

**T.C.
MİLLÎ SAVUNMA ÜNİVERSİTESİ
HEZÂRFEN HAVACILIK VE UZAY TEKNOLOJİLERİ
ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ PROGRAMI**

**MAKİNE ÖĞRENMESİ İLE E-TİCARETTE
MÜŞTERİ KAYBI TAHMİNİ**

YÜKSEK LİSANS TEZİ

**HACI ASLAN ÇAKIRDOĞAN
5181103**

**TEZ DANIŞMANI: DR.ÖĞR.ÜYESİ SABAHATTİN KEREM
AYTULUN**

**İSTANBUL
HAZİRAN 2021**

T.C.
MİLLÎ SAVUNMA ÜNİVERSİTESİ
HEZÂRFEN HAVACILIK VE UZAY TEKNOLOJİLERİ
ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ PROGRAMI

MAKİNE ÖĞRENMESİ İLE E-TİCARETTE
MÜŞTERİ KAYBI TAHMİNİ

YÜKSEK LİSANS TEZİ

HACI ASLAN ÇAKIRDOĞAN
5181103

Tezin Enstitüye Verildiği Tarih:
Tezin Savunulduğu Tarih: 22.06.2021

Tez Oy birliği / Oy çokluğu ile başarılı bulunmuştur.

Unvan Ad Soyad

İmza

Tez Danışmanı	Dr.Öğr.Üyesi Sabahattin Kerem AYTULUN
Jüri Üyeleri	Dr.Öğr.Üyesi Mehmet YÖRÜKOĞLU
	Dr.Öğr.Üyesi Atınç YILMAZ

İSTANBUL
HAZİRAN 2021

ÖZGÜNLÜK RAPORU

Tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve ç) Sonuç kısımlarından oluşan toplam 48 sayfalık kısmına ilişkin, 01/06/2021 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan özgünlük raporuna göre, tezimin benzerlik oranı %15'tir.

Uygulanan filtrelemeler:

- 1- Kaynakça hariç
- 2- Alıntılar hariç/dâhil
- 3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Millî Savunma Üniversitesi Hezârfen Havacılık ve Uzay Teknolojileri Enstitüsü Lisansüstü Tez Çalışması Özgünlük Raporu Alınması ve Kullanılması Uygulama Usul ve Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Hacı Aslan ÇAKIRDOĞAN

01/06/2021

ETİK BEYAN

Millî Savunma Üniversitesi Enstitüleri Lisansüstü Tez Hazırlama Kılavuzu'nda yer alan kurallarına uygun olarak hazırladığım bu tez çalışmasında; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu, bildirir; aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Bu tezdeki düşünce, görüş, varsayım, sav veya tezler bana aittir; Millî Savunma Bakanlığı, Millî Savunma Üniversitesi ve Hezârfen Havacılık ve Uzay Teknolojileri Enstitüsü sorumlu tutulamaz.

Hacı Aslan ÇAKIRDOĞAN

01/06/2021



Saygıdeğer hocalarıma

ÖNSÖZ ve TEŞEKKÜR

Üzerimde emeği bulunan tüm hocalarıma, hususen bilgi ve tecrübesi ile her daim bana yol gösteren saygıdeğer danışmanım Dr. Öğr. Üyesi Sabahattin Kerem AYTULUN'a, tez süresince yardımını esirgemeyen sevgili arkadaşım Hv. Müh. Tğm. Timuçin ÇINAR'a ve hayatım boyunca benden desteğini esirgemeyen değerli aileme şükranlarımı sunarım.

İstanbul; Haziran 2021

Hacı Aslan ÇAKIRDOĞAN



İÇİNDEKİLER

Sayfa

ÖZGÜNLÜK RAPORU	
ETİK BEYANI	
İTHAF	
ÖNSÖZ VE TEŞEKKÜR	
İÇİNDEKİLER	vii
TABLO LİSTESİ	ix
ŞEKİL LİSTESİ	x
SEMBOL LİSTESİ	xi
KISALTMALAR	xii
TÜRKÇE ÖZ	xiii
İNGİLİZCE ÖZ (ABSTRACT)	xiv
1. GİRİŞ	1
1.1. Literatür Araştırması	3
2. MÜŞTERİ KAYBI TAHMİNİ	6
2.1. Müşteri Kaybı	6
2.2. Müşteri Kaybının Modellenmesi	6
3. YÖNTEM	8
3.1. Kontratsız İş Tipinde Müşteri Kaybı Tahmini	8
3.2. Müşteri Kaybı Teknikleri	9
3.2.1. RFM Analizi	9
3.2.2. Lojistik Regresyon	9
3.2.3. Destek Vektör Makineleri (SVM).....	10
3.2.3.1 RBF Çekirdeğinin Parametreleri.....	11
3.2.4. Boosting.....	12
3.3. Çapraz Doğrulama ve Hiperparametre Ayarlama.....	14
3.4. Değerlendirme Ölçütleri	16
3.4.1. Karışıklık Matrisi	16
3.4.1. Kesinlik-Duyarlılık Eğrisi	17
3.4.2. Alıcı İşletim Karakteristiğinin Altındaki Alan	18
3.5. Yazılım ve Kütüphaneler.....	19
4. DENEYSEL ÇALIŞMA	20
4.1. Veri Kümesi.....	20
4.2. Açıklayıcı Veri Analizi	22
4.2. Hedef Değişkenin Oluşturulması ve Öznitelik Çıkarımı	23
4.2. Modelleme	25
4.2.1. Bölme Noktasının Model Başarısına Etkisi.....	29
5. SONUÇ	31

KAYNAKÇA		32
ÖZGEÇMİŞ		34



TABLÖLAR LİSTESİ

	Sayfa
Tablo 3.1: Çekirdek Fonksiyonları	11
Tablo 4.1: Veri Kümesinde Bulunan Öznitelikler ve Açıklamaları	20
Tablo 4.2: Veri Kümesinin Son Hali.....	25
Tablo 4.3: Yeni Değişkenlerin Tanımlayıcı İstatistikleri	25
Tablo 4.4: Model Değerlendirme	27



ŞEKİLLER LİSTESİ

	Sayfa
Şekil 1.1: Küresel E-ticaret Hacmi, Trilyon Dolar.....	2
Şekil 1.2: İnternette Sipariş Alan Perakende İşletmelerin Oranı	2
Şekil 3.1: Veri Kümesini Dönemlere Ayırma.....	9
Şekil 3.2: İkili DVM.....	11
Şekil 3.3: RBF çekirdeği ile Destek Vektör Sınıflandırıcıları	12
Şekil 3.4: Gradient Boosting Algoritmasına Basit Bir Örnek.....	13
Şekil 3.5 : K-kat Çapraz Doğrulama	15
Şekil 3.6 : Model eğitiminde çapraz doğrulama ve üst parametre ayarlama.....	16
Şekil 3.7 : Karışıklık Matrisi.....	17
Şekil 3.8: Kesinlik-Duyarlılık Eğrisi.....	18
Şekil 3.9 : ROC Eğrisi ve AUC	19
Şekil 4.1: Çalışmanın genel çerçevesi	21
Şekil 4.2: Her Bir Sütuna Ait Dolu Değer Sayısı.....	22
Şekil 4.3: Eksik Değerler Çıkarıldıktan Sonra Her Bir Sütuna Ait Dolu Değer Sayısı	23
Şekil 4.4: Modellerin ROC Eğrisi.....	27
Şekil 4.5: Modellerin Kesinlik-Duyarlılık Eğrisi.....	28
Şekil 4.6: Modellerin Karışıklık Matrisleri.....	28
Şekil 4.7: Öznitelik Önemleri	29
Şekil 4.8: Modellerin Yüzdebirlik – Doğruluk Grafiği	30

SEMBOL LİSTESİ

B	: Bölme noktası
C	: Düzenleştirme parametresi
γ	: Çekirdek katsayısı
k	: Çekirdek fonksiyonu



KISALTMALAR

RFM	: Recency, Frequency and Monetary
SVM	: Support Vector Machines
SVC	: Support Vector Classifier
DVM	: Destek Vektör Makineleri
NBD	: Negative Binomial Distrubition
NOI	: Number of Items
NOC	: Number of Cancelled Invoices
ROC	: Receiver Operating Characteristic
AUC	: The Area Under The Receiver Operating Characteristic
RBF	: Radial Basis Function

ÖZ

Makine Öğrenmesi ile E-Ticarette Müşteri Kaybı Analizi

Hacı Aslan ÇAKIRDOĞAN

Millî Savunma Üniversitesi, Hezârfen Havacılık ve Uzay Teknolojileri Enstitüsü
İstanbul, Haziran, 2021

Dijitalleşmenin her geçen gün artmasıyla neredeyse her iş internet üzerinden yapılır hale gelmiştir. Son yıllarda yaşanan küresel salgının neticesi olarak insanların evde kalmaya zorlanmasıyla da dijitalleşme iyice ivme kazanmıştır. Her türlü e-ticaretin artmasıyla çevrimiçi alışveriş önemli ölçüde gelişmiştir. E-ticaret pazarının büyümesiyle sahadaki aktörler ve rekabet de artmıştır. Mevcut müşterileri elde tutmak ve yenilerini çekmek e-ticaret siteleri için zorlu bir hal almıştır. Bir müşteriyi elde tutmanın yeni bir müşteri kazanmaktan ekonomik olarak daha avantajlı olduğu göz önüne alınırsa müşteri kaybı oranını düşürmek şirketler için temel iş hedefleri arasındadır. Sadık bir müşteri tabanı oluşturmak için şirketlerin, müşteri odaklı ve etkili pazarlama stratejileri oluşturması gerekir. Bununla beraber çevrimiçi müşterilerin şirketlere kazandırdığı büyük bir avantaj var: Çok miktarda tüketici verisi. E-ticaret siteleri topladıkları tüketici verilerini analiz ederek ayrılmaya meyilli müşterileri önceden tespit edebilirler. Böylece müşterileri kaybetmemek için gerekli aksiyonları da alabilirler. Bu çalışmada bir çevrimiçi perakende satış sitesi için kontratsız iş tipinde müşteri kaybı tahmini modeli geliştirilmiştir. Modelde çoğunlukla müşteri sınıflandırılmasında kullanılan RFM analizine dayanan değişken seçimi yapılmıştır. Ayrıca tahminleyici değişkenlerin modele katkısını da test ettik. İşin yapısından dolayı veri kümesinde hedef değişken olmadığından veri kümesini kalibrasyon ve gözlem olmak üzere iki kısma ayırdık. Daha sonra müşterileri her iki dönemde aktiflik durumuna göre kayıp veya aktif olarak sınıflandırarak hedef değişkeni belirlemiş olduk. Tahminleme Boosting, Lojistik Regresyon ve SVM algoritmaları ile yapılarak algoritmaların performansı kıyaslanmıştır.

Anahtar Sözcükler : Müşteri Kaybı Tahmini, Makine Öğrenmesi, E-Ticaret, RFM Analizi

Bilim Kodu : 90618

Sayfa Sayısı : 48

Tez Danışmanı : Dr.Öğr.Üyesi Sabahattin Kerem AYTULUN

ABSTRACT

Customer Churn Prediction In E-Commerce Using Machine Learning

Hacı Aslan ÇAKIRDOĞAN

National Defense University, Hezârfen Aeronautics and Space Technologies Institute
Istanbul, June, 2021

In an age where almost everything in life happens on the internet. As a result of the global epidemic in recent years, digitalization has gained momentum as people are forced to stay at home. With the increase of all types of e-commerce, online shopping has improved significantly. With the growth of the e-commerce market, the actors and competition in the field have also increased. Retaining existing customers and attracting new ones has become a challenge for e-commerce sites. Considering that retaining a customer is economically more advantageous than acquiring a new customer, reducing the churn rate is among the main business goals for companies. To build a loyal customer base, companies need to create customer-focused and effective marketing strategies. However, there is one big advantage that online customers bring to companies: A large amount of consumer data. By analyzing the consumer data they collect, e-commerce sites can detect customers who tend to leave in advance. Thus, they can take the necessary actions to avoid losing customers. In this study, a customer churn prediction model in the non-contractual business type has been developed for an online retailer. Variable selection was made based on RFM analysis, which is mostly used in customer segmentation, in the model. We also tested the importance of predictive variables to the model. Since there is no target variable in the data set due to the nature of the business, we divided the data set into two parts as calibration and observation. Then, we determined the target variable by classifying the customers as churn or active according to their activity status in both periods. The prediction was made with Boosting, Logistic Regression, SVM algorithms and the performance of the algorithms was compared.

