Makine Öğrenmesi Yöntemleri Kullanılarak Müşteri Kaybı Tahmini: Karşılaştırmalı Bir Analiz

Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis

Hamdullah Karamollaoğlu
Computer Engineering
Düzce University
Düzce, Turkey
hkaramollaoglu@gmail.com

İbrahim Yücedağ
Computer Engineering
Düzce University
Düzce, Turkey
ibrahimyucedag@duzce.edu.tr

İbrahim Alper Doğru
Computer Engineering
Gazi University
Ankara, Turkey
iadogru@gazi.edu.tr

2021 6th International Conference on Computer Science and Engineering (UBMK) | 978-1-6654-2908-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/UBMK52708.2021.9558876

***Öz*—Müşteri kaybı analizi, özellikle telekomünikasyon, finans, sigortacılık vb. sektörlerde, çeşitli nedenlerle aldığı hizmeti (aboneliği) iptal etme eğilimi gösteren müşterilerin önceden tahmin edilerek bu iptal işleminin önüne geçilmesi için gerekli operasyonel adımların belirlenmesi işlemidir. Bu çalışmada, telekomünikasyon sektöründe aboneliği iptal etme eğilimi gösteren müşterilerin tespiti işlemi için kaggle.com'dan elde edilen iki ayrı verisetinden yararlanılmıştır. İlgili verisetleri üzerinde makine öğrenmesi yöntemlerinden Lojistik Regresyon, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, Destek Vektör Makineleri, Adaboost, Çok Katmanlı Algılayıcılar ve Naive Bayes metotları uygulanarak analiz işlemi gerçekleştirilmiştir. Her iki veriseti üzerinde gerçekleştirilen müşteri kaybı analizinde en başarılı yöntemin Rastgele Orman metodu olduğu görülmüştür.**

Anahtar Sözcükler—müşteri kaybı analizi; makine öğrenmesi; telekomünikasyon.

***Abstract*— Customer churn analysis is the process of predicting customers who tend to cancel the service (subscription) they receive for various reasons, especially in sectors such as telecommunications, finance and insurance, and determining the necessary operational steps to prevent this cancellation. The study used two separate datasets from kaggle.com to identify customers who tend to unsubscribe in the telecommunications industry. The analysis process was carried out by applying machine learning methods such as Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Support Vector Machines, AdaBoost, Multi-Layer Sensors and Naive Bayes methods on the relevant datasets. It was seen that the most successful method in the customer loss analysis performed on both datasets was the Random Forest method.**

Keywords—customer churn analysis; machine learning; telecommunication.

I. GİRİŞ

Müşteri kaybı analizi, özellikle telekomünikasyon, finans ve sigortacılık gibi sektörlerde çeşitli nedenlerle ürünü veya hizmeti kullanmaya son vererek aboneliğini iptal etme eğiliminde olan müşterilerin önceden tahmin edilmesi ve bu iptal işleminin önüne geçilmesi için gerekli operasyonel adımların belirlenmesi işlemidir. Hizmet sektöründeki rekabet ortamında yeni müşteri kazanılması kadar, mevcut müşterilerin hizmet almayı sürdürmelerinin sağlanması da önem arz etmektedir. Firmalarca yeni müşteri kazanmak için ayrılan pazarlama, altyapı, promosyon, reklam çalışmalarının maliyeti mevcut müşteriyi elde tutmanın maliyetinden çok daha fazladır [1].

Müşteri kayıpları, gönüllü ve gönülsüz kayıp olmak üzere ikiye ayrılmaktadır. Gönüllü kayıp, bir müşterinin hizmet aldığı hizmet sağlayıcıdan, daha fazla avantaj sağladığını düşündüğü başka bir hizmet sağlayıcıya kendi isteği ile geçiş yapmasıdır. Gönülsüz kayıp ise; bir müşterinin kendi isteği dışında oluşan çeşitli ekonomik ve çevresel faktörler nedeniyle hizmet sağlayıcıdan aldığı hizmetin sonlanması durumudur [2]. Gönülsüz müşteri kayıpları, önüne geçilmesi çok zor veya imkansız durumlar nedeniyle meydana geldiği için, müşteri kaybı analizi içerisinde ele alınmamaktadır.

Müşterilerin kişisel özellikleri, hizmet kullanımındaki çeşitli alışkanlıkları ve davranışları ile ilgili elde edilen veriler, çeşitli veri madenciliği ve istatistiki yöntemleri ile analiz edilerek, aldığı hizmeti sonlandırma potansiyeli olan müşteriler tespit edilebilmektedir [3].

Bu çalışmada, telekomünikasyon sektörü için müşteri kaybı analizi ele alınmıştır. Bu amaçla, kaggle.com'dan elde edilen ve müşterilerin bazı kişisel özellikleri ve alınan hizmeti kullanım durumları ile ilgili veriler içeren iki adet verisetinden yararlanılmıştır. İlgili verisetleri üzerinde makine öğrenmesi yöntemleri kullanılarak müşteri kaybı analizi gerçekleştirilmiştir.

