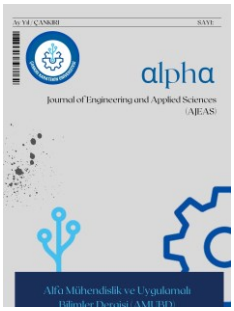


Alpha Journal of Engineering and Applied Sciences
(2024) 2/ 1: 1 – 16
Alfa Mühendislik ve Uygulamalı Bilimler Dergisi
(2024) 2/1: 1 – 16

Araştırma Makalesi

Gönderi	*Kabul*	*Yayın*
30/01/2024	*18/03/2024*	*30/04/2024*

Müşteri Terk Verisi Üzerinde Artırma Yöntemlerinin Performans Karşılaştırması

Başak Ceren SEÇİK GÖÇER [a,*], İbrahim EMİROĞLU [a]

[a] Matematik Mühendisliği Bölümü, Kimya-Metalurji Fakültesi, Yıldız Teknik Üniversitesi, İstanbul, TÜRKİYE
* Sorumlu yazar e-mail adresi: basakcerensecik@yahoo.com

Özet

Müşteri kaybı analizinde öngörü oluşturmak için günümüzde veri madenciliği ve makine öğrenmesi modelleri sıklıkla kullanılmaktadır. Müşteri kaybı analizi sayesinde, işletmeler müşterileri şirketi terk etmeden veya ürünlerini kullanmayı bırakmadan önce bazı çıkarımlarda bulunabilir ve müşteri terk oranını düşürerek hem kârlarını hem de müşteri memnuniyetini artırabilir. Bu analizleri yapmanın birçok yolu bulunmaktadır. Kural tabanlı modeller geliştirilebilir, çeşitli makine öğrenmesi modelleriyle tahminler yapılabilir. Bu makale çalışmasında kaggle.com sitesinde açık erişime sunulan 7043 gözlem 57 değişkenden oluşan veri seti üzerinde makine öğrenmesi modelleri kurularak analizler gerçekleştirilmiştir. Müşterilerin verileri kullanılarak yapılan bu analiz sonucunda hangi müşterinin Telekom şirketinin müşterisi olarak kalmaya devam edeceği, hangi müşterinin şirketi terk edeceği tahmin edilmiştir. Terk etme durumu üzerinde hangi özelliklerin önemli olduğu tartışılmıştır. Makine öğrenmesi modelleri olarak Hafif Gradyan Artırma (Light GBM), Aşırı Gradyan Artırma (XGBoost), CatBoost ve Gradyan Artırma (Gradient Boosting) metotları kullanılmış ve bu artırma metotları arasındaki performanslar değerlendirilmiştir. Veri seti yeniden örnekleme teknikleri uygulanarak veri dengeli hale getirilmiştir. Doğruluk, F1 - Skoru ve duyarlılık metrikleri baz alınarak model başarıları ölçülmüştür. Doğruluk ve F1- Skoru metrikleri değerlendirildiğinde model performansları arasında anlamlı fark bulunmazken, duyarlılık metriği değerlendirildiğinde en iyi performans Aşırı Gradyan Artırma (XGBoost) modeli tarafından 0,949 oranıyla sağlanmıştır.

Anahtar kelimeler: *Müşteri kayıp analizi, Telekomünikasyon, Sınıflandırma, Makine öğrenmesi*

Performance Comparison of Boosting Methods on Customer Churn Data

Abstract

Data mining and machine learning models are frequently used today to generate insights in churn analysis. Through churn analysis, businesses can make inferences before their customers leave the company or stop using their products, and can increase both profits and customer satisfaction by reducing customer churn. There are many ways to perform these analyses. Rule-based models can be developed and predictions can be made with various machine learning models. In this article, machine learning models were built and analyzed on a data set consisting of 7043 observations and 57 variables, which is publicly available on kaggle.com. As a result of this analysis using customers' data, it was predicted which customers will continue to be customers of the telecom company and which customers will leave the company. It is discussed which features are important on the churn situation. Light GBM, XGBoost, CatBoost and Gradient Boosting methods are used as machine learning models and the performances between these boosting methods are evaluated. The data set is balanced by applying resampling techniques. Model performance was assessed based on accuracy, F1-Score and sensitivity metrics. When metrics accuracy and F1-Score were evaluated, no significant difference was found in the model performances. However, when metric sensitivity was assessed, the best performance was achieved by the Extreme Gradient Boost (XGBoost) model with a ratio of 0.949.

