T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TELEKOM SEKTÖRÜNDE SOSYAL AĞ ANALİZİ

Ülkü Fatma GÜRSOY

YÜKSEK LİSANS TEZİ

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Danışman

Prof. Dr. Muhammet KURULAY

Eş Danışman

Doç. Dr. Mehmet Sıddık AKTAŞ

Temmuz, 2021

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TELEKOM SEKTÖRÜNDE SOSYAL AĞ ANALİZİ

Ülkü Fatma GÜRSOY tarafından hazırlanan tez çalışması 05/07/2021 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Muhammet KURULAY Yıldız Teknik Üniversitesi Danışman	Doç. Dr. Mehmet Sıddık AKTAŞ Yıldız Teknik Üniversitesi Eş-Danışman

Jüri Üyeleri

Prof. Dr. Muhammet KURULAY, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Banu DİRİ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Elif Segah ÖZTAŞ, Üye
Karamanoğlu Mehmetbey Üniversitesi _____

Danışmanım Prof. Dr. Muhammet KURULAY ve eş danışmanım Doç. Dr. Mehmet Sıddık AKTAŞ sorumluluğunda tarafımca hazırlanan Veri Madenciliği Makine Öğrenmesi Algoritmaları ile Telekom Sektöründe Sosyal Ağ Analizi 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.

Ülkü Fatma GÜRSOY

İmza

Aileme

ve

arkadaşlarıma

TEŞEKKÜR

Bu çalışma boyunca beni yönlendiren, yardım ve katkılarda bulunan değerli hocalarım Prof. Dr. Muhammet KURULAY, Doç. Dr. Mehmet Sıddık AKTAŞ'a sonsuz teşekkürlerimi sunarım.

Bu çalışmayı gerçekleştirmemde bana fırsat sağlayan Intellica'ya da teşekkürü bir borç bilirim.

Ayrıca hayatımın her alanında beni destekleyen ve cesaretlendiren sevgili aileme çok teşekkür ederim.

Ülkü Fatma GÜRSOY

İÇİNDEKİLER

SİMGE LİSTESİ vii

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

TABLO LİSTESİ x

ÖZET xi

ABSTRACT xiii

1 GİRİŞ 1

1.1 Literatür Özeti ... 1

1.1.1 Müşteri Terk Tahmininin Önemi ... 1

1.1.2 Öznitelik Seçimi ve Önemi ... 2

1.2 Tezin Amacı ... 2

1.3 Hipotez ... 3

1.4 Organizasyon Yapısı ... 3

2 TEMEL KAVRAMLAR VE LİTERATÜR ÇALIŞMASI 4

2.1 Sosyal Ağ Analizi ... 4

2.1.1 Sosyal Ağ Analizi Metrikleri ... 5

2.1.2 Arama Kaydı Verileri ve Kullanım Alanları ... 6

2.1.3 Alt Topluluk Bulma Algoritmaları ... 7

2.1.4 Denetimli Öğrenme Algoritmaları ... 9

2.1.5 Görselleştirme Araçları ... 10

3 ARAŞTIRMA PROBLEMİ 12

3.1 Problemin Tanımı ... 12

4 METODOLOJİ 15

4.1 Telekom Öznitelikleri ile Müşteri Terk Tahmin Modeli ... 15

4.1.1 Öznitelik Vektörlerinin Oluşturulması ... 17

4.1.2 Veri Ön İşlemesi ... 20

4.1.3 Özniteliklerin Seçimi ... 20

4.1.4 Eğitim-Test Veri Setlerinin Seçimi ... 20

4.1.5 Müşteri Terk Tahmin Modelinin Oluşturulması ... 20

4.2 SNA Öznitelikleri ile Müşteri Terk Tahmin Modeli ... 21

4.2.1 Özniteliklerin Oluşturulması ... 22

4.2.2 Özniteliklerin Bir Araya Getirilmesi ile Elde Edilen Veri Seti Üzerinde Veri Ön İşleme Çalışmalarının Yapılması 32
4.2.3 Müşteri Terk Etme Davranışı Modelinin Oluşturulması 32
4.2.4 Müşteri Terk Etme Davranışı Tahmini 32
4.3 Tutundurma Modelleri 33
4.3.1 Veri Ön İşlemesi 33
4.3.2 Eğitim-Test Veri Seti Seçimi 33
4.3.3 FP-Growth ve PrefixSpan Modellerinin Oluşturulması 33
5 UYGULAMA DETAYLARI **35**
5.1 Kullanılan Teknolojiler 35
5.2 Verinin Tanımlanması 36
5.3 Değerlendirme Metrikleri 39
5.3.1 Doğruluk (Accuracy) Değeri 39
5.3.2 Kesinlik (Precision) Değeri 40
5.3.3 Duyarlılık (Recall) Değeri 40
5.3.4 F1-Skor Değeri 40
6 DENEY SONUÇLARI VE DEĞERLENDİRME **41**
6.1 Sonuçlar ve Değerlendirme 41
7 SONUÇ VE ÖNERİLER **58**
KAYNAKÇA **60**
TEZDEN ÜRETİLMİŞ YAYINLAR **64**

SİMGE LİSTESİ

C_m	EdMot algoritması tarafından keşfedilen tüm toplulukların kümesi
C_l	Label Propagation algoritması tarafından tespit edilen tüm toplulukların kümesi

KISALTMA LİSTESİ

AUC	Area Under the Curve (Eğrinin Altındaki Alan)
BTS	Base Transreceiver Station (Baz istasyon alıcı-vericisi)
CDR	Call Detail Record (Arama Detay Kaydı)
CRM	Customer Relationship Management (Müşteri İlişkileri Yönetimi)
CSP	Communication Service Providers (İletişim Hizmet Sağlayıcıları)
EdMot	Edge Enhancement Approach for Motif-Aware Community Detection (Motif Duyarlı Topluluk Tespiti için Kenar Geliştirme Yaklaşımı)
FN	False Negative – Yanlış Negatifler
FP	False Positive – Yanlış Pozitifler
KNN	K-Nearest Neighbors (K-En Yakın Komşu)
LR	Logistic Regression (Lojistik Regresyon)
RF	Random Forest (Rastgele Orman)
ROC	Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)
SNA	Social Network Analysis (Sosyal Ağ Analizi)
TN	True Negative – Gerçek Negatifler
TP	True Positive – Gerçek Pozitifler
TTFB	Traditional Telecom Features Based (Geleneksel Telekom Özelliklerine Dayalı)
VAS	Value Added Service (Katma Değer Hizmet Servisi)
XGBoost	Extreme Gradient Boosting (Aşırı Gradyan Artırma)

ŞEKİL LİSTESİ

Şekil 2.1 Ground Truth algoritma yapısı 9
Şekil 4.1 Müşteri terk tahmini iş süreci yazılım mimarisi 15
Şekil 4.2 SNA modeli için önerilen sistem mimarisi 21
Şekil 4.3 Örnek alt grup müşteri terk davranış oranı tespiti 24
Şekil 4.4 Örnek alt grup popüler düğümler arasındaki müşteri terk davranış oranı 25
Şekil 4.5 Örnek alt grup etkileşimde olduğu düğüm sayısı yüksek olanlar arasındaki müşteri terk davranış oranı 26
Şekil 4.6 Örnek alt grup en çok konuşma süresine sahip düğüm için müşteri terk davranışı 27
Şekil 4.7 Örnek alt grup popüler düğümler arasında en çok konuşma süresine sahip düğüm için müşteri terk davranışı 28
Şekil 4.8 Örnek alt grup toplam konuşma süresi yüksek düğümler arasındaki müşteri terk davranış oranı 29
Şekil 4.9 CDR alt topluluk ağ grafiği 31
Şekil 4.10 CDR alt topluluk ağ grafiği genel 31
Şekil 4.11 Örnek alt grup ağ grafiği 32
Şekil 6.1 Karmaşıklık matrisi 43
Şekil 6.2 ROC eğrisi 44
Şekil 6.3 Kaldırma eğrisi 44
Şekil 6.4 Kümülatif kazanç eğrisi 45
Şekil 6.5 Doğruluk sonuç grafiği 46
Şekil 6.6 EdMot modeli karmaşıklık matrisi 48
Şekil 6.7 EdMot modeli ROC eğrisi 48
Şekil 6.8 EdMot modeli kaldırma eğrisi 49
Şekil 6.9 EdMot modeli kümülatif kazanç eğrisi 49
Şekil 6.10 Label Propagation modeli karmaşıklık matrisi 51
Şekil 6.11 Label Propagation modeli ROC eğrisi 52
Şekil 6.12 Label Propagation modeli kaldırma eğrisi 52
Şekil 6.13 Label Propagation modeli kümülatif kazanç eğrisi 53
Şekil 6.14 F1-skor sonuç grafiği 55

TABLO LİSTESİ

Tablo 4.1 Model tarafından seçilen telekom öznitelik kategorileri.....17
Tablo 4.2 SNA tabanlı telekom öznitelikleri.....30
Tablo 5.1 CDR veri seti.....37
Tablo 5.2 EdMot öznitelik veri seti.....38
Tablo 6.1 Geleneksel makine öğrenmesi müşteri terk tahmini test seti sonuçları.....42
Tablo 6.2 EdMot müşteri terk tahmini test seti sonuçları.....47
Tablo 6.3 Label Propagation müşteri terk tahmini test seti sonuçları.....50
Tablo 6.4 Ground Truth müşteri terk tahmini test seti sonuçları.....54
Tablo 6.5 FP-growth birliktelik kuralları.....56
Tablo 6.6 Prefix-Span sıralı birliktelik kuralları.....57

ÖZET

Veri Madenciliği Makine Öğrenmesi Algoritmaları ile Telekom Sektöründe Sosyal Ağ Analizi

Ülkü Fatma GÜRSOY

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Muhammet KURULAY

Eş-Danışman: Doç. Dr. Mehmet Sıddık AKTAŞ

Telekom sektörü içerisinde müşteri terk etme davranış tahmininin yapılması, aktif olarak üzerinde çalışılan bir araştırma konusudur. Müşteri terk etme davranışının başarılı bir şekilde önceden tahmin edilmesi için, genellikle müşterilerin telekom altyapısını kullanma davranışlarından elde edilen öznitelikler kullanılmaktadır. Öte yandan, müşterilerin kendi aralarında yaptıkları iletişim verilerinin oluşturduğu karmaşık ağlar ve bu ağlardan elde edilecek öznitelikler de, müşteri terk etme davranışını etkileyebilecek niteliktedir. Bu araştırma kapsamında, telekom altyapıları üzerinde müşterilerin etkileşimi sonucunda oluşan karmaşık ağlar üzerinde, sosyal ağ analizi (social network analysis - SNA) teknikleri kullanılarak, müşteri terk etme davranışı tahmininde kullanılabilecek öznitelikler önerilmektedir. Önerilen bu öznitelikler hesaplanarak, telekom altyapısına ait özniteliklerle beraber modele dahil edilmiştir. Böylece müşteri terk tahmin modelinin başarısının arttırılması hedeflenmiştir. Bunun yanı sıra, müşteri terk etme davranışını tahmin edebilecek bir veri analizi iş süreci yöntemi önerilmektedir. Önerilen yöntemin bir prototip uygulaması geliştirilmiş ve müşteri

terk etme davranış tahminindeki başarısı, deneysel çalışmalarla karşılaştırılmalı olarak değerlendirilmiştir. Bu çalışma kapsamında, bir telekom sektörü firmasına ait anonimleştirilmiş veri seti kullanılmıştır. Elde edilen sonuçlar, önerilen yöntemin başarılı tahminler yapabildiğini ve kullanılabilir bir yöntem olduğunu ortaya koymaktadır. Geliştirilen prototip uygulamada, müşteri terk etme davranışı tahminindeki performans ve tutundurma modellerinin performansı, doğruluk metriklerine dayalı olarak incelenmiştir. Elde edilen sonuçlar, önerilen iş süreci yazılım platformunun kullanılabilir olduğunu göstermektedir.

Anahtar Kelimeler: Makine öğrenmesi, sosyal ağ analizi, telekomünikasyon, müşteri terk analizi, çağrı detay kaydı

ABSTRACT

Social Network Analysis in Telecom Industry with Data Mining Machine Learning Algorithms

Ülkü Fatma GÜRSOY

Department of Mathematical Engineering

Master of Science Thesis

Advisor: Prof. Dr. Muhammet KURULAY

Co-advisor: Assoc. Prof. Dr. Mehmet Sıddık AKTAŞ

In the telecommunication industry, the prediction of customer churn behavior is a subject of active research. Features derived from customers' use of telecom infrastructure are often used to predict customer churn behavior. On the other hand, the complex networks created by the communication data between the customers and the features to be obtained from these networks can also affect customer churn behavior. Within the scope of this research, features, which are used to predict customer churn behavior by using Social Network Analysis (SNA) techniques on complex networks formed as a result of customer interaction on telecom infrastructures, are proposed. These proposed features are calculated and included in the model together with the features of the telecom infrastructure. Thus, it is aimed to increase the success of the customer abandonment prediction model. In addition to that, a data analysis workflow method that can predict customer churn behavior is suggested. A prototype application of the proposed method was developed and its success in predicting customer churn behavior was evaluated with experimental studies. In this study, an anonymized data set

belonging to a telecom industry firm is used. The results obtained show that the proposed method can make successful predictions and is usable. In the developed prototype application, the performance of customer churn behavior and promotion model prediction was examined based on accuracy metrics. The results show that the proposed business process software platform is available.

Keywords: Machine learning, social network analysis, telecommunication, customer churn analysis, call detail record

1
GİRİŞ

1.1 Literatür Özeti

Dünya çapında 2018 yılına ait yapılan çalışmalar, internet kullanıcı sayısının 4.021 milyar ile dünya nüfusun %53'ünü ifade ettiğini göstermektedir. İnternet kullanıcı sayısında her yıl ortalama %7 oranlarında bir artış görülmektedir. 2018 yılı raporlarına göre; sosyal medya kullanıcı sayısı 3.196 milyar ile nüfusun %42'si olarak tespit edilmiştir. Bu sayı yıllık ortalama %13 civarında artmaktadır. 2018 yılında cep telefonu kullanıcı sayısının ise 5.135 milyar olduğu ve toplam nüfusun %68'ine denk geldiği görülmüştür. Ayrıca yapılan araştırmalar yıllık %4 oranında bir artışın olduğunu göstermektedir. Tüm bu verilere göre; dünya nüfunun çoğunun cep telefonuna sahip ve aktif kullanıcı olduğu söylenebilir [1]. Bu veriler ışığında telekomünikasyon sektörünün dünyada ve Türkiye'de büyümeye devam edeceğini söyleyebiliriz. Türkiye'de ve dünyada faaliyet gösteren kurumların gelişen teknolojiyle birlikte pazardaki rekabet oranının arttığı ve müşterilerini kaybetmemesi içinde veri modellerindeki analitik yapının çok önemli olduğu görülmektedir. Günümüzde telekomünikasyon sektöründe globalde faaliyet gösteren kurumların, ihtiyaç duyduğu alanlarda verinin anlamlandırılması, analiz yapabileceği şekilde raporlanması gerekmektedir. Bunun yanı sıra, mevcut müşterilerini kaybetmemek için neler yapılması gerektiğinin belirlenmesi ve/veya kaybetme olasılığı için öngörüde bulunulması gerekmektedir.

1.1.1 Müşteri Terk Tahmininin Önemi

Telekom operatörleri, değerli müşterilerini korumak ve Müşteri İlişkileri Yönetimini (Customer Relationship Management - CRM) geliştirmek zorunda olduğundan, müşteri kayıp tahmini telekom sektöründe hayati önem taşımaktadır [2]. Tüm sektörlerde sıklıkla kullanılan terk etme davranışı analizi (churn analizi), veri madenciliğinin önemli uygulamalarından biridir. Kuruluşlar, terk etme

davranışı tahmini yaparak, hizmet sağlayıcısını değiştirme olasılığı en yüksek olan müşterileri tahmin edebilir ve müşterilerini elde tutmak için pazarlama stratejileri ile kampanyalar oluşturabilirler [3]. Terk etme davranışı tahmini sadece telekom sektörü için önemli değildir. Müşteri hizmetlerini içeren tüm sektörlerde kullanılmaktadır. Özellikle, telekom sektörü dışında bankacılık sektörü ile ilgili çalışmalarda da oldukça yaygın kullanılmaktadır [4].

1.1.2 Öznitelik Seçimi ve Önemi

Özellik seçimi, bir veri kümesinden ilgili özellikleri seçmek için çok önemli bir adımdır. 2019 yılında yapılan bir çalışmada, telekom sektöründe terk etme davranışı tahmini (churn prediction) yapılırken kullanılan özniteliklerin telekom verisinden elde edilen özniteliklerden oluştuğu görülmektedir [2]. 2020 yılında yapılan başka bir çalışmada, telekom veri setinden elde edilen değişkenler kategoriler halinde değerlendirilmiştir. Bunlar; demografik veriler, şirketteki yaşam döngüsüne ilişkin veriler, her müşterinin finansal gücü hakkında bilgiler ve abonelerin rakip ağların müşterileri ile etkileşimine ilişkin veriler [5] şeklinde özetlenebilir. Geçmişte telekom sektöründe yapılan terk etme davranışı tahmini çalışmaları, genelde telekom sektörü ile ilgili özniteliklere dayalı yapılmıştır. Yapılan çalışmalar incelendiğinde, müşterilerin birbirlerinden bağımsız olduğu ve birbirlerinden etkilenmediği varsayılarak müşteri terk davranışı tahmin modelleri oluşturulduğu görülmektedir.

1.2 Tezin Amacı

Bu çalışma kapsamında, telekom sektörüne yönelik müşteri terk etme davranışı tahmini üzerine bir araştırma yapılmaktadır. Arama Detay Kaydı (Call Detail Record-CDR) verisi ile oluşturulan karmaşık ağlar üzerinden bilgi çıkarımı yapan bir yöntem önerilmektedir. Karmaşık ağlar üzerinde, sosyal ağ analizi teknikleri uygulanarak, elde edilen öznitelikler, müşteri terk etme davranışı tahmininde kullanılmaktadır. Hem sosyal ağ analizi hem de telekom özellikleri kullanarak telekom kullanıcılarının terk etme davranışını tahmin etmek için bir metodoloji öneriyoruz. Böylece müşteri terk etme tahmininde kişilerin birbirlerinden bağımsız olduğunu kabul eden geleneksel makine öğrenmesi modellerinden farklı

olarak, sosyal ağ analizi ile müşterilerin birbirleriyle olan etkileşimleri göz önünde bulundurularak terk etme davranışları analiz edilecektir. Elde edilen sonuçların geleneksel makine öğrenmesi yöntemlerinde tespit edilen başarı oranlarından yüksek olması beklenmektedir.

1.3 Hipotez

Bu çalışma kapsamında, insanların birbirleriyle konuşma sürelerinden oluşan karmaşık ağ üzerinde sosyal ağ analizi teknikleri kullanılarak müşteri terk etme davranış tahminin de faydalı olabilecek öznitelikler çıkartılabileceğini düşünmekteyiz. Varsayımımız, bir alt grup içerisinde etkili bir aktör eğer o alt grubu terk etme davranışı gösterirse, bu davranıştan o ağdaki diğer kişiler etkilenecektir. Biz de bu varsayımları doğrulamak için bir yöntem öneriyoruz.

