T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜNDE ABONELİKLERİNİ İPTAL EDECEK MÜŞTERİLERİN YAPAY ÖĞRENME YÖNTEMLERİ İLE TAHMİN EDİLMESİ

Ayşe ŞENYÜREK

YÜKSEK LİSANS TEZİ

Endüstri Mühendisliği Anabilim Dalı

Sistem Mühendisliği Programı

Danışman

Doç. Dr. Selçuk ALP

Haziran, 2019

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOMÜNİKASYON SEKTÖRÜNDE ABONELİKLERİNİ İPTAL EDECEK MÜŞTERİLERİN YAPAY ÖĞRENME YÖNTEMLERİ İLE TAHMİN EDİLMESİ

Ayşe ŞENYÜREK tarafından hazırlanan tez çalışması çalışması 24.07.2019 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı, Sistem Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Selçuk ALP

Yıldız Teknik Üniversitesi

Danışman

Jüri Üyeleri

Doç. Dr. Selçuk ALP, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Taylan ENGİN, Üye
Bandırma Onyedi Eylül Üniversitesi _____

Prof. Dr. Nihan Çetin Demirel, Üye
Yıldız Teknik Üniversitesi _____

Danışmanım Doç. Dr. Selçuk ALP sorumluluğunda tarafımca hazırlanan Telekomünikasyon Sektöründe Aboneliklerini İptal Edecek Müşterilerin Yapay Öğrenme ile Tahmin Edilmesi 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.

Ayşe ŞENYÜREK

İmza

Aileme

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Doç. Dr. Selçuk Alp'e ve bu süreçte beni cesaretlendirip destek olan sevgili aileme teşekkürü bir borç bilirim.

Ayşe ŞENYÜREK

İÇİNDEKİLER

SİMGE LİSTESİ

P	Bir olayın gerçekleşme olasılığı
β0	Bağımsız değişkenler sıfır değerini aldığında bağımlı değişkenin değeri (sabiti)
βi	Bağımsız değişkenin regresyon katsayısı
Xi	Bağımsız değişkenler
m	Rastgele Orman yöntemi için tekil ağaç sayısı
k	Rastgele oramn yönteminde her ağaç için rassal seçilen açıklayıcı değişken sayısı
N	Gözlem sayısı
C	Lojistik regresyon yönteminde model karmaşıklığı
λ	YSA için ağırlık sönümleme parametresi

KISALTMA LİSTESİ

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ARPU	Average Revenue per User
AUC	Area Under Curve
BT	Bilişim Teknolojileri
BTK	Bilişim Teknolojileri Kurumu
CART	Classification and Regression Tree
CDR	Call Detail Record
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relation Management
DN	Doğru Negatif
DP	Doğru Pozitif
DVM	Destek Vektör Makinaları
GBM	Gradient Boosting Machine
GLM	Generalized Linear Models
GSM	Global System for Mobile Communications
GSYH	Gayri Safi Yurtiçi Hasıla
IoT	Internet of Things
LEM2	Learning from Examples Module, version 2
M2M	Machine to Machine
MAPE	Mean Absolute Percentage of Error
MNT	Mobil Numara Taşınabilirlik
MoU	Minutes of Usage

OTT	Over the Top
ROC	Receive Operating Curves
SMS	Short Message Service
SVM	Support Vector Machine
YN	Yanlış Negatif
YP	Yanlış Pozitif
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

Telekomünikasyon Sektöründe Aboneliklerini İptal Edecek Müşterilerin Yapay Öğrenme Yöntemleri İle Tahmin Edilmesi

Ayşe ŞENYÜREK

Endüstri Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Selçuk ALP

Günümüz iş dünyasının rekabet ortamında firmaların ayakta kalabilmesi için aboneliklerini sonlandıracak müşterilerin tahmini (churn analizi) oldukça önemli bir hale gelmiştir. Müşteri kaybının tahmini için firmalar çalışmalar yapmaktadır. Yapay öğrenme uygulamaları da bu konuda etkin bir şekilde kullanılmaktadır. Pazarın doymuş olması ve firmalar arası rekabetin büyüklüğü nedeniyle ayrılacak müşterilerin analiz ve tahmini telekomünikasyon sektöründe yoğun bir şekilde yapılmaktadır. Bu sayede firmalar değerli müşterileri belirleme, bu müşteriler için özel kampanyalar yapma ve müşterilerin deneyimini iyileştirme fırsatı yakalamaya çalışmaktadırlar. Literatürde, firmalar için yeni müşteri elde etmek yerine mevcut müşterileri elde tutmanın önemli ölçüde daha erişilebilir ve daha az maliyetli olduğu birçok çalışmada belirtilmiştir. Müşterilerin abonelik süreleri ne kadar uzun olursa firmalar için getirisinin o kadar yüksek olacağı düşünülmektedir. Telekomünikasyon sektöründeki firmalar için uzun dönemli sözleşme, kampanya veya tarifeler kısa dönem müşteri ilişkilerine göre daha tercih edilir durumdadır. Sadık müşterilerin daha değerli olmasından ötürü sadakatin oluşturulabilmesi için müşteri kaybı tahmini uygulanmaktadır.

Yapay öğrenme, mevcut verileri ve eski deneyimleri kullanarak tahminlerde bulunmayı sağlayan algoritma uygulamalarıdır. Bu uygulamalarda amaç algoritmanın belirli bir ölçüte göre başarıyı artıracak şekilde programlanmasıdır. Yapay öğrenmenin başarılı uygulamaları günümüzde her alanda karşımıza çıkmaktadır. Konuşma ve yazı tanıma, finans sektöründe kredi risk hesapları, şoförsüz araçlar, sahtekarlık (fraud) tespiti, borsa, biyoinformatik, robotik, ulaştırma, tıbbi tanı gibi bir çok alanda kendini göstermektedir. Bu bağlamda son yıllarda yapay öğrenme temelli yöntemler müşteri kaybı tahmini için oldukça popüler olmuştur.

Bu çalışmanın amacı telekomünikasyon sektöründe aboneliklerini iptal edebilecek abonelerin tahmininin yapıldığı modeli belirlemektir. Bu doğrultuda veri seçilmesi, ön hazırlığının yapılması, kullanılacak yapay öğrenme yöntemi, performans kriterleri, ölçümleme işlemlerinin belirlenmesi amaçlanmıştır. Bunun için lojistik regresyon, yapay sinir ağları yöntemi, random forest yöntemi (rastgele orman yöntemi) ve boosting yöntemlerine (sırayla eğiterek iteleme) göre potansiyel iptal abonelerinin tahmini yapılmıştır. Çalışma sonuçları incelendiğinde müşteri kaybı tahmininde boosting yönteminin diğer yöntemlerden daha doğru ve başarılı sonuçlar verdiği görülmüştür. Müşteri kaybına neden olan en önemli etkenlerin başında kontrat bitimine kalan süre, müşterin ne kadar süredir operatörün abonesi olduğu, yakın çevresinin hangi operatörü tercih ettiği ve şebeke kalitesi olduğu görülmüştür.

Anahtar kelimeler: Müşteri kayıp analizi, Telekomünikasyon, Müşteri ilişkileri yönetimi, Yapay öğrenme

ABSTRACT

Churn Prediction in Telecommunication Sector with Machine Learning Methods

Ayşe ŞENYÜREK

Department of Industrial Engineering

Master of Science Thesis

Advisor: Assoc. Prof. Dr. Selçuk ALP

In the competitive environment of today's business world, in order to survive the churn analysis has become quite important. Companies are making efforts to estimate customer churn. Machine learning applications are also used effectively in this regard. Due to the fact that the market is saturated and the size of the competition among the firms, the analysis and prediction of the customers will be made intensively in the telecommunication sector. In this way, companies are trying to identify valuable customers, make special campaigns for these customers and improve the experience of customers. In the literature, many studies have reported that it is significantly more accessible and less costly to retain existing customers than to obtain new customers for firms. The longer customers' subscription periods, the higher the earnings for the companies. Long-term contracts, campaigns or tariffs for the companies in the telecommunications sector are more preferable than short-term customer relations. Since loyal customers are more valuable, churn prediction is applied to create loyalty.

Machine learning is an algorithm application that provides estimations using existing data and old experiences. In these applications, the goal is to program the

algorithm to increase the success according to a certain criterion. The successful applications of machine learning are seen in every field today. Speaking and writing recognition, credit risk calculations in the financial sector, non-driver vehicles, fraud detection, stock market, bioinformatics, robotics, transportation, medical diagnosis is manifested in many areas such as. In this context, machine learning-based methods have become very popular in recent years for estimating customer churn.

The aim of this study is to construct a model in which the subscribers are able to cancel their subscriptions in the telecommunication sector. In this context, it was aimed to select data, to prepare the preliminary preparation, to use machine learning method, performance criteria and measurement processes. According to logistic regression, artificial neural network, random forest and boosting method, potential churn subscribers were estimated. When the results of the study are examined, it is seen that the boosting method gives more accurate and successful results than the other methods. The most important factors causing customer churn was the period remaining until the end of the contract, tenure, which operator preferred the close relatives and the quality of the network.

Keywords: Churn analysis, Telecommunication, CRM, Machine learning

1
Giriş

1.1 Literatür Özeti

Literatür incelendiğinde yapay öğrenmenin, popülerliği son yıllarda artan bir araştırma alanı olduğu görülmektedir. Ancak müşteri ilişkileri yönetiminin önemli bir konusu olan müşteri kaybı analizi (churn analizi) hakkında yapay öğrenme odaklı sınırlı sayıda araştırma bulunmaktadır. Bu araştırmalarda birçok farklı öğrenme yöntemi sınanmıştır. Bunların başında regresyon modelleri, karar ağaçları, rastgele orman (random forest), destek vektör makinesi (support vector machine), bayes ağları (bayesian network), k-en yakın komşu (k-nearest neighbors) gibi yöntemler var. Bir çok araştırmada hibrit yöntemler de kullanılmıştır. Ancak en popüler teknik, yapay sinir ağları, komite modeller ve hibrit çalışmalar olduğu görülmektedir.

Telekomünikasyon sektöründe yapılmış son çalışmalardan biri olan Coussement ve arkadaşlarının çalışmasında (2017) churn analizi öncesinde yapılan bir süreç olan veri setinde yapılan işlemlere dair incelemeler yapmıştır. Müşteri churn tahmininin, analistler için çeşitli karar noktaları içerdiği, bu karmaşık yapının CRISP-DM (Cross-Industry Standard Process for Data Mining) tarafından işi anlama, veriyi anlama, veri ön işleme, modelleme, değerlendirme ve geliştirme olmak üzere 6 aşamaya ayrıldığı belirtilmiştir. Çalışmalarında bu aşamalardan veri ön işlemeye odaklanmışlardır. Veri setlerindeki ayrık ve devamlı değişkenlerin formatını uygun hale getirerek ve müşterilere dair doğru öznitelikler seçerek analiz performansının artığını vurgulamışlardır. Çalışmada veri ön işleme işlemleri, veri indirgeme ve veri hazırlama aşaması olarak iki kısıma ayrılmıştır. Veri indirgeme yöntemlerinin amacı analizi hangi özelliklerin etkilediğini belirleyip etkisi olmayan özelliklerin modelden çıkarılmasını sağlamak böylelikle veri boyutunu azaltmaktır. Veri hazırlama yöntemleri ise değişkenlerin uygun formatlara dönüştürülmesi için gereklidir. Veri ön işlememin değer dönüşümü ve sunumu olarak iki adımı vardır. Çalışmada bu iki

adım üzerine analizler yapılmıştır. Kullanılan veri setinde 30,104 müşteri vardır. Değişkenlerin 156'sı kategorik 800'ü sürekli değişkendir. Çalışmada eğitim, seçim ve test olmak üzere sırayla veri setinin %50, %20 ve %30'luk kısımları alınarak kullanılmıştır. Veri ön işleme işlemlerinin etkisiyle birlikte Lojistik Regresyon modeli için churn tahmini başarısı ölçülmüştür. Veri ön işleme işlemlerinin tahmin başarısını %34'e kadar çıkarabildiği görülmüştür. Ayrıca çalışma sonuçlarından bir diğeri de lojistik regresyon algoritmasının yapay sinir ağları ve destek vektör makinası gibi yöntemlere göre daha hızlı olduğu görülmüştür [1].

Amin ve arkadaşlarının (2017) çalışması, referans veri tabanı üzerinde uygulanan Kapsamlı Algoritma (Exhaustive Algorithm), Genetik Algoritma, Kaplama Algoritması (Covering Algorithm) ve LEM2 Algoritması olan kural üretme yöntemlerini kullanarak müşteri davranışını tahmin etmeyi amaçlamaktadır. Yöntemlerin ölçümü için Kaba Set Sınıflandırması (Rough Set Classification) uygulanmıştır. Bu işlem sonucunda, Genetik Algoritma en iyi müşteri kaybı olasılığı oranını vermiştir [2].

Kaynar ve arkadaşlarının (2017) çalışmasında, destek vektör makineleri, Naif Bayes ve çok katmanlı yapay sinir ağları kullanılarak 3 model elde edilmiştir. Çalışmalarında 4667 adet müşteriden alınan bilgiler kullanılmıştır. 21 tane öznitelik çıkarılmıştır. Veri setinde hem kaybedilmiş müşteriler hem sadık müşteriler vardır. Veri setinin rastgele seçilmiş %75'i öğrenme verisi, kalan %25'I test verisi olarak kullanılmıştır. Çalışmada uygulanmış 3 yöntem içerisinden modelleme tahmin başarısı en yüksek olan yöntem %92.35 ile yapay sinir ağlarıdır. İkinci yöntem %87.15 başarı ile Naif Bayes yöntemidir. Üçüncü ise %77.89 başarı ile DVM yöntemidir. Ayrıca Naif Bayes yöntemi hassasiyet açısından en iyi sonucu vermiştir. Yapay Sinir Ağları ve Naif Bayes yöntemleri çalışma için beklenen başarılı sonuçları verirken, Destek Vektör Makineleri beklenenden düşük performans göstermiştir. Destek vektör makinası yönteminin daha düşük başarı göstermesinin nedenleri olarak veri setindeki bazı öznitelikler ve örnek sayısının yetersizliği ön görülmüştür [3].

Vafeiadis ve arkadaşları (2015) YSA, destek vektör makineleri, karar ağaçları, Naif Bayes ve lojistik regresyon gibi sık kullanılan churn tahmin teknikleri ve telekomünikasyon endüstrisindeki performans sınıflandırıcılarını kullanarak bu algoritmaların performanslarının değerlendirilmesi üzerine çalışmışlardır. Karşılaştırmalı sonuçlar, telekomünikasyon endüstrisindeki kayıp tahmini için boosted SVM olarak adlandırılan yöntemin en iyi sonucu verdiğini göstermiştir [4].

Abbasimehr ve arkadaşlarının çalışması (2014), komite öğrenmesi uygulamalarının bireysel temel öğreniciler için üç performans göstergesi adına, yani AUC, hassasiyet ve özgüllük açısından önemli bir gelişme getirdiğini göstermektedir. Boosting, diğer tüm yöntemler arasında en iyi sonuçları vermiştir. Bu sonuçlar, komite öğrenmesi yöntemlerinin müşteri kayıp tahmini işleri için en iyi aday olabileceğini göstermektedir [5].

Kim ve arkadaşlarının çalışmasında (2014) müşteri kişisel verilerini ve CDR verisini içeren telekomünikasyon firmasından alınan veri seti kullanılmıştır. Kullandıkları yöntem lojistik regresyon ve çok katmanlı algılayıcılardır. Veri seti %9.7'si churn etmiş müşterilerden oluşan 89.412 adet örnek müşteridir. Ağ analizi kullanan önceki çalışmaların aksine, ağ değişkenini bir yayılım süreci olan SPA'dan oluşturulmuş ve modeli eğitmek için geleneksel kişisel değişkenlerle birleştirilmiştir. Bu şekilde etkin bir yaklaşım geliştirmişlerdir [6].

Keramati ve arkadaşları (2014) çalışmalarında performanslarını karşılaştırmak için karar ağaçları, yapay sinir ağları, K-en yakın komşular ve destek vektör makinesi gibi veri madenciliği sınıflandırma tekniklerini kullanmışlardır. İranlı bir mobil operator şirketinin verilerini kullanarak, bu teknikleri birbirleriyle kıyaslayarak bazı önde gelen farklı veri madenciliği yazılımları arasında bir paralellik göstermişlerdir. Tekniklerin davranışlarını incelemek ve özelliklerini bilmek için bazı değerlendirme ölçütlerinin değerinde önemli iyileştirmeler yapan bir hibrit yöntem önermektedirler. Önerilen yöntem sonuçları, Geri Çağırma ve Duyarlık için %95'in üzerinde duyarlığın elde edilebildiğini göstermiştir. Bunun dışında, veri setindeki etkili özniteliklerin çıkarılması için yeni bir yöntem tanıtılmış ve deneyimlenmiştir. Ek olarak, en etkili öznitelik setini çıkarmak için yeni bir boyutsallık azaltma yöntemi tanıtmışlardır. Kullanım sıklığı, toplam şikayet sayısı

ve kullanım sürelerinin etkili öznitelikler olduğu gösterilmiştir. Ayrıca, aynı veri seti üzerinde yapılan önceki çalışmanın aksine, SMS sıklığının, ücret tutarının ve hizmet türünün en az etkili öznitelikler olduğununu göstermişlerdir [7].

