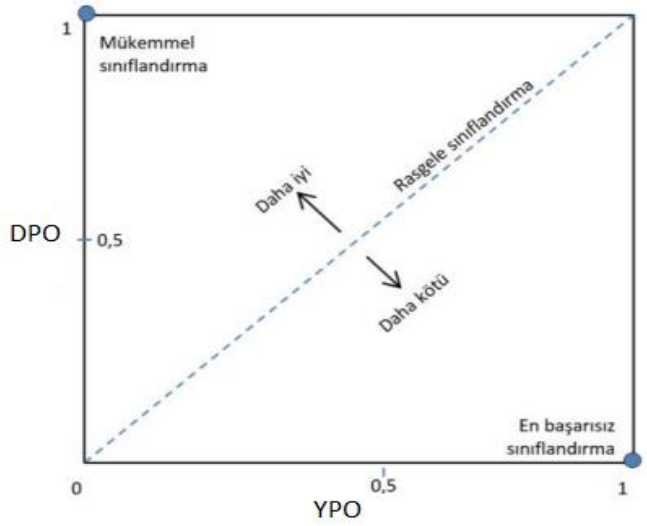
$$Kesinlik = DP/(DP + YP) = DP/\hat{n}^{+} \quad (2.7)$$

F1 Skoru: Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır.

$$F1\,Skoru = 2 * (Kesinlik * Duyarlılık) / (Kesinlik + Duyarlılık) \quad (2.8)$$

Sınıflandırıcı modelin başarılı performans sağlaması için, doğruluk oranı, doğru pozitif oranı (duyarlılık), doğru negative oranı (belirleyicilik), kesinlik ve F1 skoru değerinin 1'e yakın çıkması beklenir.

ROC Eğrisi: Farklı kesim değerleri için y-ekseninde DPO, x-ekseninde ise YPO değerleri kullanılarak oluşturulan grafiktir. Diğer bir deyişle, ROC eğrisi üstünde yer alan her nokta, farklı kesim değerlerine karşılık gelen duyarlılık ve (1-belirleyicilik) değerlerini gösterir. ROC eğrisi, sınıflandırma modelinin sınıfları birbirinden ne kadar iyi ayırdığını gösterir. Eğri altında kalan alan (AUC), model doğruluk oranını temsil eder. AUC değerinin 0.5'e eşit olması, sınıflandırmanın rassal olarak yapıldığını gösterir. AUC değeri 1'e yaklaştıkça modelin uyumu o kadar iyidir [14], [15]. Şekil 2.1'de ROC uzayı verilmiştir.



Şekil 2. 1 Roc uzayı [16]

2.3 Dengesiz Veri Problemi

Sınıflandırmada dengesiz veri problem, bir sınıftaki gözlem sayısının diğer sınıflardaki gözlem sayısına kıyasla daha fazla ya da az olmasıdır. Gözlem sayısı

daha fazla olan sınıf, çoğunluk sınıf (majority class); daha az olan sınıf ise azınlık sınıf (minority class) olarak adlandırılır.

Dengesiz veri setlerinde, sınıflandırma algoritmalarıyla yapılan çalışmalarda elde edilen doğruluk değerinin, azınlık olan sınıf için gerçeği yansıtmadığı gözlenmektedir. Bu tip çalışmalarda karmaşıklık matrisi incelendiğinde yüksek doğruluk oranları görünürken, detaya inildiğinde kesinlik ve duyarlılık değerlerinde oldukça düşük oranlar gözlenmektedir. Bu durum, çoğunluk olan sınıf lehine yanlı sınıflandırma yapılmasına sebep olmaktadır. Sınıflandırma problemlerinde sonuçları yorumlamak için birçok performans metriği geliştirilmiştir. Doğruluk oranı, sıklıkla bakılan metriklerden biridir ve örneklem kümesinde doğru sınıflandırılan gözlemlerin oranını verir. Bu metrik özellikle dengesiz veride tek başına modelin performansını açıklamak için yeterli olmaz. Modelin gerçek performansını anlayabilmek için doğruluk oranı yanında, duyarlılık, belirleyicilik, pozitif öngörü değeri, negatif öngörü değeri ve F1 skoru gibi ölçülere de bakılmalıdır.

Yanlı sınıflandırmayı önlemek için 2 temel yönteme başvurulabilir. Birinci yöntem veri seti yeniden örnekleme teknikleri ile dengeli hale getirildikten sonra dengeli veri seti üstünde sınıflandırma analizi yapılmasıdır. İkinci yöntem ise verideki dengesizlik giderilmeden ROC analizi ile en uygun kesim noktasının belirlenmesi ve sınıflama analizi yapılırken olağan olarak kullanılan 0.5 kesim noktasının, veriye en uygun kesim noktası ile değiştirilmesidir [17], [18].

2.3.1 Yeniden Örnekleme Teknikleri

Veriyi dengeli hale getirme işlemi için en çok kullanılan yeniden örnekleme yöntemleri, aşırı örnekleme (over sampling), alt örnekleme (under sampling) ve sentetik veri üretme (SMOTE) yöntemleridir.

• Aşırı örnekleme, azınlık olan sınıfın gözlem sayısının çoğunluk olan sınıfa yaklaştırılmasıdır. Bunun için, azınlık sınıftaki örnekler rassal olarak kopyalanarak sayıca çoğunluk sınıftaki örnek sayısına eşitlenir.

• Alt örnekleme, çoğunluk olan sınıfın gözlem sayısının, azınlık olan sınıfa yaklaştırılmasıdır Bunun için, çoğunluk sınıftaki örnekler rassal olarak silinerek sayıca azınlık sınıftaki örnek sayısına eşitlenir [19].

• SMOTE, bu teknikte azınlık sınıfın komşularına bakılarak sentetik örnekler oluşturulur. Böylece azınlık sınıfı aşırı örneklenmiş olur [20]. Yani SMOTE yöntemi için de bir aşırı örnekleme yöntemi denilir.

2.3.2 Uygun Kesim Noktasının Hesaplanması

Dengesiz veride sınıflandırma performansı arttırmak için kullanılan yöntemlerden biri de ROC analizidir. Lojistik regresyonda sınıflandırma yapılırken, olağan kesim değeri genellikle 0.5 olarak alınır. Bu değer dengeli verilerde iyi bir performans sergilerken, dengesiz verilerde azınlık sınıfa karşı düşük doğruluk değeri oluşturmaktadır. Karmaşıklık matrisi aracılığı ile model performansı ölçülürken belirlenen kesim noktasına bağlı kalınır. ROC analizi aracılığı ile uygun kesim noktası belirlenirken çeşitli kriterler kullanılır. Bu araştırmada, ROC analizi ile uygun kesim noktası seçilirken Youden indeksi, ve duyarlılığın kesinliğe eşit olduğunu varsayan yöntem kullanılmıştır.

Youden indeksi, genellikle J ile temsil edilir ve c herhangi bir kesim noktası olmak üzere Denklem (2.9)'da verilmiştir:

$$J(c) = [duyarlılık(c) + belirleyicilik(c) - 1] \tag{2.9}$$

Denklem (2.9)'a göre J fonksiyonuna maksimum değeri veren nokta Youden indeksine göre uygun kesim noktası olarak seçilir. Kesim noktasının değeri 0'dan 1'e doğru hareket ederken, duyarlılık değeri artar ve belirleyicilik değeri azalır. ROC eğrisi ile rassal sınıflandırmayı temsil eden çizgi arasındaki dikey uzaklığın maksimum olduğu nokta, Youden indeksine göre uygun kesim noktasıdır.