Keywords: *Customer churn analysis, Telecommunication, Classification, Machine learning*

1. Giriş

Son zamanlarda müşteriyi takip etme, elde tutma, geri kazanma gibi konular şirketler için önemli bir konu halini almıştır. Günümüzde gelişen teknolojiyle birlikte sektörde rekabetin artışı, müşterileri şirketler için daha önemli bir konuma getirmiştir. İşletmeler, teknolojiyi kullanarak daha az maliyetle daha kaliteli hizmet sunmak için çalışmalar yapmaktadırlar. Kendileri için hayati bir konumda bulunan müşterilerin karar verme, satın alma, terk etme gibi süreçlerini de mercek altına almışlardır [1].

Müşterinin şirketin bir servisini kullanmayı bırakması, üyeliğini iptal etmesi veya şirketle ilişkisini tamamen kesmesi durumuna basit bir şekilde müşteri kaybı denilebilir [2]. Şirketlerin ana gelir kaynakları müşterilerdir. Ancak artan firma sayısı ve yoğun rekabet ortamı müşteriler için çok fazla seçenek sunmakta ve cazip bir ortam oluşturmaktadır. Müşterilerin tercih yelpazesi genişlemekte ve kendilerine en kaliteli hizmeti en uygun fiyatla sunan firmaları tercih etmektedirler. Bu sebeple şirketler en değerli müşterilerini rakiplerine kaptırabilmekte ve her geçen gün müşteri kaybı şirketler için bir sorun haline gelmektedir. Bu pazar dinamiğinde işleyişlerini sürdürerek sektörde var olmaya devam edebilmeleri, müşterilerini elde tutmaları ve kârlılıklarını yükseltebilmeleri için müşteri kaybı analizi ile ayrılma olasılığı yüksek müşterileri tespit ederek gerekli aksiyonları almaları gerekmektedir [3].

Şirketlerin, yeni müşteri kazanmaları ile mevcut müşterileriyle var olan ilişkilerini geliştirip derinleştirmelerinin performansları üzerinde nasıl etki ettiğiyle ilgili araştırmalar yapılmaktadır [4]. Yapılan araştırmalar, yeni müşteri kazanmanın var olan müşteriyi elde tutmaktan en az 5 kat daha maliyetli olduğunu göstermektedir [5]. Özellikle büyük veri tabanlarına ve abonelik sistemlerine sahip sektörler (hizmet, perakende şirketleri, bankalar, telekomünikasyon şirketleri ve benzeri) direkt olarak müşteri odaklıdırlar. Çok fazla müşteri verisine sahip olan bu sektörler müşterilerini elde tutmak ve geleceğe yönelik planlamalarını yapabilmek için ileri analitik yöntemler kullanarak müşteri kayıp tahmin modelleri geliştirerek her müşteriye göre farklı aksiyonlar alabilmektedirler [6].

Bu bağlamda şirketlerin müşterilerini elde tutabilmeleri için müşteri analizi yapmalarının önemi göz önünde bulundurularak, telekomünikasyon şirketi veri seti üzerinde makine öğrenmesi ve ileri analitik modellerle müşteri kaybı analizi gerçekleştirilmiştir.

Bu çalışmanın yapısı 2. bölüm literatürde yer alan konuyla ilişkili çalışmalar, 3. bölüm çalışmada kullanılan yöntemler, 4. bölüm veri setinin analizi, 5. bölüm yöntemlerin uygulanması sonucu elde edilen bulgular, 6. bölüm elde edilen bulguların tartışılması ve sonuçlar şeklindedir.

Literatürde farklı sektör verileriyle "Müşteri Kaybı" konusu ile ilgili yapılmış birçok çalışma bulunmaktadır.

Ahmad vd. (2019), Suriye'de hizmet veren bir Telekom şirketi olan SyriaTel Telekom şirketi tarafından sağlanan verilerle çalışarak, telekomünikasyon şirketinin müşteri kaybını en aza indirmek için ileriki dönemde kaybedebileceği müşterileri makine öğrenmesi modelleri kullanarak tahmin etmişlerdir. Makine öğrenmesi modelleri olarak Karar Ağaçları, Rastgele Orman, Gradyan Artırma (GBM) ve Aşırı Gradyan Artırma (XGBoost) modellerini kullanmışlardır ve en iyi sonucu Aşırı Gradyan Artırma algoritmasını kullanarak elde etmişlerdir. Modelin performansı AUC ölçüm değeri ile ölçülmüş ve %93,3 başarı elde edilmiştir [7].