Keywords : Churn Prediction, Machine Learning, E-Commerce, RFM Analysis

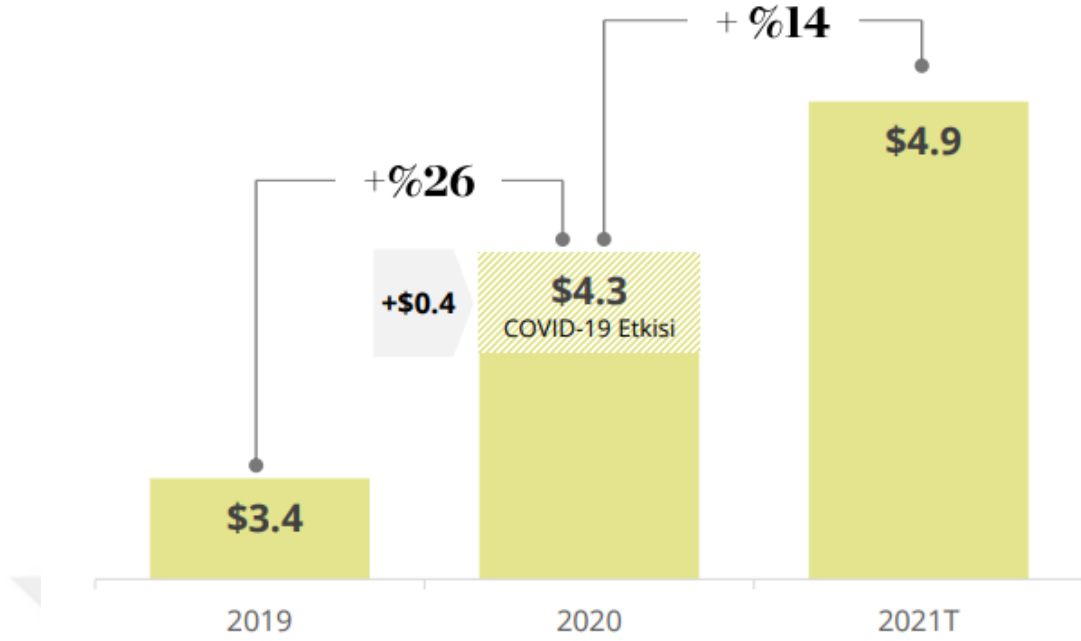
Science Code : 90618

Pages : 48

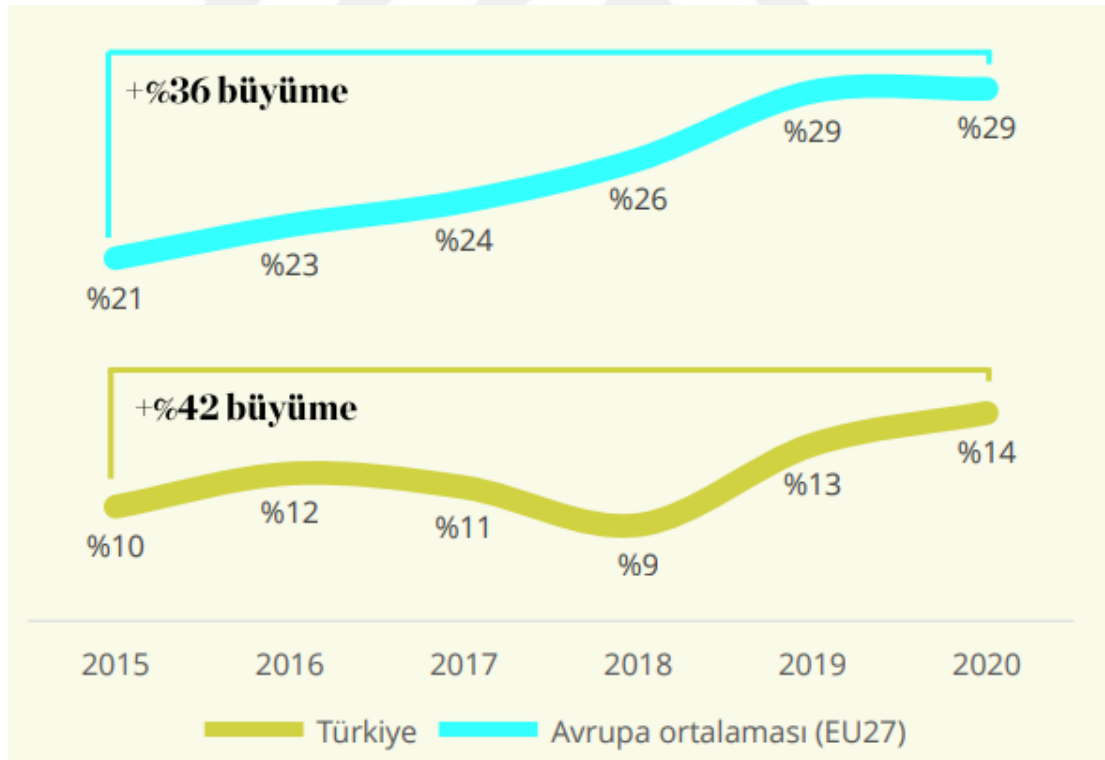
Supervisor : Dr.Öğr.Üyesi Sabahattin Kerem AYTULUN

1. GİRİŞ

Son yıllarda internete erişimin kolaylaşması ve internet teknolojilerinde yaşanan sıçramalarla beraber e-ticaret de hızla gelişmiştir. Öyle ki insanlar neredeyse her ihtiyacını internet üzerinden karşılayabilir hale gelmiştir. Özellikle 2020 yılında yaşanan küresel salgının neticesi olarak insanların evde kalmaya zorlanmasıyla e-ticarete büyüme iyice ivme kazanmıştır. Örneğin 2020 yılında dünya çapında internet kullanan her 5 kişiden 4'ü e-ticareti deneyimlemiştir. Küresel e-ticaret hacmi bir önceki yıla nazaran dünya çapında %26 artarak 4.3 trilyon dolara ulaşırken Türkiye'de %66 artarak 226 milyar lira seviyesine ulaşmıştır. E-ticaret ile talep artışına cevap vermek için yeni iş modelleri ortaya çıkmakta ve ekosistemdeki aktörler operasyonel ve dijital dönüşümler yaşamaktadır. Mesela 2020'de tüm gönderiler içerisinde e-ticaretin payı %60 olan lojistik ve taşımacılık sektöründe, mikro dağıtım gibi çeviklik ve teknoloji odaklı yeni değer önerileri gelişmeye başlamıştır. Yine aynı sektörde hacim büyüyerek %65 istihdam artışı gözlenmiştir (Bilişim Sanayicileri Derneği (TÜBİSAD), 2021). Hal böyleyken e-ticaret işletmelerinin sayısı gittikçe artmakta ve büyük ölçekte hizmet vermektedir. Böyle bir ortamda mevcut müşterileri elde tutmak ve yenilerini kazanmak e-ticaret pazarında rekabet anlamına gelir. Bu rekabette e-ticaret dünyasının ana hedefi haline gelen müşterilere kişiselleştirilmiş hizmet sunma ve müşteri memnuniyetini ve sadakatini sürekli iyileştirmede, çevrimiçi müşterilerin şirketlere kazandırdığı veri büyük avantaj sağlar. Şirketler müşterilerinin tüketim alışkanlıklarını eldeki verilere bakarak ölçümleyebilirler. Mesela müşterilerden bazılarını kaybetmişse müşterilerin terk etme nedenlerini verileri analiz ederek tespit edebilir, mevcut müşterilerin terk etme olasılıklarını bulabilir ve müşterileri kaybetmemek için gerekli aksiyonları alabilir. İşte bu örnek müşteri kaybı tahmini olarak adlandırılır.



Şekil 1.1: Küresel E-ticaret Hacmi, Trilyon Dolar.
Bilişim Sanayicileri Derneği (TÜBİSAD), 2021



Şekil 1.2: İnternette Sipariş Alan Perakende İşletmelerin Oranı.
Bilişim Sanayicileri Derneği (TÜBİSAD), 2021

Müşteri kaybı, müşteri şirketle ilişkiyi kesince gerçekleşir (Bardük, 2020). Müşteri kaybı tahmininde amaç, ayrılma ihtimali olan müşterileri tespit ederek onları elde

tutmak için birtakım tedbirler almaktır. Kayıp tahmini önemlidir çünkü bir müşteriye elde tutmak yeni bir müşteri kazanmaktan daha ucuzdur (Bagul ve diğ., 2021). Dolayısıyla şirketler müşteri sadakatini sağlayabilirlerse kaybettiklerinin yerine yenisini bulmak maliyetinden kaçınmış olurlar.

Bu tez çalışmasının başlıca hedefi, perakende sektöründe faaliyet gösteren bir şirkete ait müşterilerin satın alma alışkanlıklarından hareketle yüksek doğruluk değerine sahip bir müşteri kaybı tahmini modeli geliştirmektir. Uygulama kapsamında, bir e-ticaret şirketine ait veri kümesinden elde edilen davranışsal kullanıcı verileri analiz edilerek makine öğrenmesi algoritmaları ile ayrılmaya meyilli müşteriler tahmin edilmiştir.

Bu amaçla önce bu alanda yapılan çalışmalara değinilecek, ikinci bölümde müşteri kaybı tahmini tanımlarından ve önerilen metotlardan bahsedilecek, üçüncü bölümde veri kümesinin yapısı, veri temizleme ve veri ön işleme adımları anlatılacak, dördüncü bölümde kullanılan tahminleme modelleri detaylı olarak işlenecek ve sonuçları kıyaslanacak, beşinci ve son bölümde ise sonuç babında elde edilen bulgular değerlendirilecektir.

1.1 Literatür Araştırması

Bu tez çalışması için yapılan literatür araştırması sırasında e-ticaret sektöründe yapılmış olan müşteri kaybı tahmini araştırmaları ve uygulamaları incelenmiştir.

Shetty ve diğ. (2019) bir perakende işletmesi için müşteri kaybını tahmin etmede çeşitli algoritmaları karşılaştırmış ve müşterinin ait olduğu kümeye dayalı olarak önerilerde bulunmuştur. Her bir kayıp tahmin algoritmasının karşılaştırmalı çalışması yapılmış ve bunlar arasından Pareto / NBD seçilmiştir. Modelin girdi parametreleri, modelin ihtiyaçlarına uyacak şekilde ince ayarlanmıştır. Ürünler kategorize edilmiş ve müşteriler RMF puanlarına göre bir kümeye yerleştirilmiştir. Öneri, ait oldukları kümeye göre verilmiştir.

Guo ve Qin (2015) veri madenciliği tekniklerinden karar ağacı algoritmalarını kullanarak e-ticaret müşterilerinin temel bilgilerini analiz etmiş ve müşteri kayıplarının özelliklerini bulmuşlardır.

Jaeyalakshmi ve diğ. (2020) yapılandırılmış veriler, zorunlu özellikler, makine öğrenmesi teknikleri ile e-perakendede kayıp tahmini önermiştir. Çapraz doğrulamada daha iyi algoritmalar seçilmiştir. Topluluk algoritması kapanış

sonucudur ve kaybedilen müşteriler için uygun önleyici tedbirler önerir. Yazarlar ayrıca mevcut sistemi de incelemiştir. Yapılandırılmamış veya yarı yapılandırılmış verilerle mevcut model tahminleri yapılmıştır. İlgili çalışmalar, belirli bir makine öğrenme algoritması ile sonuçları belirtir. Çeşitli algoritmalar üzerinde karşılaştırmalı bir çalışma, doğruluğa dayalı en iyi modelle ortaya çıkar. Algoritmanın doğruluğu müşteri verilerine göre değişir.

Wu ve Meng (2016) olağan senaryoda, e-ticaret müşteri kayıp veri setlerinin dengesiz olduğunu belirtir. Kaybedilen müşterilerin sayısının, sadık tüketicileri büyük ölçüde aştığı bulunmuştur. Bu nedenle geleneksel algoritmalar çoğunluk kümesini tercih eder ve doğru sonuçlar vermez. Verileri dengelemek için geliştirilmiş SMOTE tekniği önerilmektedir. Daha sonra müşterileri sınıflandırmak için bu veriler üzerinde AdaBoost algoritması kullanılır.

Xia ve He (2018) çevrimiçi müşterilerin sözleşmesiz senaryosunu örnek olarak almış, deneysel araştırma için yerel bir e-ticaret web sitesinin işlem verilerini seçmiştir. Tek model-BP sinir ağı ve destek vektör makinesi temelinde, entegre öğrenme teorisini çevrimiçi alışveriş müşteri sınıflandırmasına uygulamıştır. Ampirik sonuçlar, birleşik tahmin modelinin isabet oranı, kapsama oranı, doğruluk oranı ve yükselme derecesi vb. konularda önemli bir gelişmeye sahip olduğunu göstermektedir.

Ahmad ve diğ. (2019) görselleştirme kullanarak bir keşfedici veri analizi, öznelilik seçimi için istatistiksel testler ve bir lojistik regresyon modeli kullanarak olası kayıp müşterileri tahmin etmek için veri madenciliği yöntemlerini uygulamışlardır. Lojistik regresyon modelinin, kayıp tahmin sürecinde daha iyi sonuçlar öngördüğü gözlemlenmiştir. Eşik değerlerini yükselterek ve çeşitli kombinasyonlarla doğru öznelilikleri seçerek daha iyi bir sonuç verecektir.