Huang ve arkadaşlarının çalışmasında (2012) veri setinden yeni öznitelikler oluşturarak ve öznitelikleri bazı özelliklerine göre gruplayarak model başarısını artırmaya çalışmışlardır. Yeni öznitelik oluşturma, veri setinin büyüklüğüne dayanır Öznitelik grupları, doğruluk ve performans açısından en verimli modeli oluşturmak için üç farklı şekilde birleştirilmiştir. Sınıflandırma işleminde lojistik regresyon, doğrusal sınıflandırma, naif bayes, karar ağacı, çok katmanlı algılayıcılar, destek vektör makineleri ve deneysel veri işleme algoritmaları kullanılmıştır. Sonuçları değerlendirmek için ROC (Receive Operating Curves) ölçütü kullanılmıştır. Veri setindeki tüm öznitelikleri kullanan destek vektör makineleri ile yapılan sınıflandırma işleminin en iyi sonuçları verdiği görülmüştür [8].

Verbeke ve arkadaşlarının çalışmasında (2012) müşteri kaybı tahmini problemleri için komite modellerinin veri madenciliğinde yaygın bir kullanımı olduğunu belirtilir. KDD 2009 veri seti de dahil olmak üzere telekomünikasyon servis sağlayıcılarından derlenen bir dizi örnek derlemişlerdir. Müşteri kaybı tahmin analizi için veri setinde hem tek hem de komite algoritmalarını uygulamışlardır. En iyi performans gösteren sınıflayıcıyı seçmek için kar temelli bir değerlendirme fonksiyonu önermişlerdir. Az sayıda değişken kullanmışlar ve sonuçların klasik değerlendirme yöntemlerinden daha iyi olduğunu bildirmişlerdir [9].

Kisioğlu ve Topçu'nun (2011) çalışma sonuçlarına göre, telekom endüstrisinde etkin bir müşteri kaybı yönetimi için, ortalama MoU (Minutes of Usage), ortalama fatura ödemesi, ara bağlantı çağrılarının sayısı ve tarife türü müşteri kayıp oranını analiz etmek için en önemli faktörlerdir [10].

1.2 Tezin Amacı

Bu çalışmanın amacı telekomünikasyon sektöründe aboneliklerini iptal edebilecek abonelerin tahmininin yapıldığı modeli kurgulamaktır. Firmalar açısından bakıldığında yeni müşteri elde etmektense mevcut müşterilerin ayrılmasını engellemek daha karlı olmaktadır. Mevcut müşterileri koruma hem zaman hem de

maliyet açışından firmalar için daha doğru bir strateji olacaktır. Özellikle pazarın doygunluğa ulaşmış olduğu ve rekabetin yoğun olduğu sektörlerde churn analizleri yoğun bir şekilde yapılmaktadır.

Pazar rekabeti göz önüne alındığında iptal oranının aylık %2.2 seviyelerinde olması bu alanda yapılacak çalışmaların ne kadar önemli olduğunu açık bir şekilde göstermektedir [11].

Aşağıda verilmiş olan rakamsal verileri de göz önünde bulundurduğumuzda iptal analizinin telekomünikasyon sektöründe önemi daha net şekilde ortaya çıkmaktadır;

- Aylık oranı %2.2 olan müşteri abonelik iptali yıllık olarak %25 müşteri abonelik iptaline karşılık gelmektedir [11].

- Müşteriyi elde tutma maliyeti yeni müşteri kazanma maliyetine kıyasla 5 kat daha düşüktür[12]. Böylece müşteriyi elde tutmak için yapılacak küçük bir iyileştirme, karlılıkta önemli bir artışa ön ayak olacaktır.

- Türkiye'de Bilgi Teknoloji Kurumunun (BTK) 2018 son çeyrek verilerine göre 14 Şubat 2018 tarihi itibariyle toplam 123.194.542 mobil numara taşıma işlemi gerçekleştirilmiştir.

1.3 Hipotez

Bu çalışmada telekomünikasyon sektöründe aboneliklerini sonlandıracak müşterilerin yapay öğrenme yöntemleriyle tespit edilebileceği varsayılmaktadır. Bu amaçla farklı yapay öğrenme yöntemleri uygulanmış ve elde edilen sonuçlar birbirleriyle kıyaslanmıştır. İçlerinden yüksek doğrulukla tahmin yapan yöntemin boosting olduğu görülmüştür. Böylece, çalışma doğru abonelere erişmek ve yeni pazarlama stratejileri geliştirmek için şirketlere olanaklar oluşturacaktır.

Bu çalışma şu şekilde düzenlenmiştir. Bölüm 1, çalışma konusu ile ilgili literatür taramasını ve tezin amacını gösterir. Bölüm 2, telekomünikasyon sektörüne genel bir bakış sunmaktadır. Bölüm 3, müşteri ilişkileri yönetiminin, müşteri kaybının ve tahmininin anlatıldığı bölümdür. Bölüm 4, yapay öğrenme yöntemlerini

açıklamaktadır. Bölüm 5, uygulamayı içermektedir. Bölüm 6'da ise çalışmada gerçekleştirilen uygulamanın sonuçlarına ve önerilere yer verilmiştir.

2
Telekomünikasyon Sektörüne Genel Bakış

2.1 Telekomünikasyon Pazarının Kısa Tarihçesi

Telekomünikasyon sektörü, telefon ve/veya internet üzerinden kullanıcılara küresel bir şekilde iletişim hizmeti veren şirketleri içermektedir. Telekomünikasyon hizmetleri, şirketlerin BT altyapısını, verilerin transferini, çağrı ve kısa mesaj servislerini sağlar.

Telgraf zamanla mobil servislere evrilmiş ve geçmişte günler süren iletişim bu sayede anında gerçekleştirilebilir olmuştur. Teknolojinin geldiği noktada şu an çok büyük miktarda veriler eşzamanlı olarak iletebilir hale gelmiştir.

Dünya telekomünikasyon sektörü 1980'lerde büyük bir evrim aşamasına gelmiştir. Bu büyük ölçekli değişimin ilk fazı, 19. yüzyıl sonlarında ve 20. yüzyılın başlarında ABD hükümeti tarafından devlete bağlı olarak "Post, Telegraph & Telecommunication" kurulması ile başlamıştır. İletişim sektörünün devlet tekeliyle sınırlandırılmaması fikri hayata geçirilene kadar uzun bir süre bu şekilde devam etmiştir. İletişim sektöründe devletin yanı sıra özel sektöründe faaliyet gösterebilmesi için, 1984 yılında "Amerikan Telefon ve Telgraf" (AT&T) kuruluşu ikiye bölünmüştür. AT&T'nin bölünmesiyle birlikte ile telekomünikasyon sektörü için yeni bir dönem başlamıştır ve bu durum ABD'deki telekomünikasyon şirketleri arasında tam rekabet sağlamıştır [13].

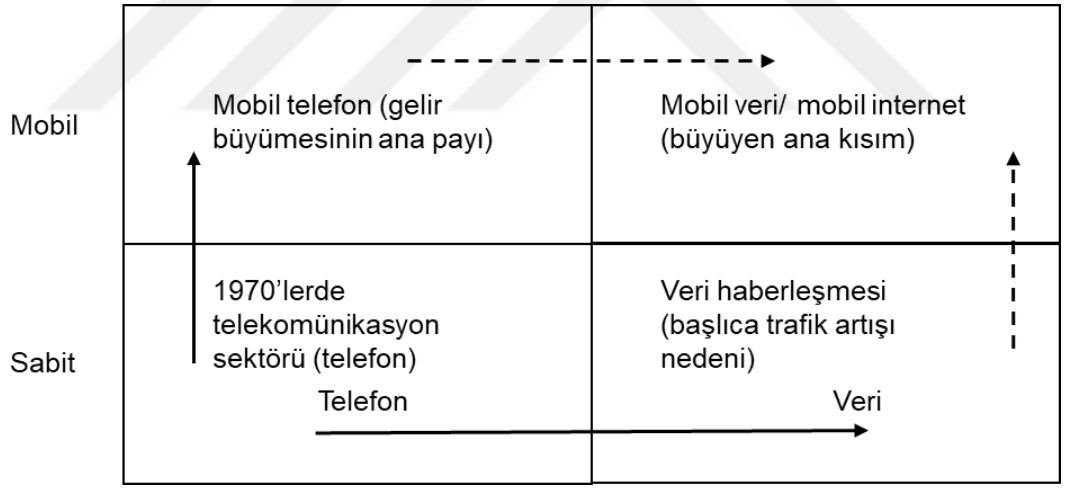
Avrupa'ya bakılacak olursa İngiltere'de telekomünikasyon piyasası 1966'ya kadar tekel konumunda olan "British Telecom"dan ibaret iken sonrasında kurum şirketleşmiş ve rekabet piyasası oluşmuştur. Telekomünikasyon sektöründe serbestleşmenin başlaması bir çok önemli sonucu beraberinde getirmiştir. Piyasaya giren telekomünikasyon firmaları potansiyel müşterileri kendilerine çekebilmek için en gelişmiş teknolojik imkanları kullanmak istemişlerdir. Bu sayede sadece özel firmalar değil, devlet kontrolünde olan firmalar da mevcut müşterilerin kaybını

engellemek için yeni teknolojileri hem ekonomik hem de teknik sebeplerle geliştirip adapte olmak mecburiyetinde kalmışlardır.

Yeniliklerin sektörde büyük bir hızla ilerlemesi, rekabetin oluşması ve dünya pazarında büyük trafik artışının olmasıyla telekomünikasyon sektörü hızla bir biçimde büyümeye evrildi. Firmaların büyümeyle ilgili beklentisi yıllık olarak, 1995 yılında 588 milyar dolardan 2000 yılında 1.06 trilyon dolara yükselmesi yönündeydi [14].

Aşağıda görüldüğü gibi, mobil telefon ve veri iletişimine yönelik iki önemli eğilime bağlı olarak, Şekil 2.1'de dört farklı bölüm tanımlanmıştır [13]:

- Ana ve gelişen segment: Sabit Telefon,
- Gelir üreten segment: Mobil Telefon,
- Piyasa üzerinde yıkıcı etkisi olan bir başka gelir artırıcı büyüme segmenti (özellikle veri trafiği): Veri İletişimi / İnternet
- Gelir açışından büyüme potansiyeli: Mobil Veri



Şekil 2.1 1970-2003 yıllarında telekomda görülen ana büyüme eğilimleri [13]

Telekomünikasyon sektörü, 2005'ten bu yana, akıllı cihazlarıyla büyük miktarda veri tüketen tüketiciler nedeniyle oldukça gelişmiştir. Ve sektör 2005 yılından bu yana fark edilmeden çeşitlenmiştir. Müşteri ihtiyaçları ile birlikte rekabet durumu tahmin edilemeyen birçok yönden farklılaşmıştır.

Sektörle ilgili çalışmaları olan EY şirketinin küresel telekomünikasyon çalışması (2015), endüstrinin geleceğe dair beklentileri ve güncel bulgularını göstermektedir:

Global telekomünikasyon çalışması temel araştırma bulguları:

- Rekabetin sertliği sektörün en büyük zorluğudur.
- OTT'ler (Over the Top uygulamaları) değişen talep senaryolarında en önde gelen ilerletici güçtür.
- Regülatif belirsizlikler, endüstrideki huzursuzlukların devam etmesine neden olmaktadır.
- Müşteri deneyimi yönetimi stratejik olarak birinci önceliktir.
- Servis seviyeleri ve servisleri kişiselleştirme müşteri deneyimi ve müşteri odaklılık için faydalıdır.
- Şebeke kalitesi firmalar için ayrışmanın kilit noktası olmaya devam etmektedir.
- İnsanların ve süreçlerin yeni etkileşimi çeviklik düzeylerinde artışa sebep olabilir.
- Pazar içi konsolidasyon, birleşme ve devralma sektörünün önde gelen faktörüdür.
- Dijital hizmetler 2020'e kadar gelir üretimini etkileyecektir.
- TV ve buluta güven oranı yüksek, ancak IoT gelir artış potansiyeli kesin değildir.

2.2 Türkiye Telekomünikasyon Piyasasının Güncel Durumu

Kuşkusuz günümüzde telekomünikasyon sektörü Türkiye'nin önde gelen sektörlerinden biridir. Türk Telekom'un (TT) analog ağını başlattığı 1986 yılından beri Türkiye'de mobil servisler mevcuttur. Ancak, o yıllarda görevlilerden oluşan yaklaşık 150.000 aboneye sahip olduğu ve zaten tam kapasitesine ulaştığı için önemsizdi. Mobil sektör gerçek anlamda 1994 yılında, Turkcell ve Telsim'in TT'yle gelir paylaşımı anlaşması yapıp faaliyete başladıkları anda başlamışlardır. Bir yıl sonra, mobil aboneler toplam telefon abonelerinin yüzde ikisine ulaşmıştır. Türkiye'nin sahip olduğu bu oran dünya ortalamasının beş yılı gerisinden gelmek demekti. 1998'in sonuna gelindiğinde mobil penetrasyon, 3,4 milyon abone ile yüzde 5'e ulaşarak ikiye katlanmıştır. Abonelerin büyümesi iki işletmecinin de her birinin GSM900 lisansı aldığı 1998 yılından sonra hızlanmıştır. Kapsamlı pazarlama

faaliyetleri ve ön ödemeli kartların piyasaya sürülmesi, 1999 yılının sonunda yüzde 12 olan penetrasyonda hızlı bir artışa yol açan ana katalizörlerdi. 2000 yılının ikinci ve üçüncü çeyreğinde, yeni müşterilerin yarısından fazlası ön ödemeli abonelerdi [15].

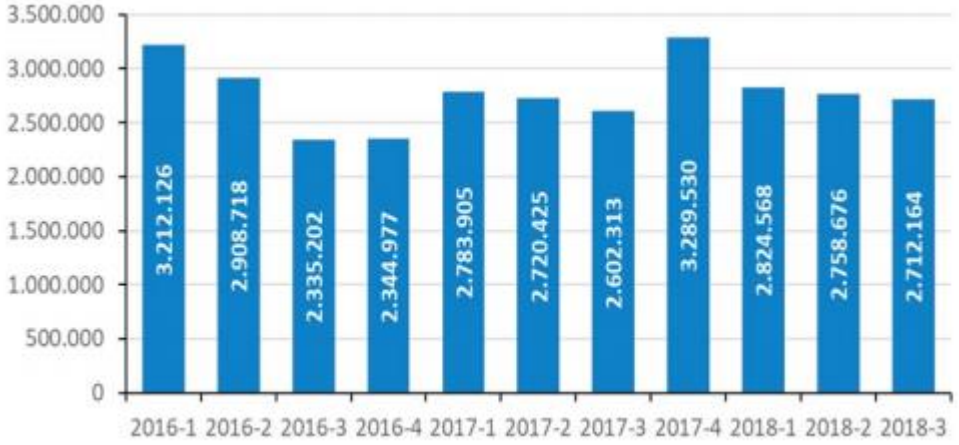
2001 yılında Turkcell ve Telsim'e iki yeni operatör katılmıştır. Bu yılda İş Bankası ve "Telecom Italia Mobile"ın sahip olduğu Aria ve Türk Telekom'un iştiraki olan Aycell, 25 yıllık GSM1800 lisansı almıştır. Aria ve Aycell, önemli operasyonel ve finansal sinerjiler oluşturmak için 2004 yılında Avea ismiyle birleşmiştir. Türk Telekom hisselerinin %55'inin özelleştirme süreci Kasım 2005'te tamamlanmıştır. Buna göre, Oger Telecom, Türk Telekom'daki %55'lik hissesiyle Avea ile birlikte Telecom Italia Mobile'ın ortak kontrolüne sahip olmuştur. Telsim, Aralık 2005'te İngiliz GSM operatörü Vodafone'a satıldı. 2015 yılında Türk Telekom Avea hisselerinin %100'üne sahip oldu. Mobil, internet, telefon ve TV hizmetlerini tek bir kanaldan sunabilmek için Avea, Türk Telekom ve TTNET markaları "Türk Telekom" tek markası altında birleştirildi [16].