Brandusoiu vd. (2016), California Üniversitesi, Bilgisayar Bilimleri departmanından temin ettikleri telekomünikasyon şirketi bilgileri yer alan veri setiyle müşteri kaybı analizi yapmışlardır. 15 sürekli, 5 kesikli ve 1 bağımlı değişken olmak üzere 21 kolon ve 3333 gözlemden oluşan veri setinde, telekom şirketinin mevcut veya gelecekteki müşterilerini ileriki dönemde kaybedebilme durumunu ileri

sınıflandırma algoritmaları kullanarak tahmin etmişlerdir. Boyut indirgemek için Temel Bileşen Analizi'ni (PCA) kullanan araştırmacılar çalışmalarında Bayes Ağları (Bayesian Networks), Destek Vektör Makineleri ve Çok Katmanlı Algılayıcı (MLP) yapay sinir ağları yöntemlerini kullanmışlardır. Kullandıkları metodolojilerde birbirlerine yakın sonuçlar elde eden araştırmacılar, üç algoritmada da %99 üzerinde başarı elde etmişlerdir [8].

Öztürk vd. (2023), bir e-ticaret şirketinde yer alan satıcıların e-ticaret uygulamasını kullanmayı bırakıp bırakmamasını tahmin ettikleri projelerinde kayıp modeli geliştirmişlerdir. Veri setinde satıcının şehri, yaptığı toplam işlem, işlemden elde ettikleri gelir, hangi sektörde yer aldığı, taksit ve vade bilgileri gibi değişkenler yer almaktadır. Araştırmacılar çalışmalarında Rastgele Orman ve Lojistik Regresyon algoritmalarını kullanmışlardır. Modelleri geliştirirken ön işleme yapmanın model üzerindeki etkisini analiz etmek için 3 farklı metot kullanmışlardır. Sırasıyla ön işleme yapmadıkları, eksik örnekleme (undersampling) ve aşırı örnekleme (oversampling) gibi yeniden örnekleme tekniklerini uyguladıkları 3 farklı yöntem kullanmışlardır. Modelin performansını analiz etmek için F1 Skoru kullanmışlardır. Çalışmalarında aşırı örnekleme kullanarak geliştirdikleri Rastgele Orman modelinin en başarılı model olduğu sonucuna ulaşmışlardır [9].

An vd. (2022), toplamda 10.000 banka müşterisi verisiyle yaptıkları, banka müşterilerinin bankanın kredi kartını kullanmayı bırakıp bırakmayacağını tahmin ettikleri kayıp analizi çalışmalarında, yaptıkları görsel analizlerle verinin dengesiz olduğunu elde etmişlerdir. Yazarlar yaptıkları özellik mühendisliği ile ana veriden yeni değişkenler elde etmişlerdir. Veride, kaybedilen müşteri sayısı edilmeyen müşteri sayısından çok daha azdır ve bu dengesizliğe yol açmaktadır. Bu sorunu çözmek için veriye Sentetik Azınlık Aşırı Örnekleme Tekniği **(**SMOTE) algoritmasını uygulayarak az olan verileri artırmışlardır. Orijinal veri setiyle SMOTE uygulanan veri setinin model üzerindeki başarılarını karşılaştırmak için iki veri setine de ayrı ayrı modelleri uygulamışlardır. Aynı zamanda hedef kodlama (target encoding) ve tek değer kodlama (one-hot encoding) olmak üzere iki farklı enkode yöntemi uygulayan araştırmacılar toplamda 4 farklı veri setine Lojistik Regresyon, Gradyan Artırma ve Yapay Sinir Ağları modellerini uygulamışlardır. Çalışmalarının sonucunda SMOTE algoritmasının model başarısını artırmadığını, tek değer kodlamanın hedef kodlamaya göre daha etkili olduğunu ve en başarılı modelin Lojistik Regresyon olduğunu belirtmişlerdir [10].

Sayed vd. (2018), Kaggle websitesinde açık erişimli olarak sunulan ve müşteri demografik bilgileri, kredi skoru, kullanım süresi bilgisi, kredi kartı olup olmadığı, maaş bilgisi ve şirketin müşterisi olmayı bırakıp bırakmadığı gibi bilgilerin yer aldığı 14 kolon, 10.000 satırdan oluşan banka müşteri verisiyle yaptıkları müşteri kayıp analizi çalışmalarında bağımlı değişken kolonunda yaptıkları analizde şirketin müşterisi olmayı bırakanların sayısının bırakmayanlara oranla çok az olduğunu belirtmişlerdir. Veriler arasındaki bu dengesizliği gidermek için eksik örnekleme (under sampling) yaparak bırakmayan sınıfındaki verileri azaltmışlardır. Apache Spark içinde yer alan daha eski makine öğrenmesi kütüphanesi olan Mlib ile daha yeni makine öğrenmesi kütüphanesi olan Spark ML'i karşılaştırdıkları çalışmalarında kayıpları tahmin etmek için Karar Ağacı algoritması kullanmışlardır. Sonuç olarak, Spark ML'in model performans başarısının daha yüksek olduğu sonucunu elde etmişlerdir [11].