Kumar ve Chandrakala (2016) sadece bankacılık sektöründe değil, aynı zamanda müşteri katılımına büyük ölçüde bağlı olan diğer sektörlerde de, araştırmacılar tarafından kayıp tahmini için kullanılan en popüler makine öğrenimi algoritmalarını inceler. Burada incelenen her kayıp tahmin modeli, düşük doğruluk ve tahmine sahiptir.

Wu ve diğ. (2010) kümeleme analizi, karar ağacı, sinir ağı vb. gibi müşteri kaybı araştırmalarına birçok veri madenciliği tekniği uygular, bir e-ticaret müşteri kaybı modeli kurar ve müşteriyi elde tutmayı etkileyen faktörleri analiz eder. Ankette güven ile ilgili beş ana faktör var: Web sitesinin güvenliği ve güvenilirliği,

sürdürülebilir ve istikrarlı iş, pratik ürün reklamı, şirketin gücü, ürünlerin maliyet performansı.

Buckinx ve Van den Poel (2005) bir hızlı tüketim ürünleri perakendecisinden alınan gerçek hayat verilerini kullanarak, gelecekteki kısmi ayrılmanın başarılı bir şekilde tahmin edilebileceğini göstermiştir. Üç sınıflandırma tekniğini - lojistik regresyon, ARD sinir ağları, rastgele orman - kullanarak davranışsal olarak sadık müşteriler tarafından kısmi ayrılmayı tahmin etmek için bir model oluşturur. Düzenli bir ziyaret örüntüsü sergileyen yüksek sıklıkta alışveriş yapanların kısmi ayrılmasına odaklanmak, sözleşme dışı ortamlarda tam ayrılmanın tanımlanamaması sorununun üstesinden gelebilir. Sınıflandırma doğruluğu (PCC) ve alıcı işletim karakteristik eğrisi altındaki alan (AUC), sınıflandırıcı performansını değerlendirmek için kullanılır. Alternatif sınıflandırma teknikleri arasında performans açısından önemli farklılıklar yoktur.

Tamaddoni ve diğ. (2015) iki ayrı çevrimiçi perakendeciden alınan iki veri setinden yararlanarak optimal modelleme yaklaşımını belirlemek için çeşitli parametrik ve parametrik olmayan kayıp tahmin tekniklerinin performansını değerlendirir. Sonuçlar çoğu durumda, yani değişen müşteri taban boyutu, satın alma sıklığı ve müşteri kaybı oranı gibi, parametrik olmayan bir yöntem olan boosting tekniğinin diğer yöntemlerden daha iyi performans sağladığını gösterir. Çok küçük örneklem boyutlarında ise Pareto/NBD'nin daha iyi performans gösterir. Son olarak, müşteriler arasındaki kayıp oranları çok düşük olduğunda, lojistik regresyon diğer modellerden daha iyi performans gösterirken, düşük ve orta kayıp oranları için boosting tercih edilir.

2. MÜŞTERİ KAYBI TAHMİNİ

2.1 Müşteri Kaybı

Müşteri kaybı, müşterilerin şirketle iş yapmayı bırakma veya ayrılma eğilimi olarak tanımlanır (Kamakura ve diğ., 2005). Yeni müşteri kazanmak var olanı elde tutmaktan hem zor hem de maliyetlidir. Kayıplar şirket gelirini düşüreceğinden büyümenin önünde engeldir. Şirketler müşteri kayıp oranını takip ederek müşteriyi elde tutmak için yöntemler geliştirir. Ayrılmaya meyilli müşteriler indirim, ürün ve hizmet değişikliği gibi pazarlama faaliyetleri ile hedeflenir. Çevrimiçi perakende pazarının gelişimi sektördeki rekabeti artırmıştır. E-ticaret sektöründe müşteri kayıp oranı yüksektir. Bu durum işletmeleri kayıp oranını nasıl en aza indireceklerini düşünmeye sevk eder. Nitekim toplanan tüketim verisi müşterilerin gelecekteki satın alma eğilimlerini tahmin etmek için avantaj sağlar (Bagul ve diğ., 2021).

2.2 Müşteri Kaybının Modellenmesi

Literatürde müşteri kaybının tahminlenmesi, olasılık ve veri madenciliği modelleri olmak üzere iki geniş metodolojik kategoriye ayrılır. Ehrenberg (1959) tarafından sunulan olasılık modelleri, bireysel müşterilerin gözlemlenen davranışlarını modellemek ve gelecekteki davranışlarına ilişkin tahminlerde bulunmak için basit olasılık dağılımları kullanır. Ehrenberg'in (1959) çalışması, sırasıyla Pareto/NBD ve BG/NBD modellerini geliştirmek için Schmittlein ve diğ. (1987) ve Fader ve diğ. (2005) tarafından daha da genişletildi. Veriler arası örüntüleri bulmak için büyük veri kümelerinin analizi olan veri madenciliği, çeşitli sektörlerde kayıp tahmini için kullanılmıştır. Yine klasik yöntemlere alternatif olarak makine öğrenmesi modelleri de literatürde kullanılmıştır (Tamaddoni ve diğ., 2015).

Tamaddoni ve diğ. (2015) iki ayrı çevrimiçi perakendeciden alınan iki veri seti üzerinde yapmış olduğu çalışma boosting modelinin diğer modellere üstünlüğünü göstermiştir. Bu çalışmada bulguların genelleştirilebilirliğini iyileştirmek için, müşteri taban boyutu, satın alma sıklığı ve müşteri kaybı oranı gibi bir veri setinin temel özelliklerinin etkisi dikkate alınmıştır. Boosting, tüm müşteri satın alma sıklığı

seviyeleri için diğer yöntemlerden daha iyi performans gösterir. Çok küçük örneklem boyutlarında ise Pareto/NBD'nin daha iyi performans gösterir. Son olarak, müşteriler arasındaki kayıp oranları çok düşük olduğunda, lojistik regresyon diğer modellerden daha iyi performans gösterirken, düşük ve orta kayıp oranları için boosting tercih edilir. Bu bulgular, hangi modeli ne zaman kullanacağımızı müşteri tabanının özelliklerine dayalı olarak önermemize olanak tanır.



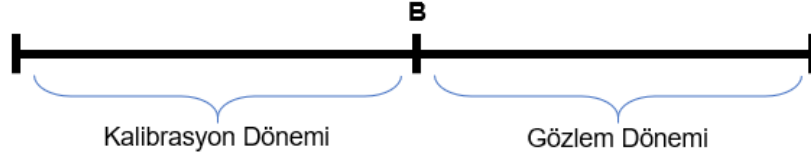
3. YÖNTEM

3.1 Kontratsız İş Tipinde Müşteri Kaybı Tahmini

Müşteri kaybı tahmini modellerini kontratlı ve kontratsız olarak ikiye ayırmak mümkündür. Kontratlı iş tipinde müşteri kontratı sonlandığında kayıp müşteri olarak sınıflandırılır. Fakat kontratsız iş tipinde müşteri kaybını tanımlamanın zorlukları vardır. Müşteriler ve şirketler arasında bir sözleşme olmadığından müşterileri kayıp olarak tanımlayacak kesin bir ölçüt yoktur. Bu nedenle kayıp müşteri ile uykudaki müşteriyi - iki işlem aralığındaki müşteriyi - kesin olarak ayırt etmek de mümkün değildir. Bununla beraber ikili sınıflandırma modeli oluşturmak amacıyla, hedef değişkenin çıkarılabilmesi için bir kayıp kriteri tanımlanmalıdır. Böyle bir durumda, veri kümesi periyotlara ayrılarak müşterilerin bu periyotlardaki aktiflik durumuna göre sınıflandırma yapılır.

Buckinx ve Van den Poel (2005), bu probleme veri kümesini “kalibrasyon” ve “gözlem” olmak üzere 1:1 oranında iki periyoda ayırarak çözüm sunmuşlardır. Kalibrasyon dönemi davranışsal veriyi analiz etmek içindir. Gözlem dönemi ise kayıp kararını vermek içindir. Müşteri kaybı kararı da şöyle verilir: Eğer müşteri kalibrasyon ve gözlem periyotlarında en az birer kez alışveriş yapmışsa “sadık müşteri” olarak sınıflandırılır; kalibrasyon döneminde alışveriş yapmış fakat gözlem döneminde hiç alışveriş yapmamışsa “kayıp müşteri” olarak sınıflandırılır.

Bu yöntemde dikkat edilmesi gereken bir husus vardır. Gözlem döneminin kısa olması durumunda alışveriş aralığı uzun olan aktif müşteriler kayıp olarak sınıflandırılabilir. Gözlem döneminin uzun olması durumunda ise kayıp müşteriler aktif olarak sınıflandırılabilir. Tamaddoni ve diğ. (2015) bu durumun önüne geçmek için müşterilerin ortalama alışveriş aralıklarının 99. yüzdebirliğinin kullanılmasını önermiştir. Bu çalışmada bu öneri ile beraber farklı yüzdebirlikler de hesaplanıp algoritma başarısı kıyaslanmıştır.



Şekil 3.1: Veri Kümesini Dönemlere Ayırma.

3.2 Müşteri Kaybı Teknikleri

Müşteri kaybı tahmini bir ikili sınıflandırma problemi olduğundan herhangi bir sınıflandırma algoritması ile uygulanması mümkündür. Fakat bu çalışmada, literatürde uygulanmış ve en iyi sonucu vermiş algoritmaları yani DVM, Boosting ve Lojistik Regresyon algoritmalarını kullanacağız (Tamaddoni ve diğ., 2015).

3.2.1 RFM Analizi

RFM (Recency, Frequency, Monetary) analizi, müşteri segmentasyonu için bir pazarlama modelidir. Müşteri davranışına dayanır. Müşterileri, işlem geçmişlerine göre yani ne kadar yakın zamanda, ne sıklıkta ve ne kadar satın aldıklarına göre gruplandırır. RFM analizi, iş ilişkisini bırakma olasılığı yüksek olan müşterileri belirlemek için müşterileri çeşitli kümeler ayırmaya yardımcı olur (Bagul ve diğ., 2021).

RFM'i oluşturan maddeler şöyledir:

- i. Recency(Güncellik): Son işlemden bu yana geçen süre.
- ii. Frequency(Sıklık): İşlem sıklığı.
- iii. Monetary(Parasallık): Ödenen toplam tutar.

Buckinx ve Van den Poel (2005) RFM değişkenlerinin müşteri kaybı tahminlemesinde en iyi tahminleyici değişkenler olduğunu söyler. Bundan ötürü RFM analizi bu çalışmada öznitelik çıkarımı için kullanılmıştır.

3.2.2 Lojistik Regresyon

Lojistik regresyon hedef sınıfların olasılıklarını tahmin eden, yaygın olarak kullanılan güçlü ve sağlam bir istatistiksel modelleme tekniğidir. Adından da anlaşılacağı üzere lojistik regresyon tıpkı doğrusal regresyon modeli gibi parametrik, yapısal ve açıklayıcı değişkenlerle bağımlı değişkenler arasında lineer ilişkiyi

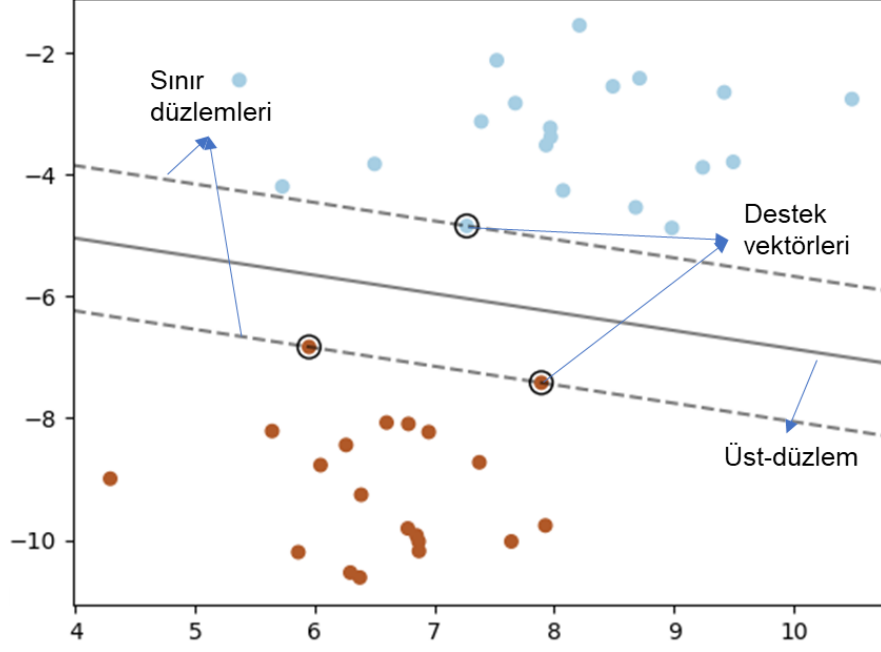
öngören bir model biçimidir. Lojistik regresyon ikili sınıflandırma problemlerini tahmin etme için kullanılır. Problemin konusu olan olayın olma ihtimali P ile tanımlanırsa lojistik regresyon aşağıdaki lineer modeli çözer:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (3.1)$$

Lojistik regresyon, açıklayıcı değişkenler üssel temel fonksiyonlar ile dönüştürülmediği sürece doğrusal karar sınırları oluşturur. Bu da modelin tahminleme performansı düşünüldüğünde en büyük dezavantajlarından birisidir. Çünkü gerçek dünyada birçok problem doğası gereği doğrusal olmayan karar sınırlarını gerektirmektedir (Şenyürek ve Alp, 2019).