Günümüze gelecek olursak güncel durumda BTK 2018 yılı üçüncü çeyrek Pazar verilerine göre telekomünikasyon sektöründe faaliyet gösteren işletmelerin net satış gelirleri 16 milyar TL'dir [17]. Bu rakam ise yıllık olarak Türkiye GSYH'nin yaklaşık %2.5'u yapmaktadır. Güncel verilere göre Türkiye'deki mobil abone sayısı yaklaşık 80.6 milyon, mobil penetrasyon oranı ise %99.8'dir. Makineler arası iletişim (M2M) ve 0-9 yaş aralığındaki nufüsü hesaptan çıkardığımızda ise mobil penetrasyon oranının %113 olduğu görülmektedir. Ön ödemeli mobil genişbant abone sayısı 25 milyon, faturalı mobil genişbant abone sayısı ise 36 milyon sayılarına erişmiştir. Aylık mobil kullanım süresi ile (460 dk) Türkiye, Avrupa ülkeleri ile kıyaslandığında 1. Sıraya yerleşmektedir [17].

Türkiye'deki telekomünikasyon sektörünün genel durumu şu şekilde özetlenebilir : Telekomünikasyon sektöründe Turkcell, Vodafone ve Türk Telekom olmak üzere üç adet operatör bulunmaktadır. 2018 yılı üçüncü üç aylık dönem itibarıyla abone sayısına bakıldığında Turkcell'in %43.3, Vodafone'un %30.9 TT Mobil'in ise %25.8'lik paya sahip olduğu görülmektedir. Pazar payları incelendiğinde ise Turkcell'in pazar payının %41.4, Vodafone'un pazar payının %36.7 ve TT Mobil'in

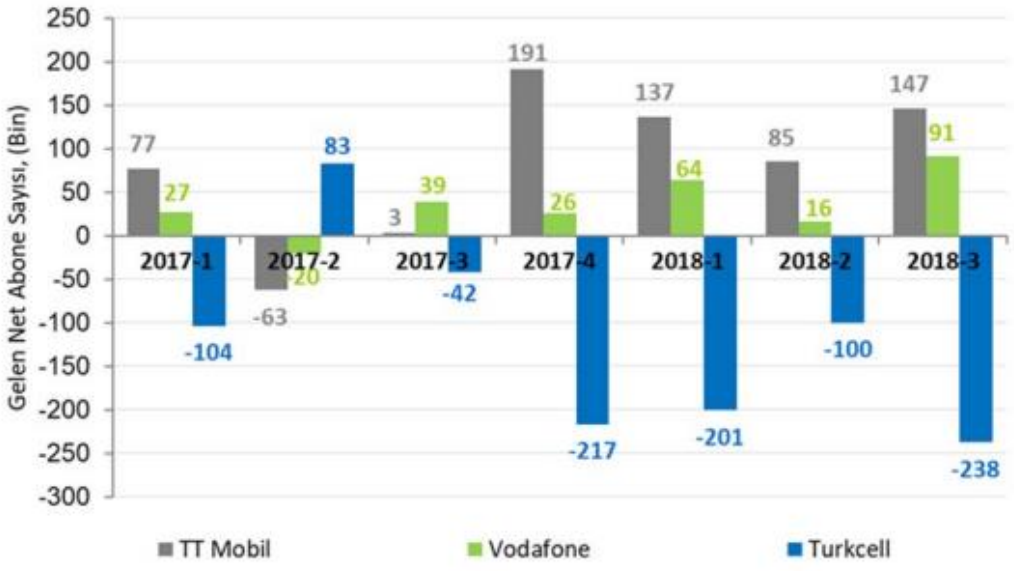
pazar payının ise %21.9 seviyelerinde olduğu görülmektedir. 2018 yılı üçüncü çeyreği itibariyle TT Mobil abonelerinin %56.5'i, Vodafone abonelerinin ise %56.7'si, Turkcell abonelerinin ise %54.5'i, faturalı abonelerden oluşmaktadır [17].

2016 yılının 1. Çeyreği itibariyle operatörler arası geçişlerde mevcut numarayı koruyarak taşıma olanağı başlamıştır. 14 Kasım 2018 tarihi itibariyle toplam 120.186.821 mobil numara taşıma işlemi gerçekleştirilmiştir. Türkiye'de 2018 3. çeyreğinde yaklaşık 2.7 milyon abone numarasını taşımıştır.



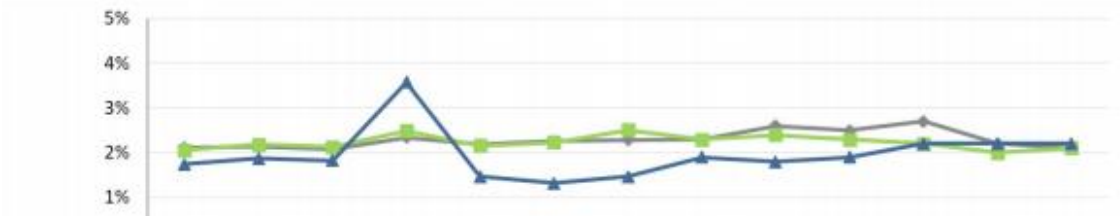
Şekil 2.2 Toplam Mobil Numara Taşıma Sayıları [17]

Şekil 2.3'te numara taşınabilirliği ile mobil işletmecilere gelen net abone sayıları üçer aylık dönemler halinde gösterilmektedir. Mobil numara taşınabilirliği (MNT) hizmeti ile 2018 yılı üçüncü üç aylık dönemde TT Mobil yaklaşık 147 bin abone, Vodafone yaklaşık 69 bin abone kazanırken, Turkcell ise yaklaşık 238 bin abone kaybetmiştir.



Şekil 2.3 MNT Kapsamında Mobil İşletmecilerin Net Gelen Abone Sayıları [17]

Şekil 2.4'te 2018 yılının üçüncü çeyreğini de kapsayacak şekilde son 12 ay için aylık bazda mobil işletmecilerin abone kayıp oranlarına yer verilmektedir. 2018 yılı Eylül ayı itibarıyla TT Mobil, Vodafone ve Turkcell'in abone kayıp oranları sırasıyla %2,1, %2,1 ve %2,2 olarak gerçekleşmiştir. Abone kayıp oranı işletmeciler tarafından kaybedilen müşterilerin miktarını ölçmek için kullanılan bir orandır. Abone kayıp oranı belli bir dönemde işletmeciden aldığı hizmeti sona erdiren abonelerin sayısının o dönemdeki mevcut ortalama abone sayısına bölünmesiyle hesaplanmaktadır [17].



	Eylül 17	Ekim 17	Kasım 17	Aralık 17	Ocak 18	Şubat 18	Mart 18	Nisan 18	Mayıs 18	Haziran 18	Temmuz 18	Ağustos 18	Eylül 18
TT Mobil	2,1%	2,1%	2,1%	2,3%	2,2%	2,3%	2,3%	2,3%	2,6%	2,5%	2,7%	2,2%	2,1%
Vodafone	2,1%	2,2%	2,1%	2,5%	2,2%	2,2%	2,5%	2,3%	2,4%	2,3%	2,2%	2,0%	2,1%
Turkcell	1,7%	1,9%	1,8%	3,6%	1,5%	1,3%	1,5%	1,9%	1,8%	1,9%	2,2%	2,2%	2,2%

Şekil 2.4 Mobil İşletmecilerin Abone Kayıp Oranları (Churn Rate) [17]

3
Müşteri Kaybı Tahmini

3.1 Müşteri İlişkileri Yönetimine (CRM) Genel Bakış

CRM, açılımı "Customer Relation Management" olan müşteri ilişkileri yönetimini anlatmak için yaygın bir şekilde kullanılan bir kısaltmadır. 90'lı yıllardan itibaren bu kavrama ilgi artmıştır ve iş dünyasında sıklıkla kullanılan bir kavram haline gelmiştir.

CRM genel olarak müşterilerle ilişkinin uzun vadeli ve karlı olmasına destek olan stratejiler ve süreçlerin tamamıdır [18]. Daha farklı bir şekilde tanımlanacak olursa CRM'in bir davranış, işe katılmış değer, iş yapış ve yaklaşım şekli ve tüm bunların müşteriyle ilişkisi olduğu söylenebilir. Hem müşterilerin zihninde organizasyona dair bir resim oluşmasını ve bu resmin gelişmesini hem de şirketin pazar payının artmasını sağlayan yöntemlerdir [19]. CRM, yeni müşteri elde etmek, mevcut müşterileri elinde tutmak, şirket karını artırmak için farklı iletişim kanallarıyla müşteri davranışlarını anlamak ve müşteriyle etkileşime geçmek için kullanılan bir yaklaşımdır [20].

Tüm bu tanımlardan yola çıkarak müşteri ilişkileri yönetiminin organizasyon çıkarı için müşteri etkileşiminde kullanılan tüm strateji, yöntem ve süreçler olduğu söylenebilir. Böylece müşteri algısı değiştirebilir ve şekillendirebilir, müşteri bilgisi daha fazla önem kazanacağı müşteri ile daha yakın ilişkiler kurulabilir ve müşterinin satın alma davranışları değiştirilebilir. Müşteri ilişkileri yönetiminde başarılı olmak firmalar için rekabet üstünlüğü yaratmaktır. Ayrıca iyi bir müşteri ilişkileri yönetimi müşteri memnuniyeti ve müşteriyi elde tutma oranlarını artırmak demektir.

CRM uygulamaları organizasyonların satın alma süreçlerinde, harcama ve yatırım miktarlarında ve yaşam döngülerini uzatma konularında müşteri sadakatini ve karlılığını değerlendirmesine yardımcı olur. CRM uygulamaları hangi ürün ve hizmetlerin müşteriler için daha iyi olduğu, müşterilerle nasıl iletişim kurulması gerektiği, müşterilerin sevdikleri renkler, müşterilerin kıyafet bedenleri vs gibi

çoğaltılabilecek bir çok soruya cevep verebilir. Özellikle müşterilerin daha iyi bilgi ve özel muamele almanın yanı sıra zamandan ve paradan tasarruf sağladıkları inancından da yararlanırlar. Ayrıca, şirketle iletişim kurmak için kullanılan kanal veya yöntem ne olursa olsun (internet, çağrı merkezleri, satış temsilcileri veya satıcı vb.) müşteriler aynı tutarlı ve verimli hizmet alırlar. Tablo 3.1, CRM'in kurum genelinde müşteri verilerini paylaşarak ve yenilikçi teknolojiyi uygulayarak sağladığı bazı faydalara kısa bir genel bakış sunmaktadır[21].

Tablo 3.1 CRM'in faydaları [21]

Müşteri veri paylaşımı sonuçlanan organizasyon:	CRM yenilikçi teknoloji:
Üstün müşteri hizmetleri seviyeleri	Self servis ve İnternet uygulamaları için müşterinin yeteneğini arttırır
Çapraz satış ve satış için fırsatlar	Kişiselleştirilmiş iletişim ve gelişmiş hedefleme ile mevcut ve yeni müşterileri çeker
Müşterilerin alışkanlıkları ve tercihleri hakkında geniş bilgi	Müşteri ve tedarikçi ilişkilerini bütünleştirir
Müşterinin entegre ve eksiksiz görünümü	Ortak ve benzersiz müşteri modellerini analiz etmek için ölçümler oluşturur
Segmentlere ve bireysel müşterilere geliştirilmiş hedefleme	
Verimli çağrı merkezleri / servis merkezleri	

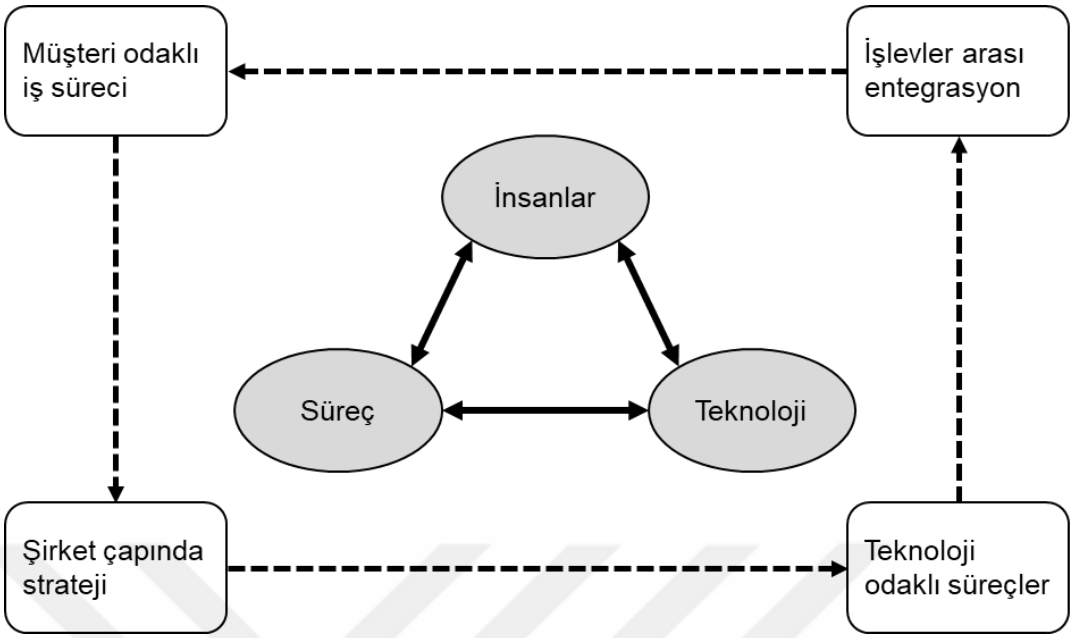
Kurumsal, müşteri odaklı, teknolojiyle bütünleşik, işlevsel bir organizasyon bağlamında, insanların, süreçlerin ve teknolojinin üç temel boyutunu birleştiren bir CRM uygulama modeli, Şekil 3.1'de sunulmuştur.

Berry ve Linoff'un [22] Şekil 3.2'de gösterildiği üzere müşterileri 5 ana gruba ayırdıkları görülmektedir.

1. Olası Müşteriler: Hedef pazar içerisinde olan fakat henüz kazanılmamış müşteriyi belirtmektedir.

2. Cevap Veren Müşteriler: Olası müşterilerin içerisinde olup ilgisi ve dikkati çekilebilmiş müşterilerdir. Bu müşteriler genellikle anket, form veya bilgilendirme için doğrudan iletişim gibi herhangi bir yol ile iletişime geçen müşterilerdir.

3. Yeni Müşteriler: Bir sözleşme ile taahhüt altına girmiş müşterilerdir.



Şekil 3.5 CRM uygulama modeline bir örnek [21]

4. Köklenmiş Müşteriler: Daha derin ve gelişmiş bir ilişkinin kurulduğu müşteri segmenti olup genellikle firmadaki süresi uzun olan ve daha fazla satışın yapılabildiği müşterilerdir.

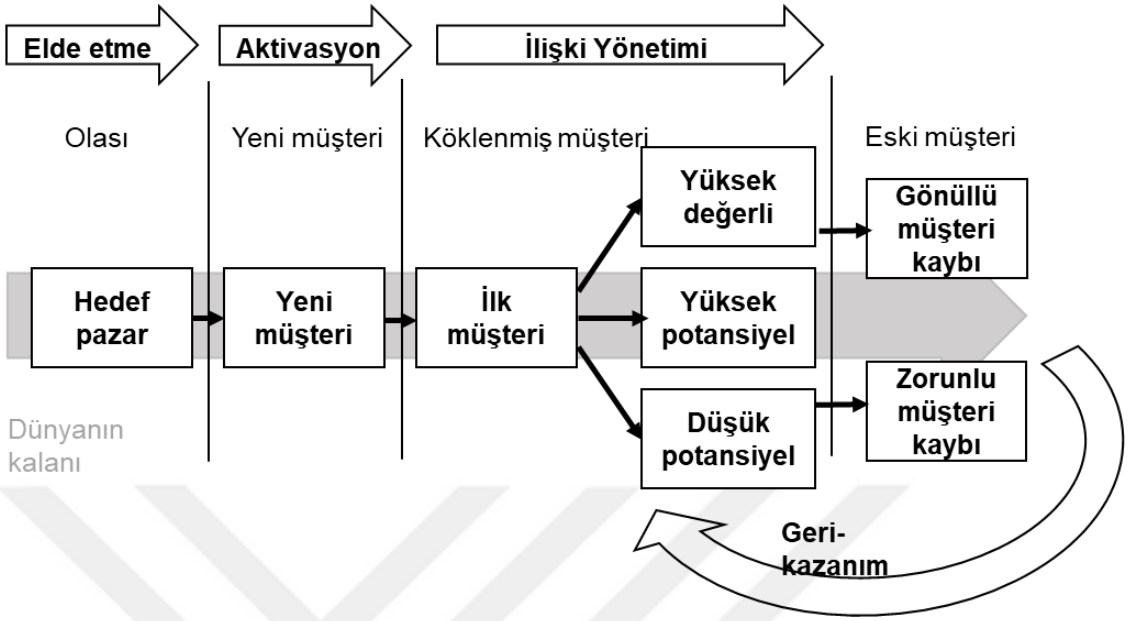
5. Eski Müşteriler: Firma ile bağını koparmış olan artık müşteri olmayan gruptur. Bu müşteriler ya gönüllü olarak ayrılmışlardır (müşteri rakip firmaya geçmiş olabilir veya müşteri ürün ile artık ilgilenmiyordur), ya zorla ayrılmaları sağlanmıştır (faturaların ödenmemesi durumu), ya da beklenen bir ayrılma işlemi (taşınma gibi zorunlu haller doğrultusunda) gerçekleşmiştir.