Duyarlılık ve kesinlik değerleri, yalnızca azınlık sınıf olan 1 sınıfının doğru tahmin edilmesiyle ilgili metrikler olduğu için dengesiz veri probleminde kullanılan en önemli metriklerin başında gelirler. ROC eğrisine benzer bir şekilde, farklı kesim değerleri için x ekseninde kesinlik ve y ekseninde duyarlılık değerleri karşılık gelecek şekilde duyarlılık-kesinlik eğrisi çizdirilebilir. Bu eğri üzerinde, duyarlılığın kesinliğe eşitleyen nokta, uygun kesim noktası olarak seçilir [15], [21], [22].

2.4 Lojistik Regresyon Analizi

Lojistik regresyon, makine öğrenmesinde sınıflandırma problemlerini çözmek için sıklıkla kullanılan bir yöntemdir. Bu yöntem, ikili sınıflandırma problemlerinde bir veri noktasının belirli bir sınıfa ait olma olasılığını tahmin etmek için kullanılır. Lojistik regresyon, geniş bir uygulama alanına sahip olup, tıbbi araştırmalardan pazarlama analizine kadar birçok alanda başarıyla kullanılmaktadır.

Lojistik regresyon, temel olarak bir lineer regresyon modelinin bir sigmoid fonksiyonuna (logit fonksiyonu olarak da bilinir) uygulanmasıyla elde edilir. Sigmoid fonksiyonu, çıktıyı 0 ile 1 arasında bir olasılık değeri olarak sınırlar ve bu şekilde sınıflandırma yapılabilir. Lojistik regresyonun temel amacı, giriş özelliklerine dayalı olarak bir veri noktasının belirli bir sınıfa ait olma olasılığını tahmin etmektir [23].

Model, örneklem kümesinin geçmiş hareketlerinden öğrenir ve ona göre tahmin geliştirir. Bağımlı değişken Y ve n adet bağımsız değişkenler vektörü $\mathbf{X}^T = (X_1, \dots, X_n)$ ile gösterilsin. Bağımlı değişkenin $Y = 0$ ve $Y = 1$ şeklinde etiketlenen iki farklı kategoriye sahip olduğunu varsayalım. X bağımsız değişkenler vektörünün aldığı değer bilindiği durumda Y'nin 1 olma olasılığı; yani pozitif sınıfa ait olma olasılığı $P(Y = 1|\mathbf{X} = \mathbf{x})$ ile gösterilir. Denklem (2.10)'da çoklu lojistik regresyon modeli verilmiştir:

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{2.10}$$

X bağımsız değişkenleri ile $P(Y = 1|\mathbf{X} = \mathbf{x})$ olasılığı arasındaki ilişki lineer değildir. Grafiği S biçiminde eğriseldir. Denklem (2.10)'da verilen modele

$$g(\mathbf{x}) = \ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{1-P(Y=1|\mathbf{X}=\mathbf{x})}\right) \tag{2.11}$$

dönüşümü uygulandığında çoklu lojistik regresyon modeli, parametrelerine göre lineer bir modele dönüşür ve lineer model Denklem (2.12)'de verilmiştir:

$$g(\mathbf{x}) = \ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{1-P(Y=1|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{2.12}$$

Model parametreleri en çok olabilirlik yöntemi ile tahmin edilir. Modelin uyum iyiliği testi için Nagelkerke R^2 ve Cox-Snell R^2 kullanılır. Ayrıca, modelin uyum iyiliği için Hosmer ve Lemeshow testi uygulanır.

Sınıflandırma analizi yapılırken her gözlem değeri için tahmin edilen $P(Y = 1|xi)$ olasılıkları, uygun kesim noktası ile kıyaslanır. Eğer $P(Y = 1|xi)$ olasılığı uygun kesim değerinden büyük ise ise bu gözlem değerine karşılık olarak gelen yi değeri 1; aksi durumda ise 0 şeklinde sınıflandırılır [24], [25].

3
UYGULAMA

Bu bölümde, bir perakende e-ticaret sitesinde kadın tekstil ürünleri satışı hacminde yaşanan azalmayı arttırmak amacıyla, kadın tekstil ürünleri kategorisi özelinde satın alma eğilimi modeli çalışılacaktır. Satın alma eğilimi modeli lojistik regresyonun olağan parametre değerleri ile kurulacak, sonrasında veri manipülasyonları ile daha iyi performe edecek şekilde geliştirilmeye çalışılacaktır. Modeller, Python programı kullanılarak oluşturulmuştur.

Bu uygulama ile dengesiz veri ile kurulan modellerin iyi performe edebilmesi için %50-%50 veri dengeleştirmenin gerekliliği araştırılacaktır. Sınıflandırma modellerinin, sınıfları öğrenecek kadar yeterli veri sağlanması ile, doğruluk harici farklı metriklerin takibi veya tahmin olasılık alt sınırının değiştirilmesi ile daha iyi performe edip etmediği araştırılacaktır.

3.1 Araştırmanın Veri Seti

E-ticaret sitesinden 1 yıllık veriler alınarak ve 50.000 gözlemli veri seti oluşturulmuştur. 50.000 müşteri için bağımlı ve bağımsız değişkenler üretilmiştir.
Bağımlı değişken:1 ay içinde kadın kategorisinden alışveriş yapıp yapmama durumu, her bir müşteri alışveriş yapma durumuna göre 1, yapmama durumuna göre 0 olarak etiketlenmiştir.
Bağımsız Değişkenler: Üretilen bağımsız değişkenler içerisinden, hedefi temsil gücü yüksek, birbiri arasında multikorelasyon olmayan 9 değişken model için seçilmiştir. Seçilen bağımsız değişkenler aşağıdaki gibidir.

- **trx_l1m:** Müşterinin son bir ayda kaç alışveriş yaptığı
- **trx_kadin_l12m:** Müşterinin son 1 yılda kadın kategorisinden kaç alışveriş yaptığı
- **trx_aksesuar_l12m:** Müşterinin son 1 yılda aksesuar kategorisinden kaç alışveriş yaptığı
- **trx_cocuk_l3m:** Müşterinin son 3 ayda çocuk kategorisinden kaç alışveriş yaptığı

- **adet_kadin_l3m:** Müşterinin son 3 ayda kadın kategorisinden aldığı ürün sayısı
- **ciro_spor_l3m:** Müşterinin son 3 ayda spor kategorisinde ne kadar para harcadığı
- **kategori_cnt_l12m:** Müşterinin son 1 yılda kaç değişik kategoriden alışveriş yaptığı
- **dongu:** Müşterinin kaç günde bir alışveriş yaptığı
- **kadin_dist_urun_view:** Müşterinin son 1 ayda kadın kategorisinden ürün görüntüleme sayısı

Tablo 3.1'de, veri setine ait bazı betimsel istatistikler sunulmuştur.