Saleh vd. (2023), ilk 3 veri seti açık erişimli son veri seti Aalborg Üniversitesi öğrencilerine yapılmış anket verilerinden oluşan 4 veri seti üzerinde Rastgele Orman, AdaBoost, Lojistik Regresyon, Aşırı Gradyan Artırma Sınıflandırıcısı ve Karar Ağacı olmak üzere 5 makine öğrenmesi algoritması kullanarak Danimarka telekomünikasyon şirketleri için müşteri kaybına sebep olabilecek özellikleri bulmaya çalışmışlardır. Makine öğrenmesi algoritmaları farklı veri setleri üzerinde farklı performanslar göstermiş olup birbirinden farklı önemli özellikler bulmuşlardır. IBM Telco veri seti için Lojistik Regresyon ve Rastgele Orman en iyi performansı göstermişlerdir. Maven veri seti için Aşırı Gradyan Artırma ve Rastgele Orman optimum performans gösterirken Cell2Cell veri setinde Aşırı Gradyan Artırma, Rastgele Orman ve Karar Ağacı modelleri en iyi performansı göstermişlerdir.

Son olarak AAU veri setinde Rastgele Orman ve Aşırı Gradyan Artırma en yüksek performansları sağlamışlardır. Ayrıca AAU veri seti için dengesiz durumu gidermek adına sentetik azınlık aşırı örnekleme tekniği de uygulanmıştır [12].

Çelik vd. (2023), R programlama dilini kullanarak kaggle.com web sitesinden aldıkları dengesiz telekomünikasyon veri seti üzerinde kayıp analizi çalışmaları yapmışlardır. Öncelikle bu veri setiyle kayıp analizi yapmak için Destek Vektör Makineleri (SVM) modeli uygulamışlardır ve Destek Vektör Makineleri modelinin dengesiz veri setlerinde performans başarılarının düşük olduğu bilgisini elde etmişlerdir. Dengesizliği gidermek için aşırı örnekleme yapan araştırmacılar, bu durumun modelin performans başarısının arttığını gözlemlemişlerdir. Çalışmalarını daha da iyileştirmek için topluluk öğrenme modeli metotlarını denemişlerdir ve başarının önemli ölçüde arttığını kaydetmişlerdir [13].

Verma (2020), Banka verilerini kullanarak müşteri kayıp oranını belirlemek için çeşitli istatistiksel yöntemler ve makine öğrenmesi tahmin modelleri geliştirmiştir. Veri seti dengesiz olduğu için eksik yeniden örnekleme metodunu kullandığı çalışmasında model başarılarını ölçmek için doğruluk, AUC, Gini katsayısı ve ROC eğrisi gibi çeşitli model karşılaştırma metriklerini kullanmıştır. En yüksek başarıyı gösteren model %78 doğrulukla Rastgele Orman modeli olmuştur. Müşterinin yaşı, ortalama borç tutarı, ortalama işlem sayısı ve meslek kodu gibi değişkenlerin kayıp tahmin modeli için önemli değişkenler olduğu bilgisini kaydetmiştir [14].

Bu alanda yapılmış çalışmalar göz önünde bulundurulduğunda, her iki müşteri sınıfı (terk eden ve terk etmeyen) üzerinde performans değerlendirmesi yapan doğruluk, AUC, F1- Skoru gibi metriklerin kullanıldığı görülmüştür. Bu çalışmada ise sadece artırma yöntemlerinin performansları kıyaslanmış aynı zamanda sadece bir sınıfa (terk eden) odaklanan duyarlılık metriğinin de önemli olduğu gösterilmek istenmiştir.

2. Materyal Metod

Bu bölümde, makale çalışmasının deney aşamasında kullanılan sınıflandırma algoritmaları, modelin performansını değerlendirmek için kullanılan başarı metrikleri ve veri setindeki dengesizliği gidermek için uygulanan yeniden örnekleme yöntemleri hakkında bilgi verilmektedir.

2.1. Gradyan Artırma Algoritması

Gradyan Artırma Algoritması (GBM), artırma (boosting) metotlarının temelini oluşturan bir makine öğrenmesi algoritmasıdır. Regresyon ve sınıflandırma problemlerinde kullanılabilen bu algoritmanın çalışma prensibi zayıf tahmin edicilerin yinelemeli şekilde bir araya gelerek güçlü bir öğrenen haline getirilmesidir. Genellikle karar ağaçları olan bu zayıf öğrenenler ilk modelde bir tahmin üretir ve bu tahmin sonucunda gerçek değerle arasındaki hata miktarı hesaplanır. Bu hata miktarları için yeni bir kolon oluşturulur. Bu kolon kendinden sonra gelen ağaç için yeni bir özellik olmaktadır. Bundan sonraki karar ağacının görevi kendinden önceki ağaçtan elde edilen bu hataları düşürerek minimuma indirmektir. Bu şekilde oluşturulan bütün birbirine bağlı ağaçlar topluluğu güçlü bir öğreniciyi temsil eder. Gradyan Artırma Algoritması istatistiksel bir yöntemden ziyade tekrarlamalı bir yöntemdir. Tekrarlamalı süreçle birlikte her adımda azalan hatalar modelin daha doğru ve performanslı tahminler yapması ile sonuçlanır [15]. Gradyan Artırma Algoritması kendinden sonra oluşturulan diğer güçlü modellerin temelini oluşturması sebebiyle makine öğrenmesi alanında araştırmalar ve sektördeki uygulamalarda sıkça kullanılmaktadır [16]. Gradyan Artırma Algoritması, veri kümelerinde tahmin performansı bakımından başarılı olması sebebiyle bu çalışmada kullanılan algoritmalarından biri olarak tercih edilmiştir.