Müşteri tanımlarının detayları sektöre göre farklılık gösterebilir. Telekomünikasyon sektörü için müşteri tanımları değerlendirilecek olursa;

- Olası müşteriler mevcutta rakip firmalardan hizmet alan ve firmanın gelecekte hizmet verebileceği bütün müşterilerdir.

- Cevap veren müşteriler çağrı merkezini hizmet, kampanya, tarife gibi bilgileri almak için firma web sitesine girmiş, çağrı merkezini aramış veya bayi kanallarıyla iletişime geçmiş durumda olan tüm müşterilerdir.

- Yeni müşteriler hizmet almak için satış kanallarının herhangi biri yoluyla başvuruda bulunmuş ve hizmeti açılarak faturalama yapılabilen müşterilerdir.



Şekil 3.6 Müşteri Yaşam Döngüsü Süreci [22]

- Kazanılan müşteriler mevcuttaki ve daha fazla kar edilebilen müşterilerdir. Böylelikle up-sell (pahalı tarife satışı) ve cross-sell (cihaz veya destekleyici hizmet satışı) yapılarak daha fazla ürün satılabilen müşterilerdir.

- Eski müşteriler rakip operatöre geçmiş, hizmetini kapatmış veya borçtan iptal durumunda kapatılmış müşterilerdir.

3.2 Müşteri Kaybı (Churn)

Lazarov ve Capota literatürde aşağıda gösterildiği gibi üç tür müşteri kaybı tanımlanmıştır [23]:

- **Aktif Kayıp Müşteri (Gönüllü):** Mevcut operatörü bırakmayı tercih eden ve başka bir operatöre geçmek isteyen müşteriler. Abonelikleri sonlandırma örnek olarak mobil telefon cihaz modelini yükseltme / düşürme isteği, şebeke kalitesi faktörü, rakip pazarlama faaliyetleri, yönetmelikler vb. sayılabilir.
- **Pasif Kayıp Müşteri (Gönüllü olmayan):** Fatura borcu nedeniyle şirket tarafından atılan müşteriler.

- **Rotasyonel Kayıp Müşteri (Sessiz):** Sözleşmesi önceden bildirimde bulunmaksızın beklenmedik bir şekilde şirket veya müşteriler tarafından sonlandırılmış müşteriler.

Müşteri kaybı ile sadakat arasında doğrudan bir bağlantı vardır. Günümüzün iş dünyasında, fiyat indirimleri sadakat için yeterli değildir. Ürüne yeni bir değer katıldığı takdirle müşterilerin daha sadık hale gelmesi beklenir. Müşteri kaybı analizi için esas amaç, aboneliklerini iptal edecek muhtemel müşterileri tespit etmek ve geri kazanma çabalarının toplam maliyetini hesaplamaktır.

3.3 Müşteri Kaybı Yönetimi

Müşteri kaybı analizi çeşitli sektörlerde yapılmaktadır. Müşterilerin davranışlarına ve hareketlerine göre çevik bir yapıda olması dinamik aksiyonlar gerektirir. Dolayısıyla analitik olarak ele alınması gereken bir konudur. Müşterilerin davranışları ve aldıkları aksiyonlar şirketler için müşteri kaybı analizinde önemli bir ipucudur.

Telekomünikasyon sektöründe müşteri sadakatini artırmak için kullanılan stratejilerden en önemlisi müşterileri ön ödemeli aboneliklerden faturalı aboneliklere geçirmektir. Bu stratejinin diğer önemli nedeni faturalı müşterilerde abone başına düşen ortalama gelirin (ARPU) ön ödemeli müşterilerinkine göre daha yüksek olmasıdır. Literatürde yapılan çalışmalar gösteriyor ki müşterilerin abonelik tipi ile gelirinin, eğitim düzeyinin, yaşının, cinsiyetinin, ailesinin eğitim düzeyi, aile gelirinin vb. arasında ilişki mevcuttur [24]. Bu sebeple ön ödemeli abonelerde fiyat değişikliklerine olan tolerans faturalı abonelere göre düşüktür.

Yapılan araştırmaların diğer sonuçları, akranların marka değiştirme konusundaki davranışları, müşterilerin sosyal çevresi, mobil şebeke kalitesi ile karşılaştırıldığında müşterilerin operator değişikliği kararları üzerinde çok daha önemli bir etki yarattığını göstermektedir.

Müşteri kaybı tahminin temeli daha önce kaybedilmiş müşteriler hakkında bilgi içeren tarihsel verilere dayanır. Mevcut müşteriler ile daha önce aboneliklerini sonlandırmış müşteriler kıyaslanır. Olası kaybedilecek müşteriler, sınıflandırmanın

önceki kaybedilmiş müşterilere benzer şekilde önerdiği müşteriler olarak tanımlanmaktadır.

4
Yapay Öğrenme Teknikleri

4.1 Yapay Öğrenmeye Genel Bakış

Yapay öğrenme hem istatistik biliminin hem de bilgisayar biliminin konusudur. Bu alan daha çok yakın dönemde duyulmaya başlansa da istatistik biliminin bu alandaki çalışmaları 1950'li yıllara dayanmaktadır. O yıllarda alanın sadece akademiyle sınırlı kalmasının ve endüstride karşılık bulamamasının sebebi geliştirilen algoritmaların etkin ve hızlı bir şekilde çalıştırılabileceği bilgisayar yazılım ve donanımlarının bulunmamasıdır. Dolayısıyla, yapay öğrenme hakkında söylenebilecek ilk şey bu alandaki algoritmaların faydalı olabilmesi için büyük miktarda veri üzerinde çalışmaları ve güçlü donanım ve yazılımlar gerektirmeleridir. 1980'den sonra bilgisayar bilimi alanındaki gelişmeler bu algoritmaların pratikteki kullanımlarını kolaylaştırmış ve yaygınlaştırmış ve bu alanı günümüzde popüler bir araştırma ve uygulama mecrası haline getirmiştir.

Bilindiği üzere bilgisayarlar kendilerine ilgili tanımlamalar yapılmadığı sürece hiç bir problem çözemezler. Örneğin, bir bilgisayar yazılımından kahve pişirmesini isteniyor ise adım adım bütün işlemleri tanıtılması gerekmektedir. Tahmin edileceği üzere kahve pişirmek oldukça basit problemdir. Muhtemelen bir yazılıma kahve pişirmeyi birkaç adımda tanıtabilirsiniz. Fakat kahve pişirmek kadar basit olmayan problemler de vardır. Tıbbi teşhis problemi bu tarz problemlere uygun bir örnektir. Bir hastanın kanser olup olmadığı konusunda karar verecek bir yazılım yazmak kahve pişirecek bir yazılım yazmak kadar kolay olmayacaktır. Bu tarz problemler çok fazla parametreye tabiidir ve bu kadar fazla parametreyi hesaba katacak belirleyici (deterministik) bir yazılım yazmak günün sonunda işin içinden çıkılamayacak bir hale gelecektir. Belirtildiği gibi, kanser teşhisi konusunda belirleyici bir yazılım yazılamasa da araştırmacının elinde önemli bir avantaj bulunmaktadır; bu avantaj veridir. Yani bir çok kişinin tahlillerini, demografik bilgilerini, tıbbi bilgilerini ve hastalığı taşıyıp taşımama bilgilerini içeren geniş bir

örneklem elde bulunmaktadır. Dolayısıyla bu örneklem kanser hastalığı ile onun belirleyicileri arasındaki bir çok ilişkiyi içermekte ve araştırmacıya kanser teşhisini deterministik bir yazılım yazılmaksızın istatistiksel yöntemlerle yapabilme imkanı sağlamaktadır. Yapay öğrenme algoritmalarının temel olarak yaptıkları şey budur. Kanser örneğine dönülürse bu algoritmalar örneklemdeki örüntüleri ortaya çıkararak araştırmacıya bir kişinin kanser olma olması olasılığını verebilmektedir. Buna istatistik de örüntü tanıma da (pattern recognition) denilmektedir. Burada göz önünde bulundurulması gereken konu yapay öğrenme algoritmaları belirli bir hata payı dahilinde genelleme yapabilme kabiliyetine sahip algoritmalardır. Bu durum algoritmanın deterministik olmamasından (istatistiksel olmasından) ve bir örneklem üzerinde çalışmasından kaynaklanmaktadır.

Yapay öğrenme literatürde kullanım alanı ve problemin tipine göre 3 ana başlıkta incelenmektedir [25]. Bu 3 ana alan aşağıdaki gibidir:

- Gözetimli Öğrenme (Supervised Learning)
- Gözetimsiz Öğrenme (Unsupervised Learning)
- Pekiştirmeli Öğrenme (Reinforcement Learning)

4.1.1 Gözetimli Öğrenme

Gözetimli öğrenmede araştırmacının elindeki örneklem tahmin etmeye çalıştığı problem hangi koşullarda hangi çıktıyla sonuçlandığını belirten gözlemleri içermektedir. Örneğin, daha önce belirtilen kanser teşhisi problemine dönülecek olursa araştırmacı hangi özelliklere sahip kişilerin kanser olup olmadığını bilmektedir. Araştırmacı bu bilgiye sahip olduğu için bu yapay öğrenme tekniğine gözetimli öğrenme denilmektedir. Gözetimli öğrenmede değişkenler ikiye ayrılmaktadır. Tahmin edilmesi hedeflenen değişkene (kanser teşhisi örneğinde hastanın kanser olup olmadığı değişkeni) literatürde tepki değişkeni (response variable) denilmektedir. Geriye kalan değişkenlere ise açıklayıcı değişkenler (descriptive variable) denilmektedir. Bu tarz modellerde temel mekanik şu şekilde işler: Açıklayıcı değişkenler ile tepki değişkeni arasında anlamlı bir istatistiksel ilişki olduğu varsayılmakta, model ise bu ilişkiyi tanımlayan rassal süreci veya ortak

dağılımı (joint distribution) açıklamaktadır. Gözetimli öğrenme teknikleri tepki değişkeninin niteliğine göre iki başlıkta incelenmektedir:

- Sınıflandırma (Classification)
- Bağlanım (Regression)

4.1.1.1 Sınıflandırma

Sınıflandırma problemlerinde tepki değişkeni sonlu (finite) sayıda değer alabilir. Örneğin, gereksiz (spam) e-posta tespiti problem düşünüldüğünde tepki değişkeni iki değerden birisini alabilir. Bir elektronik posta ya spamdir ya da gerçek bir maildir. Benzer şekilde batık kredi tespiti problemin de kredi ya sağlıklı bir kredidir ya da batık bir kredidir. Verilen her iki örnekte de tepki değişkeni iki değerden birini aldığından dolayı bu tarz problemlere ikili sınıflandırma (binary classification) problemleri denir. Bu araştırmanın konusu olan churn problemi de bir ikili sınıflandırma problemidir. Günümüzde pratikte çalışılan problemlerin çoğu ikili sınıflandırma problemi olsa da ikiden fazla sınıfın olduğu problemlerde mecvuttur. Örneğin, müşterinin farklı ürünler içerisinden hangi ürünü tercih edeceğini belirleme probleminde ikiden fazla sınıf bulunabilmektedir (üç farklı araba modelinin tercih edileceği bir problem gibi). Bu şekildeki problemlere de çoklu sınıflandırma (multi classification) problemleri denilmektedir.

Bir sınıflandırma probleminin çıktısı, gözlemin problemin sınıflarına hangi olasılıkla ait olduğudur. Örneğin, bir kanser teşhis probleminde yeni bir gözlemin çıktısı %72 ihtimalle kanser hastalığı taşıma ve %28 ihtimalle kanser hastalığı taşımama olabilir. Sınıf olasılıklarının toplamı daima 1'i vermelidir.

Endüstride sınıflandırma problemine örnek teşkil edebilecek bir çok problem vardır. Kredi riski, sahtekarlık tespiti, churn tahmini, ürün alma eğilimi, poliçe riski vb. problemler sınıflandırma problemlerine örnektir.

Sınıflandırma problemi için en çok kullanılan algoritmalar; karar ağaçları, genelleştirilmiş lineer modeller (lojistik regresyon), k-en yakın komşu, yapay sinir ağları (ANN), destek vektör makinesi (SVM), komite öğrenmeleri, doğrusal ayırtaç analizi (linear discriminant analysis), bayesci kestirim modelleridir.

4.1.1.2 Bağlanım

Regresyon problemlerinde söz konusu problemin tepki değişkeni sonsuz sayıda değerden birini alabilmektedir. Regresyon problemlerini sınıflandırma problemlerinde ayıran temel husus budur. Bir örnekle açıklamak gerekirse döviz fiyatlarının tahmini problemi ele alınabilir. Bu problemde tepki değişkeni bir fiyattır ve bu fiyat gerçek sayılar kümesinden herhangi bir değeri alabilir. Sınıflandırma problemlerinde olduğu gibi tepki değişkeni sonlu sayıda değerden birini almadığı için modelin performansı yanlış sınıflandırma hatası (misclassification error) gibi olasılıksal bir büyüklük yerine ortalama mutlak yüzde hata (MAPE: mean absolute percentage error) gibi istatiksel büyüklüklerle ölçülmektedir.

Endüstride regresyon problemlerine örnek oluşturabilecek bir çok problem bulunmaktadır. Enerji fiyat ve talep tahmini, finansal varlık fiyat ve talep tahmini, taşınır-taşınmaz malların fiyat ve talep tahmini bu problemlere örnek olarak verilebilir. Regresyon problemlerini çözmek için kullanılan algoritmaların çoğunluğu sınıflandırma problemleri için kullanılan algoritmalarla aynıdır. Karar ağaçları, lineer regresyon modelleri, yapay sinir ağları (ANN), komite öğrenmeleri, destek vektör makinesi (SVM) modelleri örnek verilebilir.

4.1.2 Gözetimsiz Öğrenme

Gözetimsiz öğrenmede gözetimli öğrenmede olduğu gibi açıklanmak istenen bir tepki değişkeni yoktur. Bu tarz öğrenme biçiminde örneklem, genelde gözlemler ile ilgili açıklayıcı değişkenleri içerir. Gözetimsiz öğrenmede amaç tahmine dayalı bir sınıflandırma veya öngörüde bulunmak değil sadece ve sadece örneklem içerisindeki örüntüleri ortaya çıkarmaktadır. Gözetimsiz öğrenmenin en yaygın kullanım alanı öbekleme (clustering) problemleridir. Örneğin bir perakande işletmecisinin müşterilerini belirli sayıda segmentlere ayırması bir öbekleme uygulamasıdır. Gözetimsiz öğrenmede bu öbekleme uygun algoritmalar kullanılarak örneklemdeki örüntülere göre yapılır. Bu uygulamalar sonucu oluşturulan öbekler şirketlerin CRM süreçlerinde kullanılır.

Endüstride gözetimsiz öğrenme problemlerine örnek olabilecek problemler müşteri öbeklemesi, görüntü sıkıştırma, belge öbekleme ve biyoinformatik alanında genlerin öbeklenmesidir [25].

4.1.3 Pekiştirmeli Öğrenme

Pekiştirmeli öğrenme daha çok robotik alanında uygulanmaktadır. Pekiştirmeli öğrenmede amaç belirli bir hedefe ulaşmak için gerekli olan doğru adımların öğrenilmesidir. Pekiştirmeli öğrenmeyi gözetimli öğrenmeden farklı kılan temel unsur budur. Çünkü gözetimli öğrenmede hedef tek adımda öğrenilir. Aksine pekiştirmeli öğrenmede bir hedef, kurallar ve o hedefe giden farklı politikalar vardır. Pekiştirmeli öğrenmeye en isabetli örnek oyun oynamaktır. Örneğin, dama gibi bir oyunun az sayıda ve herkes tarafından anlaşılabilir basit kuralları olmasına rağmen oyunu kazanmaya götürecek hamleler ustalık gerektirecek seviyede zor olabilir. Pekiştirmeli öğrenme oyunu kazanmaya götürecek stratejileri öğrenmek için uygulanan yöntemler bütünüdür. Endüstriden bir örnek olarak robotlar belirli görevlerini yerine getirmeyi pekiştirmeli öğrenme vasıtasıyla gerçekleştirirler.