Çalışmanın ikinci bölümünde müşteri kaybı analizine ilişkin literatür taraması yapılmıştır. Üçüncü bölümde çalışmada kullanılan verisetleri üzerinde makine öğrenmesi yöntemlerinin uygulanması sonucunda elde edilen bulgular paylaşılmıştır. Son bölümde ise elde edilen sonuçlar üzerinde değerlendirmeler yapılmış ve gelecek çalışmalardan bahsedilmiştir.

II. LİTERATÜR TARAMASI

Literatür incelendiğinde, müşteri kaybı analizi ile ilgili çalışmaların genellikle telekomünikasyon, finans, sigortacılık, e-ticaret vb. alanlarda yapıldığı görülmektedir. Müşteri kaybı analizi ile ilgili çalışmalardan bazıları aşağıda sunulmuştur.

Koca vd. tarafından yapılan çalışmada [4] mobil bir sadakat uygulaması yardımıyla elde edilen ve parakende sektörü hakkında 2745, restoran hakkında 15739 ve e-ticaret hakkında 1111 adet veri içeren bir veriseti üzerinde Lojistik Regresyon ve Yapay Sinir Ağları yöntemleri kullanılarak müşteri kayıp tahmini gerçekleştirilmiştir. Sonuçta %90 oranında bir doğruluk oranı ile sınıflandırma işlemi gerçekleştirilmiştir.

Wanchai tarafından yapılan çalışmada [5] telekomünikasyon sektöründe müşteri kaybı analizi

978-1-6654-2908-5/21/$31.00 ©2021 IEEE

gerçekleştirilmiştir. Bu amaçla içerisinde müşteri özellikleri ve alışkanlıklarının mevcut olduğu ve müşterilerin hizmetten ayrılma eğilimlerine göre sınıflandırılmış 262500 veri içeren bir veriseti kullanılmıştır. İlgili veriseti üzerinde C4.5 algoritması uygulanmış ve sonuçta %93,7 oranında bir doğruluk oranı ile sınıflandırma işlemi gerçekleştirilmiştir.

Khan vd. tarafından müşteri kaybı analizi çalışmasında Pakistan'da yer alan bir telekomünikasyon şirketinden elde edilen verilerle oluşturulmuş bir veriseti kullanılmıştır. Bu verisetinde müşterilere ait çeşitli demografik veriler ile fatura bilgilerin yer aldığı 26 adet nitelik bulunmakta ve veriseti toplamda 20468 adet veriden oluşmaktadır. Yapay Sinir Ağları kullanılarak yapılan analiz çalışmasında %79'luk bir sınıflandırma başarımı elde edilmiştir [6].

Göy vd. tarafından yapılan internet servis sağlayıcısı için müşteri kaybı analizi çalışmasında [7], 2016-2017 yılları arasında TurkNet firmasına abone olan kullanıcılara ait 38 adet özelliği içeren, müşterilerin hizmetten ayrılma eğitimlerine göre sınıflandırılan ve toplam 20000 adet veriden oluşan bir veriseti kullanılmıştır. İlgili veriseti üzerinde çeşitli makine öğrenmesi yöntemleri uygulanarak yapılan analiz çalışmasında en başarılı yöntemin %93,6 oranında bir doğrulukla sınıflandırma işlemi gerçekleştiren Rastgele Orman yöntemi olduğu görülmüştür.

Kaynar vd. tarafından yapılan çalışmada [8], müşteri kaybı analizi amacıyla çağrı merkezinden elde edilen müşteri kayıtlarından derlenerek oluşturulan, 21 adet özellik ve 4667 adet veri içeren verisetinden yararlanılmıştır. İlgili veriseti üzerinde makine öğrenmesi yöntemlerinden Naive Bayes, Destek Vektör makineleri (DVM) ve Yapay Sinir Ağları (YSA) yöntemleri uygulanmıştır. Söz konusu çalışmada, hizmetten ayrılma eğilimi gösteren müşterilerin tespiti işlemini %91,35'lik doğruluk oranı ile gerçekleştiren YSA'nın diğer yöntemlerden daha başarılı olduğu görülmüştür.

Yıldız vd. tarafından bireysel emeklilik şirketinin sistemine dahil olan müşteriler kullanılarak yapılan müşteri kaybı çalışmasında [9], şirket veritabanı kullanılarak oluşturulan ve müşterilerin karakteristik özelliklerini içeren 75 adet özniteliğin yer aldığı ve yaklaşık 80000 adet veriden oluşan verisetinden yararlanılmıştır. İlgili veriseti üzerinde Karar Ağacı, Rastgele Orman, Bayes ve XGBoost algoritmaları kullanılarak yapılan analizde, sistemden ayrılma eğilimi gösteren müşterilerin tespitinde en başarılı yöntemin %95,96'lık doğruluk oranı ile XGBoost algoritmasının olduğu görülmüştür.