Tablo 3. 1 Veri seti için betimsel istatistikler-1

	trx_11m	trx_kadin_l12m	trx_aksesuar_l12m	trx_cocuk_l3m	adet_kadin_l3m
Ortalama	0.62	1.03	0.09	0.22	0.81
Std Sapma	0.94	2.40	0.42	0.58	2.26
Min	0.00	0.00	0.00	0.00	0.00
0.25 %	0.00	0.00	0.00	0.00	0.00
0.5 %	0.00	0.00	0.00	0.00	0.00
0.75 %	1.00	1.00	0.00	0.00	1.00
Max	46.00	81.00	28.00	16.00	94.00

Tablo 3. 2 Veri seti için betimsel istatistikler-2

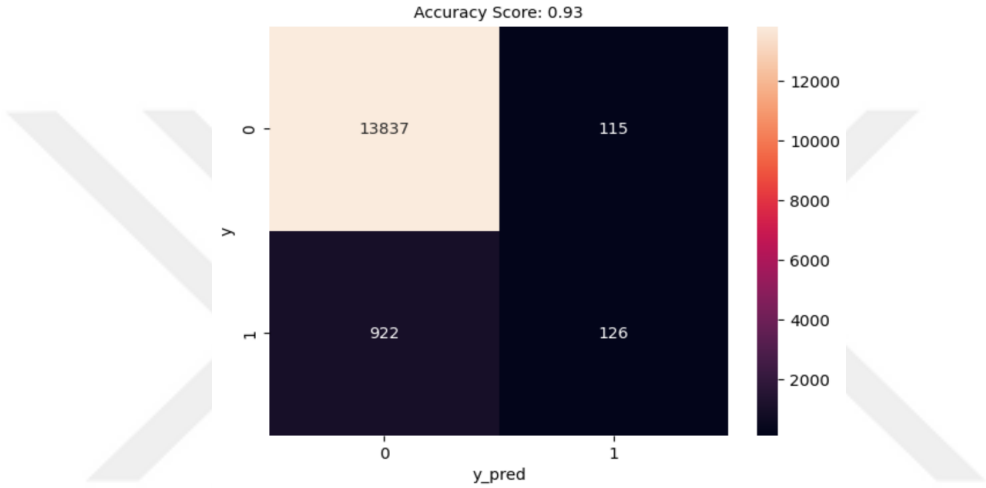
	ciro_spor_l3m	kategori_cnt_l12m	dongu	kadin_dist_urun_view
Ortalama	132.15	1.55	80.01	2.36
Std Sapma	302.69	0.98	117.19	17.11
Min	0.00	0.00	0.00	0.00
0.25 %	0.00	1.00	0.00	0.00
0.5 %	0.00	1.00	39.00	0.00
0.75 %	170.40	2.00	105.00	0.00
Max	17215.40	10.00	725.00	1142.00

Verideki hedef oranı, yani kadın tekstili kategorisinden 1 ay içinde satın alım yapan müşteri oranı %7 olarak hesaplanmıştır. 50.000 müşteri içerisinde sadece %7'sinin alışveriş yapması çalışmada dengesiz veri problemi olduğunu göstermektedir.
Verinin %70'i yani 35.000 gözlem eğitim veri seti olarak, %30'u yani 15.000 test veri seti olarak ayrılmıştır.

3.2 Model Performansları

İlk olarak, veri üzerinde hiçbir manipülasyon yapılmadan olağan parametrelerle basit bir lojistik regresyon modeli kurulmuştur. Bu modelin üzerine, veri manipülasyonları ve dengesiz veri problem çözümleri kullanılarak model performansındaki değişimler karmaşıklık matrisi kullanılarak incelenmiştir.

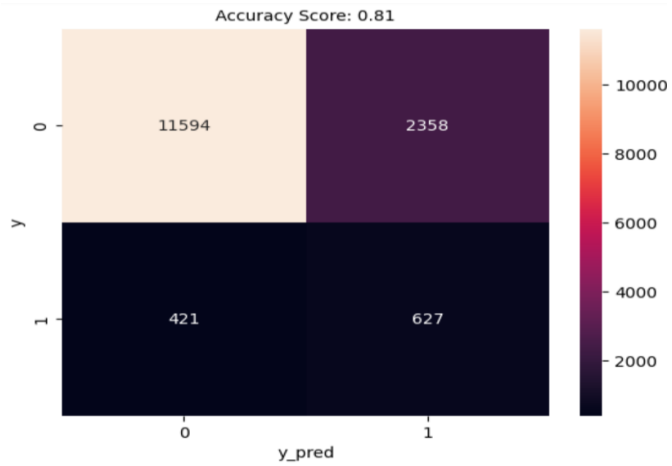
Ham şekildeki dengesiz veri seti üzerinde hiçbir manipülasyon yapılmaksızın lojistik regresyon modeli kurulmuştur. Ham veri seti için test veri setinde elde edilen karmaşıklık matrisi Şekil 3.1'deki gibidir.



Şekil 3. 1 Ham veri seti için karmaşıklık matrisi

Modelin karmaşıklık matrisi sonucuna göre, %93.1'lik bir doğruluk oranı ile modelin iyi performans sağladığı gözlenmektedir. Detaylara indiğimizde duyarlılığın %12 olduğu görülmektedir. Veri setinde 0 sınıfı %93 oranında, 1 sınıfı ise %7 oranında bulunduğu için, model 0 sınıfına göre yanlı eğitilmiş ve çoğunluk sınıfını tahminlerken başarılı olduğu için modelin doğruluk değeri yüksek çıkmıştır. Aslında, model azınlık sınıfını tahmin etmede oldukça başarısızdır.

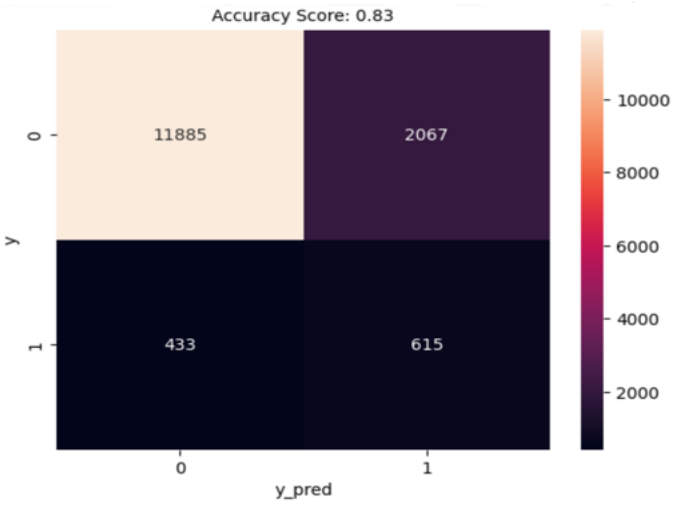
Veriyi dengeli hale getirmek için öncelikle yeniden örnekleme tekniklerinden SMOTE yöntemi denenmiştir. SMOTE yöntemi ile azınlık sınıfa benzer sentetik veriler üretilerek hedef oranı %50 olacak şekilde veri manipüle edilmiştir. Şekil 3.2'de SMOTE yönetime ait karmaşıklık matrisi gösterilmiştir.



Şekil 3. 2 SMOTE yöntemine ait karmaşıklık matrisi

SMOTE algoritması ile birlikte sentetik veri üretilerek 0 ve 1 sınıfındaki veri sayıları birbirine eşit olacak şekilde veri dengelenmiştir. Dengeli veri üzerinde lojistik regresyon modeli kurulmuştur. Model doğruluk oranı, ham veri seti ile kurulan modele göre %93.1'den %81.5'e inmiş, ama tahmin edilmek istenilen sınıftaki doğruluk değeri artmıştır. Duyarlılık oranı ise ham veri seti ile kurulan modele göre %12'den %59.8'e çıkmıştır.