Bundan sonraki kısımda bu çalışmada kullanılan 4 farklı yönteme yer verilmiştir:

4.2 Lojistik Regresyon

Lojistik regresyon, doğrusal regresyonun daha geniş bir kümesi olan genelleştirilmiş doğrusal modeller (GLM) ailesine üye bir tahminleme modelidir. 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 (binary classification) problemlerini tahmin etme için kullanılır. Problemin konusu olan olayın olma ihtimali P ile tanımlanırsa lojistik regresyon aşağıdaki lineer modeli çözer.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (4.1)$$

Burada k, bağımsız değişkenlerin toplam sayısını ifade etmektedir. Bu model, model katsayılarını belirlemektedir. Belirlenen model katsayılarıyla hesaplanan p değeri bir olasılık değer olup 0 ile 1 arasında olmak zorundadır. p aşağıdaki gibi ifade edilir:

$$p = \frac{1}{1+\exp[-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)]} \quad (4.2)$$

Bu fonksiyona literatürde sigmoid fonksiyonu adı verilmektedir.

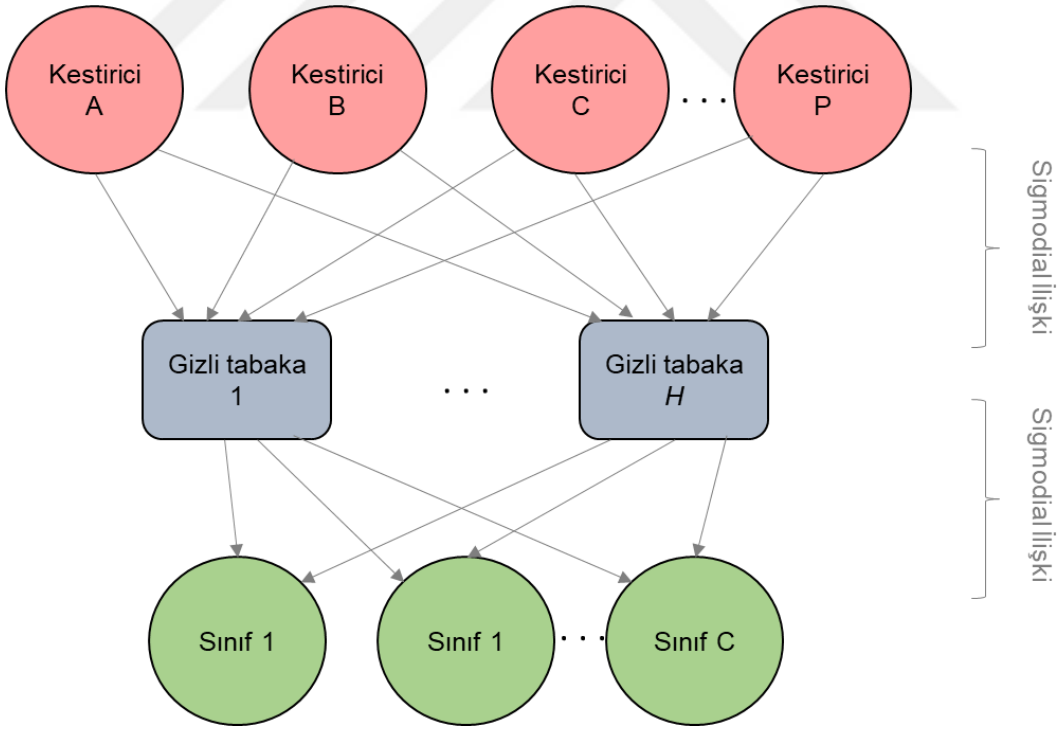
Lojistik regresyon, açıklayıcı değişkenler üssel temel fonksiyonlar ile dönüştürülmediği sürece doğrusal karar sınırları oluşturur. Bu da modelin tahminleme performansı düşünüldüğünde en büyük dezavantajlarından birisidir. Çünkü gerçek dünyada birçok problem doğası gereği doğrusal olmayan karar sınırlarını gerektirmektedir. Bunun yanında açıklayıcı değişkenlerin eğim parametrelerinin istatistiksel olarak anlamlı olup olmadığına olanak tanıdığı için istatistiksel çıkarım problemlerine uygundur. Tahminleme başarısı lojistik regresyondan daha iyi olan birçok modelde istatistiksel çıkarım yapmak mümkün değildir [26].

4.3 Yapay Sinir Ağları

Yapay sinir ağları modelleri sınıflandırma problemleri için uygulandığında lojistik regresyon yönteminde olduğu gibi açıklayıcı değişkenlerin doğrusal kombinasyonlarını kullanarak problemin konusu olan olayın olma olasılığını tahminlemeye çalışır. Ve yine lojistik regresyonda olduğu gibi aktivasyon fonksiyonu olarak sigmoid fonksiyon en yaygın olarak kullanılmaktadır. Aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanmayan yapay sinir ağları modelleri de vardır. Bunlardan en bilineni, aktivasyon fonksiyonu olarak basamak (step) fonksiyonlarını kullanan algılayıcı (perceptron) adıyla bilinen algoritmalardır. Bu iki yaklaşımın performansı çoğu problemde birbirine yakın olmasına rağmen step fonksiyonunun süreksiz olması parametre tahmininde çeşitli teknik zorluklar çıkardığı için sürekli olan sigmoid fonksiyonunun daha cazip ve yaygın hale gelmesine yol açmıştır. Yapay sinir ağları lojistik regresyonun aksine lineer olmayan karar sınırlarını da hesaplayabilmektedir. Bunu da yapay sinir ağının çok katmanlı yapısı sağlamaktadır. Daha açık bir ifadeyle yapay sinir ağlarındaki gizli tabaka bunu

olanaklı kılmaktadır (Şekil 4.1). Bu modelin parametreleri genellikle geri hata yayılımı (error back propagation) algoritmasıyla hesaplanmaktadır [27].

Özü itibariyle yapay sinir ağları da parametrik modellerdir. Dolayısıyla ağdaki düğüm sayısı ve açıklayıcı değişken sayısı arttıkça modelin hesaplaması gereken parametre sayısı kontrolsüz bir şekilde artacak ve modelde aşırı öğrenme (overfitting) ve çok boyutluluk problemleri sıklıkla görülecektir. Bu yüzden bu tarz modellerin en büyük dezavantajı bu olup çok hassas bir kalibrasyon işlemi gerektirmektedir. Genellikle bu modellerde açıklayıcı değişkenler çok hassas bir şekilde çeşitli algoritmalar kullanılarak dönüştürülür. Ve eğer farklı modeller biraraya getirilip model ortalaması alınmıyorsa overfittingi engellemek için genellikle ağırlık sönümlemesi (weight decay) gibi regülarizasyon yöntemleri kullanılmaktadır. Bütün bu işlemler modelin hesaplama yükünü artırmaktadır. Dolayısıyla sonraki bölümlerde bahsedeceğimiz parametrik olmayan ağaç tabanlı modellere göre churn gibi problemler için daha düşük genelleme performansı göstermektedirler.

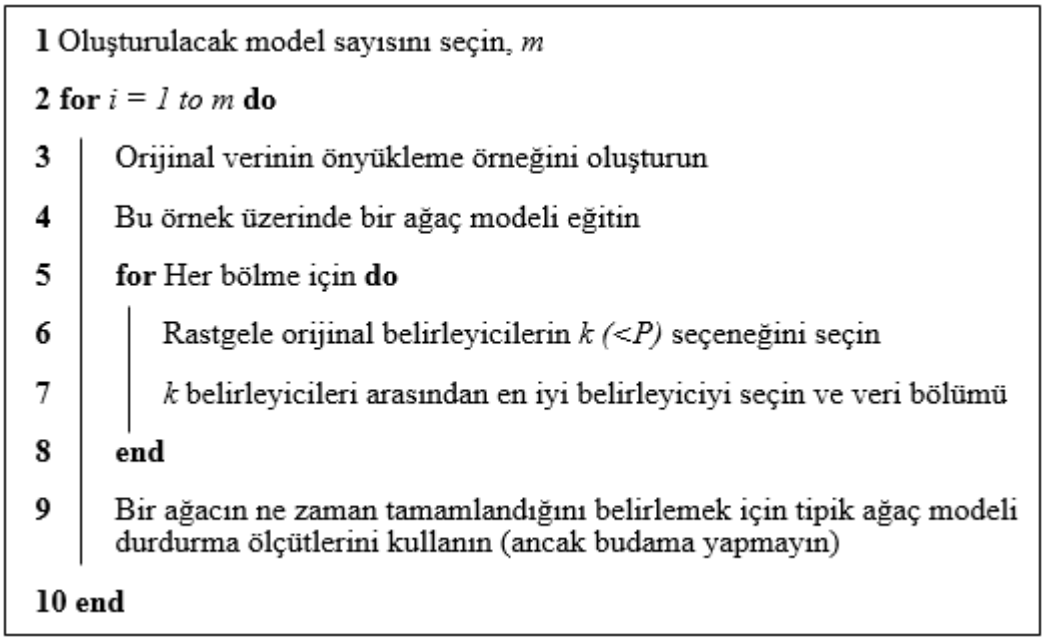


Şekil 4.7 Sınıflandırma için bir gizli katmanlı yapay sinir ağları yapısı [26]

4.4 Rastgele Orman (Random Forest)

Random forest bir komite öğrenmesi algoritmasıdır. Komite öğrenmesi algoritmaları birden fazla model oluşturarak bu modellerin ortak kararından faydalanırlar. Random forest algoritması temel öğrenici olarak genellikle CART (Classification and Regression Tree) algoritmasını kullanır. Random forest'ın genelleme performansı oldukça yüksektir. Bunu da varyans-yanlılık ilişkisinde varyansı çok başarılı şekilde düşürmesine borçludur. Modelin mekanizması kısaca şöyle çalışır: Model temel amaç olarak birbirinden bağımsız yüzlerce karar ağacı üretmeyi hedefler. Bu tekil karar ağaçlarının istatistiksel olarak birbirinden bağımsız olması modelin performansı açısından çok önemlidir. Eğer istatistiksel bağımsızlık yeterince sağlanamazsa modelin hedeflediği varyansdaki düşüş yeterli olmayabilir. Bu da genel model performansının düşük olmasına yol açacaktır. Random forest algoritması tekil ağaçların birbirinden bağımsızlığını modele iki tür rassallık katarak sağlamaktadır. Zaten modelin adı da modelin getirdiği bu rassallık unsurlarından kaynaklanmaktadır. Model her ağacı oluştururken varolan bütün gözlemlerin belirli bir alt kümesini rassal olarak seçer. Bu modelin getirdiği birinci tür rassallıktır. Bu sayede modelin oluşturduğu her bir tekil ağaç eldeki örneklemenin birbirinden bağımsız alt kümeleri üzerinden öğrenme yapacaktır. Fakat teorik ve ampirik çalışmalar göstermiştir ki bu tarz bir rassallık tekil ağaçların biribirinden yeterince bağımsız olmasını sağlamamaktadır. Özellikle yeterince büyük örneklemlerde tekil ağaçların ağacın köküne en yakın olan dallanmaları birbirine çok yakın çıkmaktadır. Bu yüzden random forest algoritması ikinci tür bir rassallıktan faydalanır. Bu ikinci rassallık her bir tekil ağaç için açıklayıcı değişkenlerin sadece belirli bir rassal alt kümesini kullanarak sağlanabilir. Böylelikle her bir tekil ağaç açıklayıcı değişkenlerin sadece belirli bir kısmını kullanarak modeli öğrenebilecektir. Bu durumda amaçlandığı gibi ağaçların birbirinden yeterince bağımsız olmasını sağlamaktadır. Bu mekanizma ile random forest algoritması oldukça düşük bir hata payına ulaşabilmektedir [28]. Random forest algoritmasının kalibrasyon parametreleri oldukça basittir. Genelleme performansı üzerinde etkisinin en yüksek olduğu gözlenen kalibrasyon parametreleri, modelin oluşturacağı tekil ağaçların sayısı (m) ve her bir ağaç için kaç tane açıklayıcı değişkeni rassal olarak seçeceğidir (k) [29]. Random forest

algoritması temel öğrenici olarak CART algoritmasını kullanması ve karar verici olarak modeller ortalaması kullanması sebebiyle tahminleme algoritmalarında karşılaşılan en ciddi problemler olan overfitting, yüksek boyutluluk ve uç değerler gibi problemlere oldukça dayanıklıdır. Bu yönüyle düşünüldüğünde diğer komite modelleriyle beraber en az ön veri işleme gerektiren modellerden bir tanesidir. Bu durum genel tahmin performansının yüksekliği ile beraber bu modelleri oldukça cazip hale getirmektedir [26].



Şekil 4.8 Temel Random Forest [26]

4.5 Boosting

Boosting algoritmaları da random forest algoritmalarında olduğu gibi gibi birer komite öğrenmesi algoritmalarıdır. Boosting yöntem olarak herhangi bir tekil sınıflandırma algoritmasıyla beraber kullanılabilmesine karşın CART algoritması istatistiksel özelliklerinden dolayı boosting için en uygun algoritma olup yaygın olarak kullanılır. Çünkü CART algoritması tipik bir düşük yanlılık ve yüksek varyans algoritmasıdır. Boosting algoritması da random forest algoritması gibi model varyansını düşürmeyi amaçlar. Böylelikle CART gibi düşük yanlılık ve yüksek varyans özelliklerine sahip bir sınıflandırma algoritması ile uygulandığında oldukça arzu edilen düşük yanlılık ve düşük varyanslı tahminler oluşturan bir komite

öğrenmesi algoritmasına dönüşebilmektedir. Boosting temel olarak örneklem içerisindeki gözlemlere ağırlıklar vererek çalışır [30]. Temel öğrenici olarak kullanılan CART modeli zayıf öğrenici olarak tabir edilir. Zayıf öğrenicinin tanımı rastgele karar veren bir öğreniciden marjinal seviyede daha iyi performans gösteren öğrenicidir. Bu özelliği ile boosting aslında birçok zayıf öğreniciyi biraraya getirerek kuvvetli bir öğrenici yaratmayı amaçlamaktadır. Algoritma iteratif bir algoritmadır. İterasyonda bir önceki ağacın yanlış tahminlediği örneklere 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 herbir 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. Bu modellerde temel öğrenici olarak CART algoritmasını kullandıkları için tıpkı random forest gibi overfitting, yüksek boyutluluk ve uç değerlere karşı oldukça dirençlidirler. Bu modellerin en etkin kalibrasyon parametreleri iterasyon sayısı, rassal olarak seçilecek örneklem alt kümesi ve ağaç derinliğidir (interaction depth) [31]. Mevcut teorik özelliklerinden dolayı random forestla beraber bu modellerin de churn ve benzeri problemler için en uygun modeller olduğunu değerlendirilmiştir.

1 Bir sınıfı+1 değerinde ve diğerini -1 değerinde gösterelim

2 Her örneğe aynı başlangıç ağırlığını verelim (1 / n)

3 for K = 1 to K **do**

4 Ağırlıklandırılmış örnekleri kullanarak zayıf bir sınıflandırıcı oluşturun ve k. modelinin yanlış sınıflandırma hatasını hesaplayın (err_k)

5 k. kademe değerini ln ((1 - err_k) / err_k) olarak hesaplayın.

6 Yanlış öngörülen örneklere daha fazla ağırlık vererek ve doğru tahmin edilen örneklere daha az ağırlık vererek örnek ağırlıklarını güncelleyin.

7 end

8 Yükseltilmiş sınıflandırıcının her bir örnek için tahminini, k. aşamasını k. model tahminiyle çarparak ve bu miktarları k'ye ekleyerek hesaplayın. Bu toplam pozitifse, örneği +1 sınıfında, aksi takdirde -1 sınıfında sınıflandırın.

Şekil 4.9 İkili sınıf problemleri için AdaBoost algoritması [26]

1 Tüm tahminleri örneklem log olasılıklarıyla başlat. $f_i^{(0)} = log\frac{p}{1-\hat{p}}$

2 for *iteration* $j = 1 \ldots M$ **do**

3 Artıkları hesaplayın (yani gradyan) $z_i = y_i - \hat{p}_i$

4 Eğitim verilerini rastgele örnekleyin.

5 Kalıntıları sonuç olarak kullanarak rastgele altküme bir ağaç modeli yetiştirin.

6 Pearson kalıntılarının terminal düğüm tahminlerini hesaplayın:

$$r_i = \frac{\frac{1}{n}\sum_1^n (y_i - \hat{p}_i)}{\frac{1}{n}\sum_1^n \hat{p}_i(1-\hat{p}_i)}$$

7 Mevcut modeli $f_i = f_i + \lambda f_i^{(j)}$ kullanarak güncelleyin.

8 end

Şekil 4.10 Sınıflandırma için basit gradient boosting [26]

4.6 Performans Değerlendirme Yöntemi

Churn tahmini gibi karmaşık veri madenciliği problemlerinde herhangi br sınıflandırma yöntemi hemen her probleme uygulanabilir halde değildir. Çoğu yapay öğrenme algoritmaları farklı modelleri oluşturmak için farklı değerlerle ve ayarlamalarla çalıştırılır. Bu modeller oluşturulduktan sonra, isabetli bir tahmin yapmak ve en iyi modeli seçmek için modellerin karşılaştırılmaları gerekir. Algoritmanın modelleri ne kadar iyi tahmin edip karşılaştırabildiğini bize söyleyen bir skorlama ölçütüne ihtiyaç duyulmaktadır.