Müşteri kaybı analizi ile ilgili çalışmalar incelendiğinde, çalışmaların genellikle makine öğrenmesi yöntemleri kullanılarak gerçekleştirildiği görülmektedir.

III. MATERYAL VE METOT

Bu çalışmada, müşteri kaybı analizi için kaggle.com'dan elde edilen 'Telco Customer Churn' [10] ve 'Customer Churn Prediction 2020' [11] isimli verisetlerinden yararlanılmıştır. Makine öğrenmesi yöntemleri bu verisetleri üzerinde ayrı ayrı uygulanarak abonelikten ayrılma eğilimi gösteren müşterilerin tespiti gerçekleştirilmiştir. Elde edilen sonuçlara göre makine öğrenmesi yöntemlerinin başarım performansları karşılaştırmalı olarak sunulmuştur.

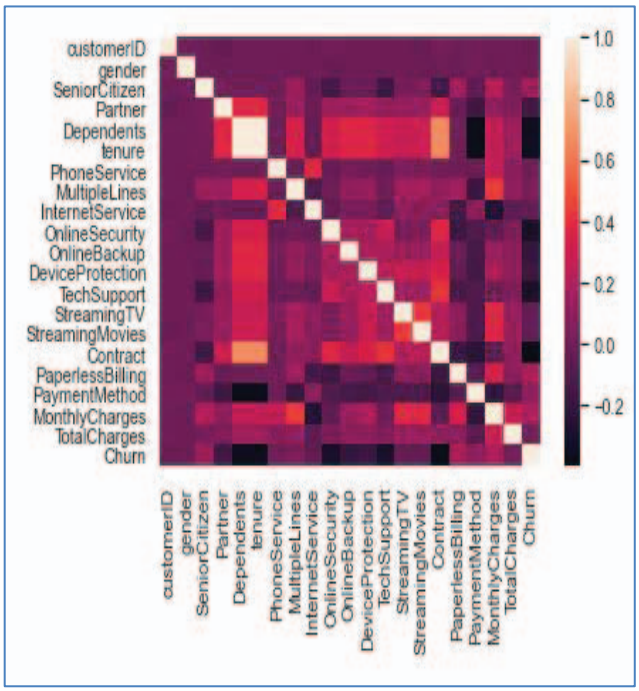
A. Çalışmada Kullanılan Verisetleri

1) Telco Customer Churn Veriseti: Kaggle.com'da yer alan bu verisetinde 7043 adet veri yer almaktadır. Verisetindeki her bir veri, müşterilerin karakteristiği hakkında bilgi veren 20 adet öznitelik ve 1 adet sonuç değerinden oluşmaktadır. Verisetinin içeriğinde yer alan verilerin içerik özniteliklerine ve sonuç sınıfına ait bilgiler Tablo 1'de görülmektedir.

TABLO I. 'TELCO CUSTOMER CHURN' VERİSETİ İÇERİK BİLGİLERİ

Öznitelik	Açıklama
customerID	Müşteri ID
Gender	Cinsiyet
SeniorCitizen	Kıdemli müşteri mi?
Partner	Ortaklık
Dependents	Muhtaçlık (kullanım mecburiyeti)
tenure	Abonelik süresi (ay)
PhoneService	Telefon hizmeti
MultipleLines	Çoklu hat
InternetService	İnternet hizmeti
OnlineSecurity	Çevrimiçi güvenlik
OnlineBackup	Çevrimiçi yedekleme
DeviceProtection	Cihaz koruma
TechSupport	Teknik destek
StreamingTV	TV yayını aboneliği
StreamingMovies	Film kanalları aboneliği
Contract	Sözleşme süresi
PaperlessBilling	Kağıtsız faturalandırma
PaymentMethod	Fatura ödeme şekli
MonthlyCharges	Aylık Ödeme tutarı
TotalCharges	Toplam ödeme tutarı
Churn	Ayrılma durumu

Tablo I'de yer alan özniteliklerin birbirleri arasındaki kolerasyon ilişkisini gösteren Spearman kolerasyon matrisine ait ısı haritası grafiği Şekil 1'de görülmektedir.