1.4 Organizasyon Yapısı

Tezin organizasyon yapısı şu şekildedir. Bölüm 1' de literatür özeti, tezin amacı ve hipotezi verilmektedir. Sosyal Ağ Analizi (Social Network Analysis-SNA) metrikleri ile müşteri terk davranışlarının tahmin edilebilmesi için çözülmesi gereken araştırma soruları Bölüm 2' de sunulmaktadır. Bölüm 3' te temel kavramlar ve detaylı literatür çalışması ele alınmaktadır. Metodoloji başlığı altında, önerilen çalışma ve yöntemler, yapılan geliştirmeler Bölüm 4' te anlatılmaktadır. Bölüm 5' te önerilmiş olan metodoloji kullanılarak gerçekleştirilmiş olan uygulamanın detayları yer almaktadır. Uygulama sonuçlarının değerlendirilmesi ve doğruluk testleri Bölüm 6'da yapılmaktadır. Bölüm 7' de elde edilmiş olan sonuçlar ışığında Bölüm 2' de verilen araştırma sorularına cevaplar verilmekte ve gelecekte yapılması planlanan çalışmalar anlatılmaktadır.

2
TEMEL KAVRAMLAR VE LİTERATÜR ÇALIŞMASI

2.1 Sosyal Ağ Analizi

Modern çağda, telekomünikasyon tüm dünyada bilgi yayılımı için temel sektörlerden biri haline gelmiştir [6]. Dijital hizmetlerin büyümesi kaçınılmazdır ve telekom operatörleri (CSP / İletişim Hizmet Sağlayıcıları), bu çağda müşterilere bağlanabilirliği sağlayan en önemli katılımcılardan biridir. 2007-2017 döneminde, CSP abone sayısı ve toplam gelir miktarı önemli ölçüde artış göstermiştir [7].

Bu çalışmada, telefon numaralarının birbirleriyle konuşma ilişkilerini gösteren, düğümlerin müşterileri ve kenarların müşterilerin birbirleriyle iletişim kurma aksiyonları olduğu çizge yapılarını temel alan, sosyal ağ analizi tabanlı analitik modeller geliştirilecektir.

Geleneksel makine öğrenmesi tabanlı veri madenciliği yöntemleri, çok iyi sonuçlar verseler de analize tabi varlıkların birbirlerinden bağımsız olduğunu varsaymaktadır. Eğer eldeki veri, çizge veri yapılarıyla temsil edilebilecek şekilde ise (Örneğin; müşterilerin birbirleriyle ilişkilerini gösterecek nitelikte ise) sosyal ağ analizi ve karmaşık ağlar alanındaki algoritmalarla verileri analiz ederek faydalı bilgiler çıkartılabilir, böylece geleneksel makine öğrenmesi algoritmalarının başarımları arttırılabilir. Ayrıca sosyal ağ analizi yöntemleriyle varlıklar arası ilişkilere odaklanarak geleneksel makine öğrenmesi algoritmalarıyla yapılamayacak gelişmiş analizler yapılabilmektedir.

Sosyal ağ analizi, hava durumu tahmini ve sosyal hesaplama dahil olmak üzere farklı alanlarda yanlış bilgi tespiti [8], gizlilik politikası tespiti [9], veri kökü oluşturma [10] ve güven hesaplaması [11] gibi çeşitli amaçlar için yararlı bilgileri elde edebilmek adına grafik verilerini analiz eden bazı çalışmalardan ortaya çıkmaktadır. Bu önceki çalışmalardan farklı olarak, bu çalışma kapsamında grafik verileri analiz edilmekte ve topluluk tespiti için bilgi çıkartılmaktadır. Sistemlere

odaklanan, web hizmetleri hakkındaki durum bilgisini veya sosyal ağlarda yayınlanan verilerle ilgili provenans (veri parçasının kaynağı) bilgisini yöneten ve analiz eden çalışmalar bulunmaktadır [12-13]. Bu çalışmalardan farklı olarak, telekom şirketi ile kullanıcıları arasındaki etkileşim verilerinin yönetimini kapsam dışı bırakmaktayız. Bu çalışmada, kayıp analizi için yararlı bilgiler elde edebilmek için yalnızca grafik verilerini analiz etmeye odaklanıyoruz.

2.1.1 Sosyal Ağ Analizi Metrikleri

Köprü analizi ve Web'de arama ile yakından ilişkili olan iki tür sosyal ağ analizi olan merkezi ve prestij metrikleri üzerinde durulmuştur. Github'dan sağlanan, Derece Merkezilik (Degree Centrality), Yakınlık Merkezilik (Closeness Centrality), Öz Vektör Merkezilik (Eigenvector Centrality) metriklerini veren algoritmalar python kodlarına uyarlanarak metrik sonuçlarına ulaşılmıştır.

2.1.1.1 Derece Merkezilik Ölçütü (Degree Centrality)

Derece Merkezilik metriğini Python'ın NetworkX paketindeki degree_centrality sınıfını kullanarak gerçekledik.

Derece Merkezilik Ölçütü bir düğümün iletişimde olduğu komşularına hangi ölçüde bağlı olduğunu ifade eden temel ölçütlerden biridir. Freeman 1979 yılında merkezilik ölçütünü bir düğümü (node) diğer düğümlere bağlayan kenar (edge) sayısı şeklinde belirtmiştir. Sosyal ağ analizi dikkate alınırsa, burada daha çok kişi ile iletişime geçen müşterilerin, merkezilik değerleri de daha yüksek olarak tespit edilecektir. Derece Merkezilik ölçütü büyük olan düğümler, ağ içerisindeki diğer üyelere göre merkezi konumda olan önemli düğüm olarak nitelendirilirler. Genellikle sosyal ağ analizlerinde kullanıcının iletişim halinde olduğu bağlantı sayısı ne kadar çoksa o kadar güçlüdür ve en aktif kullanıcı olarak adlandırılmaktadır [14].

2.1.1.2 Yakınlık Merkezilik Ölçütü (Closeness Centrality)

Uzaklık kavramına dayalı olan bir diğer ölçüt ise Yakınlık Merkeziliği (Closeness Centrality) metriğidir. Bu metrik, bir kullanıcın ağda bulunan diğer kullanıcılara olan mesafesini incelemektedir. Buna göre, birden çok düğüme yakınlık gösteren

bir düğüm bağımsız şekilde hareket etme konusunda sorun yaşıyor olabilir. Freeman'e göre, ağdaki düğümün bağımsızlığı yakınlık ölçütü ile belirlenebilmektedir. Uzaklık, düğümler arasındaki en kısa mesefanın uzunluğunu (kenar sayısını) vermektedir. Önem değeri yüksek olan düğümlerin diğer düğümlere yakın olacağı varsayılmaktadır [14].

Yakınlık Merkeziliği ölçütünü Python'ın NetworkX paketindeki closeness_centrality sınıfını kullanarak gerçekledik.

2.1.1.3 Özvektör Merkezilik Ölçütü (Eigenvector Centrality)

Özvektör Merkezilik metriği, yalnızca bir kullanıcıya ait olan merkeziliği değil, kullanıcının iletişim kurduğu tüm kullanıcıların merkeziliğini de bulan bir ölçüttür.

Bu ölçüt, düğümün ağ üzerindeki önem ve etkisini ifade etmektedir. Bir düğümün önemli düğümler ile olan bağlantısının, önemsiz kabul edilen diğer bağlantılarından çok daha fazla etkiye sahip olacağı savunulmaktadır. İletişim kurulan kişilerin merkezi bir konumda bulunması, o kişinin de merkezi konuma yükseleceğini göstermektedir. Düğüme ait komşulukların merkezilik değerleri toplanarak hesaplanmaktadır [14].

Diğer merkezilik metrikleri gibi, Özvektör Merkezilik metriği de Python dilinde NetworkX paketi kullanılarak uygulanmıştır.

2.1.2 Arama Kaydı Verileri ve Kullanım Alanları

Arama Detay Kaydı (Call Detail Records - CDR) verileri, bir abonenin telefon kullanımıyla alakalı bilgilerini içermektedir. Her kayıt, aramayı başlatan abonenin kimliğini, aramaya karşılık gelen abonenin kimliğini, aramanın yönünü, abonenin bağlı olduğu baz istasyonunun alıcı-verici (BTS) kimliğini, zaman damgasını ve aramanın süresini içerir. CDR verileri, kullanıcının davranışını analiz etmek için kullanılabilir [15]. CDR verileri kullanılarak, karmaşık ağ oluşturulabilir. Burada, karmaşık ağ üzerindeki bir düğüm (node) müşteriyi temsil ederken, düğümler arasındaki ilişkiler ise müşteriler arasındaki konuşma verilerini temsil etmektedir. Karmaşık ağ üzerinde, sosyal ağ analizi (SNA) teknikleri kullanılarak, alt gruplar

tespit edilebilir. Tespit edilen alt gruplar üzerinden bilgi çıkarımları yapılarak, farklı uygulamalar geliştirilebilir. Çizge veri yapılarıyla temsil edilebilecek şekilde olan varlıklar (Örneğin; müşterilerin birbirleriyle ilişkilerini gösterecek nitelikte ise) sosyal ağ analizi ve karmaşık ağlar alanındaki algoritmalarla, analiz edilebilir. Bu şekilde, faydalı bilgiler çıkartılabilirse, geleneksel makine öğrenmesi algoritmalarının, veriden anlam çıkarımı konusundaki başarımları arttırılabilir. Ayrıca sosyal ağ analizi yöntemleriyle varlıklar arası ilişkilere odaklanılarak, geleneksel makine öğrenmesi algoritmalarıyla yapılamayacak gelişmiş analizler yapılabilir.

Yapılan çalışmalarda CDR verisinin, veri kümelerinde dolandırıcılık tespiti için kullanıldığı tespit edilmiştir [16]. Son zamanlarda, cep telefonu arama detay kaydı verileri (CDR), insan nüfusu dağılımlarını tahmin etmek için de kullanılmıştır [17]. Ayrıca, ceza soruşturmalarında kullanılan en popüler bilgi kaynaklarından birisidir. Polis memurlarının, anahtar aktörleri ve aralarındaki ilişkileri hızlı bir şekilde belirlemelerine olanak tanımaktadır [18]. Başka bir çalışmada ise, anormallik tespiti ve telefon trafik tahmini için CDR verilerinin kullanıldığı görülmektedir [15]. Bu çalışmada, önceki çalışmalardan farklı olarak Sosyal Ağ Analizini, müşteri terk etme davranışı analizi için kullanmaktayız.

2.1.3 Alt Topluluk Bulma Algoritmaları

Topluluk tespit teknikleri, çeşitli amaçlar için yaygın olarak kullanılmaktadır. Bunlar arasında SNA yaygın bir uygulamadır. Örtüşen toplulukları belirlemek için sosyal tabanlı meta-sezgisel optimizasyon algoritmaları kullanılmıştır [19-20]. Dahası, tanınmış topluluklar diğer operasyonların performanslarını iyileştirmekle ilgilenebilirler. Örneğin, topluluk keşfi, en iyi bilgi yayıcıların belirlenmesinde düşük doğruluk dereceli algoritmaları artırmak için kullanılır [21].

SNA ve diğer uygulamalarda, görselleştirme araçları kendi bireysel varyasyonları ile yapılmıştır. Hiyerarşik topluluk yapısı ve düğümleri görselleştirmesi gerçekleştirilirken [22], görselleştirmeden önce kümeleme için Newman'ın modülerlik optimizasyon algoritmasının bazı nüanslarını kullanan araçlar da vardır [23-24]. CDR verileri, çeşitli modellerin analizinde etkili olmuştur ve

uygulaması son derece artmaktadır. Bu kullanımlar arasında kentsel algılama ve planlama [25-26], trafik mühendisliği [27-28], enerji tüketimini tahmin etme [29], hem CDR verilerini kullanarak gelişmiş kayıp tahmini hem de topluluk algılama [30] sayılabilir.

Provenans (veri parçasının kaynağı) alanı gibi farklı araştırma alanlarında grafik veri analizi incelenmiştir [31-35]. Provenans grafiği esas olarak veri kökenini anlamak için kullanılır. Kaynak araştırmasında yapılan önceki çalışmalardan farklı olarak, bu çalışmada, alt ağları tanımlamak için grafik verilerini analiz etmeye odaklanmaktayız.

CDR veri seti üzerinde grup tespit algoritmaları ile alt grupları tespit edebilmek için üç algoritma kullanılmıştır. Bu algoritmalar aşağıda maddeler halinde açıklanmaktadır.

2.1.3.1 Label Propagation Algoritması

Bu algoritma; ağdaki her düğüm için, topluluk üyeliğini belirten bir etiket eklemektedir. Ağdaki bir düğüm, belirli bir etikete sahip maksimum komşu sayısına göre bir topluluğa katılır. Bu nedenle, etiket ağda hızla yayılır ve sonunda her düğüme bir etiket atanır. Sonuç olarak, aynı etikete sahip düğümler bir topluluğun üyesi olarak birlikte kümelenir. Bu algoritmanın zaman karmaşıklığı $O(m+n)$'dir; burada n ve m, düğüm ve kenar sayısıdır [36].

2.1.3.2 EdMot (Edge Enhancement Approach for Motif-Aware Community Detection) Algoritması

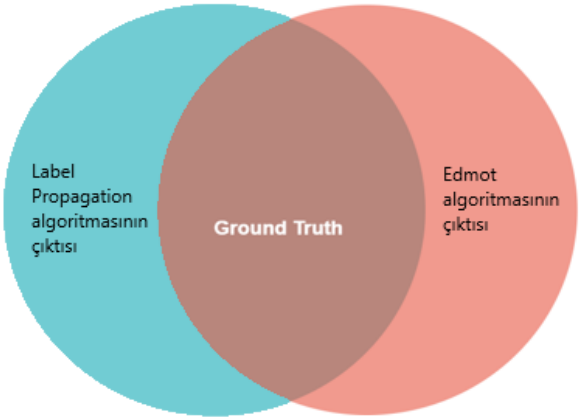
Bu yöntem temelde diğer motif temelli yaklaşımların ihmal ettiği ölçü iyileştirmeleriyle birlikte çalışmaktadır. Yani, izole düğümlerden kaynaklanan parçalanma problemi şeklinde özetlenebilir. Bu algoritma, bölümlenmiş her modülden bir grup türetip orijinal ağı yeniden bağlayarak bir çözüm getirmektedir. Bu çözümün, üst düzey topluluk tespitinin kalitesini artırdığı gösterilmiştir [37].

2.1.3.3 Ground Truth Algoritması

Label Propagation ve EdMot algoritmalarında ortak olan toplulukları bulan algoritmadır. Bu algoritmada, Label Propagation ve EdMot algoritmaları, CDR verisi üzerinde çalıştırılmış ve iki algoritma için de alt gruplar tespit edilmiştir. Elde edilen alt gruplardan, her iki algoritmanın da kesişimi olan ortak alt gruplar, sonuç olarak üretilmiştir [38]. Ground Truth algoritmasında, farklı alt topluluk algılama algoritmaları arasındaki anlaşmayı tanımlamamıza izin veren ([39] 'de kullanılan) bir doğruluk ölçütü kullanılmaktadır.

$$Anlaşma\ (C_l, C_m) = \frac{|(C_l \cap C_m)|}{|(C_l \cup C_m)| - |(C_l \cap C_m)|} * \ 100 \qquad (2.1)$$

(2.1) eşitliğinde görüldüğü gibi C_l Label Propagation tarafından tespit edilen tüm toplulukların kümesini, C_m ise EdMot tarafından keşfedilen tüm toplulukların kümesini göstermektedir. Burada bu iki yöntemin alt grupları içerisinde eşik değeri olarak en az 10 düğümden oluşan topluluklara bakılmıştır ve benzer olan topluluklar belirlenmiştir. Sonuç olarak elimizde üç farklı yöntem ile bulunmuş alt gruplar elde edilmiş oldu. Ground Truth yönteminin çalışma mantığı Şekil 3.1'de gösterilmektedir.



Şekil 2.1 Ground Truth algoritma yapısı

2.1.4 Denetimli Öğrenme Algoritmaları

Müşteri terk etme davranışı tahmini üzerinde yapılan bir çalışmada, farklı makine öğrenme modellerinin kullanıldığını görmekteyiz. Bu çalışmada, çok fazla

öznitelik kullanılarak, model tarafından öğrenilecek parametrelerin sayısı artırılmıştır. Modellerin veriden aşırı öğrenmesi (overfitting) durumunun oluşup oluşmadığını analiz etmek için iki farklı makine öğrenmesi algoritması kullanılmıştır: Lojistik Regresyon (LR) ve Rastgele Orman (RF) algoritmaları [40]. Bir başka çalışmada, terk etme davranışı tahmini için AdaBoost, Extra Trees, KNN (K-En Yakın Komşu), Yapay Sinir Ağları (Neural Network) ve XGBoost (Extreme Gradient Boosting - Aşırı Gradyan Artırma) gibi çeşitli modeller kullanılmıştır. Hassasiyet ve geri çağırma açısından bu modeller kendi arasında karşılaştırıldığında, XGBoost modelinin en yüksek performansı verdiği görülmüştür [41]. Bu araştırma kapsamında, SNA metrikleri kullanılarak elde edilen özniteliklerin, müşteri terk etme davranışı tahminindeki başarısı irdelenmektedir. Araştırmamız kapsamında, makine öğrenmesi modeli olarak, XGBoost algoritması kullanılmaktadır.

2.1.4.1 XGBoost Algoritması

XGBoost (Extreme Gradient Boosting - Aşırı Gradyan Artırma) algoritması, çok sayıda ağaç oluşturmak için ağaçları birleştiren ve daha yüksek tahmin için sonuçların ortalamasını alan, denetimli ve toplu bir öğrenme yöntemidir. Ağaçlar genellikle birbiri ardına inşa edilir. Burada hatalar önceki ağaçta düzeltilebilir ve sonraki ağaçta tahmin doğruluğunun iyileştirilmesine odaklanılabilir [42].

2.1.5 Görselleştirme Araçları

Mevcut verimizden ağ modülünü oluşturmak için görselleştirme araçları kullanılmıştır. D3.js, Javascript veri görselleştirme kütüphanesidir. Javascript kullanarak veri görselleştirme işlemlerini gerçekleştirebileceğimiz bir kütüphanedir.

2.1.5.1 Cytoscape

Cytoscape, moleküler etkileşim ağlarını ve biyolojik yolları görselleştirmek ve bu ağları ek açıklamalar, gen ekspresyon profilleri ve diğer durum verileriyle entegre etmek için açık kaynaklı bir yazılım platformudur. Cytoscape başlangıçta biyolojik araştırma için tasarlanmış olmasına rağmen, şimdilerde karmaşık ağ analizi ve

görselleştirme için genel bir platformdur. Çekirdek dağıtımı, veri entegrasyonu, analizi ve görselleştirmesi için temel özellikler sunar. Projemizde Cytoscape ve D3.js Javascript kütüphanelerini kullanarak arayüz ve görselleştirme işlemlerinin yapılması planlanmaktadır.

3
ARAŞTIRMA PROBLEMİ

3.1 Problemin Tanımı

Telekom sektöründeki şirketler, müşterilerini kaybetmemek için terk davranış tahmini yapma zorunluluğundadır. Müşteri terk davranış analizi için önerileri, makine öğrenmesine dayalı tahmin yöntemleriyle geliştirmek, telekom alanında öncelikli bir hedef haline gelmiştir. Bu kapsamda, müşterilerin terk davranış eğilimlerini tahmin edecek makine öğrenmesi yöntemlerine dayalı veri analizi iş süreçlerine ihtiyaç duyulmaktadır. Müşterilerin terk davranış eğiliminin yüksek performanslı ölçülebilmesi için kullanılacak olan öznitelikler oldukça önemlidir.