Test veri kümesine yeni bir örnek geldiğinde, model bu veriyi kendi algoritmasına göre tahmin eder. Tahmin doğruysa (test verisetindeki ile aynı değerde), doğru olarak sayılır. Tahmin yanlışsa, yanlış olarak sayılır. Eğer tahmin negatif (churn etmemiş) ve doğruysa tahmin **doğru negatif**, eğer tahmin pozitif (churn etmiş) ve doğruysa tahmin **doğru pozitif** olarak sayılır. Tahmin negatif ancak gerçek sınıf pozitifse **yanlış negatif**, tahmin pozitif ancak gerçek sınıf negatifse **yanlış pozitif** veya yanlış alarmdır. İkiye iki hata matrisi (confusion matrix), bir sınıflayıcıdan ve Tablo 4.1'deki gibi bir dizi test örneğinden oluşturulabilir. Bu matris, genellikle birçok ölçümün temelidir.

Tablo 4.2 Hata Matrisi

	Gerçek kayıp müşteriler	Gerçek sadık müşteriler
Tahmin edilen kayıp müşteriler	doğru pozitif	yanlış pozitif
Tahmin edilen sadık müşteriler	yanlış negatif	doğru negatif

Yapay öğrenme algoritmalarını değerlendirmek için kullanılan hata matrisine dayanan en yaygın performans ölçümlerinin tanımları aşağıda verilmiştir:

$$yprate = \text{Yanlış pozitif oran} = \frac{YP}{N} \tag{4.3}$$

$$dprate = \text{Doğru pozitif oran} = \frac{DP}{N} \tag{4.4}$$

$$Duyarlık\ (Precision) = \frac{DP}{DP+YP} \tag{4.5}$$

$$Geri\ çağırım\ (Recall) = \frac{DP}{DP+YN} \tag{4.6}$$

$$Hatasızlık\ (Accuracy) = \frac{(DP+DN)x100}{DP+DN+YP+YN} \quad (4.7)$$

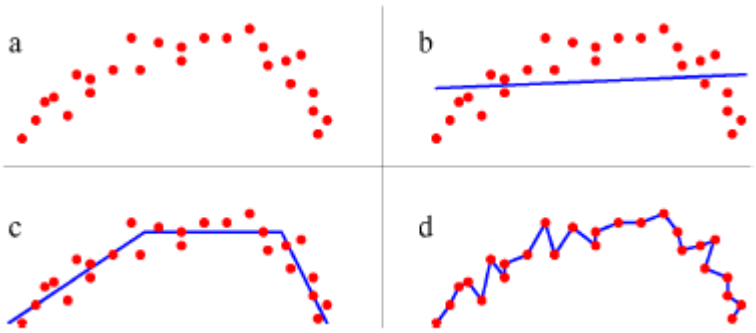
$$F\text{-}ölçüsü = \frac{2\ x\ Duyarlık\ x\ Geri\ çağırım}{Duyarlık+Geri\ çağırım} \quad (4.8)$$

Burada;

P = Gerçek pozitif sayısı,
N = Gerçek negatif sayısı,
DP = Doğru pozitifler,
DN = Doğru negatifler,
YP = Yanlış pozitifler,
YN = Yanlış negatifler'dir.

Farklı sınıflayıcıları değerlendirmek için kullanılabilecek bazı ölçütlere örnek olarak hatasızlık, F-ölçüsü, kaldırma tablosu (lift-chart), ROC alanı verilebilir. Her biri performansı değerlendirmede farklı bir yaklaşımı temsil eder. Bir ölçü seçmeden önce, performansın uygulamamız için ne anlama geldiğini netleştirmemiz gerekir. Performans, veri özelliklerine ve problem alanına göre değişir [32]. Bu çalışmada hangi performans ölçütünün ne sebeple kullanıldığı bir sonraki bölümde detaylı verilmiştir.

Açıklanması gereken diğer iki pratik konu aşırı öğrenme (over-fitting) ve eksik öğrenmedir (under-fitting). Bu terimlerle hem yapay öğrenme yöntemlerini incelerken hem de bir sonraki bölümlerde karşılaşılacaktır. Aşırı öğrenme ve eksik öğrenme ile yalnızca churn tahmininde değil tüm sınıflandırma problemlerinde sıkça karşılaşılır. Aşırı öğrenme, sınıflandırma eğitimi veri setinde çok fazla uzmanlaştığında gerçekleşir. Bu, genellikle sınıflandırma modelinin veya yapısının aşırı karmaşık olması durumunda meydana gelir, bu nedenle aşırı parametreleme nedeniyle verilerdeki gürültüden ciddi şekilde etkilenir. Bu nedenle, testte veya yeni veri setinde kötü performans gösterir. Aslında verilerde bulunmayan kalıpları keşfeder. Diğer mesele aşırı öğrenmenin tam tersi olan eksik öğrenmedir. Etkin sınıflandırma modeli verilere uygun değildir. Çünkü, eğitim verileri için bile çok genel veya basittir. Eksik öğrenmenin keşfedilmesi daha kolaydır, çünkü sınıflayıcı eğitim verileri üzerinden zayıf bir performans gösterir. Şekil 4.2, bu sorunlara bir örnektir. Şekilde veri setinin veri noktaları (a), eksik öğrenme modeli (b), fit model (c) ve aşırı öğrenme modeli (d) gösterilmektedir.



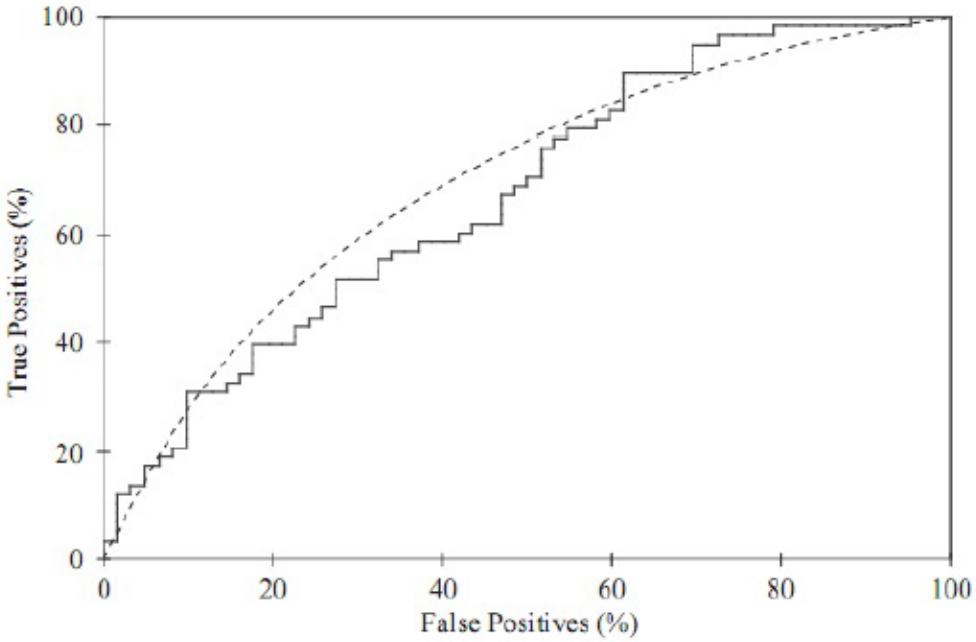
Şekil 4.5 Eksik öğrenme, fit ve aşırı öğrenme modeli [33]

Doğru pozitif oran, churner olarak sınıflandırılan churner sayısı (doğru pozitifler) ile gerçek churner sayısı *(Pozitif = DP + YN)* arasındaki orandır. Yanlış pozitif oran, gerçekte churner olmayan ancak tahminde churn kabul edilen kişi sayısı (yanlış pozitif) ile gerçek churner olmayanların sayısı *(Negatif = YP + DN)* arasındaki orandır. *dprate* ve *yprate* tek başlarına yeterli değildir. Duyarlık ve geri çağırma, dengesiz veri setlerinde kullanılabilecek iki önlemdir. Duyarlık şu soruyu cevaplar: "Churn olarak sınıflandırılan müşteriler arasında, gerçekte kaç tane churner var?". Geri çağırma ise tüm churnerların arasında kaç kişinin doğru bir şekilde tahmin edebileceğini yanıtlar. Duyarlık ve geri çağırma arasında daima bir denge vardır. Yüksek duyarlık düşük geri çağırma ile sonuçlanır. F-ölçüsü hem duyarlığı hem de geri çağırmayı dikkate alır. Bu nedenle, sınıf dengesizliğinden muzdarip olan problemler için daha iyi bir ölçümdür. Churn tahmini alanı ve veri setimiz için en uygun önlem, bir sonraki alt bölümde tartışılacak ROC alanıdır. Telekomünikasyonda yaygın olarak kullanılmaktadır, çünkü bu önlem hem doğru hem de yanlış pozitif oranları dikkate almaktadır. ROC alan skoru sınıf oranlarından bağımsızdır. Bir sonraki bölümde daha detaylı anlatılan bu nedende, çalışmanın performans ölçümü için ROC/AUC ölçütü seçilmiştir.

4.6.1 ROC/AUC

ROC (Receiver Operating Characteristic) çizelgesi iki boyutlu bir çizimdir. ROC eğrileri, sınıf dağılımı veya hata değerleri dikkate alınmaksızın bir sınıflandırıcının performansını gösterir. Dikey eksendeki doğru pozitif oranına karşılık (hassasiyet) yatay eksen üzerinde yanlış pozitif oranı (özgüllük) çizilir [34]. Şekil 4.3 bir ROC

eğrisinin nasıl göründüğünü göstermektedir. Grafikteki her adım eşiği değiştirerek oluşan bir noktadır. ROC eğrisindeki her nokta, belirli bir karar eşiğine karşılık gelen bir hassasiyet / özgüllük çiftini temsil eder. Bu grafiği kullanarak, hassasiyet ve özgüllük arasındaki en uygun denge noktası tespit edilebilir. ROC analizi ayrıca, herhangi bir eşikten bağımsız olarak bir sınıflandırıcının öngörme kabiliyetini değerlendirme şansı sağlar. AUC (area under curve) denilen ROC eğrisinin altındaki alan, çeşitli sınıflandırıcıların doğruluğunu karşılaştırmak için ortak bir önlemdir. ROC, bir yöntemin örnekleri doğru şekilde sınıflandırma yeteneğini değerlendirir. Bu yaklaşıma göre, en büyük AUC'ye sahip sınıflandırıcı daha iyi kabul edilecektir. Bir sınıflandırıcının AUC'si 1'e yakınsa, doğruluğu daha yüksek demektir.



Şekil 4.6 ROC eğrisi örneği [34]

5
Uygulama

5.1 Verinin Tanımlanması

Çalışmada bir telekomünikasyon firmasından alınan bir senelik müşteri verisi kullanılmıştır. Alınan örneklem tamamen maskelenmiş olup abonelerin kimliği çıkarılabilir durumda değildir[1]. Veriler aylık bazda ölçülmüş olup çalışmamızın konusu olan churn de aylık olarak incelenmiştir. Örnek vermek gerekirse bu çalışmada bir müşteri takip eden 30 günlük fatura döneminin herhangi bir anında başka bir operatöre geçmişse çalışmada kullanılan hesaplama yöntemine göre churn kabul edilmektedir. Bu çalışmada hattını tamamen kapatmış veya borcu nedeniyle aboneliği kapatılmış müşteriler dahil edilmemiştir. Dolayısıyla incelediğimiz problem rakibe geçiş yapmakla sınırlandırılmıştır. Literatürdeki ismiyle aktif (gönüllü) müşteri kayıpları incelenmiştir.

Kullandığımız veri her ay için 2 milyon abone içeren örneklemlerden oluşmaktadır. Bu örneklemlerin ana kütleyi temsil ettiğinden emin olmak için rastgele örnekleme (random sampling) kullanılmıştır. Örneklemin ana kütleyi temsil kabiliyeti güven aralığı içerisinde olduğu ANOVA (Analysis of Variance) testi ile kontrol edilmiştir.

Tüm örneklem göz önünde bulundurulduğunda aylık churn oranı %1.4 olup faturalı abonelerin churn oranı önödemeli abonelere kıyasla büyük oranda düşüktür. Faturalı abonelerde ortalama aylık churn %0.9 iken önödemelilerde bu oran %1.8 gözükmektedir. Önödemeli aboneler örneklemin %44'ünü, faturalı aboneler ise %56'sını oluşturmaktadır (Tablo 5.1). Faturalı müşterilerdeki churn oranının önödemelileri göre düşük olmasının sebebi faturalı aboneler ile firma arasında sözleşmesel ilişki bulunmasından ötürüdür. Dolayısıyla faturalı abonenin operatörü terketmesinde ciddi değiştirme giderleri (switching cost) vardır. Ancak dataya aylık olarak bakıldığında aylık bazda churn oranlarının ciddi dalgalanma sergilediklerini

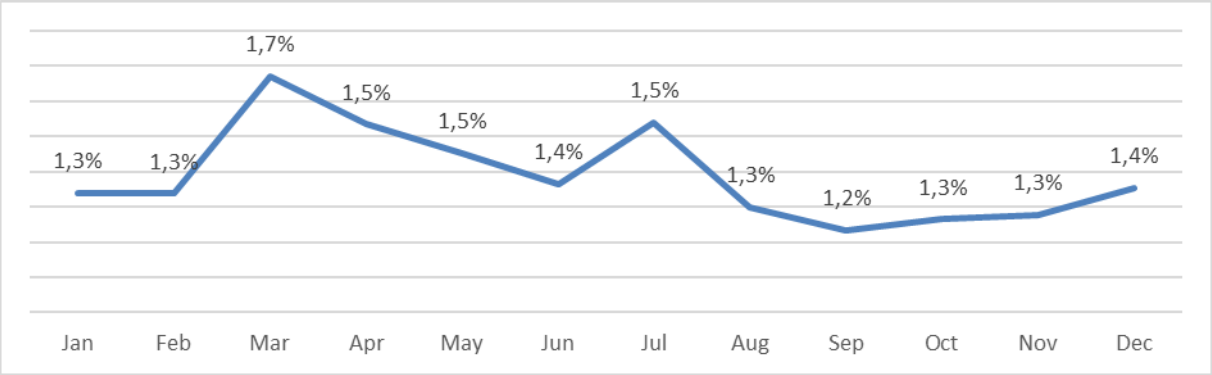
[1] Kişisel verilerin korunması kanunu (KVKK)

görüyoruz (Şekil 5.1). Bu da churn ile ilgili zaman ekseninde yapılacak kestirimleri oldukça güç hale getirmektedir. Fakat bizim analizimiz tabiatı itibariyle bir arakesit analizi olduğu için sezonsal ve zamansal etkilerin modelimizin istikrarında önemli bir sorun yaratmayacağını varsayıyoruz. Bu varsayımın test edilmesi ve modelin zaman boyutunda genişletilmesi bildirinin sonuç kısmında tartışılacaktır.

Tablo 5.3 Örneklem Özellikleri

	Toplam	Ön-ödemeli	Faturalı
Örneklem sayısı	2.000.000	880.000	1.120.000
Müşteri kayıp oranı (churn rate)	%1.4	%1.8	%0.9

Yukarıda belirttiğimiz gibi müşterinin churn davranışı oldukça seyrek görülen bir davranıştır. Bu da literatürde dengesiz veriseti (imbalance dataset) dediğimiz veri kümesindeki dengesiz dağılımı ifade eden bir olguya yol açar. Modelleme perspektifi içerisinden bakıldığında dengesiz verisetlerinin modelin genelleme kabiliyetini azaltıcı yönde etkisi vardır [26]. Dengesiz verisetleriyle mücadele etmenin literatürde bazı yolları bulunmaktadır. Bu yollardan iki tanesi oldukça sık kullanılmaktadır. Bunlardan birincisi baskın gruptaki gözlemleri azaltarak verisetini dengeli hale getirmektedir (stratified sampling). Diğer ikinci yöntem ise belirli istatistiksel yöntemler kullanarak azınlık grubu sentetik bir şekilde çoğaltmak ve veri kaybı olmadan veriyi dengeli hale getirmektir. Bu iki yaklaşımın artılarını ve eksilerini modelleme kısmında detaylı bir şekilde ele alınmıştır. Bu çalışmada daha daha sonra belirtilecek sebepler dolayı dengesiz veriseti ile devam edilmiştir.