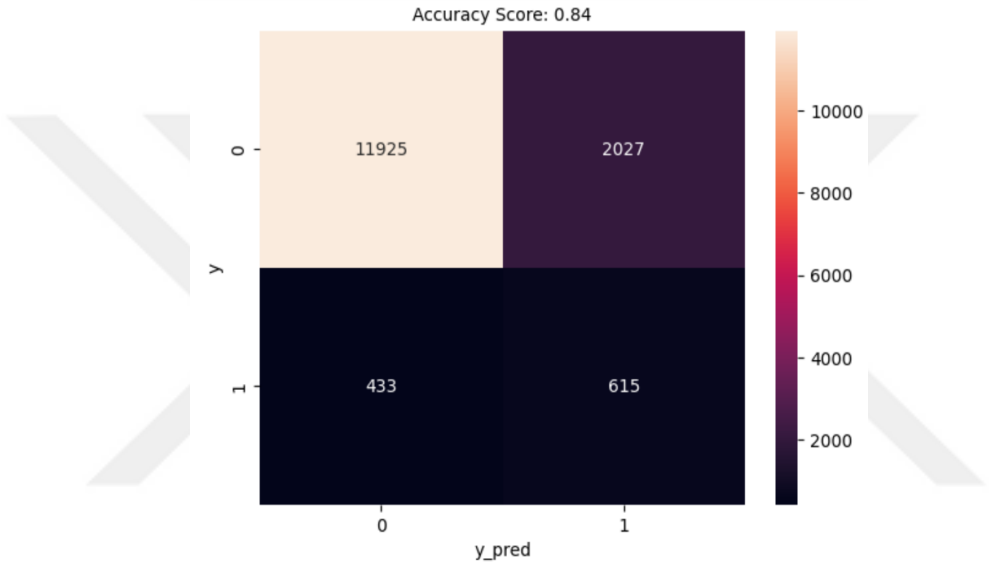
Ham veriyi dengeli hale getirmek için bir diğer yeniden örnekleme yöntemi olarak aşırı örnekleme yöntemi kullanılmıştır. Azınlık sınıfa ait gözlemler çoğaltılarak hedef oranı %50 yapılmıştır. Aşırı örnekleme yöntemi ile manipüle edilen veri üstünde kurulan modelin, test sonuçları Şekil 3.3'deki gibidir.



Şekil 3. 3 Aşırı örnekleme yöntemine ait karmaşıklık matrisi

Aşırı örnekleme algoritması azınlık sınıfındaki gözlemler çoğaltılarak 0 ve 1 sınıfındaki veri sayıları birbirine eşitlenmiştir. Dengeli veri üzerinde kurulan modelin doğruluk oranı %83.3 ve duyarlılık oranı %58.7 bulunmuştur. SMOTE yöntemi ile kıyaslandığına aşırı örnekleme yöntemi ile benzer performans elde edilmiştir.

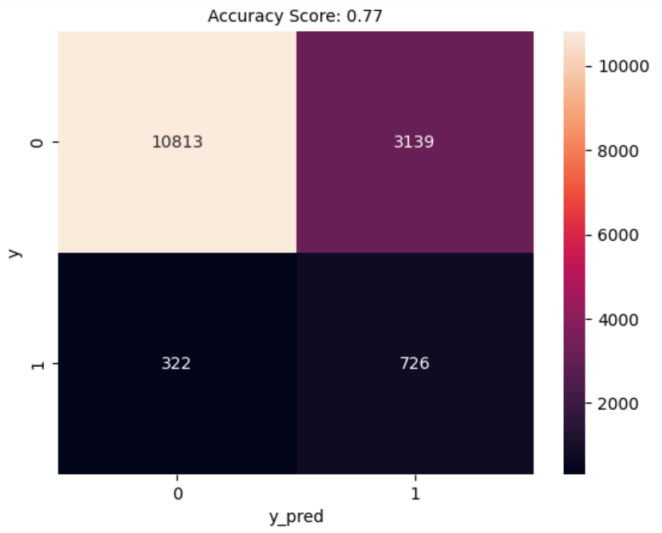
Dengesiz veri problemini çözmek için yeniden örnekleme tekniklerinden son olarak alt örnekleme yöntemine başvurulmuştur. Alt örnekleme yöntemi ile dengeli hale getirilen veri üstünde kurulan modelin, test sonuçları Şekil 3.4'deki gibidir.



Şekil 3. 4 Alt örnekleme yöntemine ait karmaşıklık matrisi

Alt örnekleme yöntemi ile birlikte çoğunluk sınıfa ait veriler silinerek, 0 ve 1 sınıfındaki veri sayıları birbirine %50-%50 olacak şekilde eşitlenmiştir. Dengeli veri üzerinde kurulan modelin doğruluk oranı %83.6, duyarlılık oranı %58.7 olarak elde edilmiştir.

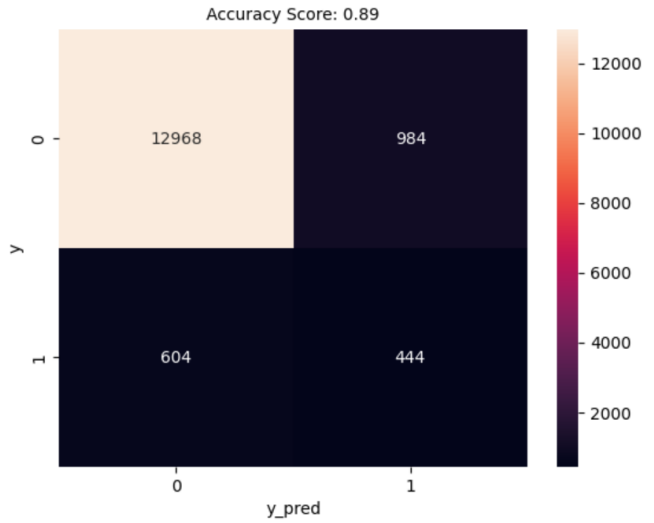
Dengesiz veri problemini çözmek için uygun kesim noktasının hesaplanması ilk olarak Youden indeksine göre yapılmış ve test sonuçları Şekil 3.5'te verilmiştir.



Şekil 3. 5 Youden indeksine göre karmaşıklık matrisi

Dengesiz veri seti üzerinde Youden indeksi kullanılarak 0.5 olan olağan kesim noktası 0.059472 olarak hesaplanıştır. $P(Y = 1|xi)$ olasılığı, 0.059472 uygun kesim değerinden büyük olan gözlemler 1 sınıfında, altında kalan gözlemler ise 0 sınıfında tahminlenmiştir. 0.5 olağan kesim noktasına göre kurulan modele göre doğruluk oranı %93.1'den %76.9'a inmiş, ancak duyarlılık oranı ise %12'den %69.3'e çıkmıştır.

Uygun kesim noktasını hesaplamak için kullanılan diğer yöntem, duyarlılık değerini kesinlik değerine eşit kılan noktanın seçilmesidir. Duyarlılığı kesinliğe eşitleyen yöntemin test sonuçları Şekil 3.6'da sunulmuştur.



Şekil 3. 6 Duyarlılığı kesinliğe eşitleyen yönteme göre karmaşıklık matrisi

Dengesiz veri seti üzerinde duyarlılığı kesinliğe eşitleyen uygun kesim noktası 0.113303 olarak hesaplanıştır. $P(Y = 1|xi)$ olasılığı, 0.113303 değerinden büyük olan gözlemler 1 sınıfında, altında kalan değerler ise 0 snıfında tahminlenmiştir. Model doğruluk oranı, Youden indekine göre %76.9'dan %89.4'e çıkmış, duyarlılık oranı ise %69.3'den %42.4'e inmiştir. Tüm modeller için performans metrikleri hesaplanmış ve Tablo 3.2'de verilmiştir.