Müşteri terk davranışının başarılı bir şekilde önceden tahmin edilmesi için, genellikle müşterilerin telekom altyapısını kullanma davranışlarından elde edilen öznitelikler kullanılmaktadır. Ancak, müşterilerin kendi aralarında oluşturdukları iletişim verilerine ait karmaşık ağlar ve bu ağlardan elde edilecek öznitelikler de, müşteri terk etme davranışını etkileyebilecek niteliktedir. Bu nedenle SNA metrikleri ile müşteri terk etme davranışının sadece geleneksel telekom özniteliklerine bağlı olarak değil, SNA yöntemleriyle tespit edilen öznitelikler ile pekiştirilerek tahmin edilmesine çalışılmıştır.

Bunun yanı sıra terk etme potansiyeline sahip olan müşterileri şirket içinde tutabilmek adına ne tür hizmetlerin sunulması gerektiğinin analiz edilmesi de büyük önem taşımaktadır. Bu gereksinimlere yanıt verebilmek için, bu çalışmada belirlediğimiz araştırma sorularını aşağıda maddeler halinde belirtiyoruz.

1) **Hangi öznitelikleri kullanarak müşterilerin terk etme davranışını belirleyeceğiz?**

Müşteri terk tahmini yaparken modele verilmesi en uygun olan özniteliklerin belirlenmesi oldukça önemli bir adımdır. Doğru tahmin oranının arttırılabilmesi

için telekom verisinden gelen hangi niteliklerin seçilmesi gerektiği ve bunlara ilave olarak sosyal ağ analizi yöntemleri kullanılarak belirlenecek olan özniteliklerin neler olabileceği en iyi şekilde tespit edilmelidir.

2) **Müşteri terk etme davranışını tespit edebilmek için nasıl bir veri analizi iş süreci yapımız olmalı?**

İş sürecinin kaç adımdan oluşacağı ve bu adımların neler olduğunun belirlenmesi gerekmektedir. Adımların hangi sıra ile gerçekleştirilmesi gerektiği tespit edilmektedir. Girdi olarak nasıl bir veri alacağı ve nasıl bir veriyi çıktı olarak vereceği belirlenmektedir.

3) **Bu öznitelikleri kullanarak yaptığımız çalışmada acaba ne oranda başarılı tahminler yapıyoruz?**

Özniteliklerin seçilmesi modelin tahmin başarısında oldukça etkili olan önemli bir adımdır. Telekom verilerden gelen öznitelikler ile yapılan tahminde elde ettiğimiz başarının, kendi belirlemiş olduğumuz Sosyal Ağ Analizi tabanlı öznitelikleri ekleyerek arttırılması planlanmaktadır.

4) **SNA metriklerini kullanarak nasıl müşteri terk etme davranışını tespit edebiliriz?**

Geleneksel makine öğrenmesi modellerinde müşterilerin birbirlerinden bağımsız bir şekilde ele alındığı bilinmektedir. Bu çalışmada bunun aksine SNA alt grup tespit algoritmaları ile müşterilerin birbirleriyle olan ilişkilerini ele alarak alt gruplar oluşturmaktayız. Oluşan bu alt gruplar içerisinde tespit edilen öznitelikler çalıştırılmakta ve her alt grup için bir sonuç elde edilmektedir. Böylece elde edilen veri seti tahmin için modele verildiğinde yüksek başarı oranları beklenmektedir.

5) **Hangi SNA metrikleri kullanılabilir?**

Hangi metriklerin kullanılması gerektiğini belirleyebilmek adına çeşitli çalışmalar yapılmıştır. Öncelikle literatürde yer alan SNA metriklerinden Merkezilik, Yakınlık ve Öz Vektör Metrikleri hesaplanmıştır. Bunun üzerine metriklerin çalışma performansları da göz önünde bulundurularak hangi metrik/metriklerin model başarısını doğrudan etkileyebileceği belirlenmiştir. Ayrıca SNA tabanlı öznitelikler

de, müşterilerin ilişkileri göz önüne alınarak yine model başarısını arttırmaya yönelik belirlenmeye çalışılmıştır.

6) **SNA metriklerini tespit edebilmek için alt topluluk tespit algoritmaları nelerdir ve nasıl çalıştırılır?**

Büyük bir ağ halinde olan verinin analiz yapılabilmesi için alt gruplara ayrılması gerekmektedir. Bu adımda üç farklı algoritma ile gruplar belirlenmektedir. Bu algoritmalar python kodlarına uyarlanarak veri seti üzerinde ayrı ayrı çalıştırılmaktadır.

7) **Terk etme potansiyeline sahip olan müşterileri tutundurmak için hangi kampanya önerilerini sunmalıyız?**

Tahmin modelinden elde edilen sonuçlar ışığında terk etme potansiyeline sahip olan müşterilere çeşitli kampanya önerilerinde bulunarak onları sistemde tutmak oldukça önemlidir. Bu sebeple müşterilerin paket alım hareketlerini içeren ayrı bir veri seti üzerinden çapraz satış analizi yapılmaktadır. Bu sayede belli bir oranda birlikte alınmış olan paketler belirlenip kural olarak adlandırılmaktadır. Bu kurallar bize terk etme olasılığı yüksek olan müşterilere en çok satın alınan ürünleri önerme imkanı verecektir.

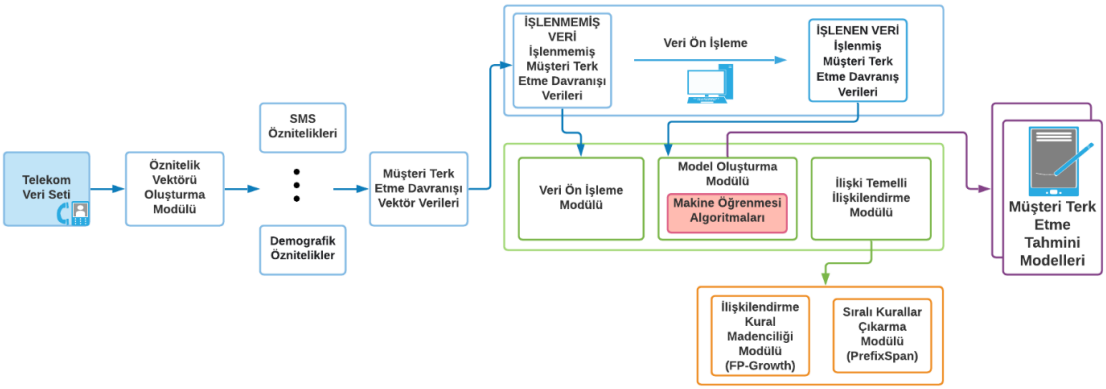
8) **Müşteriye kampanyayı hangi aksiyon zincirini takip ederek sunmalıyız?**

Çapraz satış analizi ile elde edilen kurallar bir arada alınmış olan ürünleri sırasız bir şekilde göstermektedir. Sıralı ve bir çok müşteri tarafından satın alınan ürünlerin belirlenmesi için PrefixSpan Sıralı Desen Bulma Algoritmasının kullanılması hedeflenmektedir. Bu algoritmadan elde edilecek aksiyon zincirleri ile müşterilerin almış olduğu ürünlere bakılarak, en sık tercih edilen sıralı ürün öneri olarak sunulabilecektir.

4
METODOLOJİ

4.1 Telekom Öznitelikleri ile Müşteri Terk Tahmin Modeli

Bu araştırma kapsamında, telekom sektöründeki müşterilerin davranışlarını analiz ederek, yeni gelen müşterilerin terk etme davranışlarını tespit etmek için bir çalışma yapılmıştır.



Şekil 4.1 Müşteri terk tahmini iş süreci yazılım mimarisi

Şekil 4.1'de gösterilen yazılım mimarisindeki modüllerde hangi işlemlerin yapıldığı aşağıdaki şekilde özetlenmiştir:

1) Öznitelik Vektörü Oluşturma Modülü (Feature Vector Generation Module): Bu modül kapsamında, telekom datasından elde edilen öznitelikler içerisinden gini skoru en yüksek olan öznitelikler seçilmektedir.

2) Veri Ön İşleme Modülü (Data Pre-Processing Module): Bu modülde, her model için ayrı veri temizleme, veri entegrasyonu, veri dönüştürme, veri azaltma ve veri normalizasyonu işlemleri yapılmaktadır.

3) Model Oluşturma Modülü (Model Building Module): Bu modül, müşteri terk tahmin modelinin oluşturulduğu modüldür. Makine öğrenmesi algoritmalarından XGBoost kullanılarak test veri seti üzerinde tahmin yapılması sağlanmaktadır.

4) İlişki Temelli İlişkilendirme Modülü (Relationship-Based Association Module): Veri seti üzerinden ilişki tabanlı kurallar çıkartılması bu modül kapsamında gerçekleştirilmektedir. Bu modül iki alt modülden oluşmaktadır.

5) İlişkilendirme Kural Madenciliği Modülü (Association Rule Mining Module (FP-Growth)): FP-Growth (Frequent Pattern Growth) algoritması kullanılarak, müşterilerin ürün alım işlemlerinden anlamlı kurallar elde edilmeye çalışılmıştır. Bu kurallar sonucunda terk etme davranışına sahip olan müşterilere hangi kampanya önerilerinde bulunulması gerektiği belirlenmiştir.

6) Sıralı Kurallar Çıkarma Modülü (Sequential Rules Extraction Module (PrefixSpan)): PrefixSpan algoritması kullanılarak müşterilerin sıralı şekilde aldığı paketler incelendi ve sıralı davranış kuralları tespit edilmeye çalışılmıştır. Bu sayede, terk etme potansiyeline sahip olan müşterilere sunulacak olan kampanyaların hangi sıra ile sunulması gerektiği analiz edilmiştir.

4.1.1 Öznitelik Vektörlerinin Oluşturulması

Tablo 4.1 Model tarafından seçilen telekom öznitelik kategorileri

Ana Kategoriler	Adet	Birinci Seviye Alt Kategoriler	Adet	İkinci Seviye Alt Kategoriler	Adet
SMS özellikleri	40	Giden SMS özellikleri	30	Şebeke içi SMS özellikleri	10
				Şebeke dışı SMS özellikleri	9
		Gelen SMS özellikleri	16	Şebeke içi SMS özellikleri	4
Arama özellikleri	21	Giden arama özellikleri	29	Şebeke içi arama özellikleri	10
				Şebeke dışı arama özellikleri	4
		Gelen arama özellikleri	26	Şebeke içi arama özellikleri	3
Fatura özellikleri	20	Etkinlik özellikleri	22		
		Süre özellikleri	13		
İnternet kullanım özellikleri	7				
Katma değerli hizmet özellikleri	7				

Tablo 4.2 Model tarafından seçilen telekom öznitelik kategorileri (devamı)

Gelir istatistiği özellikleri	4				
Kampanya istatistik özellikleri	8				
Dolaşım özellikleri	12				
Demografik özellikler	1				

Telekom datasına ait olan 600 adet öznitelik içerisinden en iyi sonucu veren 296 metriği seçerek bu çalışmaya katkıda bulunduk. Bu metriklerin tanımları aşağıda, kategorize edilmiş halleri Tablo 4.1' de verilmektedir.

SMS bilgilerini içeren öznitelikler: Müşterilerin mesaj ile alakalı olan tüm özniteliklerini ifade etmektedir.

Giden SMS bilgilerini içeren öznitelikler: Operatörde tanımlı müşterilerin gönderilen mesaj bilgilerini içermektedir.

Şebeke İçi (On-Net) SMS öznitelikleri: Operatörde tanımlı müşterilerin yine aynı operatörde tanımlı olan başka bir müşteriye gönderdiği mesaj bilgilerini içermektedir.

Şebeke Dışı (Off-Net) SMS öznitelikleri: Operatörde tanımlı müşterinin farklı operatörde tanımlı olan kullanıcıya gönderdiği mesaj bilgilerini içermektedir.

Gelen SMS bilgilerini içeren öznitelikler: Operatörde tanımlı müşteriye gelen mesaj bilgilerini içeren özniteliklerden oluşmaktadır.

Şebeke İçi (On-Net) SMS öznitelikleri: Operatörde tanımlı müşteriye yine aynı operatörde tanımlı olan başka bir müşteriden gelen mesaj bilgilerini içeren özniteliklerdir.

Arama bilgilerini içeren öznitelikler: Müşterilerin arama bilgilerini içeren öznitelikler ele alınmıştır.

Giden arama bilgilerini içeren öznitelikler: Operatörde tanımlı müşterilerin başlatmış olduğu arama bilgilerini içeren özniteliklerden oluşmaktadır.

Şebeke İçi (On-Net) arama öznitelikleri: Operatörde tanımlı müşterilerin yine aynı operatörde tanımlı olan başka bir müşteriyi arama bilgilerini içeren özniteliklerdir.

Şebeke Dışı (Off-Net) arama öznitelikleri: Operatörde tanımlı müşterinin farklı operatörde tanımlı olan başka bir kullanıcıyı arama bilgilerini içeren özniteliklerdir.

Gelen arama bilgilerini içeren öznitelikler: Operatörde tanımlı müşterilere gelen arama bilgilerini içermektedir.

Şebeke İçi (On-Net) arama öznitelikleri: Operatörde tanımlı müşteriye yine aynı operatörde tanımlı olan başka bir müşteriden gelen arama bilgilerini içeren özniteliklerdir.

Fatura bilgilerini içeren öznitelikler: Müşterilerin ödemeleri ile ilgili bilgileri ele alınmıştır.

Etkinlik bilgilerini içeren öznitelikler: Müşterilerin gerçekleştirmiş olduğu etkinlik bilgilerini içeren özniteliklerden oluşmaktadır.

Konuşma süresi bilgilerini içeren öznitelikler: Müşterilerin gerçekleştirmiş olduğu konuşma sürelerini içeren özniteliklerden elde edilmiştir.

İnternet kullanım bilgilerini içeren öznitelikler: Müşterilerin internet kullanım bilgileri ele alınmıştır.

Katma Değer Hizmet (Value Added Service - VAS) bilgilerini içeren öznitelikler: Müşterilerin aldığı çeşitli katma değer hizmetlerine ait bilgileri içermektedir (sesli posta, zil geri sesi, bakiye kontrolleri, kontör yükleme, SMS oylama, SMS çekilişleri vb. hizmetler).

Gelir istatistik bilgilerini içeren öznitelikler: Müşterilerden elde edilen gelir bilgileri ele alınmıştır.

Paket istatistiklerini içeren öznitelikler: Müşterilere sunulan ve tercih edilen kampanyalarla ilgili bilgiler ele alınmıştır.

Dolaşım bilgilerini içeren öznitelikler: Müşterinin, ev ağının coğrafi kapsama alanı dışında seyahat ederken otomatik olarak sesli arama yapma ve alma, veri gönderip alma veya ev veri hizmetleri dahil diğer hizmetlere erişme bilgilerini içermektedir.

Demografik veri içeren öznitelikler: Müşterilerin kişisel bilgilerini içeren öznitelikleri ifade etmektedir.

4.1.2 Veri Ön İşlemesi

Veri ön işleme adımları üç veri madenciliği modeli için de gerçekleştirilmiştir. Müşteri terk modeli için özniteliklerde %40 ve üzerinde boş (null) veri içeren kayıtlar veri setinden çıkartılmıştır. Karakter, tarih tipinde olan ve model için yararlı olmayacağı düşünülen öznitelikler de veri setinden çıkartılmıştır. Veri setinde kalan sayısal özniteliklerde boş (null) olan veriler sıfır ile doldurulmuştur. Son adım olarak özniteliklerin kendi aralarındaki korelasyon oranları doğrultusunda birbirleriyle pearson korelasyon katsayısı 0.9 üzerinde olan özniteliklerden rastgele biri elenerek, lineer ilişkisi çok yüksek olan öznitelikler veri setinden temizlenmiştir.

4.1.3 Özniteliklerin Seçimi

Öznitelik seçimi, yalnızca müşteri terk modelinde kullanılan veri seti için uygulanmıştır. 600 öznitelikten model için en faydalı olan 296 özniteliğin seçilmesi için "scikitlearn" kütüphanesinin "RFE" sınıfı kullanılmıştır. Gini skoru en yüksek 296 öznitelik seçilerek, model için kullanılacak nihai veri seti belirlenmiştir.

4.1.4 Eğitim-Test Veri Setlerinin Seçimi

Müşteri terk modeli için veri seti, %30 test veri seti ve %70 eğitim veri seti olacak şekilde belirlenmiştir.

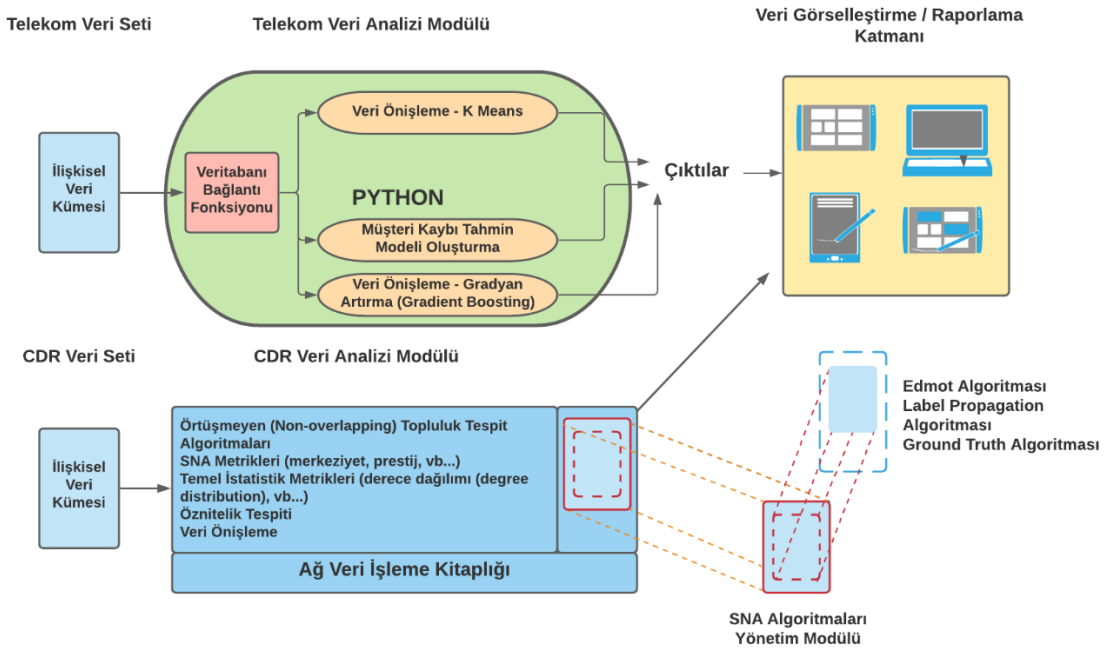
4.1.5 Müşteri Terk Tahmin Modelinin Oluşturulması

Müşteri terk modeli için son yıllarda popülerliği artmış bir model olan XGBoost modelinin kullanılması kararlaştırılmıştır. Literatürde ve ticari uygulamalarda sıkça kullanılan "XGBoost" kütüphanesinin Python versiyonu modelin geliştirilmesi için tercih edilmiştir. Diğer modellere göre hiper parametre sayısı ve

çeşidi fazla olan XGBoost modelinde, parametre optimizasyonuna ilk olarak, modelin performansına en çok etkisi olacağı düşünülen, "scale_pos_weight" ve "max_depth" parametrelerinden başlanılmıştır. Parametre optimizasyonunda, modelin test veri setindeki müşteri terk davranışını (customer churn) tahmin etme başarısı dikkate alınmıştır. Bu parametrelerle ilgili kullanılacak değerler kararlaştırıldıktan sonra diğer parametreler performans sonuçlarını maksimize edecek şekilde optimize edilmiştir.