3.2.3 Destek Vektör Makineleri

DVM, sınıflandırma, regresyon ve aykırı değerlerin tespiti için kullanılan güçlü ve çok yönlü bir gözetimli öğrenme yöntemidir. DVM, karmaşık ancak küçük veya orta ölçekli veri kümelerinin sınıflandırılması için çok uygundur (Géron, 2019). DVM, çevrimiçi perakendede müşteri kaybını tahmin etmek için başarıyla uygulanmıştır (Yu ve diğ., 2011). İkili bir sınıflandırma probleminde SVM, iki sınıf arasındaki sınır mesafesini maksimize eden üst düzlemi bulmayı amaçlar (Tamaddoni ve diğ., 2015). Bir destek vektör makinesi, yüksek veya sonsuz boyutlu bir uzayda bir üst-düzlem veya üst-düzlemler kümesi oluşturur. Sezgisel olarak, herhangi bir sınıfın en yakın eğitim veri noktalarına (fonksiyonel sınır olarak adlandırılan) en büyük mesafeye sahip olan üst-düzlem tarafından iyi bir ayırım elde edilir, çünkü genel olarak sınır ne kadar büyük olursa, sınıflandırıcının genelleme hatası o kadar düşük olur. Şekil 3.2, “destek vektörleri” olarak adlandırılan sınır düzlemlerinde üç noktayla, lineer olarak ayrılabilir bir problem için karar fonksiyonunu göstermektedir. Karar fonksiyonu için farklı çekirdek fonksiyonları kullanılabilmesi algoritmayı çok yönlü kılar (Herrero-Lopez, 2011). Bu çalışmada, algoritma çekirdeği olarak dairesel tabanlı fonksiyon (radial basis function) kullanılmıştır.



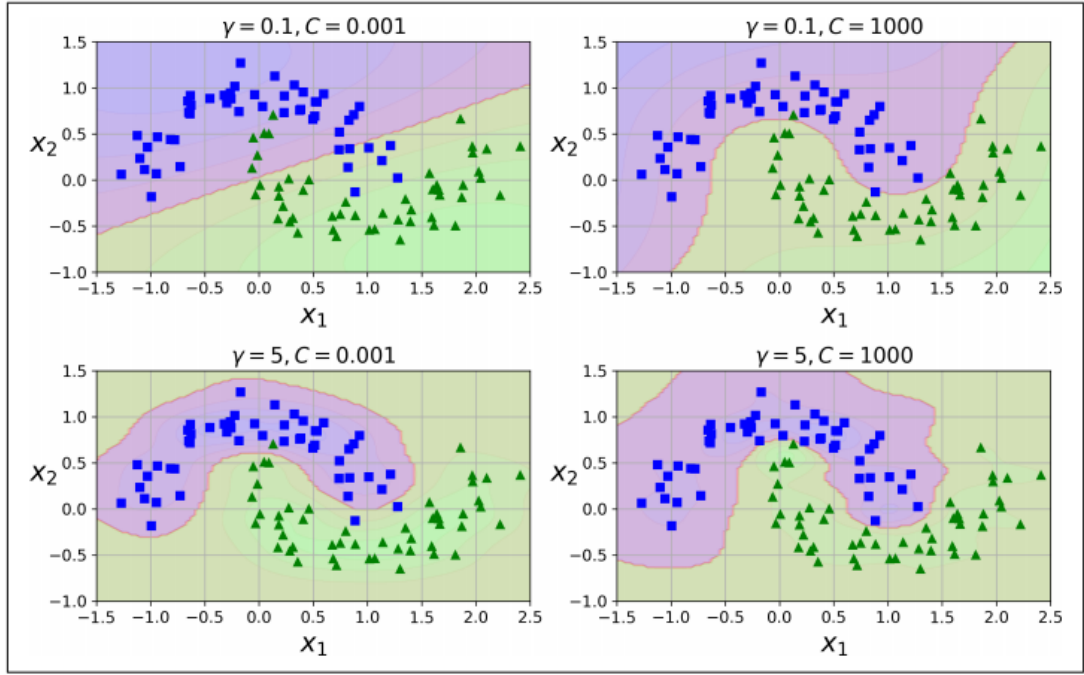
Şekil 3.2: İkili DVM.

Tablo 3.1: Çekirdek Fonksiyonları.

Doğrusal çekirdek	$k(x_i, \bar{x}_j) = \bar{x}_i \bar{x}_j$
Çokterimli çekirdek	$k(\bar{x}_i, \bar{x}_j) = (c_1 \bar{x}_i \bar{x}_j + c_2)^{c_3}$
Dairesel tabanlı çekirdek	$k(\bar{x}_i, \bar{x}_j) = e^{-\beta \ \bar{x}_i - \bar{x}_j\ ^2}$
Sigmoid çekirdek	$k(\bar{x}_i, \bar{x}_j) = \tanh(c_1 \bar{x}_i \bar{x}_j + c_2)$

3.2.3.1 RBF Çekirdeğinin Parametreleri

Dairesel Tabanlı Fonksiyon (RBF) çekirdeği ile bir DVM'yi eğitirken, iki parametre dikkate alınmalıdır: C ve gama. Tüm DVM çekirdeklerinde ortak olan C parametresi, karar yüzeyinin basitliğine karşı eğitim örneklerinin yanlış sınıflandırılmasını değiştirir. Düşük C, karar yüzeyini pürüzsüz hale getirirken, yüksek C, tüm eğitim örneklerini doğru bir şekilde sınıflandırmayı amaçlar. gama, tek bir eğitim örneğinin ne kadar etkisi olduğunu tanımlar. gama ne kadar büyükse, diğer örneklerin etkilenmesi için o kadar yakın olması gerekir (Géron, 2019).

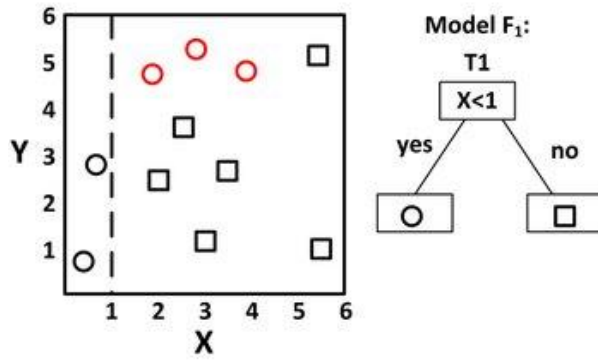


Şekil 3.3: RBF çekirdeği ile Destek Vektör Sınıflandırıcıları.

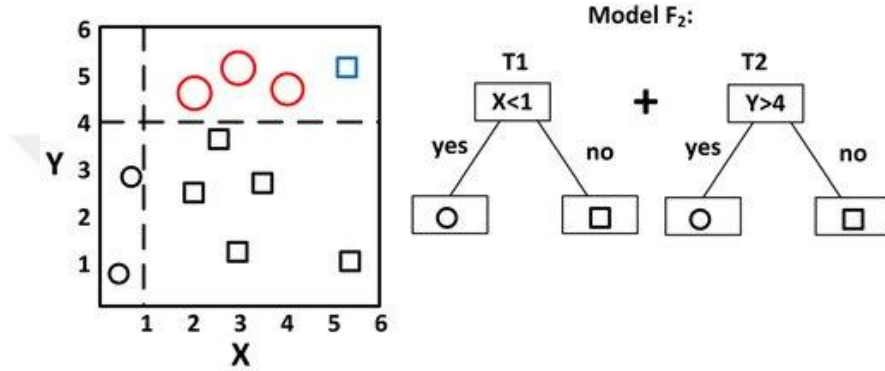
3.2.4 Boosting

Boosting algoritmaları birer topluluk öğrenmesi algoritmalarıdır. Algoritma iteratif bir algoritmadır. İterasyonda bir önceki ağacın yanlış tahminlediği örneklerle daha yüksek ağırlık vererek ilerler. Böylelikle model her iterasyonda tahmin edilmesi zor örnekleri ayırt edebilmek için zorlanır. Bu şekilde iterasyonlar ilerledikçe tahmin edilmesi zor olan örneklerin arkasındaki kural setleri öğrenilmeye başlanır. Modelin oluşturduğu her bir tekil ağacın tekil genelleme performansına göre bir ağırlığı vardır. Yeni örneklerin tahmini her bir tekil ağaç kararının bu ağırlıklara göre ortalaması ile verilir. Genel mekanizması bu şekilde olan boosting algoritmalarının birçok varyasyonu bulunmaktadır. Bunlardan en çok bilinenleri AdaBoost, Gradient Boosting Machines (GBM) ve XGBoost algoritmalarıdır (Şenyürek ve Alp, 2019). Topluluk öğrenmesi algoritmaları içerisinde boosting tekniği üstün kayıp tahmini yeteneği nedeniyle popülerdir (Lemmens ve Croux, 2006).

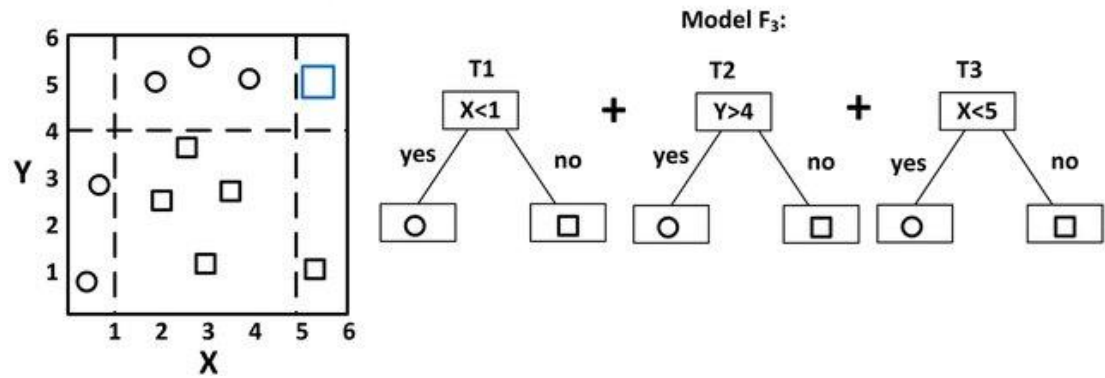
Iteration 1



Iteration 2



Iteration 3



Şekil 3.4: Gradient Boosting Algoritmasına Basit Bir Örnek.
Zhang ve diğ. (2018)

3.3 apraz Doğrulama ve Hiperparametre Ayarlama

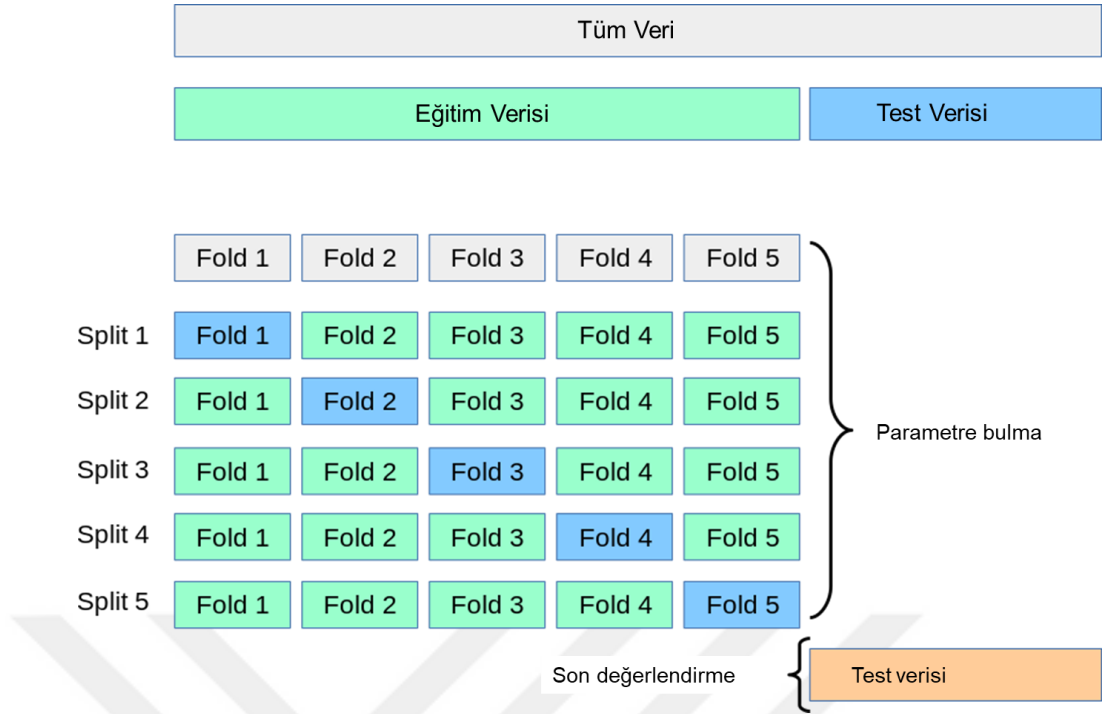
Bir tahmin fonksiyonunun parametrelerini öğrenmek ve aynı veriler üzerinde test etmek metodolojik bir hatadır: Az önce gördüğü örneklerin etiketlerini tekrar eden bir model mükemmel bir puana sahip olur, ancak henüz görülmeyen veriler üzerinde faydalı herhangi bir şeyi tahmin edemezdi. Bu duruma aşırı öğrenme (overfitting) denir. Bundan kaçınmak için bir makine öğrenimi deneyi gerçekleştirirken mevcut verilerin bir kısmını test seti olarak tutmak yaygın bir uygulamadır.