Şekil 1. Spearman kolerasyon matrisine ait ısı haritası

Şekil 1'de görüldüğü gibi, sonuç sınıf "Churn" ile en güçlü pozitif yönlü kolerasyon ilişkisi "PaperlessBilling" (0.191) ve "MonthlyCharges" (0.183) arasındadır. "Churn" ile en güçlü

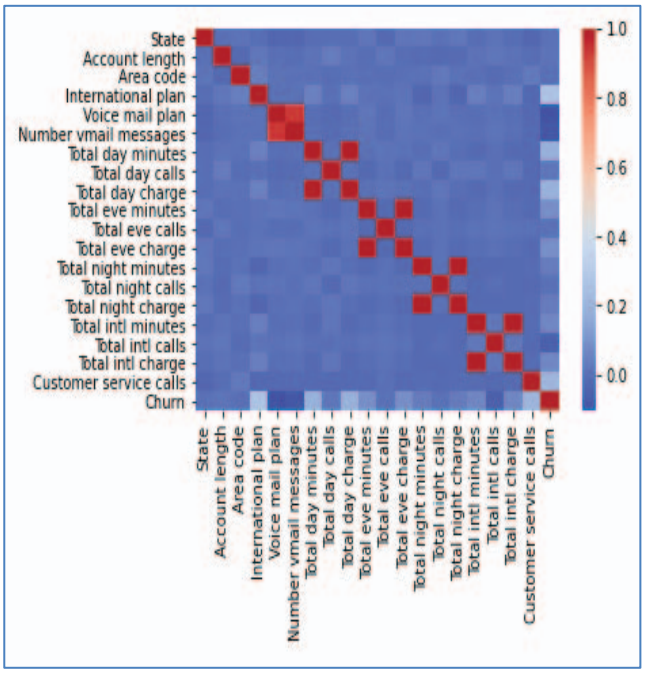
negatif yönlü kolerasyon ilişkisi ise “Dependents” (-0.352) ve “tenure” (-0.352) arasındadır.

2) Customer Churn Prediction Veriseti: Kaggle.com’dan elde edilen bu verisetinde 3333 adet veri yer almaktadır. Bu verisetinde yer alan verilerin 2850 tanesi ‘normal’ 483 tanesi ‘churn’ olarak etiketlenmiştir. Verisetindeki her bir veri, müşterilerin karakteristiği hakkında bilgi veren 19 adet öznitelik ve 1 adet sonuç değerinden oluşmaktadır. Verisetinin içeriğinde yer alan verilerin içerik özelliklerine ve sonuç sınıfına ait bilgiler Tablo II’de görülmektedir

TABLO II. ‘CUSTOMER CHURN PREDICTION’ VERİSETİ İÇERİK BİLGİLERİ

Öznitelik	Açıklama
State	İkamet edilen yer bilgisi
AccountLength	Hesabın aktif olduğu toplam süre
AreaCode	İkamet ettiği yere ait alan kodu
InternationalPlan	Uluslararası abonelik planı bilgisi
VoiceMail Plan	Sesli mesajlaşma kullanım durumu
NumberOfVoiceMailMessages	Kullanılan sesli mesaj sayısı
TotalDayMinutes	Günlük aramalarda geçiririlen toplam dakika
TotalDayCalls	Günlük gerçekleştirilen arama sayısı
TotalDayCharge	Günlük konuşmalara ait ücret tutarı
TotalEveningMinutes	Akşam saatlerinde aramalarda geçirilen toplam dakika bilgisi
TotalEveningCalls	Akşam saatlerinde gerçekleştirilen toplam arama sayısı
TotalEveningCharge	Akşam saatlerindeki aramalara ait toplam ücret tutarı
TotalNightMinutes	Gece saatlerindeki aramalarda geçirilen toplam dakika
TotalNightCalls	Gece saatlerinde gerçekleştirilen toplam arama sayısı
TotalNightCharge	Gece saatlerindeki aramalara ait ücret tutarı
TotalInternationalMinutes	Uluslararası aramalara ait toplam dakika bilgisi
TotalInternationalCalls	Toplam uluslararası arama sayısı
TotalInternationalCharge	Uluslararası aramalara ait ücret tutarı
TotalCustomerServiceCalls	Müşteri hizmetleri aramalarının sayısı
Churn	Abonelikten ayrılma durumu

Tablo II’de yer alan özellik bilgilerine ilişkin elde edilen Spearman kolerasyon matrisine ait ısı haritası grafiği Şekil 2’de görülmektedir.



Şekil 2. Spearman kolerasyon matrisine ait ısı haritası

Şekil 2’de görüldüğü gibi, sonuç sınıf olan “Churn” ile en güçlü pozitif yönlü kolerasyon ilişkisi “International Plan” (0.259) ve “Customer Service Calls” (0.208) arasındadır. “Churn” ile en güçlü negatif yönlü kolerasyon ilişkisi ise “Voice Mail Plan” (-0.102) ve “Number vmail messages” (-0.805) arasındadır.

B. Çalışmada Kullanılan Makine Öğrenmesi Yöntemleri

1) Lojistik Regresyon: Lojistik Regresyon, bağımlı değişkenin kategorik şekilde iki durumlu süreksiz değer olarak ölçüldüğü ve bağımlı değişkeni belirleyen bir veya daha fazla bağımsız değişkenin olduğu durumlarda bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi açıklamak için kullanılan istatistiksel bir yöntemdir [12].