Tablo 3. 3 Tüm modeller için performans metrikleri

	Ham Data	SMOTE	Aşırı Örnekleme	Alt Örnekleme	ROC-Youden İndex	ROC Kesinlik=Duyarlılık
DP	126	627	615	615	726	444
DN	13.837	11.594	11.885	11.925	10.813	12.968
YP	115	2.358	2.067	2.027	3.139	984
YN	922	421	433	433	322	604
YPO	6,6%	3,0%	3,1%	3,1%	2,3%	4,3%
Doğruluk	93,1%	81,5%	83,3%	83,6%	76,9%	89,4%
Hata	6,9%	18,5%	16,7%	16,4%	23,1%	10,6%
Kesinlik	52%	21%	23%	23%	19%	31%
Duyarlılık	12,0%	59,8%	58,7%	58,7%	69,3%	42,4%
F1- Skor	19,6%	31,1%	33,0%	33,3%	29,6%	35,9%

Tablo 3.2'ye göre, doğruluk oranı değeri en yüksek model ham veri seti üzerinde kurulan modeldir. Bunun nedeni, modelin çoğunluk sınıfını iyi tahmin etmesinden kaynaklandığı; aslında amacına uygun kullanılamayacak bir model olduğu bilinmektedir. Ham veri seti haricinde, doğruluk oranı (%89.4) en yüksek olan model, duyarlılık değerini kesinlik değerine eşitleyen uygun kesim noktası ile çalışam modeldir. Yeniden örnekleme yöntemleri içerisinde duyarlılığı (%59.8) en çok artırabildiğimiz yöntemin SMOTE yöntemidir. Dengesiz veriyi manipüle etmeden uygun kesim noktasını hesaplayarak en iyi duyarlılığı (%69.3) elde ettiğimiz modelin Youden İndeks kriterine göre elde edilen model olduğunu görüyoruz. Tablo 3.2'ye göre, yeniden örnekleme teknikleri ve uygun kesim noktasının hesaplanması yöntemlerine göre elde edilen modeller ile doğruluk ve duyarlılık metriklerinde elde edilen sonuçları iyileştirmek için hibrit yöntem denenecektir. SMOTE ve Youden indeks kriteri yöntemleri ile hibrit bir model çalıştırarak sonuçlar iyileştirilmeye çalışılacaktır.

Hibrit model için öncelikle veri seti SMOTE yöntemiyle 0.5 örneklem oranı ile dengeli hale getirilmiştir. Modelin iyi çalışabilmesi için gerçekten verinin %50-

%50 dengeli olması mı gerektiğini test edebilmek için, farklı örnekleme oranlarını denenmiştir. SMOTE tekniği ile çeşitli örneklem oranları ile yeniden dengelenen veri seti üzerinde Youden indeks kriterine göre uygun kesim noktası hesaplanmıştır. Elde edilen uygun kesim noktası ile kurulan lojistik regresyon analizi sonucunda elde edilen model performans metrikleri Tablo 3.3 ve Tablo 3.4'de sunulmuştur.

Tablo 3. 4 Örneklem oranı 0.1-0.5 için Smote-Youden indeks model metrikleri

	SMOTE (0.1)	SMOTE (0.1) + Youden	SMOTE (0.2)	SMOTE (0.2) + Youden	SMOTE (0.3)	SMOTE (0.3) + Youden	SMOTE (0.4)	SMOTE (0.4) + Youden	SMOTE (0.5)	SMOTE (0.5) + Youden
DP	156	758	246	760	312	776	371	843	409	774
DN	13.805	10.339	13.659	10.509	13.446	10.277	13.251	9.261	13.093	10.271
YP	147	3.613	293	3.443	506	3.675	701	4.691	859	3.681
YN	892	290	802	288	736	272	677	205	639	274
YPO	6,4%	2,1%	5,7%	2,1%	5,3%	1,9%	4,9%	1,5%	4,6%	2,0%
Doğruluk	93,1%	74,0%	92,7%	75,1%	91,7%	73,7%	90,8%	67,4%	90,0%	73,6%
Hata	6,9%	26,0%	7,3%	24,9%	8,3%	26,3%	9,2%	32,6%	10,0%	26,4%
Kesinlik	51,5%	17,3%	45,6%	18,1%	38,1%	17,4%	34,6%	15,2%	32,3%	17,4%
Duyarlılık	14,9%	72,3%	23,5%	72,5%	29,8%	74,0%	35,4%	80,4%	39,0%	73,9%
F1- Skor	23,1%	28,0%	31,0%	28,9%	33,4%	28,2%	35,0%	25,6%	35,3%	28,1%

Tablo 3.3'de örneklem oranı 0.1'den başlanarak 0.1 artım ile 0.5'e kadar denenmiş ve ardından Youden indeks yöntemine göre uygun kesim noktası hesaplanarak lojistik regresyon modeli kurulmuştur. Test veri seti performans sonuçlarına göre, SMOTE ve Youden indeks yöntemlerini hibrit ettiğimizde oluşturulan modellerin duyarlılık değerleri artmıştır. En yüksek duyarlılık değeri %80.4, örneklem oranı 0.4 olan SMOTE yöntemi ile dengelenen veriye Youden indeks kriteri uygulanması ile bulunmuştur.

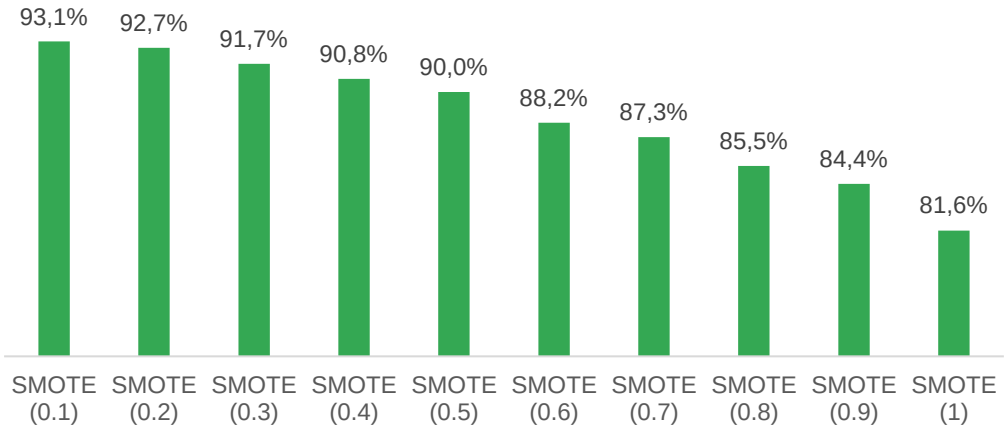
Tablo 3.4'te ise benzer şekilde örneklem oranı 0.6'dan başlanarak 0.1 artım ile 1'e kadar denenmiş veri seti için Youden indeks kriteri ile uygun kesim noktası hesaplanmış ve lojistik regresyon modeli kurulmuştur.