Tahminciler için farklı ayarları (hiperparametreler) değerlendirirken, parametreler tahmin edici en iyi şekilde performans gösterene kadar ince ayar yapılabileceğinden, test setinde hala aşırı öğrenme riski vardır. Bu şekilde, test seti hakkındaki bilgiler modele sızabilir ve değerlendirme metrikleri artık genelleme performansı hakkında rapor vermez. Bu sorunu çözmek için, veri kümesinin bir başka parçası da "doğrulama kümesi" olarak düzenlenebilir: eğitim, eğitim kümesinde devam eder, ardından doğrulama kümesinde değerlendirme yapılır ve deney başarılı görüldüğünde , test seti üzerinde son değerlendirme yapılabilir. Bununla birlikte, mevcut verileri üç kümeye bölerek, modeli öğrenmek için kullanılabilecek örnek sayısını büyük ölçüde azaltırız ve sonuçlar, (eğitim, doğrulama) küme çifti için belirli bir rastgele seçime bağlı olabilir.

Bu soruna bir çözüm, apraz doğrulama (cross validation) adı verilen bir prosedürdür. Nihai değerlendirme için bir test seti hala tutulmalıdır, ancak apraz doğrulama yaparken doğrulama setine artık ihtiyaç yoktur. K-katlı apraz doğrulama (k-fold cross validation) olarak adlandırılan temel yaklaşımda, eğitim seti k tane daha küçük kümeye bölünür. K “katlarının” her biri için aşağıdaki prosedür izlenir:

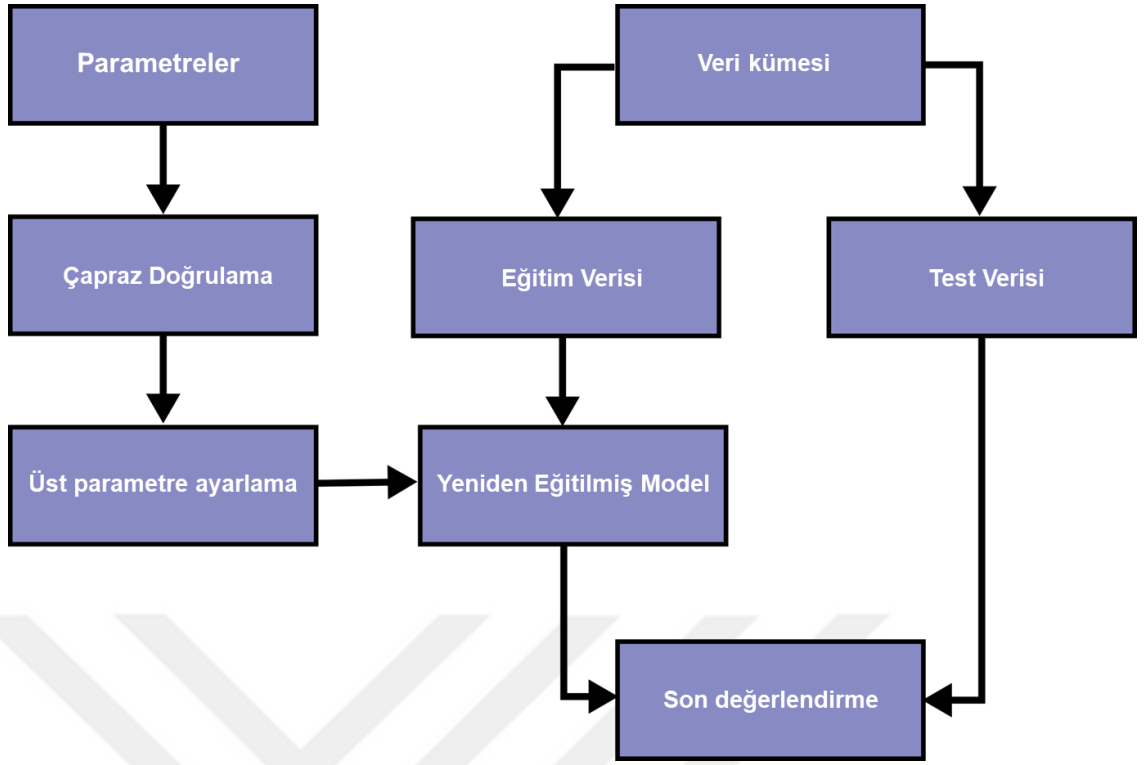
- Model, eğitim verisi olarak katların k-1'i kullanılarak eğitilir
- Elde edilen model, verilerin geri kalan kısmında doğrulanır

K-kat apraz doğrulama tarafından rapor edilen performans ölçüsü, daha sonra döngüde hesaplanan değerlerin ortalamasıdır.



Şekil 3.5 : K-kat Çapraz Doğrulama.

Makine öğrenmesi algoritmalarının uygulanması için model oluşturmada genellikle iki tür parametre eğitilir ve kullanılır. Daha iyi performans için modele geçirilen hiper parametreler ve yürütme aşamasında algoritma tarafından öğrenilen diğerleri. Hiperparametreler, tahmin ediciler içinde doğrudan öğrenilmeyen parametrelerdir. Hiper parametreler rastgele seçilebilir ve daha iyi bir model için test edilebilir, ancak deneyerek en iyi parametre kombinasyonunu aramak ve bulmak hesaplama açısından ayrıntılı ve zaman alıcıdır. Çalışmamızda hiper parametrelerin seçiminde grid arama yöntemi kullanılmıştır. Belirtilen hiper parametrelerin her kombinasyonu, grid arama yönteminde modeli oluşturmak için kullanılır ve ardından model belirtilen kriterlere göre değerlendirilir. Grid aramada, parametre değerleri matris formunda tanımlanır ve bu parametrelerin kombinasyonları en iyi performansı elde etmek için bir model için değerlendirilir.



Şekil 3.6 : Model Eğitiminde Çapraz Doğrulama Ve Üst Parametre Ayarlama.

3.4 Değerlendirme Ölçütleri

Sınıflandırma problemleri için karışıklık matrisi (confusion matrix), doğruluk (accuracy), alıcı işletim karakteristiğinin altındaki alan (the area under the receiver operating characteristic - AUC) gibi birkaç değerlendirme kriteri vardır. Bu çalışmada değerlendirme ölçütü olarak literatürde en çok tercih edilen doğruluk, AUC ve kesinlik-duyarlılık eğrisi (precision-recall curve) ölçütleri kullanılmıştır.

3.4.1 Karışıklık Matrisi

Karışıklık matrisi (confusion matrix), sınıflandırma performansını ölçmek için kullanılan araçlardan biridir ve hem ikili hem de çok sınıflı problemler için kullanılır. Bu tablo, model çıktısından tahmin edilen değerlere karşı gerçek girdi verisi sınıfındaki değerleri gösterir. Kesinlik, duyarlılık, F1-değeri gibi diğer performans ölçülerinin hesaplamaları, karışıklık matrisindeki değerlerden yapılır. Öğrenme algoritması ile maksimum tahmin doğruluğunun sağlanabilmesi için gerçek pozitif ve gerçek negatif değerlerin maksimize edilmesi hedeflenir (Tanveer, 2019).

		Gerçek	
		1	0
Tahmin	1	TP	FP
	0	FN	TN

Şekil 3.7: Karışıklık Matrisi.

Doğruluk, toplam gözlemlerden doğru tahmin edilen gözlemlerin bir ölçüsüdür.

Karışıklık matrisi değerleri kullanılarak doğruluk şu şekilde hesaplanabilir:

$$\text{Doğruluk} = \frac{TP+TN}{TP+FP+FN+TN} \quad (3.2)$$

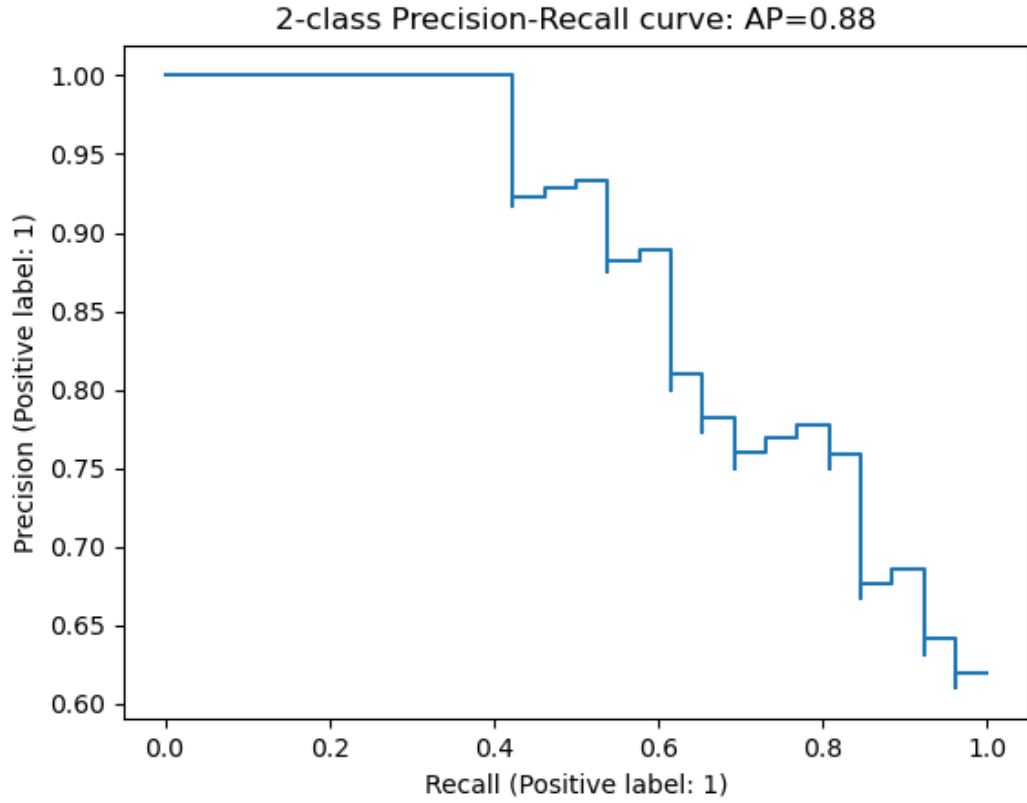
$$\text{Kesinlik} = \frac{TP}{TP+FP} \quad (3.3)$$

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \quad (3.4)$$

3.4.2 Kesinlik-Duyarlılık Eğrisi

Doğruluk ve AUC, öğrenme modellerinin performans değerlendirmesi ve karşılaştırması için yaygın olarak kullanılır ancak dengesiz veri setinde, alternatif bir ölçü olan kesinlik-duyarlılık eğrisi kullanılması tavsiye edilir. Kesinlik-duyarlılık eğrisi, olasılık eşik değerleri kullanılarak pozitif tahmin değeri (yani kesinlik) ile gerçek pozitif oran (yani duyarlılık) arasındaki dengedir. Dolayısıyla kesinlik-duyarlılık eğrisi, azınlık sınıfının doğru tahminini dikkate alır. Kesinlik ve duyarlılık arasındaki denge, farklı eşik değerleri test edilerek elde edilebilir, çünkü eşik değerinin düşürülmesi, duyarlılık değerini (gerçek pozitif sayısı) artıracak ve kesinlik değerini azaltacaktır (Tanveer, 2019). Ortalama Kesinlik (Average Precision - AP), her bir eşikte elde edilen kesinlik değerlerinin ağırlıklı ortalaması gibi bir grafiği özetler.