2.2. Aşırı Gradyan Artırma Algoritması

Aşırı Gradyan Artırma Algoritması (XGBoost), 2016 yılında Chen ve Guestrin tarafından geliştirilmiş, Gradyan Artırma Algoritması metodunun daha etkili hale getirilmiş versiyonudur. Hata oranı düşük olmakla birlikte sınıflandırma ve regresyon problemlerinde yüksek performansla çalışan karar ağacı tabanlı bir algoritmadır [16]. Artırma algoritması olduğu için temelinde karar ağacı yer almaktadır. Çok hızlı hesaplama yapabilmesi ve iyi bir tahmin başarısı olması sebebiyle uygulamalarda ve araştırmalarda sıkça kullanılmaktadır.

2.3. Hafif Gradyan Artırma Algoritması

Hafif Gradyan Artırma Algoritması (Light GBM) veri setleri üzerinde etkili şekilde çalışan bir karar ağacı algoritmasıdır. Microsoft şirketi tarafından geliştirilip açık kaynaklı olarak sunulan bu makine öğrenmesi algoritması, Gradyan Artırma Algoritması'nı temel alarak oluşturulmuştur. Karar ağaçlarında seviye odaklı (level wise) ve yaprak odaklı (leaf wise) olmak üzere iki farklı yöntem kullanılır. Gradyan Artırma Algoritması seviye odaklı büyüme gösterirken Hafif Gradyan Artırma Algoritması yaprak odaklı bölünme göstermektedir. Dolayısıyla diğer algoritmalar aynı seviyedeki yapraklardaki bilgilerden bağımsız olarak tüm yaprakları bölerek başarı metriklerini artırmaya çalışırken Hafif Gradyan Artırma Algoritması aynı seviyedeki yapraklarda en çok bilgi içeren yaprağı bölüp bilgi içermeyen yapraklara dokunmayarak başarı metriğini artırmaya çalışır. Bu da hem zamandan tasarruf etmeye hem de daha az bellek kullanımına yarar sağlar. Çok sayıda veri üzerinde yapılan deneyler Hafif Gradyan Artırma Algoritması'nın geleneksel Gradyan Artırma Algoritması'na göre 20 kat daha hızlı olduğunu ve neredeyse benzer doğruluğu elde ettiğini göstermektedir [17]. Bu sebeple Hafif Gradyan Artırma Algoritması büyük veri setlerinde, gerçek zamanlı tahminlerde ve birçok veri bilimi yarışmasında da odak haline gelmiştir. Bu çalışmada da hızından ve performansından faydalanmak için Hafif Gradyan Artırma Algoritması'na yer verilmiştir.

2.4. CatBoost

Aşırı Gradyan Artırma Algoritması ve Hafif Gradyan Artırma Algoritması makine öğrenmesi modellerine alternatif olarak 2017 yılında Yandex şirketi tarafından geliştirilmiş hem sınıflandırma hem de regresyon probremlerinde kullanılan, kategorik verilerle güçlü performans gösteren bir gradyan artırma algoritmasıdır. Kategorik verilerin yer aldığı karmaşık veri setlerinde üstün performans başarısı göstermekte ve veri ön işlemeye duyulan ihtiyacı azaltarak verileri doğrudan işleyebilmekte ve bu sayede daha basit ancak daha güçlü bir model performansı sağlamaktadır [18]. Veri ön işleme sırasında diğer makine öğrenmesi algoritmalarının çalışma prensibine göre makine için daha anlamlı olması amacıyla kategorik verileri nümerik verilere çevirmek gerekmektedir. Ancak CatBoost algoritmasında kategorik veriler enkode edilmeden işlenebildiği için böyle bir dönüşüme ihtiyaç bulunmamaktadır. Yüksek öğrenme hızı, aşırı öğrenmeye karşı dayanıklılığı ve karmaşık veri setleri üzerinde etkili çalışmasıyla diğer makine öğrenmesi algoritmaları arasında ön plana çıkmaktadır [19].