4.2 SNA Öznitelikleri ile Müşteri Terk Tahmin Modeli

Şekil 4.2'de SNA özniteliklerinden yola çıkarak veri analizi iş sürecinin nasıl gerçekleştirildiği anlatılmaktadır. Bu araştırmanın bir parçası olarak, CDR verisi üzerinden bir ağ oluşturuldu ve bu ağ üzerinde alt topluluk algılama algoritmaları çalıştırıldı. Böylelikle alt grupta yer alan müşterilerin birbirlerini etkileyeceği varsayımına dayalı olarak özniteliklerin tespiti yapıldı.



Şekil 4.2 SNA modeli için önerilen sistem mimarisi

Bu araştırma kapsamında belirlediğimiz SNA özelliklerini hesaplamak için her alt toplulukta da algoritmalar çalıştırıldı. Daha sonra telekom verilerinden elde ettiğimiz öznitelikler ve SNA öznitelikleri bir arada kullanılarak veri analizi

yapıldı. Her bir müşteriye ait öznitelik vektörlerinin (davranış öznitelik vektörleri) birleştirilmesi sonucu müşteri davranış matrisi elde edildi.

Araştırmamız kapsamında kullanılan telekom veri seti üzerinde müşteri terk etme davranışı etiketleri bulunmaktadır. Etiketli müşteri davranış matrisi, eğitim ve test veri seti olarak ayrılmıştır. Eğitim veri seti üzerinde sınıflandırma algoritması çalıştırılarak, makine öğrenmesi modeli elde edilmiştir. Bu model kullanılarak yeni müşterilerin terk etme davranışı tahmin edilmiştir.

Şekil 4.2'de gösterilen CDR Veri Analizi Modülü, CDR veri seti üzerinde, SNA tabanlı özniteliklerin oluşturulması için geliştirdiğimiz modülü temsil etmektedir. Aynı şekilde Telekom Veri Analizi Modülü, telekom verilerinden özniteliklerin çıkartılması için geliştirilen iş süreci modülünü temsil etmektedir. Aşağıda bu modülleri kullanarak geliştirdiğimiz veri analizi iş sürecini maddeler halinde açıklıyoruz.

4.2.1 Özniteliklerin Oluşturulması

4.2.1.1 Telekom Verilerinden Müşteri Davranışı Modeli Oluşturulması Amaçlı Önerdiğimiz Öznitelikler

Telekom verilerinden elde ettiğimiz ve başarılı bir makine öğrenmesi gerçekleştirilebilmesi için önerdiğimiz öznitelikleri aşağıda sıralıyoruz: SMS öznitelikleri; mesajlaşma bilgilerini, Giden SMS öznitelikleri; gönderilen mesaj bilgilerini, On-Net SMS öznitelikleri; şebeke içi mesaj bilgilerini, Off-Net SMS öznitelikleri; şebeke dışı mesaj bilgilerini, Gelen SMS öznitelikleri; gelen mesaj bilgilerini, Arama öznitelikleri; arama bilgilerini, Giden arama öznitelikleri; gönderilen arama bilgilerini, On-Net arama öznitelikleri; şebeke içi arama bilgilerini, Off-Net arama öznitelikleri; şebeke dışı arama bilgilerini, Gelen arama öznitelikleri; gelen arama bilgilerini, Veri kullanımı öznitelikleri; internet kullanım bilgilerini açıklamaktadır. Bunun yanı sıra, Fatura öznitelikleri; ödeme bilgilerini, Katma değerli hizmet öznitelikleri; çeşitli katma değer hizmet bilgilerini (sesli posta, zil geri sesi, bakiye kontrolleri vb.), Gelir istatistik öznitelikleri; gelir bilgilerini, Kampanya istatistik öznitelikleri; sunulan kampanya bilgilerini, Dolaşım öznitelikleri; kapsama alanı dışında erişilebilen hizmet

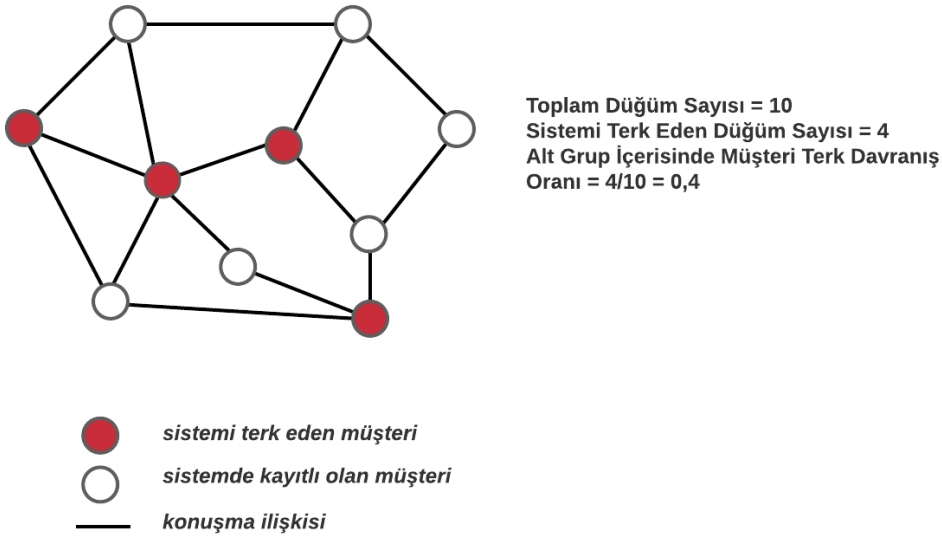
bilgilerini, Demografik veri öznitelikleri; müşterilerin kişisel bilgilerini açıklamaktadır.

4.2.1.2 SNA Teknikleri Kullanılarak Müşteri Davranışı Modeli Oluşturulması Amaçlı Önerdiğimiz Öznitelikler

Bu araştırma kapsamında, telekom müşterilerinin, CDR verisi üzerinden oluşturdukları karmaşık ağ üzerinde, sosyal ağ analizi teknikleri uygulanarak öznitelikler elde edilmiştir. Burada, alt topluluk tespit algoritmaları uygulanarak, müşterilerin kendi aralarında oluşturdukları topluluklar tespit edilmiştir. Tespit edilen topluluklardaki, müşteri terk etme davranışlarının birbirlerini etkileyeceği varsayımı yapılmıştır.

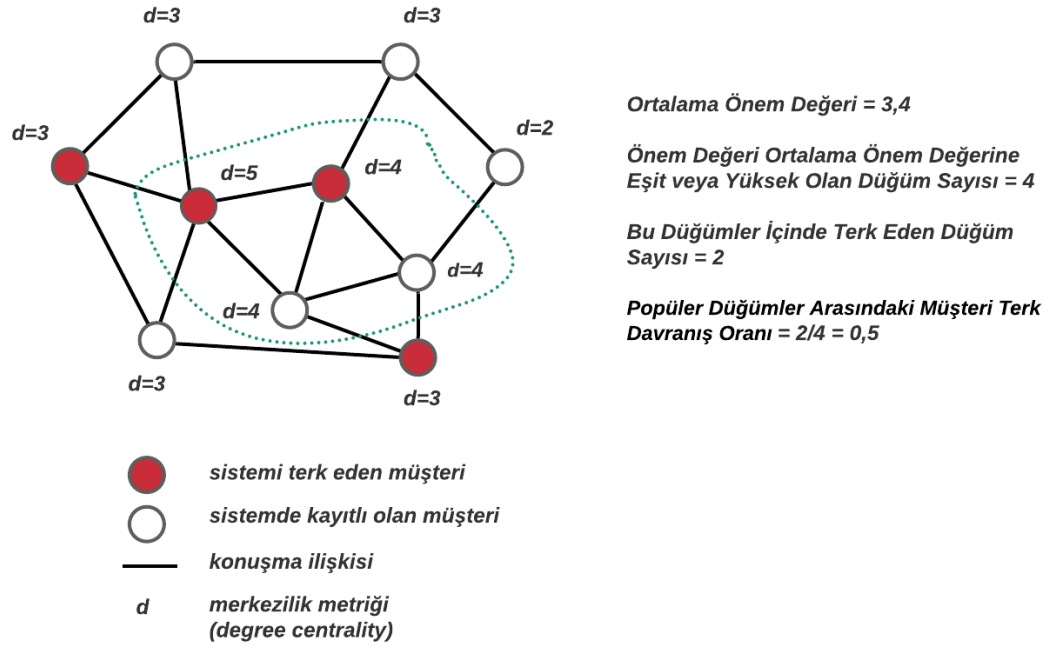
Alt topluluk tespit algoritmaları kullanılarak, herhangi bir alt topluluğun üyesi olan her müşteri için topluluk aidiyet etiketleri oluşturulmuştur. Elde edilen topluluk aidiyeti etiketlerinden oluşan yeni veri seti üzerinden öznitelikler çıkartılmıştır. Bu öznitelikleri aşağıda özetliyoruz:

Müşteri Terk Davranış Oranı (Churn-Behavior Proportion): Düğümün (node) bulunduğu alt grup içerisinde terk etme davranışı göstermiş kullanıcıların oranını tespit eden özniteliktir. Şekil 4.3'te örnek olarak bir alt grup için bu özniteliğin nasıl hesaplanıldığı gösterilmektedir. Bu örnekte, alt grup içerisinde toplam 10 müşteri bulunmaktadır ve bunların 4'ü sistemi terk etmiş müşteri olarak belirtilmiştir. Buna göre; bu alt grup içerisinde yer alan tüm düğümler için müşteri terk davranış oranı 0,4 olarak tespit edilmiştir. Bu değer bize diğer müşterilerin de bu oranda terk etme riskine sahip olduklarını göstermektedir.



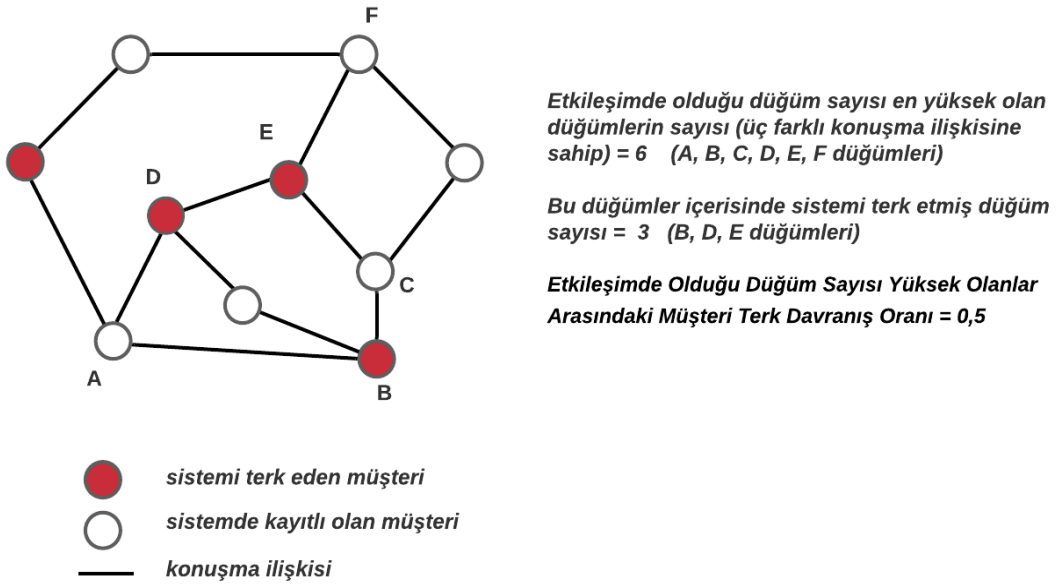
Şekil 4.3 Örnek alt grup müşteri terk davranış oranı tespiti

Popüler Düğümler Arasındaki Müşteri Terk Davranış Oranı (Churn-Behavior Proportion Among Popular Nodes): Önem değeri en yüksek (ortalama değerden daha yüksek veya eşit) kullanıcılar için terk etme davranışı göstermiş kullanıcıların oranını vermektedir. Şekil 4.4'te verilen örnek alt grup toplam 10 düğümden oluşmaktadır. Derece merkezilik (degree centrality) değerleri her düğümün üzerinde gösterilmektedir. Bu alt grup içerisinde öncelikle ortalama önem değeri tespit edilmektedir. Ardından bu ortalama önem değerine eşit veya yüksek önem değerine sahip düğümler içerisinde sistemi terk etmiş olanların oranı hesaplanmaktadır. Burada ortalama önem değeri 3,4 olarak tespit edilmiş ve bu değere eşit veya yüksek olan düğüm sayısının 4 olduğu belirtilmiştir. Bu tespit edilen 4 düğüm içerisinde sistemi terk etmiş olan 2 kullanıcı bulunmaktadır. Belirlediğimiz bu öznitelik kapsamında terk davranış oranı 0,5 olarak tespit edilmiştir.



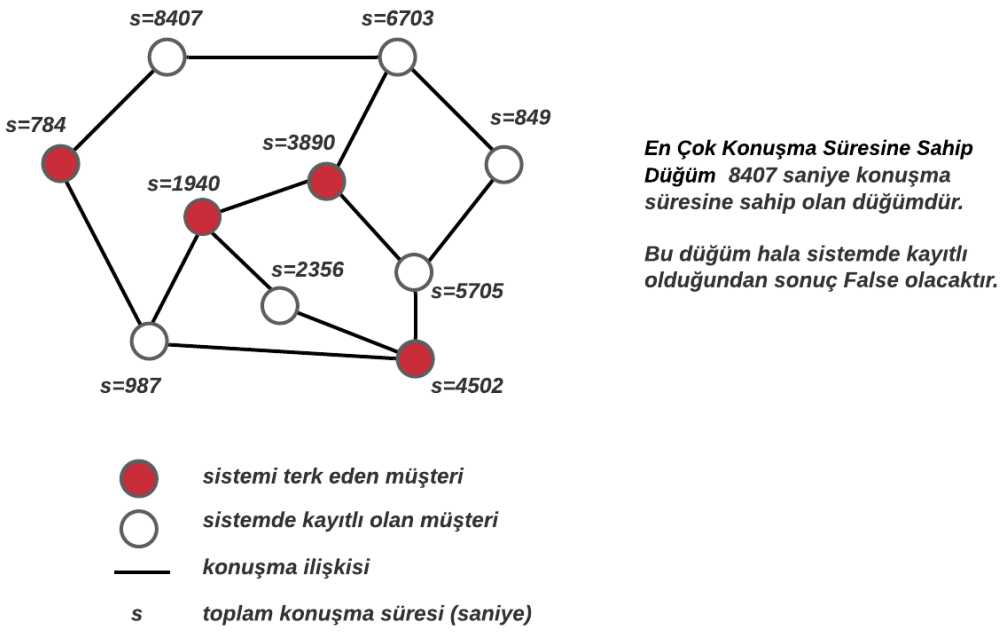
Şekil 4.4 Örnek alt grup popüler düğümler arasındaki müşteri terk davranış oranı

Etkileşimde Olduğu Düğüm Sayısı Yüksek Olanlar Arasındaki Müşteri Terk Davranış Oranı (Churn-Behavior Proportion for Highly Connected Node): Etkileşimde olduğu düğüm sayısı fazla olan ve terk etme davranışı göstermiş olan kullanıcıların oranını vermektedir. Şekil 4.5'te verilmiş olan örnekte A, B, C, D, E, F düğümlerinin alt topluluk içerisinde en çok konuşma ilişkisine (3 farklı konuşma ilişkisi) sahip oldukları görülmektedir. Bu düğümler içerisinde sistemi terk eden kullanıcılar B, D ve E düğümleri olarak tespit edilmiştir. Böylece bu öznitelik kapsamında terk davranış oranı 0,5 olarak hesaplanmıştır ve bu topluluk içerisindeki her düğüm için bu değer veri setine yeni bir özellik olarak eklenmiştir. Bu oran bize sistemdeki diğer müşterilerin 0,5 oranında sistemi terk etme riskine sahip olduklarını ifade etmektedir.



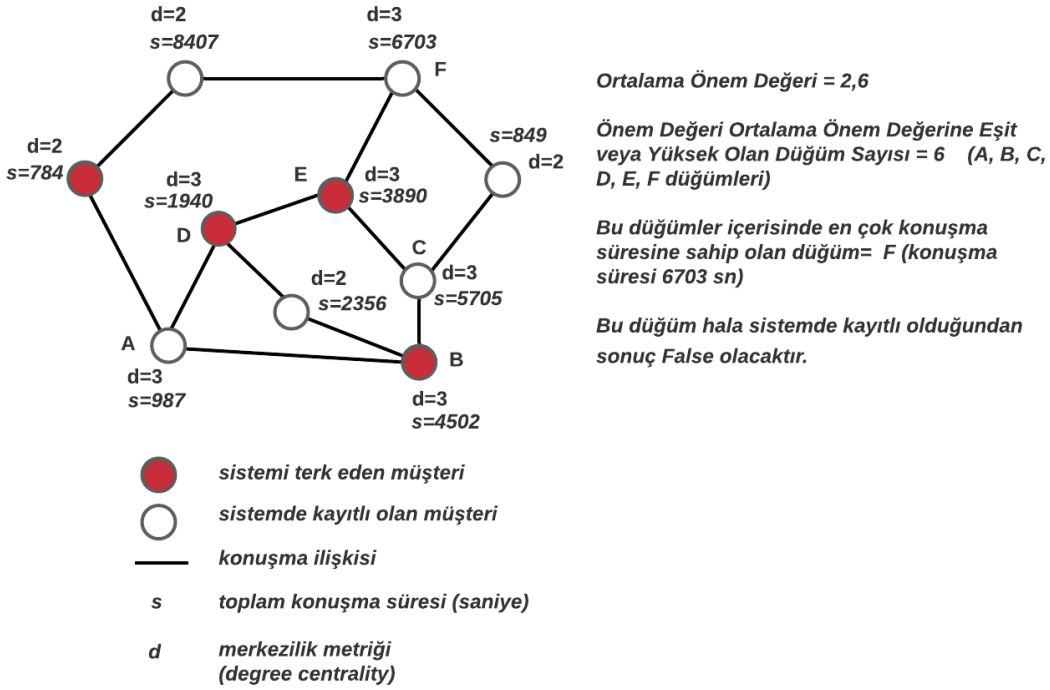
Şekil 4.5 Örnek alt grup etkileşimde olduğu düğüm sayısı yüksek olanlar arasındaki müşteri terk davranış oranı

En Çok Konuşma Süresine Sahip Düğüm İçin Müşteri Terk Davranışı (Churn-Behavior for the Talkative Node): Toplam en yüksek konuşma süresine sahip kullanıcının terk etme davranışı gösterip göstermediğini belirlemektedir. Bu öznitelik True veya False şeklinde sonuç döndürmektedir. Burada True dönmesi en yüksek konuşma süresine sahip olan kullanıcının sistemi terk ettiğini belirtirken, False bu kullanıcının hala sistemde olduğunu belirtmektedir. Sonucun true dönmesi toplulukta yer alan diğer kullanıcıların da terk etme ihtimalini arttırmaktadır. Şekil 4.6'da verilen örnek toplulukta en çok konuşma süresine sahip olan kullanıcının sistemi terk etmediği görülmektedir. Bu sebeple öznitelikten dönen sonuç false olacaktır.