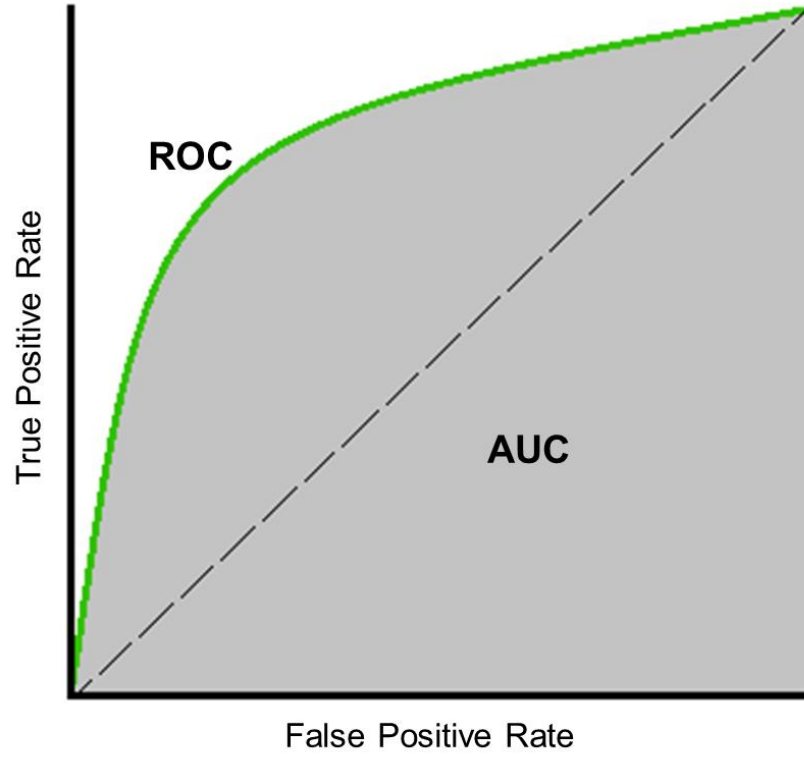
$$AP = \sum_n (Duyarlılık_n - Duyarlılık_{n-1}) Kesinlik_n \quad (3.5)$$



Şekil 3.8: Kesinlik-Duyarlılık Eğrisi.

3.4.3 Alıcı İşletim Karakteristiğinin Altındaki Alan

ROC eğrisi, ayırım eşiği değıştikçe ikili sınıflandırıcı sistemin performansını gösteren grafiksel bir çizimdir. ROC, sırasıyla x ve y eksenini boyunca gerçek pozitif orana karşı yanlış pozitif oran çizer. AUC puanı ROC eğrisinden hesaplanır ve değeri 0 ile 1 arasındadır. Eğrinin altındaki alan değeri ne kadar büyük olursa model performansı o kadar iyi olur, eğri altında alan değeri 0,5 veya altında olan bir model iyi performans göstermez olarak kabul edilir. Ancak ROC eğrilerinin, sınıf dengesizliğine karşı duyarsızlığı bu ölçüt için dezavantajdır (Tanveer, 2019).



Şekil 3.9: ROC Eğrisi ve AUC.

3.5 Yazılım ve Kütüphaneler

Bu çalışmada veri analizi ve modelleme python dilinde kodlanmıştır. Veri içe aktarma ve veri ön işlemede pandas kütüphanesi, sayısal hesaplamalar için NumPy, veri görselleştirmede matplotlib kütüphanesi, model uydurma, model seçme ve değerlendirmede scikit-learn kütüphanesi kullanılmıştır.

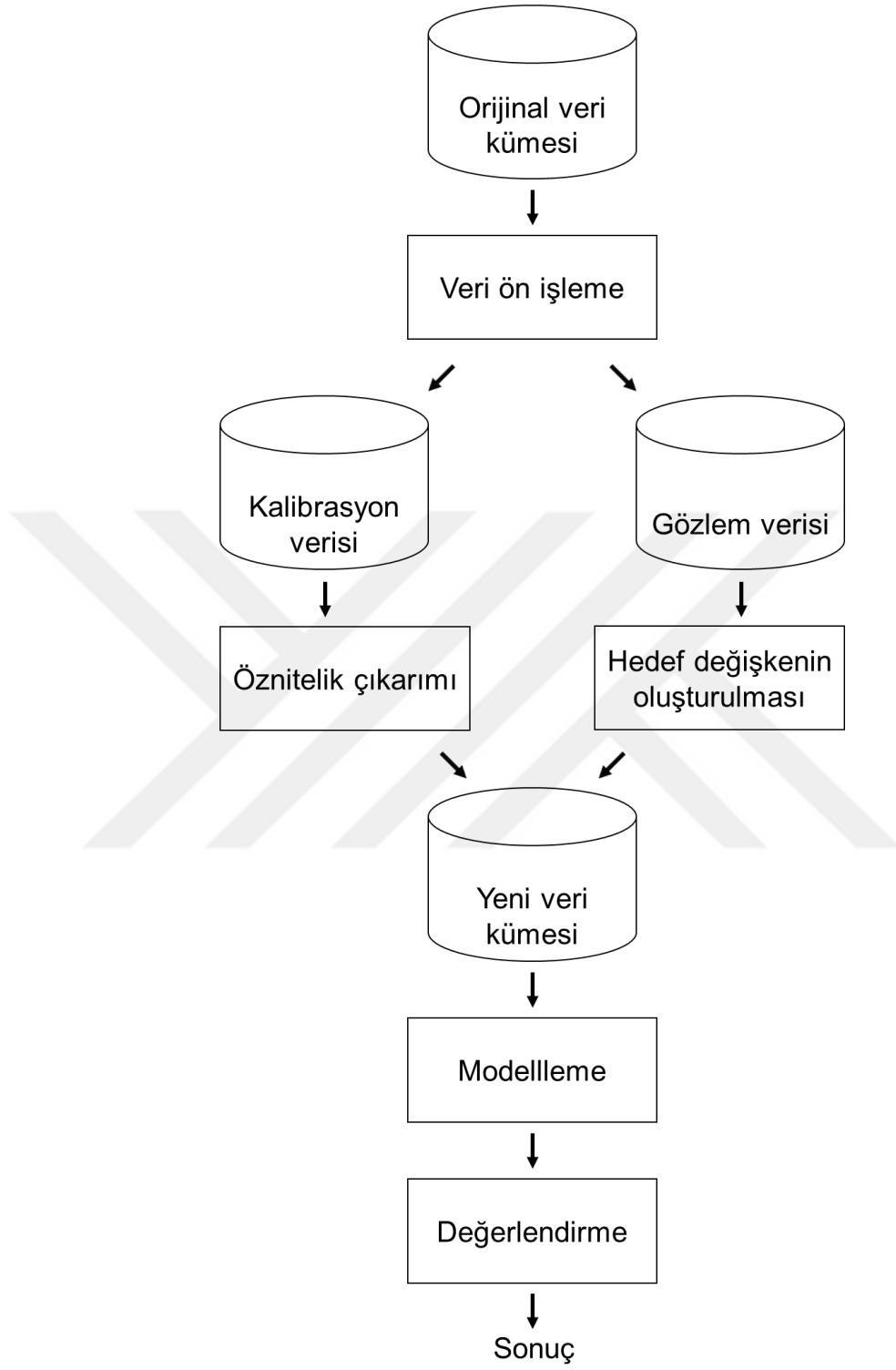
4. DENEYSEL ÇALIŞMA

4.1 Veri Kümesi

Bu çalışmada kullanılan veri kümesi, Birleşik Krallık merkezli çevrimiçi perakende hediyelik eşya mağazasına ait 01/12/2010 ve 09/12/2011 arasında gerçekleşen tüm alışveriş işlemlerini içerir. Veri kümesi, “Online Retail Data Set” başlığıyla UCI Machine Learning Repository internet sitesinde yayınlanmaktadır.

Tablo 4.1: Veri Kümesinde Bulunan Öznitelikler ve Açıklamaları.

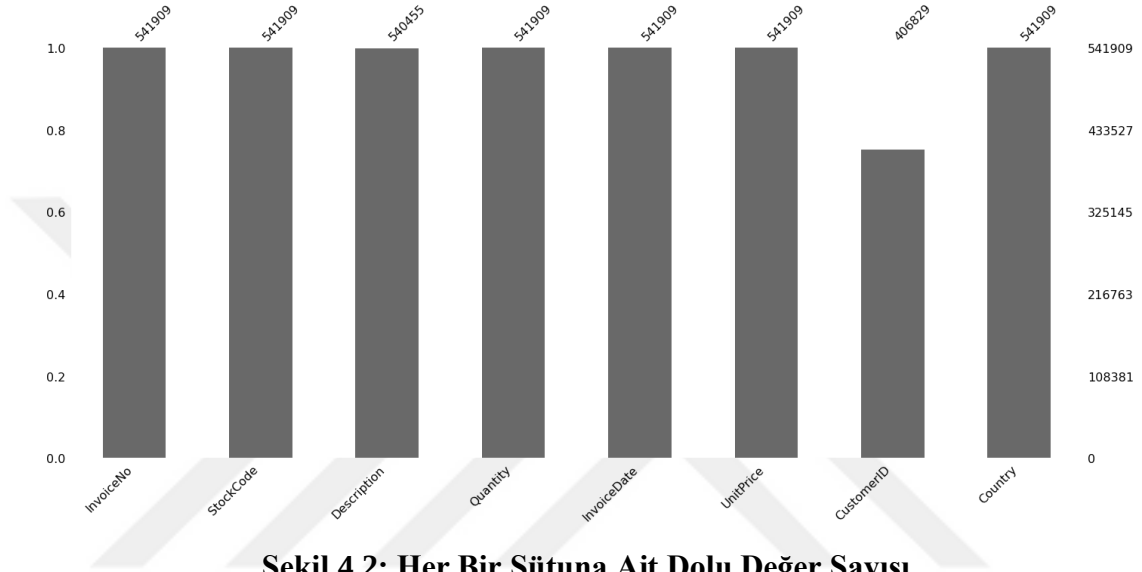
	Öznitelik	Açıklama
1	InvoiceNo	Fatura numarası. Bu kod 'c' harfiyle başlıyorsa işlemin iptal olduğunu gösterir.
2	StockCode	Ürün kodu. Her farklı ürüne benzersiz şekilde atanır.
3	Description	Ürün adı.
4	Quantity	İşlem başına her ürünün miktarı.
5	InvoiceDate	Fatura tarihi ve saati.
6	UnitPrice	Birim fiyat. Sterlin cinsinden birim başına ürün fiyatı.
7	CustomerID	Her müşteriye benzersiz olarak atanan müşteri numarası.
8	Country	Her bir müşterinin ikamet ettiği ülke adı.



Şekil 4.1: Çalışmanın Genel Çerçevesi.

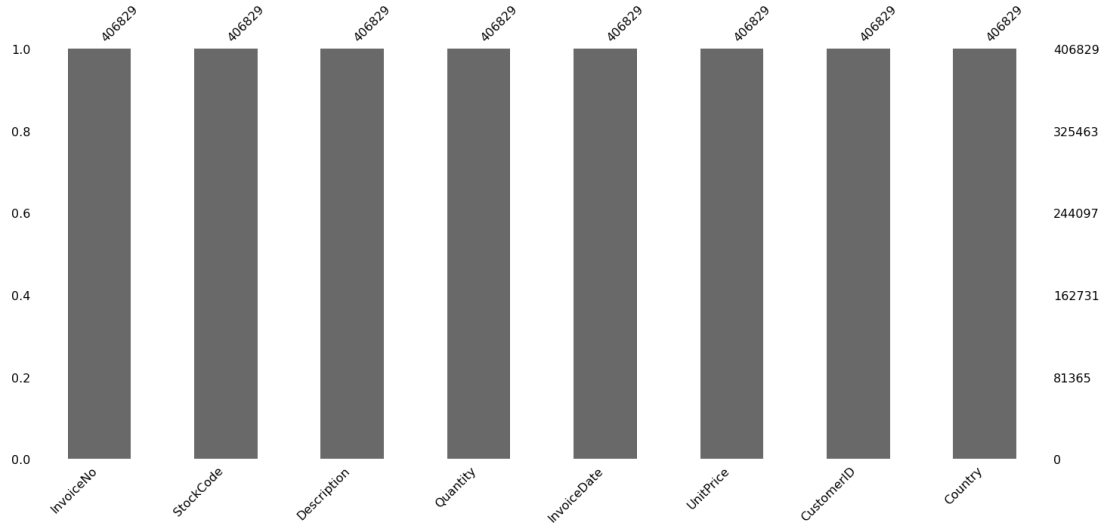
4.2 Açıklayıcı Veri Analizi

Veri kümesine ilk bakışta 8 sütun ve 541909 satırdan oluştuğu görülür ancak “Description” ve “CustomerID” sütunlarında eksik değerler vardır. Modelimizi müşteri sınıflandırma üzerine kuracağımızdan “CustomerID” sütuna denk gelen satırlar veri kümesinden çıkarılır.