2.5. Eksik Örnekleme

Dengesiz olan veri setini daha dengeli hale getirmek için yeniden örnekleme yöntemleri kullanılmıştır. Eksik örnekleme yöntemi de yeniden örnekleme yöntemlerinden biridir. Verilerin arasındaki dağılım eşitsizliğini gidermek için fazla olan veri sınıfı az olanla eşit hale getirilene kadar azaltılır ve dengeleme sağlanmış olur. Ancak verileri silmek bilgi kaybına sebep olma dezavantajı sağlayabilir.

2.6. Aşırı Örnekleme

Eksik örnekleme yöntemine benzer bir yeniden örnekleme yöntemidir. Dengesizliği dengeli hale getirmek için az olan veri sınıfı rastgele tekrarlanarak artırılır ve dengesizlik giderilir. Bu yöntemde eksik örneklemede olduğu gibi veri azaltma olmadığı için bilgi kaybına sebep olmaz, ancak verileri artırmak için var olan aynı bilgiler tekrarlandığından dolayı aşırı öğrenmeye sebep olabilmektedir.

2.7. Sentetik Azınlık Aşırı Örnekleme Tekniği

Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) de bir yeniden örnekleme tekniğidir. 2002 yılında Chawla ve diğ. tarafından tanıtılan bu yöntem de azınlık sınıfta bulunan örneklerin artırılarak dengeli hale getirilmesi için kullanılır. Aşırı örneklemede yapıldığı şekilde azınlık sınıftaki örneklerin tekrarlanarak artırılması yerine belirli işlemler sonucunda yapay örnekler oluşturularak artırılır. Yapay gözlemler, bir gözlem ve o gözleme en yakın komşusu arasındaki fark alınarak rastgele 0 ile 1 arasında herhangi bir değer ile çarpılarak o gözleme eklenmesi sonucu oluşturulmaktadır [20].

2.8. Model Performans Metrikleri

Bir makine öğrenmesi modelinin başarısını ölçmek için model başarı metrikleri kullanılır. Modelin türüne göre kullanılan başarı metrikleri değişmektedir. Sınıflandırma problemleri için karmaşıklık matrisi (confusion matrix), doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 Skoru (F1 Score), ROC Eğrisi ve AUC değeri ve benzeri metrikler kullanılmaktadır [21].

		Tahmin Edilen Sınıf: Pozitif	Tahmin Edilen Sınıf: Negatif
Gerçek Sınıf	Pozitif	Gerçek Pozitif	Yanlış Negatif
	Negatif	Yanlış Pozitif	Gerçek Negatif

Şekil 1. Karmaşıklık matrisi (Confusion Matrix)

Karmaşıklık Matrisi: Şekil 1'de verilen matris karmaşıklık matrisi olarak adlandırılır. Modelin sonucunda elde edilen tahmin verileriyle gerçek sınıf verilerini karşılaştırmak için kullanılır. Gerçekte pozitif olup modelin pozitif tahmin ettiği değerlere Gerçek Pozitif, gerçekte negatif olup modelin negatif tahmin ettiği değerlere ise Gerçek Negatif denmektedir. Bu değerler model tarafından doğru tahmin edilmiştir. Gerçekte pozitif olup model tarafından negatif tahmin edilen modellere Yanlış Negatif, gerçekte negatif olup model tarafından pozitif tahmin edilen modellere ise Yanlış Pozitif denmektedir. Bu değerler model tarafından yanlış tahmin edilmiştir.

Doğruluk: Modelin verileri ne kadar doğru tahmin ettiğini ölçmeye yarayan bir metriktir. Modelin doğru tahmin ettiği veri sayısının toplam veri sayısına bölünmesiyle bulunur.

$$Doğruluk = \frac{\text{Gerçek Pozitif+Gerçek Negatif}}{\text{Gerçek Pozitif+Gerçek Negatif+Yanlış Pozitif+Yanlış Negatif}} \quad (1)$$

Kesinlik: Model performansının kesinliğini belirten metriktir. Tahmin sonucunda pozitif etiketlenen verilerin gerçekte ne kadarının pozitif olduğunun ölçülmesidir.

$$Kesinlik = \frac{\text{Gerçek Pozitif}}{\text{Gerçek Pozitif+Yanlış Pozitif}} \quad (2)$$

Duyarlılık: Gerçekte pozitif olan verilerin model tarafından ne kadarının pozitif tahmin edildiğini ölçen metriktir.

$$Duyarlılık = \frac{\text{Gerçek Pozitif}}{\text{Gerçek Pozitif+Yanlış Negatif}} \quad (3)$$