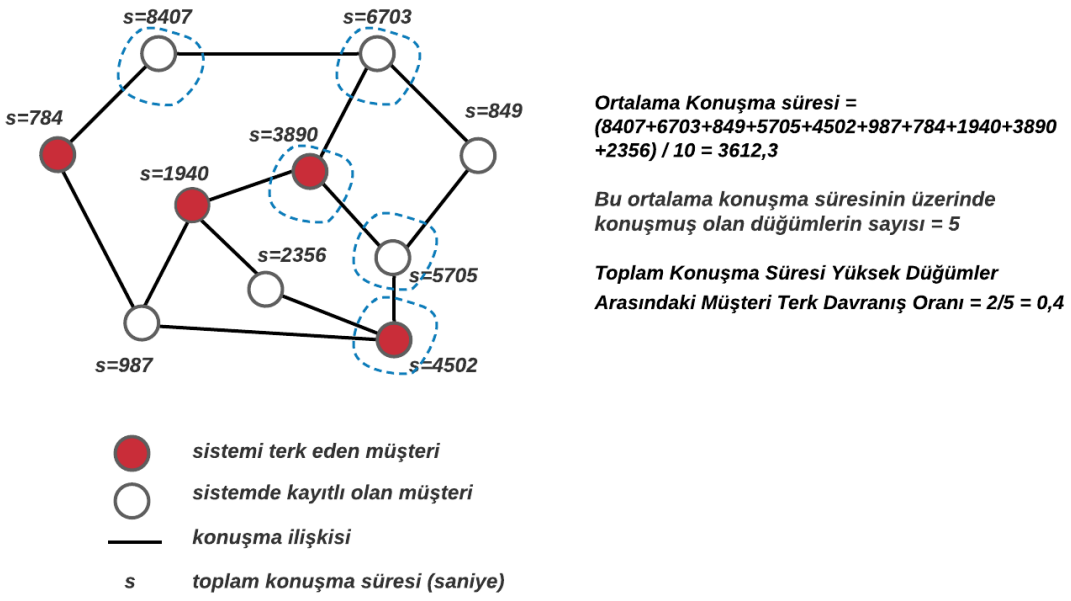
Şekil 4.6 Örnek alt grup en çok konuşma süresine sahip düğüm için müşteri terk davranışı

Popüler Düğümler Arasında En Çok Konuşma Süresine Sahip Düğüm İçin Müşteri Terk Davranışı (Churn-Behavior for the Highly Important Node): Önem değeri en yüksek (ortalama değerden daha yüksek veya eşit) kullanıcılar arasında toplam konuşma süresi en fazla olan kullanıcının terk etme davranışı gösterip göstermediğini true veya false olarak tespit etmektedir. Şekil 4.7'de verilen örnekte, ortalama önem değerine eşit veya daha yüksek değere sahip olan düğümler arasında toplam konuşma süresi en yüksek olan düğüm F düğümüdür. F düğümü hala sistemde kayıtlı olduğundan bu özniteliğin döndürdüğü sonuç False olacaktır.



Şekil 4.7 Örnek alt grup popüler düğümler arasında en çok konuşma süresine sahip düğüm için müşteri terk davranışı

Toplam Konuşma Süresi Yüksek Düğümler Arasındaki Müşteri Terk Davranış Oranı (Churn-Behavior Proportion for the Talkative Nodes): Ortalama konuşma süresinin üzerinde konuşma gerçekleştirmiş olan kullanıcılar arasında terk etme davranışı göstermiş kullanıcıların oranını vermektedir. Şekil 4.8'de verilen örnek alt grup içerisinde ortalama konuşma süresi hesaplanmıştır. Bu ortalama konuşma süresinin üzerinde konuşma gerçekleştirmiş olan düğümlerin sayısı 5 olarak tespit edilmiştir. Bu tespit edilen düğümler içerisinde sistemi terk eden düğümlerin sayısı ise 2 olarak görünmektedir. Bu nedenle, bu öznitelik kapsamında müşteri terk davranış oranı 0,4 olarak hesaplanmıştır.



Şekil 4.8 Örnek alt grup toplam konuşma süresi yüksek düğümler arasındaki müşteri terk davranış oranı

Derece Merkeziyet Metriği (Centrality Metric): Derece merkezilik ölçütü bir düğümün (node) diğer düğümlere olan kenar (edge) sayısını hesaplamaktadır.

Bu özniteliklerin hesaplanması için geliştirilen modül, Şekil 4.2' de gösterilen yazılım mimarisindeki CDR Veri Analizi modülüdür. Bu modül kapsamında farklı alt topluluk tespit algoritmaları çalıştırılarak, müşterilerin topluluk aidiyet bilgileri ve sosyal ağ analizi yapılarak derece merkezilik metriklerinin değerleri hesaplanmıştır.

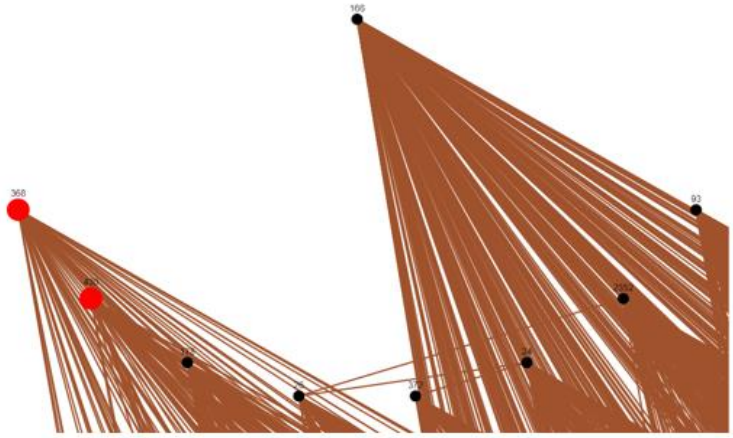
Temel Kavramlar ve Literatür Çalışması bölümünde, detaylarını açıkladığımız üç farklı alt topluluk tespit algoritması kullanılarak, alt topluluklar tespit edilmiştir.

Tablo 4.2'de maddeler halinde listelediğimiz öznitelikler, müşteri terk etme davranış modelinin tahmin başarısını arttırmaya yönelik belirlenmiştir.

Tablo 4.3 SNA tabanlı telekom öznitelikleri

SNA Tabanlı Telekom Öznitelikleri
Müşteri Terk Davranış Oranı (Churn-Behavior Proportion)
Popüler Düğümler Arasındaki Müşteri Terk Davranış Oranı (Churn-Behavior Proportion Among Popular Nodes)
Etkileşimde Olduğu Düğüm Sayısı Yüksek Olanlar Arasındaki Müşteri Terk Davranış Oranı (Churn-Behavior Proportion for Highly Connected Node)
En Çok Konuşma Süresine Sahip Düğüm İçin Müşteri Terk Davranışı (Churn-Behavior for the Talkative Node)
Popüler Düğümler Arasında En Çok Konuşma Süresine Sahip Düğüm İçin Müşteri Terk Davranışı (Churn-Behavior for the Highly Important Node)
Toplam Konuşma Süresi Yüksek Düğümler Arasındaki Müşteri Terk Davranış Oranı (Churn-Behavior Proportion for the Talkative Nodes)
Merkeziyet Metriği (Centrality Metric)

Şekil 4.9'da, CDR verisinden elde edilen bir alt topluluğa ait graph gösterilmektedir. Kırmızı ile gösterilen düğümler sistemi terk eden müşterileri ifade etmektedir.



Şekil 4.9 CDR alt topluluk ağ grafiği

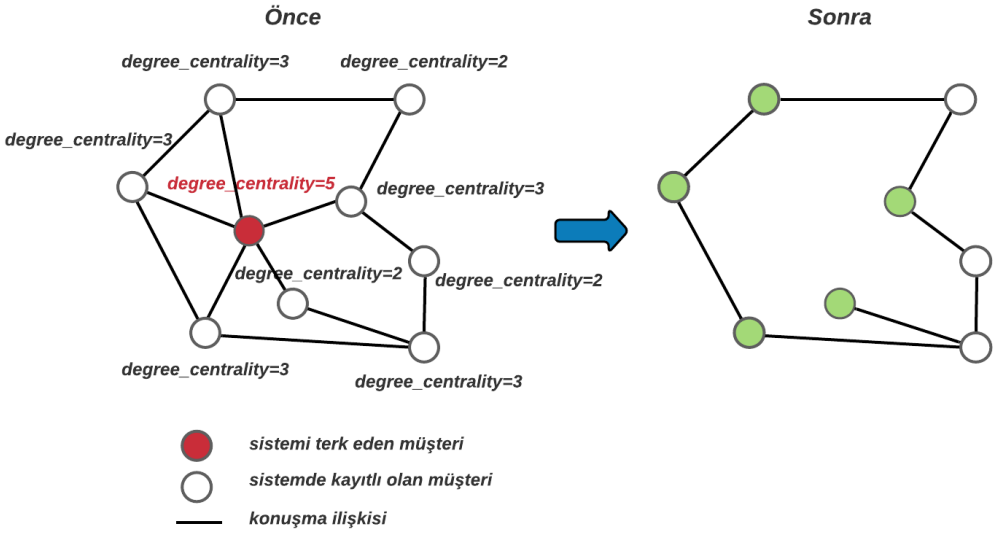
Bu kapsamdaki öznitelikler belirlenirken, etkin roldeki müşterilerin terk etme davranışı göstermesi durumunda, bundan direkt etkilenecek olan müşterilerin de terk etme davranışı göstermesinin olası olduğu göz önünde tutulmuştur. Bu varsayımın doğruluğu, deneysel çalışma ile doğrulanmıştır.

Şekil 4.10' da CDR verisinden elde edilmiş olan ağ grafiğinin genel hali gösterilmektedir.



Şekil 4.10 CDR alt topluluk ağ grafiği genel

Şekil 4.11'de resmedilen örnekte, en yüksek önem derecesine sahip olan müşteri (kırmızı fontla renklendirilmiş düğüm) sistemi terk ettiğinde, bu müşteri ile doğrudan etkileşimde olan diğer müşteriler (yeşil fontla renklendirilmiş düğümler) de terk etme davranışı gösteren olası müşteriler olarak gösterilmektedir.



Şekil 4.11 Örnek alt grup ağ grafiği

4.2.2 Özniteliklerin Bir Araya Getirilmesi ile Elde Edilen Veri Seti Üzerinde Veri Ön İşleme Çalışmalarının Yapılması

Özniteliklerin bir araya gelmesi ile elde ettiğimiz verinin modellemeye uygun hale dönüştürülebilmesi için aykırı verilerin temizlenmesi, eksik verilerin tamamlanması, veri normalizasyonu adımları gerçekleştirilmiştir.

4.2.3 Müşteri Terk Etme Davranışı Modelinin Oluşturulması

Model tarafından işlenecek olan verinin eğitim ve test veri seti olarak ikiye bölünmesi bu aşamada yapılmıştır. Veri setinin %70'i eğitim seti, geri kalan %30'u ise test seti olarak ayrılmıştır. En yüksek performansa sahip olan öğrenme algoritması bu aşamada belirlenmiştir. Yapılan araştırma sonucunda XGBoost algoritmasının en başarılı tahmin sonuçlarını verdiği gözlemlenmiş ve çalışmamızda bu algoritma kullanılmıştır. Bu aşamada, müşteri terk tahmin modelleri, eğitim verisinden oluşturulmaktadır.

4.2.4 Müşteri Terk Etme Davranışı Tahmini

Bu aşamada, sisteme sürekli olarak gelen müşteri terk etme davranışı ile ilgili vektörlerin tahmini gerçekleştirilmiştir. Bir önceki aşamada oluşturmuş olduğumuz tahmin modeli kullanılarak, müşteri terk etme davranışı tahmini

yapılmaktadır. Burada ikili karar verilmektedir (terk etme eğilimi var / terk etme eğilimi yok).

4.3 Tutundurma Modelleri

SNA yöntemleri ile müşteri terk tahmininden elde edilen sonuçlar ışığında, terk etme eğilimine sahip olan müşterilere hangi kampanyaların, hangi sıra ile sunulması gerektiği belirlenmeye çalışılmıştır. Burada müşterilerin hareket verileri incelenmiştir ve çapraz satış analizi ile birliktelik kurallarının belirlenebilmesi için FP-Growth algoritması kullanılmıştır. Ayrıca satın alınan paketlerin hangi sıra ile alındığını tespit edebilmek için, sıralı desen algoritması olan PrefixSpan algoritması tercih edilmiştir.

4.3.1 Veri Ön İşlemesi

FP-Growth modeli için tabloda tutulan veri seti; müşteri kimliği, işlem tarihi (transaction date) ve alınan ürün bilgilerini içermektedir. Veri setinin modelde kullanılabilmesi için müşteri kimliği, belirli tarih aralığında alınan ürünler virgülle ayrılmış bir formata dönüştürülmüştür.

PrefixSpan modeli için veri seti FP-Growth modeli için kullanılacak formata getirilmiş ve alınan ürünlerin sıralı şekilde olması sağlanmıştır.

4.3.2 Eğitim-Test Veri Seti Seçimi

FP-Growth ve PrefixSpan modelleri için veri seti zaman bazında eğitim ve test veri setine ayrılmıştır. İşlem tarihine (transaction date) göre belirlenen zaman diliminin yedide ikisi test seti için, kalanı eğitim veri seti için ayrılmıştır.

4.3.3 FP-Growth ve PrefixSpan Modellerinin Oluşturulması

FP-Growth ve PrefixSpan algoritmaları Python ortamında geliştirilmiştir. FP-Growth modeli için eğitim setinde destek oranı en yüksek olan 14 kural, PrefixSpan modeli için de 10 kural belirlenmiştir. Kural başarılarını ölçmek için, FP-Growth modeli eğitim setinden elde edilen bir kurala ait işlemlerden (transaction) biri test setinde gözlemlenmişse, bu kurala ait diğer işlemlerden herhangi birinin gözlemlenme durumuna göre kural geçerli veya geçersiz kabul

edilmiştir. PrefixSpan modelinin testinde ise FP-Growth modeline benzer bir yaklaşım izlenmiş ve kuralın geçerli olabilmesi için kuraldaki işlem sırasının test setinde de olması gözetilmiştir. Kuralların başarı metriği, test setindeki geçerli durumların, geçerli ve geçersiz durumların toplamına oranı olarak kabul edilmiştir.

5
UYGULAMA DETAYLARI

5.1 Kullanılan Teknolojiler

Veri ön işleme, modellerin geliştirilmesi, test edilmesi gibi uçtan uca tüm geliştirme süreçlerinde Python programlama dilinin 3.7 versiyonu kullanılmıştır. Müşteri terk modeli tek düğümlü (node) Spark 3.0.1 ortamında çalıştırılmıştır. Spark ortamı 6 GB RAM ve 4 CPU'ya sahip Ubuntu 18.04 işletim sistemli bir sanal makineye kurulmuştur. XGBoost modelinin çalıştığı Spark düğümünde varsayılan ayarlar kullanılmıştır. Müşteri terk modelinde veri ön işleme, öznitelik seçimi adımları için "Pandas" ve "Sklearn" kütüphaneleri kullanılmıştır. Modelleme kısmında "XGBoost" kütüphanesinin "Python" dili tercih edilmiştir.

Metodoloji bölümünde Şekil 4.2'de bahsettiğimiz sistem mimarisinin, aşağıda açıkladığımız teknolojileri ve donanımsal alt yapıyı kullanarak, bir prototip uygulamasını gerçekleştirdik.

CDR verisinin işlenmesi için “Amazon AWS” ortamındaki yüksek nitelikli sanal sunucular üzerinde çalışmamız yürütüldü. CDR verisinin çok büyük olması sebebiyle ve üzerinde çalıştırmayı hedeflediğimiz grup tespit algoritmalarının performanslarını en yüksek seviye de gözlemleyebilmek adına (sonuçların daha kısa sürede ve en az eforda elde edilebilmesi için), Amazon AWS ortamında, grup tespit algoritmaları çalıştırılmıştır.

Prototip uygulama, Python dili kullanılarak geliştirilmiştir. Python 3.7 versiyonu ile "GraphFrames" kütüphanesi kullanılarak SNA metrikleri çalıştırılmıştır. Derece Merkezilik (Degree Centrality) metriği için, Python'ın NetworkX paketindeki degree_centrality sınıfı kullanılmıştır.

Üç ayrı yöntem için ayrı ayrı hesaplanan öznitelik sonuçları veri ön işleme adımları uygulanarak, makine öğrenmesi algoritması için hazır hale getirildi. Geri planda

veri tabanı için MS-SQL veri tabanı kullanılmıştır. Görselleştirme için Cytoscape.js kütüphanesi kullanılmıştır.

Bu araştırma kapsamında kullandığımız grup tespit algoritmaları Python NetworkX kütüphanesi kullanılarak geliştirilmiştir. Dash kütüphanesi kullanılarak sonuçlar görselleştirilmiştir.

CDR verisi üzerinde sosyal ağ analizi yaparak alt gruplar tespit edilmiştir. Üçüncü bir grup tespit algoritması olan Ground Truth algoritması ile EdMot ve Label Propagation algoritmalarında ortak tespit edilen gruplar belirlenmiştir.

Bulunan topluluklar her yöntem için ayrı ayrı ele alınmıştır. SNA tabanlı özniteliklerin değerleri, her bir yöntem için ayrı ayrı hesaplanmıştır.

FP-Growth ve PrefixSpan modellerinin veri ön işlemesinde "Pandas" kütüphanesi kullanılmıştır.

5.2 Verinin Tanımlanması

Bu proje kapsamında kullanılan, telekom CDR veri seti, karmaşık ağ verisine dönüştürüldüğünde; 1.224.510 milyon düğüm sayısı ve 1.048.363 milyon kenar sayısı olan bir çizge veri seti oluşturmaktadır. Veriler, Orta Asya'da bulunan bir telekom şirketinden alınmıştır ve 2019 yılı Mayıs ayına ait verilerdir.

Kullanılan verinin CSV formatındaki boyutu 23,1 MB. şeklindedir. CDR veri seti Tablo 5.1' de gösterildiği gibi 4 sütundan oluşur. Bunlar; arayan kimliği (aramayı başlatan müşteri kimliği), aranan kimliği (aranan müşteri kimliği), konuşma süresi (saniye) ve müşteri terk etti mi (müşterinin ayrılıp ayrılmadığını göstermektedir (0: ayrılmadı, 1: ayrıldı)).

Burada, müşterilerin operatörü terk edip etmediği bilgisi; takip eden aydaki müşteri davranışları için (sistemde kalma / sistemi terk etme) telekom şirketinin veri tabanında eşleştirme yapılarak elde edilmiştir ve CDR verisi ile entegre edilmiştir.

Tablo 5.1 CDR veri seti

Arayan Kimliği	Ananan Kimliği	Konuşma Süresi (sn)	Müşteri Terk Etti Mi
814360	385525	5181	0
814360	833758	1297	0
138595	396126	1931	0
846785	194505	754	1
846785	443381	303	1
846785	223412	208	1
1107280	1029688	800	0
…	…	…	…

SNA özniteliklerinin her alt topluluk için hesaplatılması ile elde edilen veri seti ise Tablo 5.2' de gösterilmektedir. Tablo 5.1'de verilen CDR veri setinde alt alta verilmiş olan müşteri hareketleri her bir müşteri için tek bir satır haline getirilmiştir. Böylece, bir müşterinin en çok konuştuğu üç kişiye ait bilgiler artık yan yana görülmektedir. En çok konuşma gerçekleştirdiği kişi sayısı bazı müşteriler için bir veya iki de olabilmektedir. Ardından Tablo 5.2'de her bir müşteri için hesaplanan öznitelikler CDR veri setine eklenmiştir.