2) K-En Yakın Komşuluk (KNN): KNN metodunda, sınıflandırılmak istenen yeni bir verinin, daha önceden sınıfları belirlenmiş olan verisetindeki verilere olan vektörel uzaklığı hesaplanılarak, k sayıda en yakın komşuluk mesafesine göre değerlendirme yapılmak suretiyle hangi sınıfa dahil olacağı belirlenmektedir [13].

3) Karar Ağaçları: Karar ağaçlarında sınıflandırma modeli ağaç yapısı şeklinde oluşturularak veriler bu ağaç üzerine işlenmektedir. Bu sınıflandırmada verilere ait her nitelik bir düğüm ile temsil edilmektedir. Ağaç yapısı kök düğümden başlayıp ara düğümler tarafından dallanacağı yön belirlenerek yaprak düğüme ulaşıncaya kadar ilerlemektedir. En sonunda bir sınıfı ifade eden ve son düğüm olarak adlandırılan yaprağa ulaşılmaktadır [14].

4) Rastgele Orman: Rastgele Orman algoritması, karar ağaçlarını temel almaktadır. Tek bir ağaç yapısı kullanmak yerine çoklu ağaçlardan meydana gelen bir orman oluşturulmaktadır. Yeni bir veriyi sınıflandırmak için ormandaki her ağaca bir giriş vektörü verilmektedir. Ardından her bir ağaç, o ağaçta kullanılan tahmin değişkenlerine göre bir sonuç üretmektedir. Üretilen sonuç değerlerine göre her ağaç bir sınıfı oylamaktadır. En çok oyu

alan sınıf belirlenerek yeni veri bu sınıfa dahil edilmektedir [15].

5) Destek Vektör Makineleri: Destek vektör makineleri, sınıflandırma, regresyon analizi ve aykırı değer tespitinde kullanılan denetimli öğrenme metodudur. Bu yöntem ile bir takım lineer parametreli fonksiyonlarla modellenen verileri optimal bir şekilde sınıflara ayıracak n boyutlu bir hiperdüzlemin oluşturulmasını amaçlanmaktadır. Burada n değeri verilerin sahip oldukları özellik sayısıdır [16].

6) Adaboost: AdaBoost, ardışık bir takım zayıf sınıflandırıcıları birleştirerek güçlü bir sınıflandırıcı elde etmeyi amaçlayan bir topluluk öğrenme metodudur. Her bir sınıflandırıcı kendisinden önce gelen sınıflandırıcı tarafından yanlış sınıflandırılan eğitim örnekleri üzerinde durur. Her aşamada ilgili sınıflandırıcının doğruluk başarım oranına göre ağırlık seti güncellenmekte ve bir sonraki iterasyon için geri besleme aşamasında güncel ağırlık değerleri ile hesaplamalar yapılmaktadır. Yüksek doğruluğa sahip sınıflandırıcılar yüksek ağırlık değeri almaktadırlar [17].

7) Çok Katmanlı Algılayıcılar: Çok katmanlı algılayıcılar Yapay Sinir Ağları'nın (YSA) bir sınıfıdır. Çok katmanlı bir algılayıcı (MLP) topolojik olarak giriş katmanı, ara katman ve çıkış katmanı olmak üzere üç ana katmandan oluşmaktadır. Girdi katmanı dış dünyadan ağa veri aktarımı sağlamaktadır. Bu katman üzerinde herhangi bir hesaplama işlemi gerçekleştirilmemektedir. Girdi katmanında yer alan her bir düğüm ara katmandaki tüm düğümler ile bağlantılıdır. Ara katman, girdi katmanından gelen verilerin çeşitli hesaplama işlemleri yardımıyla işlenip çıkış katmanına iletilmesini sağlamaktadır. Ağ üzerinde bir veya daha fazla ara katman yer alabilmektedir. Çıkış katmanındaki her bir düğüm bir önceki katmanda yer alan tüm düğümler ile bağlantılıdır. Çıkış katmanı, ara katmandan gelen işlenmiş yeni verileri dış dünyaya aktarma görevini yerine getirmektedir [18].

8) Naive Bayes: Naive Bayes Metodu, Bayes teoremini esas alan bir yöntemdir. Bayes Teoremi bir olayı meydana getiren her bir etkenin, ilgili olaya olan etki olasılığının hesaplanması ve olayın meydana gelmesinde belirleyiciliği fazla olan etkenin tespit edilmesi esasına dayanmaktadır. Naive Bayes yöntemi ile bağımsız değişkenler ve ulaşılmak istenilen hedef değişken arasındaki ilişki önsel (marjinal) olasılık ve koşullu olasılık hesaplamaları yardımıyla belirlenmektedir. Sisteme sunulan yeni girdiler ilgili olasılık hesaplamalarından elde edilen değerler ışığında sınıflandırılmaktadır [19].