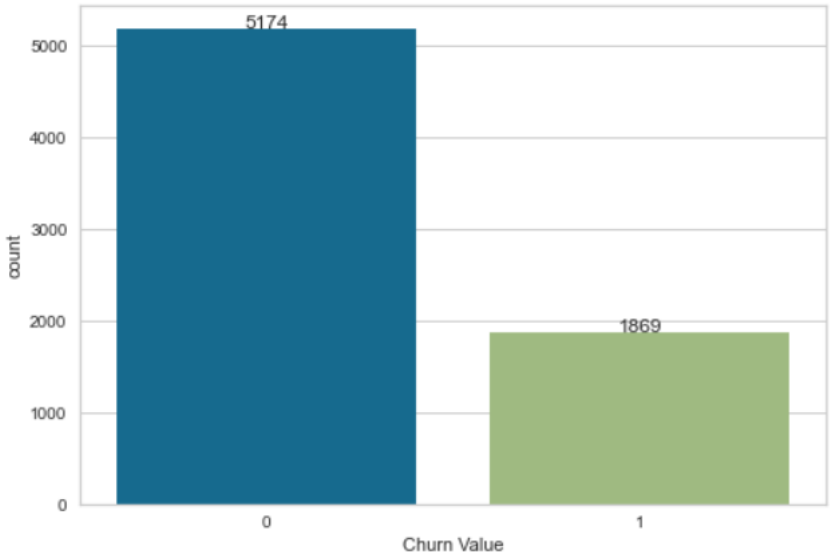
F1 Skoru: Duyarlılık ve kesinlik metriklerinin ikisini de barındıran bir metriktir. Duyarlılık ve kesinlik metriklerinin harmonik ortalaması alınarak hesaplanır.

$$F1\ Skoru = \frac{2*\text{Kesinlik}*\text{Duyarlılık}}{\text{Kesinlik+Duyarlılık}} \quad (4)$$

Model başarılarını kıyaslamak için birçok farklı yöntem mevcuttur. Bu yöntemler veri setlerinden elde edinmek istenen sonuçlara göre farklılık gösterir. Örneğin; dengesiz bir veri setinde doğruluk metriğine bakmak araştırmacıyı yanıltabilir. Çünkü dengesiz veri setlerinde doğruluk metriğinin çok yüksek çıkması modelin başarılı olduğu anlamına gelmez. Çoğunluk sınıf ile azınlık sınıf arasında doğru tahmin edilen gözlem sayısına bakmak önemlidir. Aksi halde çoğunluk sınıftaki doğru tahmin edilen büyük çoğunluk azınlık sınıftaki yanlış tahmin edilen başarısızlığın önüne geçer.

2.9. Analizler

Çalışmada kaggle.com web sitesi aracılığıyla açık erişimli olarak sunulan veri seti kullanılmıştır (https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113). ABD, California eyaletinde ikamet eden kişilerin haberleşme verilerinin yer aldığı veri setinde telekomünikasyon şirketini kullanmayı bırakacak müşterilerin tahmin edilmesi amaçlanmaktadır. Veriler, müşterilerin demografi, lokasyon, servis, statü, popülasyon ve şirketin ürünlerini kullanmayı bırakıp, bırakmama bilgilerinin yer aldığı 6 farklı csv dosyası şeklinde erişime sunulmuştur. Veri setleri ortak anahtarlar üzerinden birleştirilerek tek ve nihai bir veri seti haline dönüştürülmüştür. Nihai veri seti 57 kolon, 7043 satırdan oluşmaktadır. Bağımlı değişken olan "Churn Value" değişkeninde bu çeyreklikte şirketten ayrılan müşteriler 1, şirkette kalmaya devam eden müşteriler ise 0 değer ile etiketlidir. Şekil 2'de görüleceği üzere veri setinin %74'ü "0", %26'sı "1" verisinden oluşmaktadır. Bu durum sınıf dağılımlarının dengesiz olduğunu göstermektedir. Bu dengesizliği gidermek için metodoloji bölümünde bahsedilen farklı dengeleme teknikleri uygulanacaktır.



Şekil 2. Bağımlı değişkenin dağılımı (Distribution of the dependent variable)

Veri setinde yer alan değişkenlerin açıklamaları Tablo 1'de verilmiştir.