Tablo 3. 5 Örneklem oranı 0.6-1 için Smote-Youden indeks model metrikleri

	SMOTE (0.6)	SMOTE (0.6) + Youden	SMOTE (0.7)	SMOTE (0.7) + Youden	SMOTE (0.8)	SMOTE (0.8) + Youden	SMOTE (0.9)	SMOTE (0.9) + Youden	SMOTE (1)	SMOTE (1) + Youden
DP	463	845	499	822	554	841	588	824	626	834
DN	12.761	9.187	12.594	9.568	12.276	9.272	12.079	9.512	11.618	9.170
YP	1.191	4.765	1.358	4.384	1.676	4.680	1.873	4.440	2.334	4.782
YN	585	203	549	226	494	207	460	224	422	214
YPO	4,2%	1,5%	3,9%	1,6%	3,5%	1,5%	3,3%	1,6%	3,0%	1,5%
Doğruluk	88,2%	66,9%	87,3%	69,3%	85,5%	67,4%	84,4%	68,9%	81,6%	66,7%
Hata	11,8%	33,1%	12,7%	30,7%	14,5%	32,6%	15,6%	31,1%	18,4%	33,3%
Kesinlik	28,0%	15,1%	26,9%	15,8%	24,8%	15,2%	23,9%	15,7%	21,1%	14,9%
Duyarlılık	44,2%	80,6%	47,6%	78,4%	52,9%	80,2%	56,1%	78,6%	59,7%	79,6%
F1- Skor	34,3%	25,4%	34,4%	26,3%	33,8%	25,6%	33,5%	26,1%	31,2%	25,0%

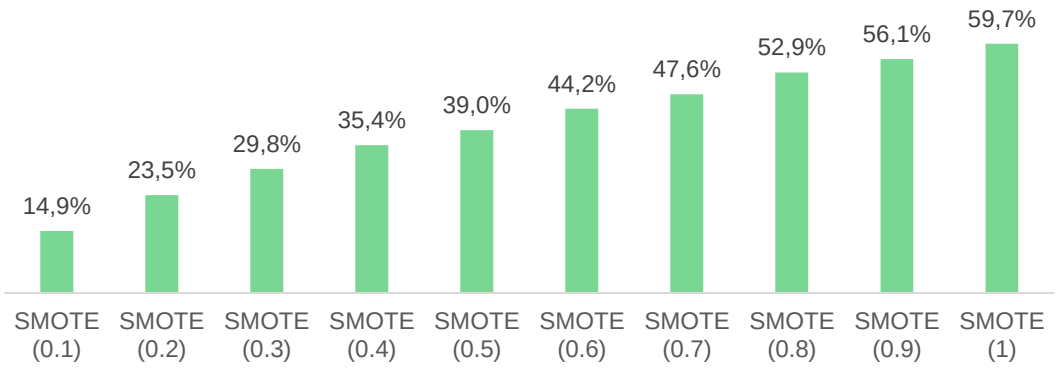
Tablo 3.4'de verilen test veri setindeki performans sonuçlarına göre, benzer şekilde SMOTE ve Youden indeks yöntemlerini birleştiren hibrit modellerde duyarlılık değerleri artmıştır. En yüksek duyarlılık değeri %80.6'dır.

Denenen hibrit yöntemlerde duyarlılık değeri artarken doğruluk oranında nispeten düşüş gözlenmiştir. Farklı örneklem oranları için SMOTE yöntemi doğruluk oranları Şekil 3.7'de verilmiştir.



Şekil 3. 7 Farklı örneklem oranları için SMOTE yönteminin doğruluk oranları

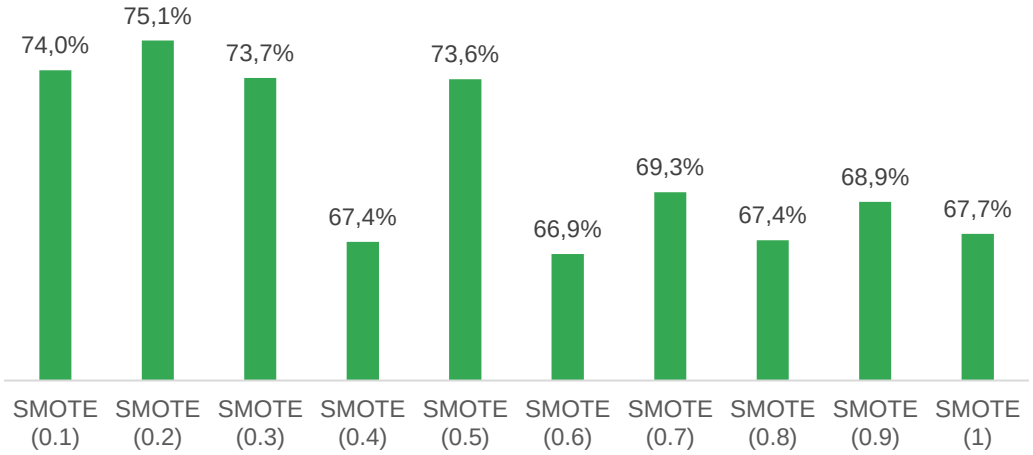
Şekil 3.7'ye göre azınlık sınıfı gözlem değerleri SMOTE örnekleme yöntemi ile arttıkça modeldeki doğruluk metriğin düştüğünü görüyoruz. Bu modelde önem verilen nokta araştırılan azınlık grubun doğru tahmini olduğu için duyarlılık metriğidir. Farklı örneklem oranları için SMOTE yönteminin duyarlılık metriği karşılaştırması Şekil 3.8'de verilmiştir.



Şekil 3. 8 Farklı örneklem oranları için SMOTE yönteminin duyarlılık oranları

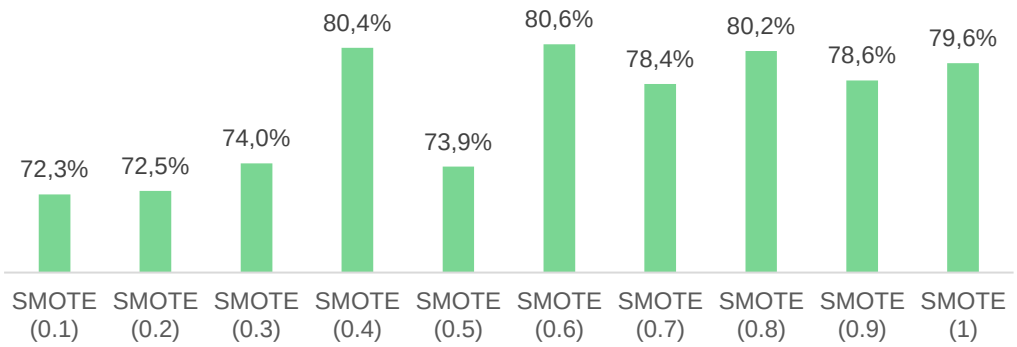
Şekil 3.8'e göre örneklem oranı 0.1'den 1'e arttıkça; yani azınlık sınıfı gözlem değerleri artıkça, duyarlılık metriğinin %14.9'dan %59.7'ye yükseldiği görülmektedir. Diğer bir ifade ile, azınlık sınıfı doğru tahminleme gücünün arttığı söylenebilir.

SMOTE yönteminde farklı örneklem oranları ile azınlık gözlem sayısı değiştirilen veri üzerinde Youden indeksine göre uygun kesim noktası hesaplanan hibrit modelin doğruluk ve duyarlılık oranları Şekil 3.9 ve Şekil 3.10'da verilmiştir.