Tablo 5.2 EdMot öznitelik veri seti

Müşteri Id	Merkeziyet Metriği	EdMot Id	Ö1	Ö2	Ö3	Ö4	Ö5	Ö6
814360	2.449962E-06	0	0.2030	0.1875	0.1788	FALSE	FALSE	0.1314
36681	2.449962E-06	0	0.2030	0.1875	0.1788	FALSE	FALSE	0.1314
892461	2.449962E-06	0	0.2030	0.1875,	0.1788	FALSE	FALSE	0.1314
1136684	1.633308E-06	0	0.2030	0.1875	0.1788	FALSE	FALSE	0.1314
976678	2.449962E-06	0	0.2030	0.1875	0.1788	FALSE	FALSE	0.1314
1135777	8.166539E-07	7194	0.0943	0	0.0943	TRUE	FALSE	0.1428
1088781	8.166539E-07	7194	0.0943	0	0.0943	TRUE	FALSE	0.1428
883071	8.166539E-07	7194	0.0943	0	0.0943	TRUE	FALSE	0.1428
1132284	0.000003	9999	0.3333	1	0.3333	FALSE	TRUE	0
803087	0.000003	9999	0.3333	1	0.3333	FALSE	TRUE	0
73649	0.000003	9999	0.3333	1	0.3333	FALSE	TRUE	0

Tablo 5.2'deki kolonların açıklamaları ise aşağıdaki şekildedir:

Müşteri Id: Aramayı başlatan kullanıcının kimliği,

Merkeziyet Metriği: Müşterinin ağdaki diğer kullanıcılar ile olan bağlantı sayısını göstermektedir. Bu değer ne kadar büyükse müşteri bulunduğu ağda o kadar önemli demektir.

EdMot Id: Verilen tablodaki veri seti EdMot algoritması çalıştırılarak elde edilmiş olan topluluklara örnektir. Burada topluluk id müşterilerin hangi alt toplulukta yer aldıklarını belirtmektedir.

Ö1: Bölüm 4'te SNA teknikleri kullanılarak müşteri davranışı modeli oluşturulması amaçlı önerdiğimiz özniteliklerden müşteri terk davranış oranını,

Ö2: Bölüm 4'te SNA için önermiş olduğumuz popüler düğümler arasındaki müşteri terk davranış oranını,

Ö3: Bölüm 4'te SNA için önermiş olduğumuz etkileşimde olduğu düğüm sayısı yüksek olanlar arasındaki müşteri terk davranış oranını,

Ö4: Bölüm 4'te SNA için önermiş olduğumuz en çok konuşma süresine sahip düğüm için müşteri terk davranışını,

Ö5: Bölüm 4'te SNA için önermiş olduğumuz popüler düğümler arasında en çok konuşma süresine sahip düğüm için müşteri terk davranışını,

Ö6: Bölüm 4'te SNA için önermiş olduğumuz toplam konuşma süresi yüksek düğümler arasındaki müşteri terk davranış oranını göstermektedir.

5.3 Değerlendirme Metrikleri

5.3.1 Doğruluk (Accuracy) Değeri

Bu değer, doğru tahmin ettiğimiz alanların toplam veri kümesine oranı ile hesaplanmaktadır. Doğruluk değeri bir modelin başarısını ölçmek için çok önemlidir ancak tek başına yeterli değildir. (5.1) eşitliğinde gösterilen *TP (True Positive – Gerçek Pozitifler):* müşterilerden sistemi terk edecek şekilde tahminde bulunulmuş olanların içerisinde sistemi gerçekten terk etmiş olanların sayısını (sistemi terk edenlerin doğru tahmin sayısı), *TN (True Negative – Gerçek Negatifler):* müşterilerden sistemi terk etmeyecek şekilde tahminde bulunulmuş olanların içerisinde sistemi gerçekten terk etmemiş olanların sayısını (sistemi terk etmemiş olanların doğru tahmin sayısı), *FP (False Positive – Yanlış Pozitifler):* müşterilerden sistemi terk edecek şekilde tahminde bulunulmuş ve bunlar içerisinde sistemi gerçekten terk etmemiş olanların sayısını (yanlış tahmin

sayısını), *FN (False Negative – Yanlış Negatifler):* müşterilerden sistemi terk etmeyecek şekilde tahminde bulunulmuş ve bunlar içerisinde sistemi gerçekten terk etmiş olanların sayısını (yanlış tahmin sayısını) göstermektedir.

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

5.3.2 Kesinlik (Precision) Değeri

Bu değer, *FP (False Positive – Yanlış Pozitifler*) değerini yani müşterilerden sistemi terk edecek şekilde tahminde bulunulmuş ve bunlar içerisinde sistemi gerçekten terk etmemiş olanların sayısını (yanlış tahmin sayısını) tahmin etmenin yüksek maliyetli olacağı hallerde önemlidir. Tahmin model seçimlerinde Kesinlik değerinin yüksek olması tercih edilmektedir. Hesaplanma formülü (5.2)'de gösterilmektedir.

$$Kesinlik = \frac{TP}{TP + FP} \tag{5.2}$$

5.3.3 Duyarlılık (Recall) Değeri

Duyarlılık ise *FN (False Negative – Yanlış Negatifler)* değerinin yani müşterilerden sistemi terk etmeyecek şekilde tahminde bulunulmuş fakat bunlar içerisinde sistemi gerçekten terk etmiş olanların sayısının (yanlış tahmin sayısının) hesaplanmasının masraflı olduğu durumlarda bize yol göstermektedir. (5.3)'te duyarlılığın hesaplanmasında kullanılan formül verilmektedir.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{5.3}$$

5.3.4 F1-Skor Değeri

Bu değer bize Kesinlik (Precision) ve Duyarlılık (Recall) değerlerinin harmonik ortalamasını vermektedir. Kesinlik ve duyarlılık değerleri arasında bir denge aramak istendiğinde F1-Skor değeri yararlı olmaktadır. (5.4)'te F1-Skor formülü görülmektedir.

$$F_1\ Skor = 2 * \frac{Kesinlik\ *\ Duyarlılık}{Kesinlik +\ Duyarlılık} \tag{5.4}$$

6
DENEY SONUÇLARI VE DEĞERLENDİRME

6.1 Sonuçlar ve Değerlendirme

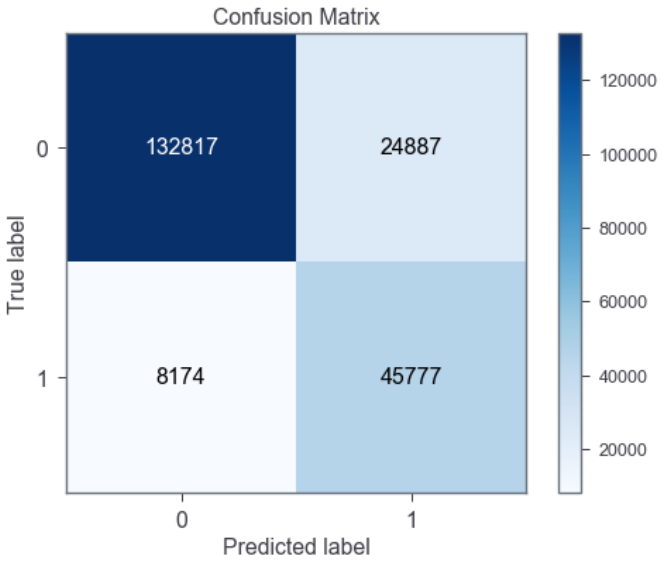
Müşteri terk tahmin modeli XGBoost algoritması ile çalıştırılmıştır ve elde edilen sonuçlar Tablo 6.1'de verilmiştir. Modelin genel tahmin başarısı (accuracy) yaklaşık olarak 0.84 olarak bulundu. Bu durum, modelin müşteri terk tahmininde %84 oranında doğru sonuç verdiğini göstermektedir. Tablo 6.1'de verilen 0: sistemde kalan müşterileri, 1: sistemi terk eden müşterileri ifade etmektedir. Support ise model için seçilen test setindeki müşteri sayılarını göstermektedir. Gerçekte sistemde kalan müşteri sayısı 157704 iken, sistemi terk etmiş müşteri sayısı ise 53951' dir. Veri setindeki toplam müşteri sayısı 211655 olup tabloda gösterilmektedir. Tabloya göre, Kesinlik ve Duyarlılık değerleri, sistemi terk edenlerin kesinlik değeri hariç, %80'in üzerinde sonuç vermektedir. Bu değerlerin yüksek olması model seçimlerinde önem arz etmektedir. Fakat bu iki değerin harmonik ortalaması olan F1-Skor değerine bakarak yorum yapmak daha sağlıklı olacaktır. Buna göre; F1-Skor değeri sistemi terk etmeyenler için (0 değeri için) 0.89 ile gayet iyi bir sonuç döndürmektedir. Yalnız sistemi terk edenler için biraz daha düşük, 0.73 değerinde bir sonuç görülmektedir. Tüm bu değerler bize, modelin ortalamanın üzerinde sonuçlar döndürerek başarılı bir tahmin yaptığını göstermektedir.

Diğer parametrelere de bakılacak olursa, modelin gayet başarılı sonuçlar verdiği görülmektedir.

Tablo 6.1 Geleneksel makine öğrenmesi müşteri terk tahmini test seti sonuçları

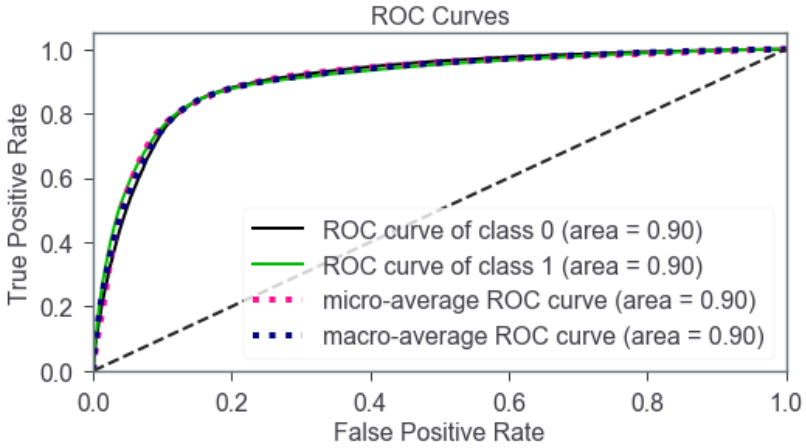
	Kesinlik	Duyarlılık	F1-Skor	Toplam Kullanıcı Sayısı
0	0.94	0.84	0.89	157704
1	0.65	0.85	0.73	53951
Doğruluk			0.84	211655
Makro Ortalama	0.79	0.85	0.81	211655
Ağırlıklı Ortalama	0.87	0.84	0.85	211655
Doğruluk: 0.8437976896364366				

Şekil 6.1'deki Karmaşıklık Matrisini yorumlayacak olursak; True Label alanı gerçek değerleri, Predicted Label ise tahmin değerlerini nitelemektedir. 0: müşterilerin sistemde kaldığını, 1: müşterilerin sistemi terk ettiğini göstermektedir. Örneğin; sistemi terk etmemiş olarak tahmin edilen ve gerçekte de sistemi terk etmemiş olanların sayısı *(TN - True Negative değeri)* 132817 olarak görülmektedir. Sistemi gerçekten terk etmiş olup doğru tahmin edilenlerin sayısı ise *(TP – True Positive değeri)* 45777 olarak görülmektedir. Bu iki değerin yüksek olması model seçiminde oldukça önem arz etmektedir.



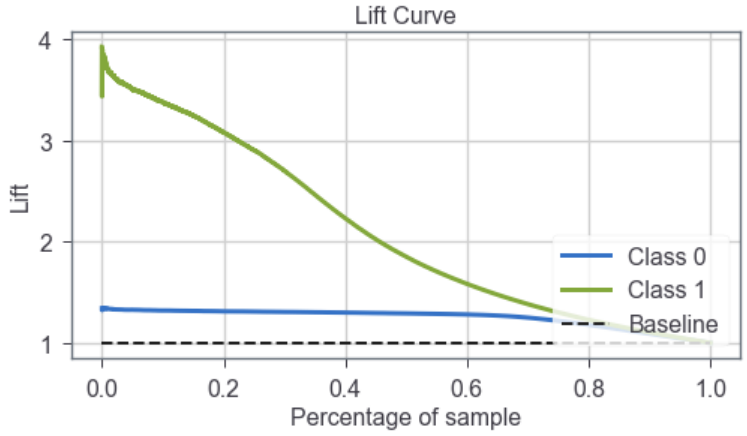
Şekil 6.1 Karmaşıklık matrisi

Modelin performans ölçümünde bakılması önemli olan bir diğer ölçüm ise ROC (Receiver Operating Characteristic - Alıcı İşletim Karakteristiği) eğrileridir. Modelin tahmin yapmada ne kadar başarılı olduğunu göstermektedir. ROC eğrisinin altında kalan alana (AUC - Area Under the Curve) bakılarak model performansını değerlendirebiliriz. Bu nedenle Şekil 6.2'de gösterilen ROC eğrisinin altında kalan alana bakılacak olursa 0.9 olarak bulunduğu görülmektedir. Eğri altında kalan alan 1'e oldukça yakın bir değere sahiptir. Bu da bize modelin tahmin başarısının yüksek olduğunu kanıtlamaktadır.



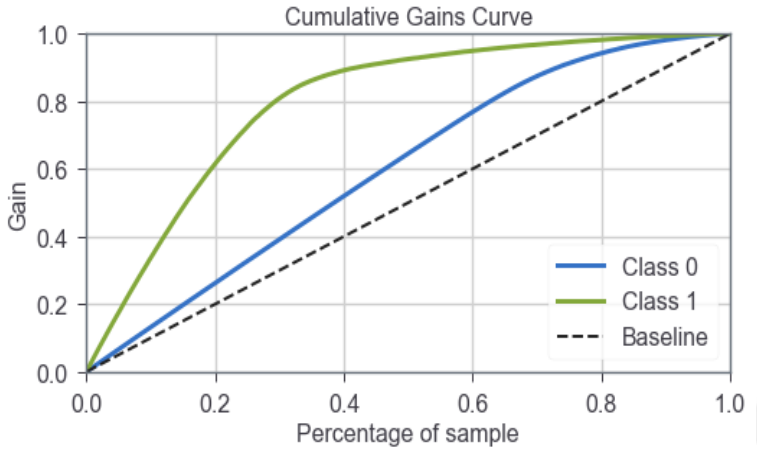
Şekil 6.2 ROC eğrisi

Model performanslarında yaygın olarak kullanılan bir diğer eğri ise Lift (Kaldırma) eğrisidir. Şekil 6.3'teki y ekseni, kullandığımız modelin rastgele tahmin etmeye oranla kaç kat daha iyi tahminde bulunduğu konusunda bilgi vermektedir. Şekil 6.3'te modelin 0.2'deki artışının 3 kat olduğu belirtilmektedir. Bu da modelin, Class 1'deki %20'lik rastgele seçimine göre müşterilerin 3 katını alabileceği anlamına gelmektedir. Örneğin, model yardımı ile müşterilerin %20'sine kampanya sunulursa, terk edenlerin %60'ı hedeflenebilir. Fakat rastgele seçim ile yalnızca terk edenlerin %20'sine odaklanılabilir. Kaldırma değerinin yüksek olması daha iyi bir modeli temsil etmektedir.



Şekil 6.3 Kaldırma eğrisi

Kümülatif Kazanç Eğrisi, model performansını değerlendiren ve sonuçları rastgele seçimler ile karşılaştıran bir değerlendirme eğrisidir. Modele göre hedef olma olasılığı en yüksek nüfusun belirli bir yüzdesi dikkate alındığında ulaşılan hedeflerin yüzdesini göstermektedir. Yatay eksenin sol kısmı modele göre hedef olma ihtimali en yüksek olan gözlemleri, sağ tarafı ise tam tersini belirtmektedir. Şekil 6.4'e göre yatay eksenin %20 noktasında en yüksek olasılığın %60 olduğu görülmektedir. Dikey eksen, hedeflerin yüzde kaçının dahil edildiğini yani kazancı gösterir. Dolayısıyla, gözlemlerin %20'si hedeflenirse, model toplam terk edenlerin %60'ının bu grupta olmasını sağlayacaktır.

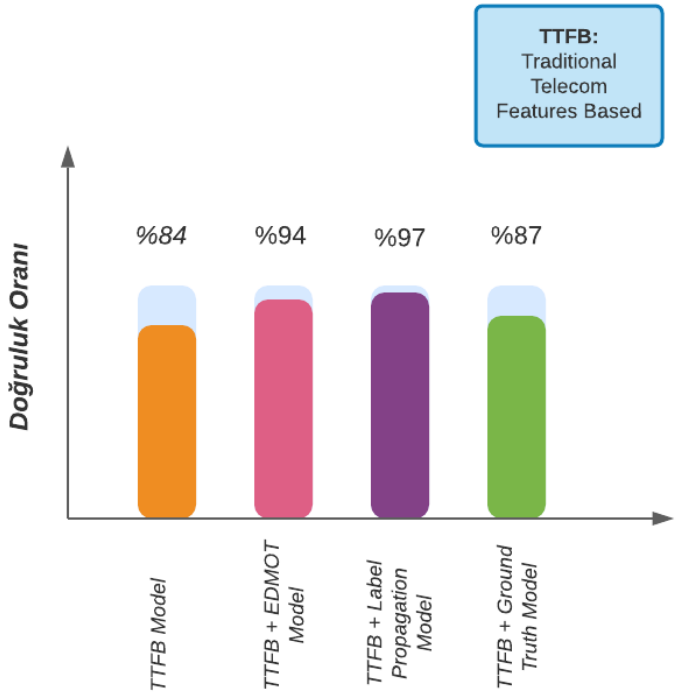


Şekil 6.4 Kümülatif kazanç eğrisi

Önerilen metodolojinin kullanabilir olduğunu göstermek amacıyla geliştirilen prototip uygulamanın performansı, daha çok Doğruluk ve F1-Skor değerleri baz alınarak analiz edilmiştir.

İlk olarak, bu araştırma kapsamında kullandığımız telekom veri setinden elde edilmiş telekom özniteliklerini kullanarak XGBoost makine öğrenme algoritması ile bir müşteri terk modeli oluşturuldu. Veri seti eğitim ve test veri seti olarak ikiye ayrıldı. Veri setimizin %70'ini eğitim seti olarak, geri kalan %30'unu ise test seti olarak belirlendi. Makine öğrenmesi algoritmasına eğitim seti verildi, böylece müşterilerin terk davranışlarını tahmin edecek model oluşturulabildi. En son da test verisini modele vererek modelin daha önce hiç görmediği müşteriler üzerinde müşteri terk davranışını tahmin etmesi sağlandı. Bu modelin tahmin doğruluk sonuçlarını Şekil 6.5'te, F1-Skor sonuçlarını ise Şekil 6.14'te TTFB (Traditional

Telecom Features Based) Model alanında görmekteyiz. Sonuçlara bakılacak olursa; doğruluk tahmini oranı yaklaşık olarak 0.84 çıkmıştır. Yani model genel olarak %84 oranında doğru tahminde bulunmuştur.



Şekil 6.5 Doğruluk sonuç grafiği

Bu çalışma kapsamında, Sosyal Ağ Analizi metrikleri kullanılarak müşteri terk davranış tahmini başarısının arttırılması hedeflenmektedir. Bunun değerlendirilmesi için şu adımlar takip edilmiştir:

İlk olarak; alt grup tespiti için EdMot algoritmasını kullanarak veri setinin alt gruplara ayrılması sağlanmış ve belirlemiş olduğumuz SNA öznitelikleri ile telekom özniteliklerini birlikte kullanarak yine XGBoost algoritması ile eğitilmiş bir makine öğrenmesi modeli oluşturulmuştur. Böylece müşterilerin birbirlerinden etkilenme durumları dikkate alınarak, modelin oluşturulması sağlanmıştır.