C. Deneysel Çalışma

Bu çalışmada 'python sklearn' kütüphanesinden yararlanılmıştır. İlgili verisetlerinde yer alan veriler %70'i eğitim ve %30'u test verisi olacak şekilde rastgele olarak ayrıştırılarak, veriler üzerinde makine öğrenmesi yöntemlerinden Lojistik Regresyon, K-En Yakın Komşu, Karar Ağaçları (k=3 seçilmiştir), Rastgele Orman, Destek Vektör Makineleri, AdaBoost (n_estimators=5, learning_rate=1 seçilmiştir), Çok Katmanlı Algılayıcılar ve Naïve Bayes metotları uygulanmıştır. İlgili makine öğrenmesi yöntemlerinin sınıflandırma performansları, Kesinlik, Duyarlılık, F1 Skoru ve Doğruluk ölçütlerine göre değerlendirilmiştir. Bu ölçütler ile ilgili hesaplamalar sırasıyla (1), (2), (3) ve (4)'teki gibidir.

$$Kesinlik\ (K) = (DP)/(DP + YP) \quad (1)$$

$$Duyarlılık\ (D) = (DP)/(DP + YN) \quad (2)$$

$$F1\ Skoru = 2 * [(K * D)/(K + D)] \quad (3)$$

$$Doğruluk = (DP + DN)/(DP + DN + YP + YN) \quad (4)$$

Burada DP (Doğru Pozitif), sınıflandırıcının doğru olarak sınıflandırdığı 'Non-Churn' (abonelikten ayrılmamış) değerlerinin sayısını, DN (Doğru Negatif) sınıflandırıcının doğru olarak sınıflandırdığı 'Churn' (abonelikten ayrılmış) değerinin sayısını, YP (Yanlış Pozitif) sınıflandırıcının yanlış olarak sınıflandırdığı 'Non-Churn' değerlerinin sayısını, YN (Yanlış Negatif) sınıflandırıcının yanlış olarak sınıflandırdığı 'Churn' değerlerinin sayısını ifade etmektedir. Kesinlik, pozitif olarak tahmin edilmesi gereken durumların ne kadarının pozitif olarak tahmin edildiğini ifade eden bir metriktir. Duyarlılık, pozitif durumların tahmin başarısını gösteren bir metriktir. F1 Score, Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır.

Tablo III'te performans ölçütlerinin hesaplanmasında yararlanılacak olan ve çalışmada kullanılan 'Telco Customer Churn' veriseti üzerinde uygulanan makine öğrenmesi yöntemlerinden elde edilen sonuçlara göre oluşturulmuş karmaşıklık matrisi görülmektedir.

TABLO III. KARMAŞIKLIK MATRİSİ (TELCO CUSTOMER CHURN)

Logistik Regresyon		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1388 (DP)	181 (YN)
	Churn	265 (YP)	279 (DN)
K- En Yakın Komşu (KNN)		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1322 (DP)	247 (YN)
	Churn	336 (YP)	208 (DN)
Karar Ağaçları		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1276 (DP)	293 (YN)
	Churn	263 (YP)	281 (DN)
Rastgele Orman		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1410 (DP)	159 (YN)
	Churn	284 (YP)	260 (DN)
Destek Vektör Makineleri (DVM)		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1567 (DP)	2 (YN)
	Churn	544 (YP)	20 (DN)
AdaBoost		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1178 (DP)	391 (YN)
	Churn	148 (YP)	396 (DN)
Çok Katmanlı Algılayıcılar (ÇKA)		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1559 (DP)	10 (YN)
	Churn	515 (YP)	29 (DN)
Naïve Bayes		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	1172 (DP)	397 (YN)
	Churn	156 (YP)	388 (DN)

Tablo III'teki değerler ışığında, (1), (2), (3), (4)'ten yararlanılarak yapılan hesaplamalara ait sonuç değerler Tablo IV'te görülmektedir.

TABLO IV. BAŞARIM PERFORMANSI (TELCO CUSTOMER CHURN)

Metot	Kesinlik	Duyarlılık	F1 Skoru	Doğruluk
Logistik Regr.	0.866	0.997	0.926	%86.5
KNN	0.897	0.951	0.923	%86.4
Karar Ağaçları	0.958	0.951	0.954	%92.2
Rastgele Orman	**0.900**	**0.994**	**0.944**	**%95.4**
DVM	0.880	0.993	0.933	%87.7
AdaBoost	0.892	0.968	0.928	%87.2
ÇKA	0.909	0.851	0.879	%79.9
Naïve Bayes	0.919	0.932	0.925	%87.1

Benzer şekilde Tablo V'te performans ölçütlerinin hesaplanmasında yararlanılacak olan ve çalışmada kullanılan 'Customer Churn Predıctıon 'veriseti üzerinde uygulanan makine öğrenmesi yöntemlerinden elde edilen sonuçlara göre oluşturulmuş karmaşıklık matrisi görülmektedir.