Şekil 5.11 Aylık churn oranının bir senelik trendi

Tahminleyici değişkenler:

Operatörden aldığımız müşteri dataları aşağıdaki bilgileri içermektedir:

- Müşterilerin demografik bilgileri: cinsiyet, yaş , hattın aktive edildiği il-ilçe bilgisi, kullandığı cihaz bilgisi.
- Müşterilerin operatör servislerinden faydalanma bilgileri: aylık data tüketim miktarı, aylık konuşma dakikası, aylık SMS kullanımı, katma değerli servisler abonelikleri, digital servisler kullanım miktarları.
- Müşterilerin kullandığı ürün bilgileri: tarife bilgisi, tarife fiyatı, tarife içeriği, tarife indirimi, tarife güncelliği.
- Müşterilerin operatör sözleşmesine dair bilgiler: müşteri taahhüt süresi, taahhüt bozdurma durumundaki ceza miktarı, operatörden taksitli alınmış cihaz bilgisi ve taksit bilgisi.
- Müşterilerin şebeke deneyimine dair bilgiler: müşterinin veri kullanım hızı (throughput), şebeke yoğunluk oranı, çağrı sayısı, çağrı kesinti oranı, müşterinin kapsama dışı geçirdiği süre, müşterinin kullandığı radyo şebeke teknoloji bilgileri.
- Müşterinin fatura ve ödeme bilgileri: müşterinin fatura bilgisi, tarife aşım ücretleri, fatura gecikme ücreti, aylık abonelikler, roaming ücretleri.
- Müşterinin yakın çevresinin özellikleri: müşterinin en çok konuştuğu kişilerin operatör bilgileri, müşterinin yakın çevresinin daha önceki aylarda churn edip etmediği.
- Müşterinin müşteri deneyimi bilgileri: müşteri mağaza ziyaret sayısı, müşterinin açtığı şikayet kayıt sayıları ve bunların sınıfları.

Bu çalışmada pazardaki diğer operatörlerden alınmış veriler olmadığı için rekabete dayalı herhangi bir tahminleyici değişken yer almamaktadır.

5.2 Öznitelik Çıkarımı

İstatistiksel modellemede (yapay öğrenme modelleri de bir çeşit istatistiksel model olarak değerlendirilebilir) çoğu zaman eldeki ham veriler tahmin etmek istediğimiz davranışı isabetli bir şekilde tahmin etmek yeterli değildir. Bu yüzden literatürde de

sıklıkla belirtildiği gibi mevcut ham veriyi kullanarak yaptığımız kestirimin gücünü artıracak başka değişkenlerin türetilmesi, mevcut değişkenlerin çeşitli matematiksel ve istatistiksel yöntemlerle dönüştürülmesi, verideki eksikliklerin yine matematiksel ve istatistiksel yaklaşımlarla tamamlanması (inputation) gerekmektektedir. Tüm bu işlemlere öznitelik çıkarımı (feature engineering) adı verilmektedir. Operatörden alınan veriler oldukça ham olduklarından mevcut halleriyle modellendiklerinde modeller beklenildiği gibi düşük performans gösterdi. Bu yüzden bu veriyi kullanarak çeşitli yeni değişkenler türetilmiştir.

Üretilen değişkenler aşağıdaki şekilde özetlenebilir:

- Müşterinin operatördeki ömrü (tenure)
- Müşteri paketine tanımlı veri kullanım hakkı (data allowance)
- Paket aşım miktarı
- Tarifenin ömrü
- Müşterinin maruz kaldığı şebeke yoğunluk yüzdesi
- Müşterinin veri indirme hızı
- Müşterinin çağrı kesinti orantı
- Müşterinin çağrı kurma başarı oranı
- Müşterinin yakın çevresinin kullandığı operatör bilgisi
- Müşteri kapsama dışında geçirdiği süre
- Müşterinin tarife kontratının bitmesine ne kadar kaldığı kaç gün geçtiği, iptal ettiğinde ödeyeceği ceza bedeli
- Müşterinin sinyalleşmesine göre lokasyon bilgisi
- Müşteri çağrı merkezi aramalarının sınıflandırılması
- Müşterinin daha önce operatör değiştirip değiştirmediği bilgisi
- Müşterinin cihaz kampanyası bilgileri
- Müşterinin telefonunun markası, modeli, piyasaya çıkış tarihi ve fiyatı
- Müşteri teknoloji kırılımlı servis kullanım bilgileri
- Müşterinin operatöre ait akıllı telefon uygulamalarını kullanma sıklığı
- Müşterinin operatöre ait sadakat kampanyalarını ne sıklıkla kullandığı
- Müşterinin promosyon kullanım miktarları

Bütün bu öznitelik çıkarımı sonunda müşterinin churn davranışını tahmin etmek üzere 100'den fazla açıklayıcı değişken elde edilmiştir. Bu süreçte doğruluğu tartışılır olan müşteri beyanına dayanan eğitim durumu, aylık gelir beyanı gibi ve birçok müşteri için mevcut olmayan bilgiler modellemenin daha isabetli olması adına çıkarılmıştır.

5.3 Modelleme ile İlgili Hususlar

Elimizdeki verinin özellikleri ve müşterinin churn davranışı göz önünde bulundurulduğunda modellemeyle ilgili dikkat edilmesi gereken bazı durumlar söz konusudur. Öncelikle müşterinin churn davranışının çok sayıda değişkene bağlı olmasıyla beraber literatürdeki genel görüş bu değişkenlerin birbiriyle etkileşimleri davranışı etkilemekte ve ortaya churn ile onu tahmin etmemizi sağlayacak değişkenler arasında doğrusal olmayan (non-linear) bir ilişki olduğunu ortaya koymaktadır. Dolayısıyla lojistik regresyon gibi değişkenler arasında doğrusal ilişkiler olduğunu varsayan modellerin böyle bir problem için tatmin edici bir performans göstermeyeceği beklenmektedir. Bu düşüncenin doğruluğunu göstermek için lojistik regresyon modelini bir referans model olarak kullanılmış ve çalışmanın bir sonraki bölümünde görüleceği üzere incelenen diğer modellere göre daha düşük performans gösterdiği gözlenmiştir. Performans kriterleri modellerin sonuçlarının açıklandığı bölümde ayrıntılı bir biçimde tartışılmıştır.

Churn modellemedeki tek sorun doğrusal olmama sorunu değildir. Churn çok boyutlu bir problemdir ve bunu tatmin edici şekilde tahmin etmek çok sayıda açıklayıcı değişken gerektirmektedir. Literatürde de sıklıkla belirtildiği gibi özellikle parametrik modellerde (yapısal modeller) değişken sayısı arttıkça modelin performansında overfitting olarak ifade edilen olgudan dolayı (Boyutsallığın overfittinge yol açtığı gözlemlenmiştir) düşüş görülmektedir. Dolayısıyla eğer parametrik bir model seçilmişse modelleme yapılmadan önce yine algoritmik yöntemlere dayanan bir öznitelik seçimi yapılmalıdır. Bu öznitelik seçimi algoritmalarının performansı ve bilimselliği hala tartışılmaktadır ve literatürde uzlaşılmış bir konu değildir. Bütün bunlar dikkate alındığında churn gibi çok fazla boyutsallık içeren ve değişkenler arasında doğrusal olmayan ilişkilerin olduğu

problemlerde parametrik ve doğrusal olmayan model tiplerinin kullanılmasının bir çok defa performans açısından çok daha başarılı olduğu gözlenmiştir. Pratikteki uygulamalar göz önünde bulundurulduğunda bu tarz modellerin endüstri standartı haline geldiği gözlemlenmektedir. Çünkü parametrik olmayan modellerin önemli bir kısmı (karar ağacı tabanlı modeller, komite modellerinin önemli bir kısmı ve Kernell tabanlı modeller bu gruba girmektedir) çok zahmetli ve yöntemi tartışmalı bir süreç olan öznitelik seçimine ihtiyaç duymamakta ve hem açıklayıcı değişkenler arasındaki hem de açıklayıcı değişkenler ve bağımlı değişken arasındaki doğrusal olmayan ilişkileri başarılı bir şekilde modelleyebilmektedir.

Ayrıca bahsedilen modellerin model kalibrasyon süreci (model tuning) diğerlerine kıyasla çok daha az sayıda nüans parametresiyle yapılmaktadır. Bir başka avantaj da bu modellerde uç değer (outliers) tespitinin yapılmasına gerek olmayışıdır. Bu teorik üstünlükler göz önünde bulundurulduğunda bu sınıfa dahil olan modellerin lojistik regresyon, yapay sinir ağları gibi parametrik modellere kıyasla daha yüksek başarı göstermesini bekliyoruz.

Bilindiği üzere modelin örnekleme olan bağımlılığı genelleme performansını düşüren bir unsurdur. Bu durum tekil model sonuçlarının varyansının bazen istemediğimiz seviyelerde yüksek olmasına sebep olmaktadır. Zaten modellemede amaç varyans ve yanlılık (bias) arasında genelleme performansını maksimize eden bir uzlaşı yakalamaktır. Komite (ensemble) modelleri yanlılık açısından bir miktar tavize karşılık model varyansını önemli oranda düşürerek modelin toplam performansını artırmaktadır. Bu teorik sonucun doğruluğu literatürde ve endüstride birçok defa gösterilmiştir. Örneğin CART ve random forest modellerinin her ikisi de karar ağacı tabanlı olmasına rağmen bir komite modeli olan random forest modeli CART modeline göre çoğu zaman daha yüksek performans götermektedir. Bütün bu unsurlar değerlendirildiğinde müşteri churn davranışını tahmin etmek için en uygun modellerin komite modelleri olduğunu düşünüyoruz.

5.4 Modelleme

Bu çalışmamızda churn tahmini için 4 farklı model çalıştırılmıştır. Bu modeller lojistik regresyon, yapay sinir ağları, random forest ve XGBoost modelleridir. Yapay

sinir ağları, random forest ve XGBoost modellerinin sonuçları hem birbirleriyle hem de referans model olarak seçtiğimiz lojistik regresyon modeliyle karşılaştırılmıştır.

Modellerin karşılaştırılabilir olması için modellerin tamamında 9 aylık abone örneklemi öğrenme datası olarak kullanılmıştır. Takip eden 3 aylık abone datası ise test datası olarak kullanılmıştır.

Bu çalışmada algoritmalar Python ortamında çalıştırılmıştır. Veri hazırlığı için SQL ve Pandas kullanılmıştır. Algoritmalar ise Scikit-Learn kütüphanesi kullanılarak çalışıtırılmıştır.

Bu modellerin ayrıntılarına girecek olursak; ilk model olan lojistik regresyon oldukça basit bir model olup bir adet kalibrasyon parametresi vardır. Bu parameter C ile ifade edilip modelin karmaşıklığını kontrol etmektedir. Bu parameter 0 ile $+\infty$ arasında değer alabilmektedir. Büyük C değerleri karmaşıklığı daha yüksek modellere, dolayısıyla overfittinge yol açmaktadır. C'nin küçük değerler alması ise daha az karmaşık bir modele ancak underfittinge sebep olmaktadır. Dolayısıyla, iyi bir modelde C iyi ayarlanmalı ve hem overfitting hem de underfittingden kaçınılmalıdır. Bu çalışmada C için şu değerler denenmiştir: {0.001, 0.01, 0.01, 1, 10, 100, 1000}. En optimal sonuç 10 değerinde alınmıştır.

Yapay sinir ağları modelinin üç temel kalibrasyon parametresi vardır. Bunlardan biri ağırlık sönümlemesi (weight decay) parametresi, diğeri ise gizli katman sayısı ve gizli katmandaki hücre sayısıdır. Fakat literatürde de sıklıkla bahsedildiği gibi gizli katman sayısı parametresi aslında pratik amaçlar için uygun bir parametre değildir. Çünkü bilindiği gibi sadece tek bir gizli katmandan oluşan yapay sinir ağı mimarisi herhangi bir doğrusal olmayan fonksiyonel formu başarıyla modelleyebilmektedir. Bu yüzden bu çalışmada bir gizli katmanlı yapay sinir ağı mimarisi seçilmiştir. Gizli katmandaki hücre sayısı için ise literatürde sıklıkla kullanılan bir başparmak kuralı olan giriş ve çıkış katmanlarındaki hücre sayılarının ortalaması kullanılmıştır. Bu da bu çalışma için 56 değerine karşılık gelmektedir. Yapay sinir ağları modelinde kalibrasyon ağırlık sönümlemesi parametresi üzerinden yapılmıştır. Ağırlık sönümleme parametresi λ ile ifade edilir. Küçük λ değerleri daha karmaşık modellere yani daha keskin karar sınırlarına sebep olur. Daha büyük λ değerleri ise daha az karmaşık ve daha yumuşak hatlı karar sınırlarına

sebep olur. Bir başka deyişle gereğinden küçük λ değerleri overfittinge, gereğinden büyük λ değerleri ise underfittinge sebebibiyet verebilir. Bu çalışmada λ için şu değerler denenmiştir: {0, 0.1, 1, 2, 10}. En optimal sonuç 2 değerinde alınmıştır.

Rastgele orman modellemesinde genelleme performansı üzerinde etkisinin en yüksek olduğu gözlenen kalibrasyon parametrelerinin, modelin oluşturacağı tekil ağaçların sayısı (*m*) ve her bir ağaç için kaç tane açıklayıcı değişkeni rassal olarak seçeceği (*k*) olduğunu belirtmiştik. *k* değerinin gereğinden büyük değerler alması tekil ağaçların birbiriyle olan korelasyonunu artıracak ve underfitting yol açacaktır. *k* değerinin gereğinden küçük değerler alması ise tekil ağaçların birbiriyle olan korelasyonunu düşürecek ve overfittinge sebep olacaktır. Bu çalışmada *k* için şu değerler denenmiştir: {1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. En optimal sonuç 20 değerinde alınmıştır. *m* değeri yani tekil ağaçların sayısı tahminleyicinin varyansını etkileyen bir parametredir. Daha yüksek *m* değerleri daha düşük tahminleyici varyansına yol açarak daha iyi bir genelleme performansına götürür. Ama literatürdeki çalışmalar göstermiştir ki belirli bir eşikten sonra *m* değerini daha fazla artırmanın modelin genelleme performansına marjinal katkısı gittikçe düşmektedir. Dolayısıyla, *m*değerini çok artırmak model performansını daha fazla artırmayacak fakat modelin çalışma süresinin uzamasına sebep olacaktır. Bu sebeplerden ötürü çalışmada *m* değeri literatürde kabul gören değer olan 1000 değerinde sabit tutulmuştur, kalibrasyon *m* parametresi üzerinden yapılmıştır.

XGBoosting algoritmasının en etkin kalibrasyon parametreleri iterasyon sayısı, rassal olarak seçilecek örneklem alt kümesi oranı ve ağaç derinliğidir (interaction depth). XGBoosting algoritmasındaki iterasyon sayısının genelleme performansı üzerindeki etkisi rastgele orman algoritmasındaki ağaç sayısı parametresinin etkisine benzemektedir. Bu sebepten dolayı iterasyon sayısını literatürde kabul gören değer olan 1000'de sabitledik. XGBoosting algoritması her iterasyonda bütün öğrenme datasının kullanıcının belirttiği oran kadarını rassal olarak olarak seçer. Bu parameter rassal olarak seçilecek örneklem alt kümesidir. 0 ile 1 arasında oransal bir değer alır. Bu parametrenin değeri 1'e yaklaştıkça overfitting davranışı gözlenir, 0'a yaklaşırsa tam tersi underfitting davranışı gösterir. Bu çalışmada rassal olarak seçilecek örneklem alt kümesi oranı için şu değerler denenmiştir: {0.1, 0.2, 0.3, 0.4,

0.5, 0.6, 0.7, 0.8, 0.9, 1}. En optimal sonuç 0.5 değerinde alınmıştır. Ağaç derinliği her iterasyonda inşa edilen ağacın kaç basamaktan oluştuğunu ifade etmektedir. Bu değer arttıkça model daha karmaşıklaşmaya başlar ve overfitting oluşur. Bu değer düştükçe underfittinge sebep olur. Bu çalışmada rassal olarak seçilecek örneklem alt kümesi oranı için şu değerler denenmiştir: {5, 7, 9, 11, 13, 15}. En optimal sonuç 9 değerinde alınmıştır.