Şekil 4.2: Her Bir Sütuna Ait Dolu Değer Sayısı.

Yeni durumda eksik değerlerin yer almadığı, 4372 farklı müşteriye ait 406829 adet işlemin yer aldığı veri kümesi elde edilir.



Şekil 4.3: Eksik Değerler Çıkarıldıktan Sonra Her Bir Sütuna Ait Dolu Değer Sayısı.

Veri kümesine miktar ve birim fiyatın çarpımından oluşan “TotalPrice” etiketiyle yeni bir değişken ekledik. Her bir müşterinin toplam kaç adet işlemi iptal ettiğini gösteren “NOC” etiketiyle bir değişken daha ekledik. Daha sonra iptal edilen toplam 8905 adet işlemi veri kümesinden çıkardık. Ayrıca veri kümesinde 3665 farklı ürünün satışının gerçekleştiği görülür. Son olarak her bir müşterinin kaç çeşit ürün aldığını ifade eden “NOI” etiketli değişken ekledik.

4.3 Hedef Değişkenin Oluşturulması ve Öznitelik Çıkarımı

Hedef değişken oluşturulurken veri kümesi kalibrasyon ve gözlem olmak üzere iki döneme ayrılır. Kalibrasyon döneminde veri analizi yapılır, gözlem döneminde müşterinin aktiflik durumuna göre kayıp kararı verilir. Buckinx ve Van den Poel’in (2005) çalışmasından hareketle periyotları 1:1 oranında ayarladık. Veri kümesi 01/12/2010 ve 09/12/2011 arası tarihleri kapsadığından 01/12/2010 – 07/06/2011 dönemini kalibrasyon; 07/06/2011 – 09/12/2011 dönemini gözlem olarak tayin ettik. Şu halde kalibrasyon dönemindeki 2763 müşteriden 918’i kayıp müşteri olarak etiketlenecektir (Kayıp müşteriler “1”, aktif müşteriler “0” etiketiyle sınıflandırıldı). Yapay öğrenme algoritmalarını kullanarak bir modeli eğitmeden önce, değişken seçimi tahmin modellerinin performansını etkileyen en önemli faktörlerden biridir.

Model kurarken eldeki ham veriler tahmin etmek istediğimiz davranışı isabetli bir şekilde tahmin etmek için yeterli olmayabilir. Bu yüzden literatürde de sıklıkla uygulandığı gibi mevcut ham veriyi kullanarak yaptığımız kestirimin gücünü artıracak başka değişkenlerin türetilmesi gerekmektedir. Bu işleme öznitelik çıkarımı adı verilmektedir (Şenyürek ve Alp, 2019). Bu adımda, literatürde kayıp tahmininde etkili sonuçlar verdiği belirtilen Recency, Frequency ve Monetary (RFM) değişkenlerine ilaveten yine daha önce denenmiş ve iyi sonuç vermiş müşterinin satın aldığı ürün çeşidi sayısı (NOI etiketiyle) ve iptal ettiği işlem sayısı (NOC etiketiyle) tahmin edici değişkenler olarak atanmıştır (Khodabandehlou ve Rahman, 2017).



Tablo 4.2: Veri Kümesinin Son Hali.

	Recency	Frequency	Monetary	NOI	NOC	Churn
CustomerID						
12346	325	1	77183.60	1	1	1
12347	246	3	1823.43	103	0	0
12348	248	3	1487.24	22	0	0
12350	310	1	334.40	17	0	1
12352	262	5	1561.81	59	10	0
...	...	...	...	...	...	...
18272	225	2	980.54	99	4	0
18273	257	1	51.00	1	0	0
18280	277	1	180.60	10	0	1
18283	200	5	535.05	263	0	0
18287	201	1	765.28	59	0	0

2763 rows x 6 columns

Tablo 4.3: Yeni Değişkenlerin Tanımlayıcı İstatistikleri.

	Recency	Frequency	Monetary	NOI	NOC	Churn
count	2763.000000	2763.000000	2763.000000	2763.000000	2763.000000	2763.000000
mean	66.016287	2.764748	1242.511832	73.743757	2.771625	0.298588
std	53.897180	4.074927	4181.054467	97.235489	8.812614	0.457722
min	1.000000	1.000000	2.900000	1.000000	0.000000	0.000000
25%	21.000000	1.000000	246.020000	19.000000	0.000000	0.000000
50%	53.000000	2.000000	480.300000	45.000000	0.000000	0.000000
75%	99.000000	3.000000	1092.755000	96.000000	2.000000	1.000000
max	188.000000	88.000000	116135.920000	1787.000000	226.000000	1.000000

4.4 Modelleme

Bu aşamada XGBoost, Histogram-based Gradient Boosting, AdaBoost, SVM ve Lojistik Regresyon algoritmalarıyla beş ayrı model kurulmuştur. Bağımsız değişkenler farklı aralıklarda olduğundan veriye normalizasyon uygulanmıştır.

İlaveten tahmin performansının tarafsız deęerlendirmesi iin veri kmesi 77:33 oranında eęitim ve test olarak alt kmelere blnd. Aşırı ęrenmenin nne gemek iin 5-kat apraz doęrulama yntemi veriye uygulandı.

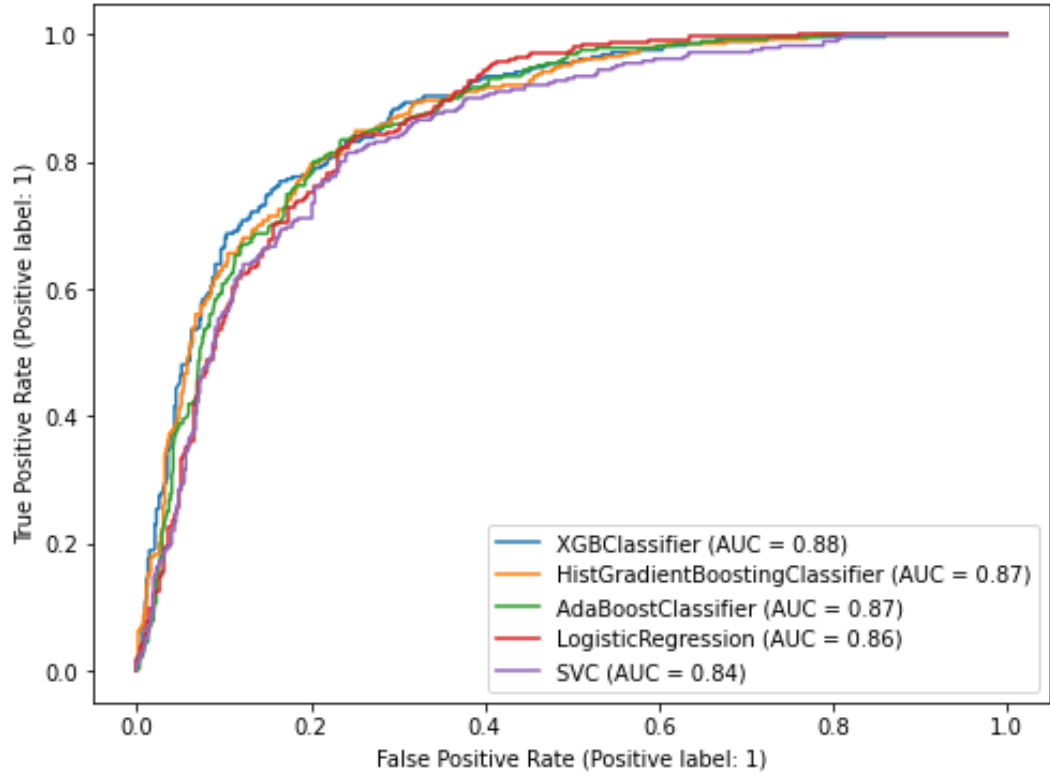
Hiperparametre ayarlama işlemi neticesinde varsayılan parametrelere ilaveten

- XGBClassifier iin (max_depth=3, learning_rate=0.1, min_child_weight=7),
- HistGradientBoostingClassifier iin (max_depth= 3),
- SVC iin (C= 10, gamma= 0.5, kernel= 'rbf') hiperparametre deęerleri optimum sonucu vermiştir.
- AdaBoostClassifier ve LogisticRegression iin varsayılan hiperparametre deęerleri en iyi sonucu vermiştir.

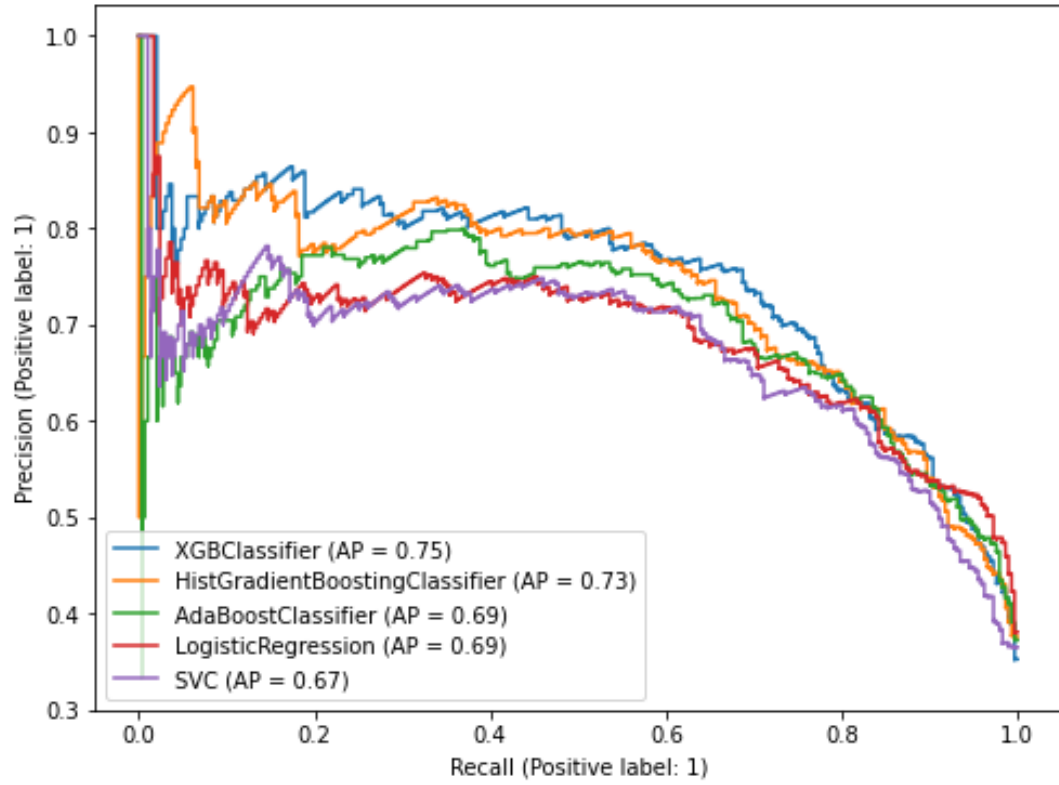
Algoritma başarıları birbirine ok yakın olmakla beraber XGBoost algoritması biraz ndedir.

Tablo 4.4: Model Değerlendirme.