TABLO V. KARMAŞIKLIK MATRİSİ (CUSTOMER CHURN PREDICTION)

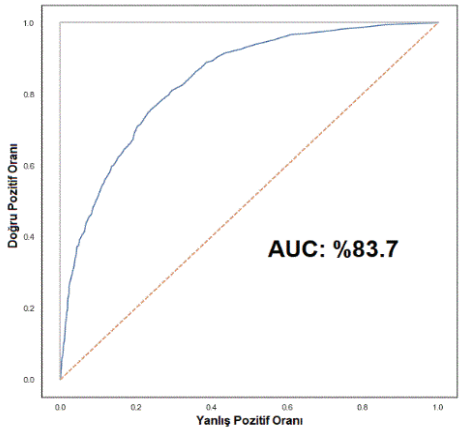
Logistik Regresyon		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	862 (DP)	2 (YN)
	Churn	133 (YP)	3 (DN)
K- En Yakın Komşu (KNN)		**Tahmin**	
		Non-Churn	Churn
Gerçek	No Churn	822 (DP)	42 (YN)
	Churn	94 (YP)	42 (DN)
Karar Ağaçları		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	822 (DP)	42 (YN)
	Churn	36 (YP)	100 (DN)
Rastgele Orman		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	859 (DP)	5 (YN)
	Churn	41 (YP)	95 (DN)
Destek Vektör Makineleri (DVM)		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	858 (DP)	6 (YN)
	Churn	117 (YP)	19 (DN)
AdaBoost		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	837 (DP)	27 (YN)
	Churn	101 (YP)	35 (DN)
Çok Katmanlı Algılayıcılar (ÇKA)		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	736 (DP)	128 (YN)
	Churn	73 (YP)	63 (DN)
Naïve Bayes		**Tahmin**	
		Non-Churn	Churn
Gerçek	Non-Churn	806 (DP)	58 (YN)
	Churn	71 (YP)	65 (DN)

Tablo V'teki değerler ışığında, (1), (2), (3), (4)'ten yararlanılarak yapılan hesaplamalara ait sonuç değerler Tablo VI'da görülmektedir.

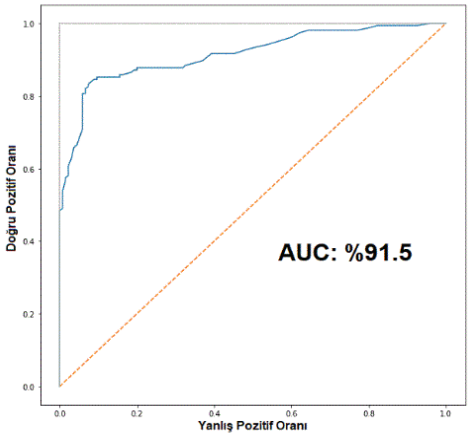
TABLO VI. BAŞARIM PERFORMANSI (CUSTOMER CHURN PREDICTION)

Metot	Kesinlik	Duyarlılık	F1 Skoru	Doğruluk
Logistik Regr.	0.839	0.884	0.860	%78.8
KNN	0.797	0.842	0.818	%72.4
Karar Ağaçları	0.829	0.813	0.820	%73.6
Rastgele Orman	**0.832**	**0.898**	**0.863**	**%79.0**
DVM	0.742	0.998	0.851	%74.4
AdaBoost	0.888	0.750	0.813	%74.5
ÇKA	0.751	0.993	0.855	%75.1
Naïve Bayes	0.882	0.746	0.808	%73.8

Tablo IV ve Tablo VI'da yer alan ve her iki veriseti üzerinde gerçekleştirilen müşteri kaybı analizine ilişkin olarak makine öğrenmesi metotlarının başarım performansları incelendiğinde, en başarılı makine öğrenmesi yönteminin 'Telco Custumer Churn' veriseti üzerinde %79 ve 'Customer Churn Prediction' veriseti üzerinde %95.4'lük doğruluk oranı sınıflandırma başarısı gösteren Rastgele Orman yöntemi olduğu görülmektedir. Random Forest metodunun başarım performansına ilişkin elde edilen ROC eğrileri Şekil 3 ve Şekil 4'te görülmektedir.



Şekil 3. ROC Eğrisi (Customer Churn Prediction)



Şekil 4. ROC Eğrisi (Telco Customer Churn)

Şekil 3 ve Şekil 4'te görüldüğü gibi Random Forest Metodu'nun ilgili verisetleri üzerindeki başarımına ilişkin ROC eğrileri altındaki alan değeri (AUC), Customer Churn Prediction veriseti için 0.837, Telco Customer Churn veriseti için ise 0.915 olarak elde edilmiştir.