Daha önceden belirtildiği gibi datada ciddi derecede sınıf dengesizliği (class imbalance) bulunmaktadır. Modelleme yaklaşımımızda bu sınıf dengesizliğini düzeltme yoluna gidilmemiştir. Bilindiği üzere istatistiksel modellerde -ki üzerinde çalıştığımız model ikili sınıflandırma problemidir- model sonuçları belirli bir gözlem şu veya bu sınıfa ait şeklinde değil belirli bir sınıfa ait olma olasılığı şeklinde yorumlanmaktadır. Bu açıdan düşünüldüğünde sınıf dengesizliğini değiştirmek aslında önsel olasılıkları değiştirmek ve modelin sonuçları olan şartlı olasılıkları anlamsız hale getirmektir. Dolayısıyla sınıf dengesizliği düzeltildiği takdirde model sonuçlarında ortaya çıkan olasılıkların gerçekçi bir yorumu bulunmamaktadır. Örneğin, elimizdeki datada aylık churn olasılığı ortalama %1.4'tür. Bu churn'ün önsel olasılığıdır. Model sonucunda herhangi bir müşterinin churn etme ihtimali %10 çıkıyorsa bu yorumlanabilir bir olasılıktır ve müşterinin elimizdeki en iyi bilgiye göre %10 ihtimalle churn edeceğini söylenebilir. Fakat datadaki sınıf düzensizliğini ortadan kaldırdığımız takdirde (bu müşteri kayıp oranının önsel olasılığını %50 olacak şekilde değiştirmek demektir) bahsedilen müşterinin model sonucunda churn etme ihtimali %70 çıkıyorsa bunu gerçek bir olasılık değerlendirmek mümkün değildir. Yani bu müşteri önümüzdeki ay %70 ihtimalle churn edecek demek yanlış bir yorumdur. Gerçek olasılıkları bilmek pratikte önemlidir. Çünkü telekomünikasyon firmalarında churn analizi sadece churnü tahmin etmek için değil müşterinin yaşam döngüsü değerini hesaplamak, gelir riskini hesaplamak için de kullanılır. Bu hesaplamaları yapabilmek için gerçek olasılıklara ihtiyaç duyulmaktadır.

Verideki sınıf dengesizliğini düzeltmediğimiz için model performanslarını ölçmek için hata matrisi analizini (confusion matrix) kullanmadık. Bunun sebebi bu tarz analizler pozitif sınıf lehine karar vermek için model sonucu olarak ortaya çıkan

olasılığın %50'den büyük olmasını gerektirir. Bu durum ise sınıf dengesizliği düzeltilmemiş veri setlerinde çok zor elde edilen bir durumdur. Dolayısıyla dengeli veri setlerini analiz ederken oldukça anlamlı ve yönlendirici çıkan recall ve precision gibi performans kriterleri anlamsız çıkmaktadır. Örneğin, elimizdeki gibi aşırı derecede sınıf dengesizliklerinin olduğu verilerde modelin performansı çok iyi olsa bile recall ölçütü genellikle 0 çıkmakta, tam tersine modelin performansı çok kötü olsa dahi accuracy ölçütü 1'e çok yakın çıkmaktadır. Genellikle dengeli veri setlerinde oldukça anlamlı olan precision ölçütü ise model performansından bağımsız 0 ile 1 arasında rassal bir seyir izlemektedir. Bu ölçütlerin güvensizliğinden dolayı sınıf dengesizliğinin aşırı olduğu veri setlerinde en yaygın ve en güvenilir performans ölçütü olarak ROC/AUC ölçütü kullanılır. Bu sebeple çalışmada ROC/AUC ölçütü kullanılmıştır.

Tablo 5.4 Karşılaştırılan modellerin ROC/AUC performansları

Model	**ROC/AUC**
Lojistik regresyon	0.611
Yapay sinir ağları	0.670
Random Forests	0.754
XGBoost	0.763

Tablo 5.2'de her modelin en başarılı kalibrasyon parametresi setine göre ROC/AUC performansları verilmiştir. Buna göre yapay sinir ağları, random forest, XGBoost modellerinin her üçü de referans model olarak seçtiğimiz lojistik regresyon modelinden daha iyi performans göstermiştir. Bunun sebebi daha önce de belirttiğimiz gibi lojistik regresyon modelinin sadece doğrusal karar sınırları oluşturabilmesi ve parametrik olması sebebiyle çok boyutluluktan zarar görmesidir. Buna karşın yapay sinir ağlarının lojistik regresyona göre daha iyi performans göstermesinin veri seti içerisindeki doğrusal olmayan ilişki desenlerini de modelleyebilmesine bağlayabiliriz. Sonuçlar en iyi performans gösteren iki modeli

random forest ve XGBoost olarak işaret etmektedir. Bunların arasında da en iyi performans XGBoost algoritmasına aittir.

6
Sonuçlar ve Öneriler

Bu çalışmada telekomünikasyon sektöründe churn davranışını tahmin etmek üzere mevcuttaki performansı en yüksek sınıflandırma algoritmalarını karşılaştırılmıştır. Algoritmalar bir telekomünikasyon operatöründen alınan 12 aylık veri üzerinde uygulanmıştır. Alınan verinin örneklem olarak büyüklüğünün böyle bir analize uygunluğu test sonucunda ortaya konulmuştur. Bu çalışma için operatörden alınan ham veriler bir araya getirilip işlenerek churn tahmininde kullanılmak üzere yüzden fazla açıklayıcı değişken üretilmiştir. Açıklayıcı değişkenlerin içeriğinin çalışma açısından ayrıca önemi vardır. Çünkü telekomünikasyon operatörleri için amaç sadece churn edecek müşteriyi tahmin etmek değil aynı zamanda churn nedenlerini anlamak, churn azaltıcı yapısal değişikliklere gitmek ve müşteriye uygun önlemler alabilmektir. Bu açıdan düşünüldüğünde çalışmada öne çıkan bazı açıklayıcı değişkenlerin operatörlere bu amaçlar yolunda da katkı sağlayacağını düşünüyoruz. Daha önce de belirtildiği gibi en optimal kalibrasyon parametreleriyle çalıştırdığımız algoritmaları kıyasladığımızda en başarılı sonuca XGBoost algoritmasının ulaştığı gözlemlenmiştir. Bu modelin bir çıktısı olan göreceli önem matrisine baktığımızda abone churn davranışını anlamada aşağıdaki önemli sonuçlarla karşılaşıyoruz. Modelimizin sonuçlarına göre özellikle faturalı müşterilerde churn edecek müşteriyi ayrıştırmada en önemli değişken müşterinin kontratının bitimine ne kadar süre kaldığıdır. Bilindiği gibi telekomünikasyon operatörleri aboneleriyle mümkün mertebede sözleşme çerçevesinde çalışmayı tercih etmektedirler. Operatörün bu tercihi temelde gelir riskini azaltmak içindir. Çünkü modern ekonomi teorisinde de görüldüğü gibi sözleşmesel ticari ilişkileri terketmenin sözleşmenin tarafları açısından ciddi bir değiştirme maliyeti vardır. Operatör – müşteri ilişkisine bakıldığında bu maliyeti müşterinin sözleşmesini erken sonlandırdığı takdirde belli bir ceza miktarını ödemek durumunda kalması şeklinde ortaya çıkmaktadır. Bu yüzden sözleşmesinin bitimine uzun süre kalan aboneler sözleşme sonu gelen abonelere göre çok daha düşük olasılıkla churn

etmektedirler. Hem faturalı hem ön-ödemeli müşterilerde görülen bir diğer önemli açıklayıcı değişken ise müşterinin operatörün ne kadar süredir abonesi olduğudur (tenure). Müşterileri abonelik sürelerine göre incelediğimizde daha eski abonelerin yeni abonelere göre çok daha düşük olasılıklarla churn ettiğini gözlemlemekteyiz. Bu durum literatürde sağ kalma önyargısı (survival bias) olarak bilinmektedir. Bu olguya göre abonelik süresi ile churn arasındaki ilişki ters yöndedir. Yani churn eğilimi düşük müşteriler churn etmedikleri için operatörde daha uzun yıllar kalma ihtimalleri daha fazladır. Modelin ortaya koyduğu bir başka churn davranışı açıklayıcısı da müşterinin yakın çevresinin hangi operatörü tercih ettiğidir. Bu açıklayıcı değişken kısaca müşterinin sıklıkla görüştüğü 10 yakın kişinin yüzde kaçının rakip operatörlerin abonesi olduğudur. Churn analizi buna göre incelendiğinde en yakınları diğer operatörlerde olan müşterilerin churn etme ihtimalinin oldukça yüksek olduğu görülmektedir. Model müşterinin churn etme ihtimalinin altyapıdan aldığı deneyimin kalitesine göre de değiştiğini göstermektedir. Örneğin modele göre şebeke yoğunluğu, hat düşmesi ve kapsama yüzdesi değişkenlerinin her üçü de önemli değişkenler arasına girmeyi başarmıştır. Bu bilgiler doğrultusunda churn eden müşteriler incelendiğinde düşük kapsama yaşayan müşterilerin, sıklıkla çağrı düşmesi yaşayan müşterilerin ve sık olarak şebeke yoğunluğuyla karşılaşan müşterilerin diğer müşterilere kıyasla daha yüksek ihtimallerle churn ettikleri gözlemlenmektedir. Model telekomünikasyon hizmetlerinin fiyatlanması konusunda da önemli içgörüler sağlamıştır. Özellikle kontratsız ve ön-ödemeli müşterilerde müşterinin mevcut tarifesine yapılan herhangi bir fiyat artırımının müşterinin churn ihtimalini önemli bir ölçüde artırdığı görülmektedir. Benzer şekilde operatörün aynı içeriğe sahip fakat daha pahalı olan eski tarifelerini kullanan müşterilerin churn etme ihtimallerinin yüksek olduğu görülmüştür. Bunun dışında modelde aşım ücreti de önemli bir değişken olarak ortaya çıkmıştır. Dolaşım ücreti, tarife aşım ücreti ve katma değerli servis ücreti gibi beklenmedik faturalarla karşılaşan müşterilerin daha yüksek ihtimallerle churn ettikleri görülmektedir.

Özetle çalışma sonucunda XGBoost algoritmasının telekomünikasyon operatörlerinde abone churn davranışını hem tahmin etmede hem de sebeplerini anlamada oldukça başarılı olduğu görülmektedir. Düşüncemize göre

telekomünikasyon operatörlerindeki abone churn davranışı rakip operatörlerle ilgili daha fazla bilgi ile beslenmediği sürece hala tam olarak anlaşılabilmiş değildir. Bu çalışma tek bir operatörün bilgileri kullanılarak gerçekleştirilmiştir. Diğer operatörlerle ilgili bilgiler modele alınarak geliştirilebilir. Bir başka gelişim alanı da churn davranışının zaman eksenindeki gelişiminin oldukça varyasyon göstermesinden kaynaklanan sorunlara odaklanmaktır. Mevcut çalışmanın konusu olan modeller zaman eksenindeki gelişimi göz ardı etmiş arakesit modelleridir. Dolayısıyla churn davranışının zaman içerisindeki gelişimini açıklayamamaktadırlar. Bunu takip eden çalışmalar sorunun bu boyutu da düşünülerek geliştirilmelidir.

Kaynakça

[1] K. Coussement, S. Lessmann ve G. Verstraeten, “A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry,” Decision Support Systems, vol. 95, pp. 27-36, 2017.

[2] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain ve K. Huang, “Customer churn prediction in the telecommunication sector using a rough set approach,” Neurocomputing, vol. 237, pp. 242-254, 2017.

[3] O. Kaynar, M. F. Tuna, Y. Görmez ve M. A. Deveci, “Makine öğrenmesi yöntemleriyle müşteri kaybı analizi,” C.Ü. İktisadi ve İdari Bilimler Dergisi, 2017.

[4] T. Vafeiadis, K. Diamantaras, G. Sarigiannidis ve K. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” Simulation Modelling Practice and Theory, vol. 55, pp.1-9, 2015.

[5] H. Abbasimehr, M. Setak ve M. Tarokh, “A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction,” The International Arab Journal of Information Technology, vol. 11, no. 6, 2014.

[6] K. Kim, C.-H. Jun ve J. Lee, “Improved churn prediction in telecommunication industry by analyzing a large network,” Expert Syst. Appl., vol. 41, no. 15, pp. 6575–6584, 2014.

[7] A.Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari ve U. Abbasi, “Improved churn prediction in telecommunication industry using datamining techniques,”Applied Soft Computing, vol. 24, pp. 994–1012, 2014.

[8] B. Huang, M. T. Kechadi ve B. Buckley, “Customer Churn Prediction in Telecommunications,” Expert Systems with Applications, vol. 39, no.1, pp. 1414-1425, 2012.

[9] W. Verbeke, K. Dejaeger, D. Martens, J. Hur ve B. Baesens, “New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach", European Journal of Operational Research, vol. 218, pp. 211-229, 2012.

[10] P. Kisioglu ve Y. I. Topcu, “Applying Bayesian Belief Network approach to customer churn analysis: a case study on the telecom industry of Turkey,“ Expert Syst.Appl., vol. 38, pp. 7151–7157, 2011.

[11] C. Wei ve I. Chiu, "Turning telecommunications call details to churn prediction: a data mining approach," Expert Systems with Applications, vol. 23, pp. 103-112, 2002.

[12] N. Kamalraj ve A. Malathi, "A Survey on Churn Prediction Techniques in Communication Sector," International Journal of Computer Applications, 2013.

[13] S. Lindmark, E. Andersson, E. Bohlin ve M. Johansson, "Innovation system dynamics in the Swedish telecom sector." Info, vol. 8, no. 4, pp.49-66, 2006.

[14] O. Ploetner, "NEGBI – Introducing New Systems In The Telecom Market, Journal of Business & Industrial Marketing," vol. 19, no. 5, pp.344-350, 2004.

[15] H. Hükmenoğlu ve H. Alperat, "GSM Sector Review," EgeYatırım Equity Research, Türkiye, (2001).

[16] Turktelekom.com.tr, URL: https://www.turktelekom.com.tr/hakkimizda/sayfalar/kilometre-taslari.aspx (Erişim zamanı; Haziran, 8, 2019).

[17] Bilgi Teknolojileri ve İletişim Kurumu Pazar Verileri 3. Çeyrek Raporu, (2018).

[18] R. Ling ve D. Yen, "Customer Relationship Management: An Analysis Framework and Implementation Strategies," Journal Of Computer Information Systems, vol. 41, no. 3, pp. 82-97, 2001.

[19] G. Roberts-Phelps, Customer Relationship Management: How to Turn a Good Business Into a Great One!, Londra: Hawksmere, 2001.

[20] R. S. Swift, Accelerating Customer Relationships: Using CRM and Relationship Technologies, New Jersey: Prentice Hall Professional, 2001.

[21] I. J. Chen ve K. Popovich, "Understanding customer relationship management (CRM): People, process and technology," Business Process Management Journal, vol. 9, no. 5, pp. 672-688, 2003.

[22] G. S. Linoff ve M.J.A. Berry, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Indianapolis: Wiley Publishing, 2011.

[23] V. Lazarov ve M. Capota, "Churn Prediction," Bus. Anal. Course TUM Comput. Sci., 2007.

[24] E. Oktay, H. Ozer ve M. S. Ozcomak ,"Ataturk Universitesi ogrencilerinin cep telefonu abonelik turunu tercih etmeleriyle ilgili faktorlerin tespiti", EKEV Akademi Dergisi, vol. 10, no. 27, pp. 275-296, 2006.

[25] E. Alpaydın, Introduction to Machine Learning, Cambridge: The MIT Press, 2010.

[26] M. Kuhn ve K. Johnson, Applied Predictive Modeling, New York: Springer, 2013.

[27] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford: Oxford University Press, 1995.

[28] L. Breiman, “Bagging Predictors,” Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

[29] L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.

[30] T. Hastie, R. Tibshirani ve J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, 2nd edition, 2008.

[31] Y. Freund ve R. Schapire, Boosting, Cambridge: The MIT press, 2012.

[32] R. O. Duda, P. E. Hart ve D. G. Stork, Pattern Classification, Wiley, 2nd edition, 1999.

[33] Y. Lee, S. Lee, I. Ivrissimtzis ve H. Seidel, “Overfitting Control for Surface Reconstruction", In Proc. Fourth Eurographics Symposium on Geometry Processing 2006, 231-235, Sardinia, Italy, (2006).

[34] I. H. Witten, E. Frank ve M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques, USA: Elsevier Inc., 3rd Edition, 2011.

Tezden Üretilmiş Yayınlar

İletişim Bilgisi: aysesenyurek@gmail.com

Konferans Bildirileri

1. A. Şenyürek ve S. Alp, "Telekomünikasyon Sektöründe Aboneliklerini İptal Edecek Müşterilerin Yapay Öğrenme İle Tahmin Edilmesi," Akdeniz 1. Uluslararası Multidisipliner Çalışmalar Kongresi Tam Metin Kitabı, 1.Baskı, Yayıncı Sertifika No: 2018/42945, Sayfa No: 329-348, 2019.