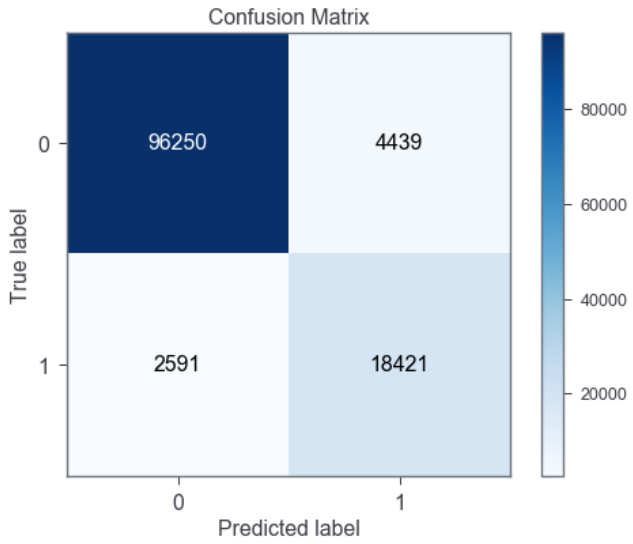
Sonuçlar Şekil 6.5 ve Şekil 6.14'te gösterilmiştir, model genel olarak %94 oranında doğru tahminde bulunmuştur. Bar grafik üzerinde TTFB Model ile TTFB + EDMOT Model sonuçlarını karşılaştırdığımızda; EdMot algoritması kullanılarak elde edilen genel başarı oranının çok daha yüksek olduğu görülmektedir. F1-Skor sonuçları da SNA metriklerinin modelin başarısını arttırdığını göstermektedir.

Modelin ayrıntılı sonuçları Tablo 6.2'de verilmektedir. Tablodaki değerlere bakıldığında bir önceki müşteri terk tahmin modelindeki değerlerden çok daha yüksek sonuçların elde edildiği net bir şekilde görülmektedir. Tüm değerlerin %80'in üzerinde tespit edilmiş olması TTFB + EDMOT modelinin oldukça iyi bir performansa sahip olduğunu ifade etmektedir.

Tablo 6.2 EdMot müşteri terk tahmini test seti sonuçları

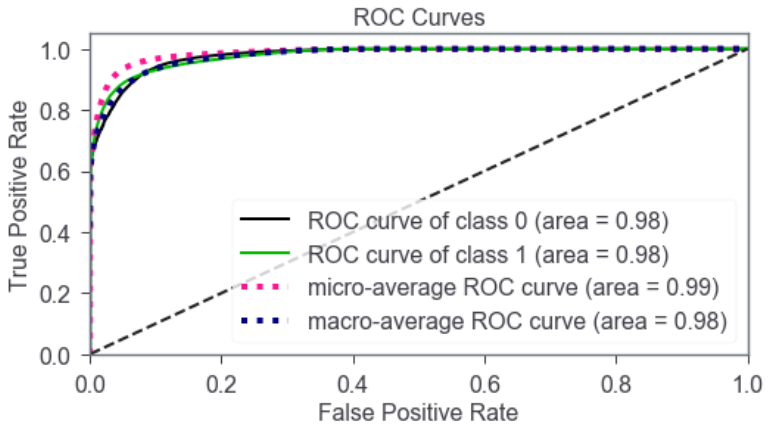
	Kesinlik	**Duyarlılık**	**F1-Skor**	**Toplam Kullanıcı Sayısı**
0	0.97	0.96	0.96	100689
1	0.81	0.88	0.84	21012
Doğruluk			0.94	121701
Makro Ortalama	0.89	0.92	0.90	121701
Ağırlıklı Ortalama	0.94	0.94	0.94	121701
Doğruluk: 0.9422354787553101				

Şekil 6.6'da verilen EdMot Karmaşıklık Matrisi, bir önceki modelde elde edilmiş olan karmaşıklık matrisi ile karşılaştırıldığında yanlış tahmin edilenlerin oranının oldukça azaldığı, doğru tahmin oranlarının da arttığı görülmektedir.



Şekil 6.6 EdMot modeli karmaşıklık matrisi

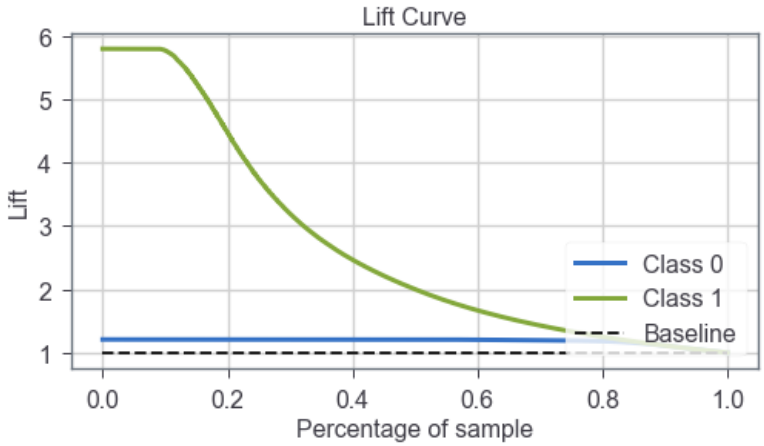
EdMot Modelinden elde edilmiş ROC eğrisi Şekil 6.7'de verilmektedir. Eğrinin altında kalan alan yaklaşık 0.98 olarak belirtilmiştir. Bu da 1'e oldukça yakın bir değerdir ve model hakkında oldukça iyi sonuçlar sunmaktadır. Bir önceki yalın müşteri terk modeline oranla bu modelin daha tercih edilebilir olduğunu vurgulamaktadır.



Şekil 6.7 EdMot modeli ROC eğrisi

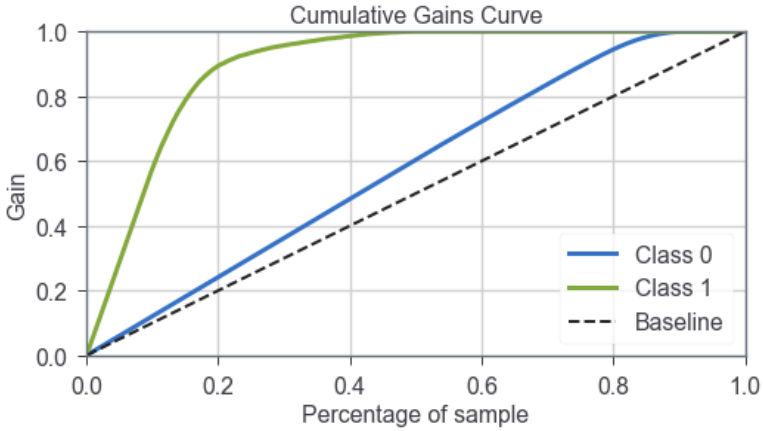
Şekil 6.8'de EdMot modeline ait kaldırma eğrisi verilmektedir. Burada bir önceki yalın terk modeline kıyasla kaldırmanın arttığı görülmektedir. Modelin 0.2'deki artışının yaklaşık 4.5 kat olduğu söylenebilir. Bu da modelin, Class 1'deki %20'lik rastgele seçimine göre müşterilerin 4.5 katını alabileceği anlamına gelmektedir.

Daha yüksek bir kaldırma söz konusu olduğundan EdMot modeli daha iyi bir model olarak nitelendirilebilir.



Şekil 6.8 EdMot modeli kaldırma eğrisi

Şekil 6.9'a göre EdMot modelinin kümülatif kazanç eğrisi, bize yatay eksenin %20 noktasında en yüksek olasılığın yaklaşık %90 olduğunu söylemektedir. Bu nedenle, gözlemlerin %20'si hedeflendiği takdirde, model toplam terk edenlerin %90'ını bu grup içerisinde bulabilecektir. Ayrıca kümülatif kazanç eğrisi yalın terk modeline oranla artış göstermiştir ve bu da EdMot modelinin toplam kazancının daha yüksek olduğunu vurgulamaktadır.



Şekil 6.9 EdMot modeli kümülatif kazanç eğrisi

Bu sonuçlardan da anlaşılacağı gibi, SNA metrikleri kullanılarak yapılan müşteri terk modellerinin başarısı daha yüksek çıkmaktadır.

İkinci olarak; alt grup tespit algoritmalarından bir diğeri olan Label Propagation algoritması ve oluşturulan özniteliklerle, makine öğrenmesi modelimiz eğitildiğinde; modelin tahmin başarısındaki doğruluk sonuçları Şekil 6.5'te, F1-Skor sonuçları ise Şekil 6.14'te verilmektedir.

Tablo 6.3'e göre Label Propagation algoritmasını kullanarak yaptığımız müşteri terk tahmini sonuçları daha önce yapmış olduğumuz EdMot algoritmasından daha başarılı sonuçlar vermiştir. Modelin doğruluk (accuracy) değeri %97 olarak bulunmuştur. Kesinlik, duyarlılık ve F1-Skor değerleri EdMot modelindeki değerlere oranla çok daha yüksek, %90'ın üzerinde, bulunmuştur. Sonuç olarak, en yüksek tahmin başarısını veren alt topluluk tespit algoritmamız Label Propagation algoritması olmuştur.

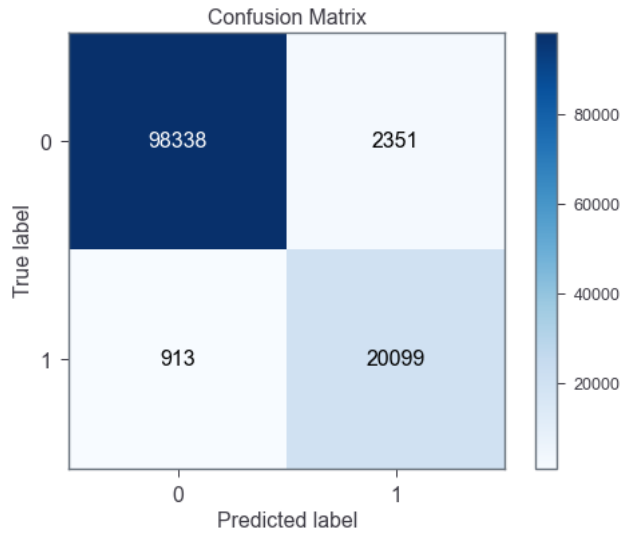
Tablo 6.3 Label Propagation müşteri terk tahmini test seti sonuçları

	Kesinlik	**Duyarlılık**	**F1-Skor**	**Toplam Kullanıcı Sayısı**
0	0.99	0.98	0.98	100689
1	0.91	0.94	0.92	21012
Doğruluk			0.97	121701
Makro Ortalama	0.94	0.97	0.95	121701
Ağırlıklı Ortalama	0.97	0.97	0.97	121701
Doğruluk: 0.9731801710750118				

Label Propagation Karmaşıklık Matrisi Şekil 6.10'da gösterilmektedir. EdMot karmaşıklık matrisi ile kıyaslandığında tahmin edilen değerler ile gerçek değerlerin aynı olduğu değerlerde artış görülmektedir. Yanlış tahmin sayıları yüksek oranda azaltılmıştır. Bu karmaşıklık matrisinde müşterinin sistemi terk ettiği yönünde tahmin de bulunulan fakat gerçekte müşterinin hala sistemde olduğunu gösteren kişi sayısı (yanlış tahmin sayısı) 2351 iken bir önceki matriste

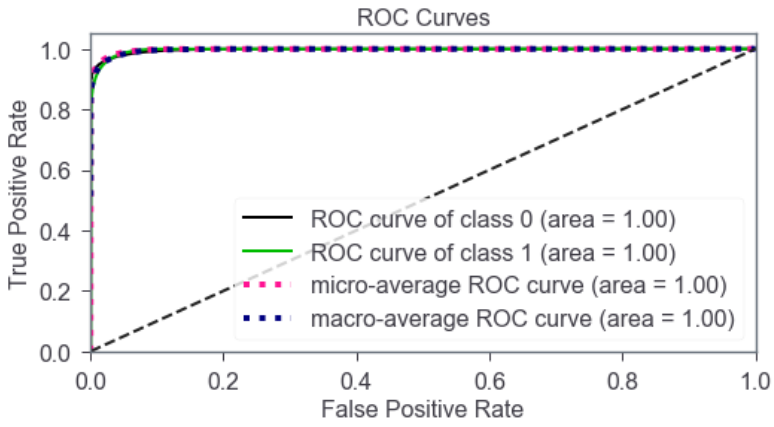
bu değer 4439 idi. Bu da bize bu modelin EdMot modelinden daha doğru tahminler yaptığını net bir şekilde ifade etmektedir.

Bunun yanı sıra model performansını daha doğru şekilde değerlendirebilmek adına Label Propagation modeli için cross validation yöntemi kullanıldı. Bu yöntem ile önceden rastgele olarak ayrılan veri seti, farklı şekillerde bölünüp doğruluk değerleri elde edilir ve elde edilen bu doğruluk değerlerinin ortalaması alınır. Böylece modelin daha doğru bir yaklaşımla ürettiği sonuca odaklanılabilir. Bu nedenle veri seti on parçaya ayrılarak cross validation yöntemi uygulandı ve model bu şekilde eğitildi. Elde edilen doğruluk değerinin 0.97 ile rastgele olarak %30 test ve %70 eğitim setine ayırdığımız modelimiz ile aynı sonucu verdiği gözlemlenmiştir. Bu da modelin bu veri seti üzerinde gerçekten başarılı sonuçlar verdiğini belirtmektedir.



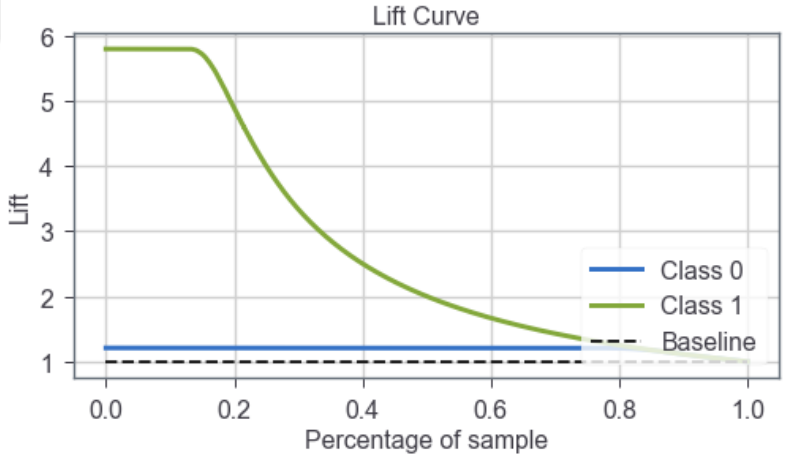
Şekil 6.10 Label Propagation modeli karmaşıklık matrisi

Üçüncü modelimiz olan Label Propagation Modeline ait ROC eğrisi Şekil 6.11'de gösterilmektedir. AUC (eğri altında kalan alan) 1 değerine ulaşmış olarak görülmektedir. Bu da modelin mükemmele yakın bir başarı elde ettiğini ifade etmektedir.



Şekil 6.11 Label Propagation modeli ROC eğrisi

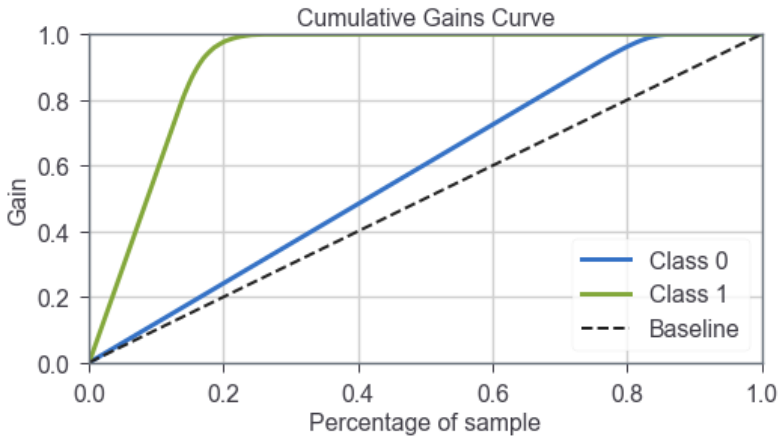
Label Propagation modeline ait Kaldırma Eğrisi Şekil 6.12'de karşımıza çıkmaktadır. Bu kez modelin 0.2'deki artışının neredeyse 5 kat olduğu ve EdMot modelindeki kaldırma değerinden daha yüksek olduğu gözlemlenmektedir. Modelimiz müşteri havuzundan rastgele %20'yi seçmekten yaklaşık 5 kat daha iyi sonuçlar vermektedir. Bu nedenle en iyi modelin Label Propagation modeli olduğu net bir şekilde söylenebilmektedir.



Şekil 6.12 Label Propagation modeli kaldırma eğrisi

Label Propagation modelinin kümülatif kazanç eğrisinin gösterildiği Şekil 6.13'e bakılacak olursa, yatay eksenin %20 noktasında en yüksek olasılığın yaklaşık %99 olduğu gözlemlenmektedir. Gözlemlerin %20'si hedeflendiği takdirde, model toplam terk edenlerin neredeyse %99'unu bu grup içerisinde bulabilecektir. Sonuç

olarak Label Propagation kümülatif kazanç eğrisi diğer modellere oranla en yüksek kazancı sağlamaktadır.



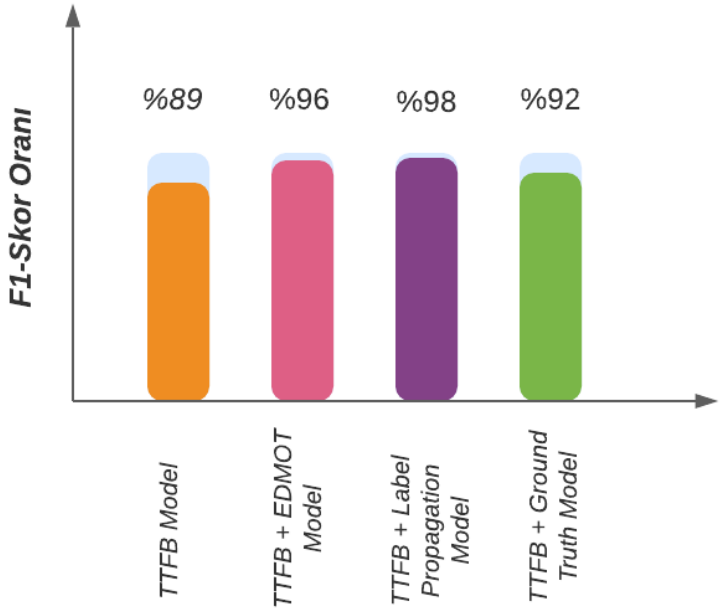
Şekil 6.13 Label Propagation modeli kümülatif kazanç eğrisi

Üçüncü olarak, çalıştırmış olduğumuz EdMot ve Label Propagation alt topluluk tespit algoritmalarına ek olarak, Ground Truth algoritması da kullanılmıştır. EdMot ve Label Propagation algoritmalarının bulduğu ortak alt grupları baz alan Ground Truth algoritması ile bir tahmin modeli daha çalıştırılmıştır ve sonuçları Tablo 6.4'de verilmiştir. Buradan elde etmiş olduğumuz doğruluk (accuracy) değeri yaklaşık olarak %87 ile Şekil 6.5'te yer alan TTFB Modele ait doğruluk sonucundan yine yüksek bulunmuştur. Aynı zamanda Şekil 6.14'te verilen F1-Skor değerleri karşılaştırıldığında da aynı sonuca varılmaktadır.