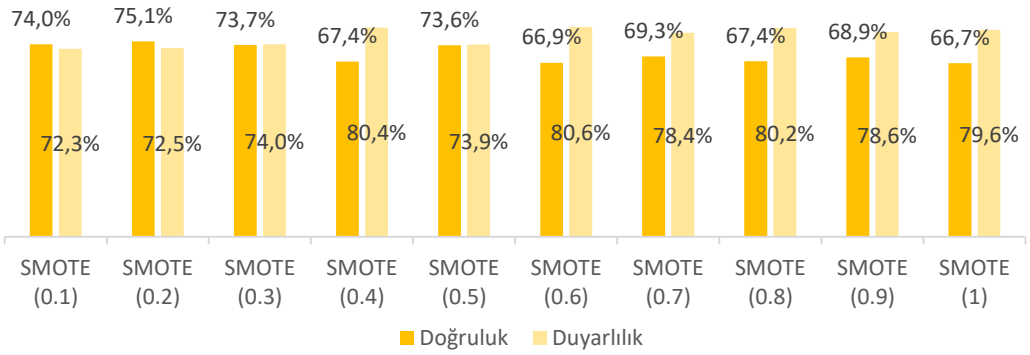
Şekil 3. 9 SMOTE ve Youden indeksi hibrit modeli için doğruluk oranları

Şekil 3.9'a göre hibrit modelde, sadece SMOTE yöntemde olduğu gibi veri sayısı artıkça azalan bir doğruluk grafiği görülmemektedir. Youden indeksini kullanılarak veriyi manüpüle etmeden doğruluk metriği (%75.1) baz alınarak en iyi performe eden model, 0.2 örneklem oranı ile sentetik veri üreten model olmuştur.



Şekil 3. 10 SMOTE ve Youden indeksi hibrit modeli için duyarlılık oranları

Şekil 3.10 incelendiğinde, 0.6 örneklem oranı ile SMOTE tekniği ile sentetik veri üreten model üzerinden Youden indeksine göre uygun kesim noktasını hesaplayan hibrit model %80.6 duyarlılık değeri ile en iyi performans sağlayan model olmuştur. Burada 0.4 örneklem modeli, 0.6 ve 0.8 örneklem modellerinin duyarlıklıklarının birbirine çok yakın olduğunu görüyoruz. Bu noktada aslında yüksek bir örneklem oranı ile azınlık sınıfı artıracak bir örnekleme yapılıp yapılmaması gerektiği sorgulanabilir. 0.4 örnekleme oranından 0.5'e geçerken duyarlılık değerinde %80.4'ten %73.9'a kadar düşüş yaşandığını görülmektedir. Şekil 3.11'de hibrit model için duyarlılık ve doğruluk metrikleri aynı grafikte sunulmuştur.



Şekil 3. 11 Hibrit model için doğruluk ve duyarlılık oranları

Şekil 3.11'de hibrit model için doğruluk ve duyarlılık oranlarını incelediğimizde, her iki metrik için en iyi sonucu veren tek model olmadığı, ancak sonuçların birbirine nispeten yakın çıktığı gözlenmiştir. Örneğin, duyarlılık metriğine göre en iyi model performans gösteren modeller, örneklem oranı 0.4 (%80.4) ve 0.6 (%80.6) seçildiğinde manipüle edilen veri üzerinden Youden indeksine göre

belirlenen uygun kesim değeri ile kurulan modellerdir. Eğer örneklem oranı 0.5 seçilip azınlık ve çoğunluk sınıftaki gözlem değerleri eşit olacak şekilde dengeli veri üzerinden Youden indeksine göre uygun kesim noktası hesaplansaydı duyarlılık değeri %73.9 elde edilirdi. Sonuç olarak, farklı örnekleme oranı ile SMOTE yöntemi ile Youden indeks yöntemini birleştiren hibrit modeller ile elde edilen performanslara göre, en iyi model performans sonucu için verinin %50-%50 dengeli olmak zorunda olmadığını söyleyebiliriz.

SMOTE yöntemine farklı örneklem oranları denendiğinde duyarlılık değeri en çok %59.7 elde edilmişti. Dolayısı ile bu sınıflandırma modeli özelinde, duyarlılığı en çok artıran yöntemin SMOTE ve Youden indeks yöntemlerinin hibrit edildiği model ile sınıflandırma yapan modeller olduğunu görüyoruz. Örneklem sayısını ne kadar artırırsak artıralım duyarlılıkta yeteri kadar artış sağlayamazken, Youden indeks kriterine göre uygun kesim noktası hesaplandığında duyarlılıktaki artışın %80.6'ya ulaştığını görebiliyoruz. Aynı zamanda doğruluktaki azalış da göze çarpmaktadır. Dengesiz veri setlerinde azınlık sınıfı iyi öğrenemediği zaman doğruluk metriği yanıltıcı sonuçlar verebilir. Bu nedenle doğruluk oranının nispeten daha düşük fakat duyarlılık oranının arttığı modellerin iyi performans gösterdiği kabul edilebilir. Yapılan araştırmada, dengesiz veri problem yer alması nedeni ile azınlık sınıfı tahmin etmede önemli olan ve artırılmak istenen metrik duyarlılık olduğu için 0.6 örneklem oranı için SMOTE yöntemi ile sentetik veri üreten ve bu veri üzerinden Youden indeksine göre hesaplanan uygun kesim noktası ile sınıflandırma yapan hibrit modelin satın alma eğilim tahmini için en iyi performans gösterdiği söylenebilir.

4
SONUÇ

Yapılan araştırmada, dengesiz veri problemi ile sınıflandırma analizi yapabilmek için veri manipülasyonu yapılmamış ham veri üzerinde kurulan bir lojistik regresyon modelinde, model değişkenleri sabit kalacak şekilde, yeniden örnekleme yöntemleri ve farklı kesim noktaları denenerek model performasları iyileştirilmeye çalışılmıştır. Literatürde sıkça belirtilen, sınıflandırma probleminde modelin iyi öğrenebilmesi ve iyi tahmin edebilmesi için %50-%50 dengeli veri olmalıdır argümanı üstünde durulmuştur. Bir e- ticaret sitesinden alınan veriler doğrultusunda satın alma eğilim tahmini modeli üzerinde çalışmalar gerçekleştirilmiştir. Amaç e-ticaret sitesine üye müşterilerin kadın giyim-aksesuar kategorisindeki satışlarını artırmaktır. Satın alma eğilimi olan müşteriler tespit edilip, bu müşterilere kampanya sunulacaktır. Model sonuçlarını karşılaştırmak için duyarlılık metriği göz önünde bulundurulmuştur. Çünkü dengesiz veri setlerinde azınlık sınıfı yeteri kadar öğrenemediği için doğruluk metriği yanıltıcı olabilmektedir. Yüksek duyarlılık, yanlış negatifleri minumuma indirerek azınlık sınıfın tespit edilme oranını artırır. Model amacına göre eğer hatalı tahminler bir maaliyet oluşturmuyorsa, doğruluğun azaldığı fakat duyarlılığın arttığı modeller seçilebilir. Bu çalışmada yanlış tahmin etme firma için bir maaliyet oluşturmamaktadır. Satın almayacak müşterileri, satın alacak olarak tahmin edip kampanya sunmak artı bir maaliyet yaratmamaktadır. Bu yüzden doğruluk metriği değil, daha çok önem verilen metrik duyarlılık metriği olmuştur. Gerçekten satın alacakları tahminlemek ve doğru zamanda doğru reklamı ve kampanyayı müşteriye sunmak satış hacmini artıracağından duyarlılık metriği yapılan tez çalışmasında ön plana çıkmıştır. Önce model üzerinde , %50-%50 sınıf dağılımı olacak şekilde, SMOTE, alt örnekleme, aşırı örneklem yöntemleri denenmiştir. Ayrı olarak olarak uygun kesim noktasını belirleyerek tahminleme yapmak için Youden indeksi ve duyarlılık değerini kesinlik değerine eşitleyen indeks yöntemleri denenmiştir. İyi çalışan örneklem yöntemlerinden ve indeks yöntemlerinden birer tane seçilerek hibrit model denemeleri yapılmıştır. Hibrit olarak çalıştırdığımız modelde, hem örnekleme oranı 0.1'den 1'e olmak üzere azınlık sınıfı için SMOTE örneklem yöntemleri