IV. SONUÇLAR VE DEĞERLENDİRME

Telekomünikasyon sektörünün rekabet ortamında yeni müşteri kazanılması kadar, mevcut müşterilerin hizmet almayı sürdürmelerinin sağlanması da bu alanda faaliyet gösteren firmaların öncelikli hedefleri arasındadır. Bu nedenle hizmet almayı sonlandırarak aboneliği iptal etmeyi planlayan müşterilerin önceden tahmin edilerek, iptal işleminin önüne geçilmesi için gerekli operasyonel işlemlerin yürütülmesi önem arz etmektedir.

Bu çalışmada, kaggle.com'dan elde edilen iki ayrı veriseti üzerinde müşteri kaybı analizi gerçekleştirilerek, aldığı hizmete sonlandırıp aboneliğini iptal etme eğiliminde olan müşterilerin önceden tahmini işlemi gerçekleştirilmiştir. Bu amaçla ilgili verisetleri üzerinde 8 ayrı makine öğrenmesi yöntemi uygulanmıştır. Gerçekleştirilen başarım performans analizleri sonucunda her iki veriseti üzerinde de en başarılı tahmin başarısı gösteren makine öğrenmesi yönteminin sırasıyla %79 ve %95.4'lük doğruluk oranıyla sınıflandırma işlemi gerçekleştiren Rastgele Orman metodu olduğu görülmüştür. Gelecek çalışmalarda, Makine Öğrenmesi ve Derin Öğrenme metotlarının birlikte kullanıldığı hibrit yöntemler ile müşteri kaybı analizinin gerçekleştirilmesi planlamaktadır.

KAYNAKÇA

[1] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty" *Journal of Business Research,* vol. 94, pp. 290-301, 2019.

[2] S.H. Iranmanesh, M. Hamid, M. Bastan, G. H. Shakouri, and M. Nasiri, "Customer churn prediction using artificial neural network: An analytical CRM application" *International Conference on Industrial Engineering and Operations Management,* pp. 2214-2226, 2019.

[3] T. Satıcı and M. Bekler, "Customer Segmentation and Customer Churn Analysis System for Insurance Companies" *28th Signal Processing and Communications Applications Conference (SIU)*, pp.1-4, 2020.

[4] Y. Koca, B. E. Söğüt and S. Madikyan,"Sadakat Programında Müşteri Kayıp Tahmini: Bir Vaka Çalışması", *Journal of Information Systems and Management Research,* vol. 1, no.1, pp. 59-66, 2019.

[5] P. Wanchai, "Customer churn analysis: A case study on the telecommunication industry of Thailand", *12th International Conference for Internet Technology and Secured Transactions (ICITST)*, 2017.

[6] Y. Khan, S. Shafiq, A. Naeem, S. Ahmed, N. Safwan,S. Hussain, "Customers churn prediction using artificial neural networks (ANN) telecom industry." Editorial Preface From the Desk of Managing Editor, vol. 10, no. 9, 2019.

[7] G. Göy, B. Kolukısa, C. Bahçevan and V. Ç. Güngör, "Ensemble Churn Prediction for Internet Service Provider with Machine Learning Techniques." *2020 5th International Conference on Computer Science and Engineering (UBMK)*, pp.248-253, 2020.

[8] O. Kaynar, M. F. Tuna, Y. Görmez and M. A. Deveci, "Makine öğrenmesi yöntemleriyle müşteri kaybı analizi." Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, vol. 18, no.1, pp. 1-14, 2017.

[9] S. Yıldız, O. Aydemir, İ. Yılmaz, A. Şay, and S. Varlı, "Customer Churn Analysis" 28th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, 2020.

[10] Telco Customer Churn. (2021, May 03) [Online] Avaiable: https://www.kaggle.com/blastchar/telco-customer-churn

[11] Customer Churn Prediction 2020, (2021, May 03) [Online] Avaiable: https://www.kaggle.com/c/customer-churn-prediction-2020.

[12] J. Tolles and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes" Jama, vol.316, no.5, pp.533-534, 2016.

[13] G. D. Cavalcanti and R. J. Soares, "Ranking-based instance selection for pattern classification. Expert Systems with Applications," vol. 150, no.113269, 2020.

[14] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology" Transactions on Systems, Man, and Cybernetics,vol. 21, no.3, pp. 660-674, 1991.

[15] S. Misra, Y. Wu, "Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking," Machine Learning for Subsurface Characterization, vol. 289, 2019.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no.3, pp.273-297, 1995.

[17] T. Hastie, S. Rosset, J. Zhu and H. Zou, "Multi-class adaboost. Statistics and its Interface," vol.2, no.3, pp.349-360, 2009.

[18] A. Arı and M. E. Berberler, "Yapay sinir ağları ile tahmin ve sınıflandırma problemlerinin çözümü için arayüz tasarımı," Acta Infologica, vol.1, no.2, pp.55-73, 2017.

[19] M. Granik and V. Mesyura, "Fake News Detection Using Naive Bayes Classifier," First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp.900-903, 2017.