Tablo 6.4 Ground Truth müşteri terk tahmini test seti sonuçları

	Kesinlik	Duyarlılık	F1-Skor	Toplam Kullanıcı Sayısı
0	0.94	0.90	0.92	100689
1	0.61	0.74	0.67	21012
Doğruluk			0.87	121701
Makro Ortalama	0.78	0.82	0.79	121701
Ağırlıklı Ortalama	0.89	0.87	0.88	121701
Doğruluk: 0.8740062858199298				

Bu sonuçlar ile birlikte, telekom sektöründe yapılan müşteri terk etme davranışı tahminlerinde, SNA metrikleri ve alt topluluk tespiti algoritmaları kullanılarak elde edilen öznitelikler ile oluşturulan makine öğrenmesi modellerinin daha başarılı sonuçlar verdiği gözlemlenmiştir.



Şekil 6.14 F1-skor sonuç grafiği

Sistemi terk etme ihtimali olan müşterilerin tahmin edilmesinin ardından, bu müşterileri sistemde tutabilmek için tutundurma modelleri oluşturulmuştur. Tutundurma modellerinde müşterilerin, çapraz satış analizi ile sıklıkla aldığı paketler belirlenmiştir. Burada birlikteliklerin tespiti için FP-Growth algoritması kullanılmıştır. Bu sonuçlara göre, müşterileri sistemde tutabilmek adına hangi kampanyaların sunulması gerektiği rahatlıkla belirlenebilecektir. Ardından yine aynı müşterilerin bu kez hangi sıra ile hangi paketleri aldıkları incelenmiştir. Burada sıralı birlikteliklerin belirlenmesi için PrefixSpan algoritması tercih edilmiştir. Kampanyalar belirlenirken müşterilerin en son aldığı paketlere bakılarak sıralı birliktelik kurallarında bu paketlerden sonra hangi ürünlerin en çok satın alındığı tespit edilebilir ve müşterilere bu ürünler sunulabilir.

FP-Growth algoritması çalıştırılarak elde edilen kurallar Tablo 6.5'te gösterilmektedir. Burada elde edilen kurallar en az ikili kurallar olacak şekilde belirlenmiştir. Doğru Tahmin alanı, test veri setinde ikinci ürünü alanlar arasında, kurallarda tahmin edilmiş olan ikinci ürünü alanların sayısını ifade etmektedir. Yanlış Tahmin ise, test veri setinde ikinci ürünü almış fakat, kuralda tahmin edilmiş olan ikinci ürünü değil beklediğimizden farklı bir ürünü alanların sayısını

göstermektedir. Başarı alanı ise, Doğru Tahmin alanının, Doğru Tahmin ve Yanlış Tahmin toplamına bölünmesi ile elde edilmiştir. Elde ettiğimiz sonuçlara göre, bulunan birinci kural %83 oranında en başarılı tahmin sonucunu vermektedir. Sonuç olarak, Tablo 6.5'te görüldüğü üzere modelimiz yüksek başarı oranlarını veren kuralları üretebilmektedir.

Tablo 6.5 FP-growth birliktelik kuralları

Birliktelik Kuralları	Doğru Tahmin	Yanlış Tahmin	Başarı
DUMMY OptOUT3 ⟶ Local Bundle 360 Min PRP	2073	397	0,8392
Shabab140 Prepaid Disc ⟶ Shabab140 Add-On Prepaid	1293	445	0,7439
FPSMS ⟶ IVRARPUDMY	1979	1701	0,5377
OSN 2M Promo ⟶ New Shabab59 Add-On Prepaid	6913	7082	0,4939
New Shabab59 Add-On Prepaid ⟶ New Shabab59 Prepaid Disc	3292	4127	0,4437
SIM CARD Replacement ⟶ IVRARPUDMY	1434	2246	0,3896
Local&Data PRPD DCL OUT ⟶ Local&Data PRPD Decile 9	1184	2498	0,3215
DUMMY DOVE Opt OUT ⟶ Internet 50 MB	1395	4427	0,2396
OSN 2M Promo ⟶ New Shabab59 Prepaid Disc	3429	10935	0,2387
OSN 2M Promo, New Shabab59 Add-On Prepaid ⟶ New Shabab59 Prepaid Disc	3292	11072	0,2291
Data Over Voice 500MB PRP ⟶ DUMMYDOVE OptOUT	1244	4474	0,2175
Inernational bndl1 optout ⟶ Int Bundl Pakistan1 CA	1047	4948	0,1746
OSN 2M Promo ⟶ NewShabab149 AddOn Prepaid	1727	12268	0,1234
OSN 2M Promo ⟶ NewShabab99 AddOn Prepaid	1510	12485	0,1078

PrefixSpan algoritması ile elde edilen sıralı kurallar listesi Tablo 6.6'da verilmektedir. Aynı şekilde Doğru Tahmin, ikinci ürünü alanlar arasında belirlemiş olduğumuz ikinci ürünü alanların sayısını vermektedir. Yanlış Tahmin, ikinci

ürünü alan fakat, beklediğimizden farklı bir ürünü alanların sayısını göstermektedir. Başarı ise, doğru tahminlerin oranını ifade etmektedir. Tablo 6.6'da verilmiş olan performans sonuçlarına bakıldığında, %92 başarı oranı ile en yüksek başarı oranına sahip birinci kuralın tespit edildiği görülmektedir. Bu sonuçlar, PrefixSpan modelinin bulmuş olduğu kuralların, test datası üzerinde başarılı olduğunu ifade etmektedir.

Tablo 6.6 Prefix-Span sıralı birliktelik kuralları

Sıralı Birliktelik Kuralları	Doğru Tahmin	Yanlış Tahmin	Başarı
OSN 2M Promo ⟶ New Shabab59 Prepaid Disc	12911	1084	0,9225
New Shabab59 Add-On Prepaid ⟶ New Shabab59 Prepaid Disc	5843	1070	0,8452
Int Bundl Pakistan Daily ⟶ Int Bundl Pakistan Daily	4614	1019	0,8191
OSN 2M Promo ⟶ New Shabab59 Add-On Prepaid	11237	2758	0,8029
Int Bundl Pakistan Daily ⟶ Inernational bndl1 optout	4044	1589	0,7179
New Shabab59 Prepaid Disc ⟶ New Shabab59 Add-On Prepaid	2720	1078	0,7161
New Shabab59 Prepaid Disc ⟶ OSN 2M Promo	2660	1138	0,7003
Inernational bndl1 optout ⟶ Int Bundl Pakistan Daily	3073	1615	0,6555
New Shabab59 Add-On Prepaid ⟶ OSN 2M Promo	4263	2650	0,6166
DUMMY OptOUT3 ⟶ Local Bundle 360 Min PRP	1188	1048	0,5313

7
SONUÇ VE ÖNERİLER

Bu çalışma kapsamında; telekom sektöründe müşteri terk davranışı tahmini problemi üzerinde çalışılmıştır. Müşteri terk etme davranışı, karmaşık ağlar üzerinde gerçekleştirilen sosyal ağ analizi, alt topluluk tespiti algoritmaları kullanılarak elde edilen öznitelikler ve telekom verisinden elde edilen öznitelikler birlikte kullanılarak oluşturulan makine öğrenmesi modeli ile tahmin edilmiştir. Bu amaçla, başarı tahminini gerçekleştirecek bir veri analizi iş süreci yazılımının mimari yapısı önerilmiştir. Önerilen mimari tasarım, CDR veri seti ve telekom verileri üzerinde, özniteliklerin çıkartılması ve bu özniteliklerle makine öğrenmesi modelinin oluşturulması amaçlı olarak geliştirilmiştir. Sadece telekom özniteliklerinin kullanıldığı ve telekom öznitelikleri ile SNA özniteliklerinin bir arada kullanıldığı modeller oluşturulmuştur ve tahmin başarıları incelenmiştir. Elde edilen sonuçlar, SNA kullanılarak yapılan tahmin modelinde daha başarılı sonuçlar üretildiğini göstermektedir.

Ayrıca bu araştırma kapsamında, tutundurma modelleri oluşturulabilmesi için bir veri analizi iş süreci yöntemi de önerilmektedir. Tutundurma modeli kapsamında, çapraz satış analizi ile en çok satılan ürünler tespit edilmektedir. Terk davranışına sahip müşterilere hangi kampanyaların sunulması gerektiğine bu şekilde karar verilebilmektedir. Ayrıca, ürünlerin hangi sıra ile müşterilere sunulması gerektiğine yönelik sıralı örüntülerin belirlendiği bir yöntem de sunulmaktadır. Çalışma kapsamında önerilen iş süreci yazılım mimarisindeki modüller ayrıntılı bir şekilde ele alınmıştır. Bu yazılım mimarisinin bir prototipi gerçekleştirilmiştir. Telekom öznitelikleri arasından öznitelik seçimi yapılmış ve modele giren öznitelikler kategorize edilmiştir.

Bu çalışmanın gelecek adımlarında müşteri terk modeli için kullanılan öznitelikler zenginleştirilecektir. Yapılan deneysel çalışmalarda kullanılan makine öğrenmesi algoritmasının dışında, farklı makine öğrenmesi algoritmaları da kullanarak,

algoritmaların bu veri seti üzerindeki başarıları karşılaştırmalı olarak irdelenecektir. Bunun yanı sıra, veri seti sayısı ve yapılan analizlerin kapsam ve çeşitliliğini de arttırmayı planlıyoruz. Tespit edilen ilişkisel tabanlı kural sayılarının artırılması ve daha karmaşık kuralların belirlenmesi ile kural seti zenginleştirilecektir. Model performanslarının analiz edilebilmesi için farklı algoritmalar kullanılarak, sonuçlar karşılaştırılacaktır.

KAYNAKÇA

[1] E. Kabaklı and Ş. Işıcık, “Review of Turkey's telecommunication sector”, Aksaray university journal of faculty of economics and administrative sciences, vol. 12, no. 1, pp. 35–42, 2020.

[2] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector”, IEEE access, vol. 7, pp. 60134–60149, 2019.

[3] U. T. S. Gursoy, “Customer churn analysis in telecommunication sector”, Istanbul university journal of the school of business administration, vol. 39, no. 1, pp. 35–49, 2010.

[4] F. Shirazi and M. Mohammadi, “A big data analytics model for customer churn prediction in the retiree segment”, International journal of information management, vol. 48, pp. 238–253, 2019.

[5] A. Dumitrache, A. A. M. Nastu, and S. Stancu, “Churn prediction in telecommunication industry: Model interpretability”, Journal of eastern Europe research in business and economics, vol. 2020, pp. 1–11, 2020.

[6] A. Kiani, “Telecom penetration and economic growth: An empirical analysis”, GCU economic journal, vol. LI, no. 1&2, pp. 105–123, 2018.

[7] R. Hendrawan, K. W. Nugroho, and G. T. Permana, “Efficiency perspective on telecom mobile data traffic”, Journal of business and economics review, vol. 5, pp. 38–44, 2020.

[8] M. J. Baeth and M. S. Aktaş, “Detecting misinformation in social networks using provenance data”, Concurrency and computation: Practice and experience, vol. 31, no. 3, 2019.

[9] M. J. Baeth and M. S. Aktaş, “An approach to custom privacy policy violation detection problems using big social provenance data”, Concurrency and computation: Practice and experience, vol. 30, no. 21, 2018.

[10] A. Tüfek, A. Gürbüz, Ö. F. Ekuklu, and M. S. Aktaş, “Provenance collection platform for the weather research and forecasting model”, in Proceedings of the 14th international conference on Semantics, Knowledge and grids, 2018.

[11] M. Riveni, T. D. Nguyen, M. S. Aktaş, and S. Dustdar, “Application of provenance in social computing: A case study”, Concurrency and computation: Practice and experience, vol. 31, no. 3, 2019.

[12] Y. Taş, M. J. Baeth, and M. S. Aktaş, “An approach to standalone provenance systems for big social provenance data”, in Proceedings of the 12th international conference on Semantics, Knowledge and grids, 2016, pp. 9–16.

[13] M. S. Aktas, G. C. Fox, and M. Pierce, “Fault tolerant high performance information services for dynamic collections of grid and web services”, Future generation computer systems, vol. 23, no. 3, pp. 317–337, 2007.

[14] C. Salman, “Yeni bir ağ merkezilik ölçütü: Göreceli kenar önemi metodu”, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2018.

[15] K. Sultan, H. Ali, and Z. Zhang, “Call detail records driven anomaly detection and traffic prediction in mobile cellular networks”, IEEE access, vol. 6, pp. 41728–41737, 2018.

[16] N. Ruan, Z. Wei, and J. Liu, “Cooperative fraud detection model with privacy-preserving in real cdr datasets”, IEEE access, vol. 7, pp. 115261–115272, 2019.

[17] G. Zhang, X. Rui, S. Poslad, X. Song, Y. Fan, and B. Wu, “A method for the estimation of finely-grained temporal spatial human population density distributions based on cell phone call detail records”, Remote sensing, vol. 12, no. 16, pp. 1–28, 2020.

[18] R. Kozik, M. Choras, M. Pawlicki, A. Pawlicka, W. Warczak, and G. Mazgaj, “Proposition of innovative and scalable information system for call detail records analysis and visualisation”, in Proceedings of the 13th international conference on computational intelligence in security for information systems, CISIS, 2020, pp. 174–183.

[19] F. Altunbey and B. Alataş, “Overlapping community detection in social networks using parliamentary optimization algorithm”, International journal of computer networks and applications, IJCNA, vol. 2, no. 1, pp. 12–19, 2015.

[20] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, “Community detection in large-scale social networks”, in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis (WebKDD/SNAKDD’07), 2007, pp. 16–25.

[21] M. S. Khan, A. W. A. Wahab, T. Herawan, G. Mujtaba, S. Danjuma and M. A. Al-Garadi, “Virtual community detection through the association between prime nodes in online social networks and its application to ranking algorithms”, IEEE access, vol. 4, pp. 9614–9624, 2016.

[22] L. Runpeng, H. Jun, and W. Xiaofan, “VCD: A network visualization tool based on community detection”, in Proceedings of the 12th international conference on control, Automation and systems, 2012, pp. 1221–1226.

[23] M. Crampes and M. Plantie, “A unified community detection, visualization and analysis method”, Advances in complex systems, vol. 17, no. 1, pp. 1–25, 2014.

[24] J. David Cruz, C. Bothorel, and F. Poulet, “Community detection and visualization in social networks: Integrating structural and semantic information”, ACM transactions on intelligent systems and technology, vol. 5, no. 1, pp. 11–26, 2013.

[25] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, “A tale of one city: Using cellular network data for urban planning”, IEEE pervasive computing, vol. 10, no. 4, pp. 18–26, 2011.

[26] F. Calabrese, L. Ferrari, and V. D. Blondel, “Urban sensing using mobile phone network data: A survey of research”, ACM computing surveys, vol. 47, no. 2, pp. 1–23, 2014.

[27] L. Alexander, S. Jiang, M. Murga, and M. C. Gonzalez, “Origin–destination trips by purpose and time of day inferred from mobile phone data”, Transportation research part c: Emerging technologies, vol. 58, pp. 240–250, 2015.

[28] O. Järv, R. Ahas, E. Saluveer, B. Derudder, and F. Witlox, “Mobile phones in a traffic flow: A geographical perspective to evening rush hour traffic analysis using call detail records”, Plos one, vol. 7, no. 11, pp. 1–12, 2012.

[29] A. Bogomolov, B. Lepri, R. Larcher, F. Antonelli, F. Pianesi, and A. Pentland, “Energy consumption prediction using people dynamics derived from cellular network data”, EPJ data science, vol. 5, no. 1, pp. 1–15, 2016.

[30] K. Kim, C. Jun, and J. Lee, “Improved churn prediction in telecommunication industry by analyzing a large network”, Expert systems with applications, vol. 41, no. 15, pp. 6575–6584, 2014.

[31] M. S. Aktaş and M. J. Baeth, “On the detection of information pollution and violation of copyrights in the social web”, in Proceedings of the 8th international conference on service-oriented computing and applications (SOCA-15), 2015, pp. 252–254.

[32] B. Dündar, M. Astekin, and M. Aktaş, “A big data processing framework for self healing internet of things applications”, in Proceedings of the 12th international conference on Semantics, Knowledge and grids (SKG-16), 2016, pp. 62–68.

[33] M. S. Aktas and M. Astekin, “Provenance aware run-time verification of things for self-healing internet of things applications”, Concurrency and computation practice and experience, vol.31, no. 3, 2019.

[34] M. S. Aktas, “Hybrid cloud computing monitoring software architecture”, Concurrency and computation practice and experience, vol. 30, no. 21, pp. 1–9, 2018.

[35] Y. Tas, M. J. Baeth, and M. Aktas, “An approach to standalone provenance systems for big provenance data”, in Proceedings of the 12th international conference on Semantics, Knowledge and grids (SKG-16), 2016, pp. 9–16.

[36] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks”, Physical review E, vol. 76, no. 3, pp. 1–11, 2007.

[37] P. Li, L. Huang, C. Wang, and J. Lai, “EdMot: An edge enhancement approach for motif-aware community detection”, in Proceedings of the 25th ACM

SIGKDD international conference on knowledge discovery and data mining (KDD'19), Association for computing machinery, 2019, pp. 479–487.

[38] K. W. Afrassa, G. Cosgun, U. F. Gursoy, E. M. Yildiz, and M. S. Aktaş, "On the community discovery methods for complex networks: A case study", in proceedings of the 2020 federated conference on computer science and information systems, ACSIS, vol. 21, 2020, pp. 473–477.

[39] M. S. Aktas and M. Kapdan, "Structural code clone detection methodology using software metrics", International journal of software engineering and knowledge engineering, vol. 26, no. 2, pp. 307–332, 2016.

[40] K. Eria and B. P. Marikannan, "Significance-based feature extraction for customer churn prediction data in the telecom sector", Journal of computational and theoretical nanoscience, vol. 16, pp. 1–4, 2019.

[41] D. Do, P. Huynh, P. Vo, and T. Vu, "Customer churn prediction in an internet service provider", in IEEE international conference on big data (Big data), Boston, MA, USA, 2017.

[42] J. Pamina, J. Beschi Raja, S. Sathya Bama, S. Soundarya, M. S. Sruthi, S. Kiruthika, V. J. Aiswaryadevi, and G. Priyanka, "An effective classifier for predicting churn in telecommunication", Journal of advanced research in dynamical and control systems, vol. 11, no. 1, pp. 221–229, 2019.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. K. W. Afrassa, G. Cosgun, U. F. Gursoy, E. M. Yildiz, and M. S. Aktaş, “On the community discovery methods for complex networks: A case study”, in Proceedings of the 2020 federated conference on computer science and information systems, vol. 21, 2020, pp. 473–477.

2. U. F. Gursoy, M. Kurulay, and M. S. Aktas, “Customer churn behaviour predicting using social network analysis techniques: A case study”, in Proceedings of the 3rd international conference on electrical, Communication and computer engineering (ICECCE), 12-13 June 2021.

3. (kabul edildi) U. F. Gursoy, E. M. Yildiz, M. S. Aktas, and M. E. Okay, “Customer churn prediction and promotion models in the telecom sector: A case study”, in IntelliSys 2021, 2-3 September 2021.