denenmiş, hem de üstüne uygun kesim noktası belirlemek için Youden indeksi çalışılmıştır. Bu hibrit modeller sonucunda, en iyi model sonucu için verinin %50-%50 dengeli olmak zorunda olmadığını söyleyebiliriz. SMOTE 0.4 ve 0.6 örnekleme oranı ile manipüle edilen veri ve Youden indeksi ile berlirlenen kesim değeri ile kurulan modellerin duyarlılık metriğinde en iyi çalışan modeller olduğu görülmüştür. Teoride en iyi tahmin için dengeli bir veriye sahip olunması gerektiği bilinirken, gerçek dünyada dengeli veri ile karşılaşmak neredeyse imkansızdır. Veriyi direkt %50-%50 dengeli hale getirme öğrenme ve test sonuçları üzerinde tutarsızlığa, gerçek veri üzerinde çalışmada yanlış ve yanlı tahmine yönlendirebilir. Dengesiz veri setleri model kurmak belirtildiği kadar büyük bir problem olmayabilir, burada aslında bakış açısını değiştirip başka metriklere yönelmek gerekebilir. %50-%50 dengeleştirdiğimiz ve dengeleştirmediğimiz türlerde de benzer duyarlılık sonuçları elde edilmiştir. Sonuç olarak, dengesiz veri setlerinde yapılacak çalışmanın amacı ve fayda maliyetine göre azınlık olan 1 sınıfını daha iyi tahmin etmek için yeniden örnekleme ve uygun kesim noktası hesaplayan yöntemleri birleştiren hibrit modeller arasından doğruluk oranının nispeten daha düşük çıktığı ancak duyarlılık oranının daha yüksek olduğu sınıflandırma modelleri tercih edilebilir.

KAYNAKÇA

[1] E. A. Greenleaf and D. R. Lehmann, "Reasons for substantial delay in consumer decision making," *Journal of Consumer Research*, vol. 22, no. 2, pp. 186–199, 1995.

[2] A. Marcus Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *ICML-2003 workshop on learning from imbalanced data sets II*, vol. 2, pp. 1-8, 2003.

[3] E. Duman, Y. Ekinci and A. Tanrıverdi, "Comparing alternative classifiers for database marketing: The case of imbalanced datasets," *Expert Systems with Applications*, vol. 39, no. 1, pp. 48-53, 2012.

[4] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve analysis for medical diagnostic test evaluation," *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627-635, 2013.

[5] B. Song, G. Zhang, W. Zhu and Z. Liang, "ROC operating point selection for classification of imbalanced data with application to computer-aided polyp detection in CT colonography," *International Journal Of Computer Assisted Radiology and Surgery*, vol. 9, no. 1, pp. 79-89, 2014.

[6] E. C. M. D. Santos, "Predictive modelling applied to propensity to buy personal accidents insurance products," *Ph.D. dissertion, Universidade Nova de Lisboa*, Portugal, 2018.

[7] S. Stubseid and O. Arandjelovic, "Machine Learning Based Prediction of Consumer Purchasing Decisions: The Evidence and Its Significance," *In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI, pp. 100-106, 2018.

[8] M. G. Munoz and P. Hedström, "Predicting Customer Behavior in E-commerce using Machine Learning," *Bachelor's Thesis, KTH Royal Institute of Technology*, Sweden, 2019.

[9] A. Gramegna and P. Giudici, “Why to buy insurance? an explainable artificial intelligence approach,” *Risks,* vol. 8, no. 4, pp. 1-9, 2020.

[10] Y. Zhang, “Prediction of Customer Propensity Based on Machine Learning,” *2021 Asia-Pacific Conference on Communications Technology and Computer Science,* ACCTCS, pp. 5-9, 2021.

[11] Y. Dani and M. A. Ginting, “Classification of Predicting Customer Ad Clicks Using Logistic Regression and k-Nearest Neighbors,” *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 1, pp. 98-104, 2023.

[12] M.E. Balaban and E. Kartal, *Veri Madenciliği ve Makine Öğrenmesi*. *2. Baskı*. İstanbul, Türkiye, Çağlayan, 2018.

[13] T.M. Mitchell, *Machine Learning. 1st edition,* New York, USA, McGraw-Hill, 1997.

[14] B.N. Alkış, “Çoklu lojistik regresyon ve k-en yakın komşu yöntemleri ile BIST 100 endeks getiri yönünün tahmini,” *Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi*, Türkiye, 2017.

[15] G. Kemalbay and B.N. Alkış, “Borsa endeks hareket yönünün çoklu lojistik regresyon ve k-en yakın komşu algoritması ile tahmini,” *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 27, no. 4, pp. 556-569, 2020.

[16] F. Bulut, İ. Kılıç and İ.F. İnce, “Beyin Tümörü Tespitinde Görüntü Bölütleme Yöntemlerine Ait Başarimlarin Karşilaştirilmasi ve Analizi,” *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, vol. 20, no. 58, pp.173-186, 2018.

[17] E. Kartal and Z. Özen, Dengesiz Veri Setlerinde Sınıflandırma. *Mühendislikte Yapay Zeka ve Uygulamaları*. *1. Baskı*, Sakarya, Türkiye, Sakarya Üniversitesi Kütüphanesi Yayınevi, 2017.

[18] M.E. Balaban and E. Kartal, *Veri Madenciliği ve Makine Öğrenmesi. 2. Baskı*, İstanbul, Türkiye, Çağlayan Yayınevi, 2018.

[19] S. Lessmann, B. Baesens, H. V. Seow and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of

research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124-136, 2015.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[21] M. J. Kelly, F. D. Dunstan, K. Lloyd and D. L. Fone, "Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods," *BMC Psychiatry*, vol. 8, no. 10, pp. 1-9, 2008.

[22] I. Unal, "Defining an optimal cut-point value in ROC analysis: an alternative approach," *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1-14, 2017.

[23] P. McCullagh and J. A. Nelder, *Generalized Linear Models. 2nd edition*, London, England, Chapman and Hall, 1989.

[24] R. Alpar, *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 5. Baskı*, Ankara, Türkiye, Detay, 2017.

[25] J. R. Hosmer, W. David, S. Lemeshow and R. X. Sturdivant, *Applied Logistic Regression. 3rd edition*, New Jersey, USA, John Wiley & Sons, 2013.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. Z. Rendeci, G. Kemalbay, "Dengesiz Veri Seti ile Bir E-Ticaret Şirketi İçin Satın Alma Eğilimi Tahmini". 7. Uluslararası Mühendislik ve Teknoloji Yönetimi Kongresi Bildiri Kitabı, ss. 931, 16-18 Nisan 2022, İstanbul, Türkiye.

Projeler

1. Z. Rendeci (Araştırmacı), G. Kemalbay (Yürütücü), "Dengesiz Veri Seti ile Bir E-Ticaret Şirketi İçin Satın Alma Eğilimi Tahmini", Yıldız Teknik Üniversitesi, FYL2022-4857 no'lu Bilimsel Araştırma Projesi.