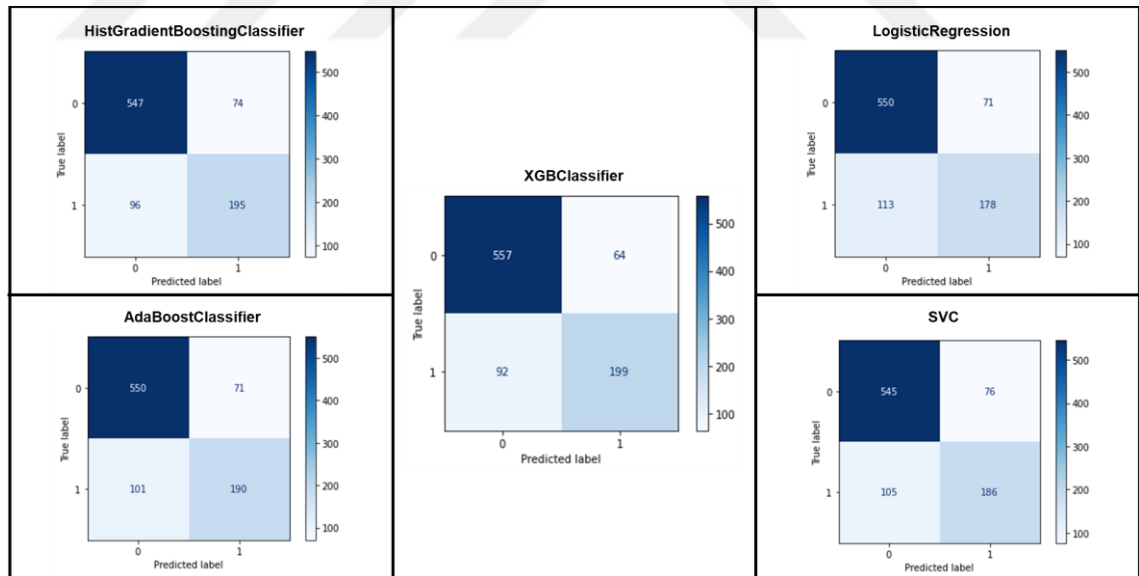
Model	Ölçüt	
XGBClassifier	Doğruluk	0,83
	AUC	0,88
	AP	0,75
HistGradientBoostingClassifier	Doğruluk	0,81
	AUC	0,87
	AP	0,73
AdaBoostClassifier	Doğruluk	0,81
	AUC	0,87
	AP	0,69
LogisticRegression	Doğruluk	0,80
	AUC	0,86
	AP	0,69
SupportVectorClassifier	Doğruluk	0,80
	AUC	0,84
	AP	0,67



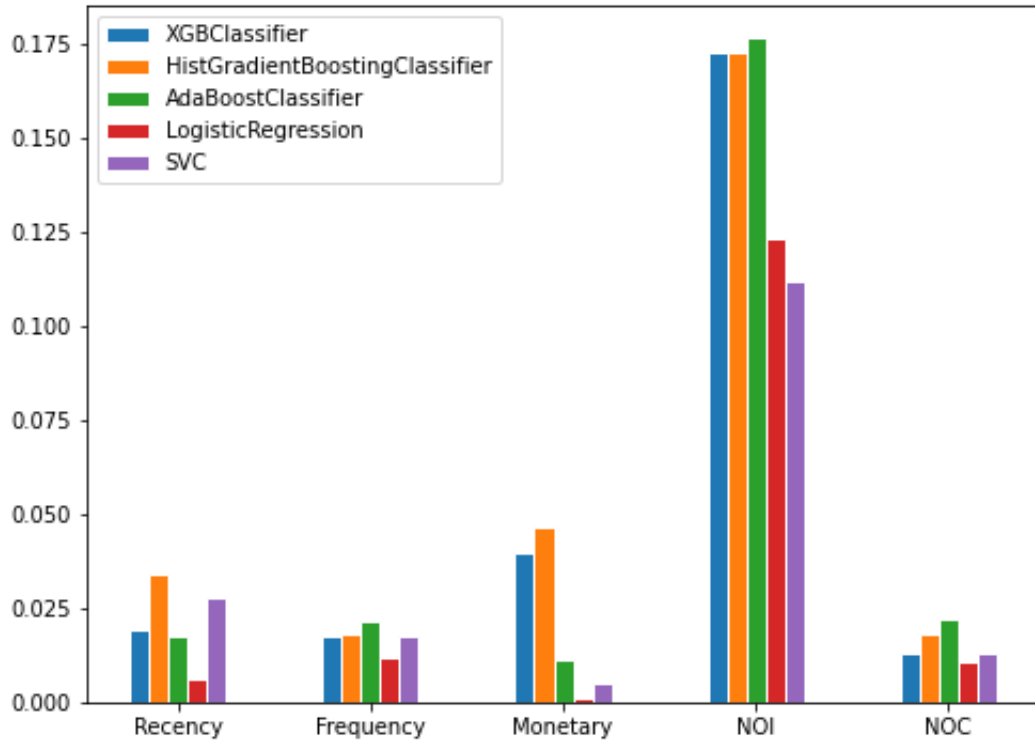
Şekil 4.4: Modellerin ROC Eğrisi.



Şekil 4.5: Modellerin Kesinlik-Duyarlılık Eğrisi.



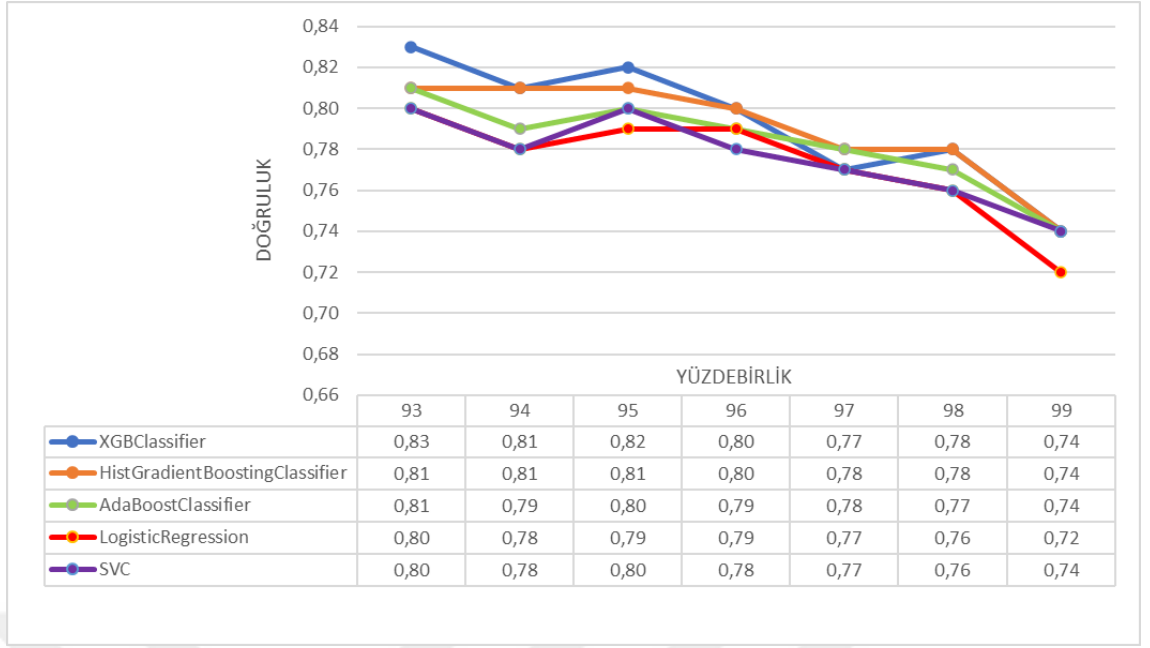
Şekil 4.6: Modellerin Karışıklık Matrisleri.



Şekil 4.7: Öznitelik Önemleri.

4.4.1 Bölme Noktasının Model Başarısına Etkisi

İlk modelde bölme noktası 1:1 oranında ayarlanmıştı. Tamaddoni ve diğ. (2015) optimum bölme noktası için müşterilerin ortalama alışveriş aralıklarının 99. Yüzdebirliğinin kullanılmasını önerir. Bu değeri modele uyguladığımızda Gradient Boosting, SVC, ve Lojistik Regresyon için doğruluk değerleri sırasıyla 0.75, 0.74, 0.72 olarak elde edildi. Görüldüğü gibi bölme noktasının 99. Yüzdebirlik olarak ayarlanması algoritma başarısını düşürdü. Bu sebeple bölme noktasını deneysel olarak yaklaşık 1:1 noktasına denk gelen 93. Yüzdebirlik ile 99. Yüzdebirlik arasındaki değerler için belirleyip model başarısını test ettik.



Şekil 4.8: Modellerin Yüzdebirlik – Doğruluk Grafiği.

5. SONUÇ

Bu çalışmada e-ticaret sektöründe faaliyet gösteren bir şirket için ayrılacak müşterileri tahmin etmek amaçlanmıştır. Bu amaçla ayrılma olasılığı en yüksek olan müşterileri tahmin edecek müşteri kaybı tahmini modeli geliştirilmiştir. Modelde sınıflandırma performansı yüksek olan GradientBoosting, DVM ve Lojistik Regresyon algoritmaları karşılaştırılmıştır.

Algoritmalar bir çevrimiçi perakende mağazasına ait yaklaşık bir yıllık veri üzerinde uygulanmıştır. Veri kümesindeki eksik veriler temizlendikten sonra ham veriler işlenerek ham verideki 8 adet öznitelikten müşteri kaybı tahmininde kullanılmak üzere 5 adet yeni öznitelik üretilmiştir. Öznitelik çıkarımında literatürde sıkça kullanılan RFM analizi tekniğinden faydalanılmıştır. Öznitelikler arasında NOI (müşterinin aldığı ürün çeşidi sayısı) tüm modellerde en fazla katkıyı sağlamıştır.

Veri kümesinde hedef değişken olmadığından bir bölme noktası tayin edip veri kümesini “kalibrasyon” ve “gözlem” olarak ikiye ayrılmıştır. Müşteri kalibrasyon döneminde aktif gözlem döneminde pasif ise “kayıp müşteri” olarak sınıflandırılır. Bölme noktasının model başarısı üzerinde doğrudan etkisi vardır. Yapılan deneysel çalışma neticesinde optimum bölme noktasının 93. Yüzdebirliğe denk geldiği görülmüştür.

Literatürdeki çalışmalara paralel olarak en başarılı sonuca Boosting algoritmasının ulaştığı gözlemlenmiştir. Bu çalışma tek veri kümesi kullanılarak gerçekleştirilmiştir. Birden fazla veri kümesi üzerinde çalışılarak modeller geliştirilebilir.

KAYNAKÇA

- Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal Big Data*, 6(28) <https://doi.org/10.1186/s40537-019-0191-6>
- Bagul, N., Berad, P., Khachane, C., & Surana, P., (2021). Retail Customer Churn Analysis using RFM Model and K-Means Clustering. *International Journal Of Engineering Research & Technology (IJERT)*, 10(3), 349-354.
- Bardük, B. (2020, Ekim). Modelling Time Statistics for Customer Churn Prediction. 2020 28th Signal Processing and Communications Applications Conference (pp. 1-4). IEEE.
- Bilişim Sanayicileri Derneği (TÜBİSAD). (2021).Ekonominin dönüştürücü gücü: e-ticaret etki analizi. Deloitte Digital.
- Buckinx, W. & Van den Poel, D. (2005). Customer Base Analysis: Partial Defection of Behaviourally Loyal Clients in a Non-contractual FMCG Retail Setting. *European Journal of Operational Research*, 164(1), 252-268.
- Ehrenberg, A. S. C. (1959). The Pattern of Consumer Purchases. *Journal of the Royal Statistical Society*, 8 (1), 26-41.
- Fader, P. S., Hardie, B. & Lee, K. L. (2005). “Counting Your Customers” the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, 24 (2), 275-284.
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.
- Guo, F., & Qin, H. (2015, Kasım). The Analysis of Customer Churns in e-Commerce Based on Decision Tree. 2015 International Conference on Computer Science and Applications (CSA) (pp. 199-203). IEEE. doi: 10.1109/CSA.2015.74.
- Herrero-Lopez, S. Multiclass support vector machine. In *GPU Computing Gems Emerald Edition*; Elsevier:Burlington, MA, USA, 2011; pp. 293–311.
- Jaeyalakshmi, M., Gnanavel, S., Guhapriya, K. S., Phriyaa, S. H., & Sree, K. (2020) Prediction of Customer Churn on e-Retailing. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(6), 5541-5545.
- Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P., Wedel, M., & Wilcox, R. (2005). Choice Models and Customer Relationship Management. *Marketing Letters*, 16 (3), 279-291.

- Khodabandehlou, S., & Rahman, M. Z. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, 19(2), 65-93. <https://doi.org/10.1108/JSIT-10-2016-0061>
- Kumar, A. S., & Chandrakala, D. (2016). A Survey on Customer Churn Prediction using Machine Learning Techniques. *International Journal of Computer Applications*, 154(10), 13-16.
- Schmittlein, D. C., Morrison, D. G. & Colombo, R. (1987). Counting Your Customers: Who Are They and What Will They Do Next?. *Management Science*, 33 (1), 1-24.
- Shetty, P., Varsha, C. M., Vadone, V., Sarode, S. & Kumar D. (2019). Customers Churn Prediction with Rfm Model and Building a Recommendation System using Semi-Supervised Learning in Retail Sector. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(1), 3353-3358.
- Şenyürek, A., & Alp, S., (2019). Telekomünikasyon Sektöründe Aboneliklerini İptal Edecek Müşterilerin Yapay Öğrenme ile Önceden tahmin Edilmesi. *Akdeniz 1. Uluslararası Multidisipliner Çalışmalar Kongresi* (pp.1-18). UBAK.
- Tanveer, A. (2019). Churn Prediction Using Customers' Implicit Behavioral Patterns And Deep Learning [Yayınlanmamış yüksek lisans tezi]. Sabancı University.
- Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. *Journal of Service Research*, 19(2), 123–141. <https://doi.org/10.1177/1094670515616376>
- Wu, H., Zhang, W., & Zhang, Y. (2010). An Empirical Study of Customer Churn in E-Commerce Based on Data Mining. 2010 International Conference on Management and Service Science (pp. 1-4). IEEE. doi: 10.1109/ICMSS.2010.5576627.
- Wu, X., & Meng, S., (2016, Ağustos). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. 2016 13th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-5). IEEE. doi: 10.1109/ICSSSM.2016.7538581.
- Xia, G., & He, Q. (2018). The Research of Online Shopping Customer Churn Prediction Based on Integrated Learning. *Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)*, 149, 259-2