Tablo 1. Veri setinde kullanılan değişken isimleri ve açıklamaları

Değişkenler	Açıklamalar
CustomerID	Her müşteriyi tanımlayan benzersiz bir kimlik
Count	Gruptaki müşteri sayısını özetlemek için gösterge tablosunda kullanılan bir değer
Gender	Müşterinin cinsiyeti: (Kadın, Erkek)
Age	Mali çeyreğin sona erdiği tarihte müşterinin mevcut yaşı (yıl olarak)
Senior Citizen	Müşterinin 65 yaş ve üzeri olduğunu belirtir: (Evet, Hayır)
Married	Müşterinin evli olup olmadığını belirtir: (Evet, Hayır)
Dependents	Müşterinin bakmakla yükümlü olduğu kişilerle birlikte yaşayıp yaşamadığını belirtir: (Evet, Hayır)
Number of Dependents	Müşteriyle birlikte yaşayan bakmakla yükümlü olunan kişilerin sayısını gösterir
Country	Müşterinin birincil ikamet ettiği ülke
State	Müşterinin birincil ikamet ettiği eyalet
City	Müşterinin birincil ikamet ettiği şehir
Zip Code	Müşterinin birincil ikamet ettiği posta kodu
Lat Long	Müşterinin birincil ikamet yerinin enlem ve boylamının birleşimi
Latitude	Müşterinin birincil ikamet yerinin enlemi
Longitude	Müşterinin birincil ikamet yerinin boylamı

Population	Posta Kodu alanının tamamı için güncel bir nüfus tahmini
Quarter	Verilerin elde edildiği mali çeyrek (ör. 3. Çeyrek))
Referred a Friend	Müşterinin bu şirkete bir arkadaşını veya aile üyesini tavsiye edip etmediğini belirtir: (Evet, Hayır)
Number of Referrals	Müşterinin bugüne kadar yaptığı yönlendirmelerin sayısını gösterir
Tenure in Months	Üç aylık dönem sonu itibarıyla müşterinin şirkette bulunduğu toplam ay tutarını gösterir.
Offer	Varsa, müşterinin kabul ettiği son pazarlama teklifini tanımlar. Değerler: (Yok, Teklif A, Teklif B, Teklif C, Teklif D, Teklif E)
Phone Service	Müşterinin şirketin ev telefonu hizmetine abone olup olmadığını belirtir: (Evet, Hayır)
Avg Monthly Long Distance Charges	Müşterinin belirtilen çeyreğin sonuna kadar hesaplanan ortalama şehirlerarası ücretlerini gösterir.
Multiple Lines	Müşterinin şirkette birden fazla telefon hattına abone olup olmadığını belirtir: (Evet, Hayır)
Internet Service	Müşterinin firmanın internet hizmetine abone olup olmadığını belirtir: (Hayır, DSL, Fiber Optik, Kablo)
Avg Monthly GB Download	Belirtilen çeyreğin sonuna kadar hesaplanan, müşterinin ortalama indirme hacmini gigabayt cinsinden gösterir
Online Security	Müşterinin şirket tarafından sağlanan ek bir çevrimiçi güvenlik hizmetine abone olup olmadığını belirtir: (Evet, Hayır)
Online Backup	Müşterinin şirket tarafından sağlanan ek bir çevrimiçi yedekleme hizmetine abone olup olmadığını belirtir: (Evet, Hayır)
Device Protection Plan	Müşterinin, İnternet ekipmanı için şirket tarafından sağlanan ek cihaz koruma planına abone olup olmadığını belirtir: (Evet, Hayır)
Premium Tech Support	Müşterinin, bekleme sürelerinin kısaltıldığı şirketten ek bir teknik destek planına abone olup olmadığını belirtir: (Evet, Hayır)
Streaming TV	Müşterinin İnternet hizmetini üçüncü taraf bir sağlayıcıdan televizyon programı yayınlamak için kullanıp kullanmadığını belirtir: (Evet, Hayır)
Streaming Movies	Müşterinin, üçüncü taraf bir sağlayıcıdan film akışı sağlamak için İnternet hizmetini kullanıp kullanmadığını belirtir: (Evet, Hayır)
Streaming Music	Müşterinin İnternet hizmetini üçüncü taraf bir sağlayıcıdan müzik yayını yapmak için kullanıp kullanmadığını belirtir: (Evet, Hayır)
Unlimited Data	Müşterinin sınırsız veri indirme/yükleme için ek bir aylık ücret ödeyip ödemediğini belirtir: (Evet, Hayır)
Contract	Müşterinin mevcut sözleşme türünü belirtir: (Aydan Aya, Bir Yıllık, İki Yıllık)
Paperless Billing	Müşterinin kağıtsız faturalamayı seçip seçmediğini belirtir: (Evet, Hayır)
Payment Method	Müşterinin faturasını nasıl ödediğini belirtir: (Bankadan Para Çekme, Kredi Kartı, Posta Çeki)

Monthly Charge	Müşterinin şirketten aldığı tüm hizmetler için geçerli toplam aylık ücretini gösterir.
Total Charges	Müşterinin belirtilen çeyreğin sonuna kadar hesaplanan toplam ücretini gösterir
Total Refunds	Müşterinin belirtilen çeyreğin sonuna kadar hesaplanan toplam iade tutarını gösterir
Total Extra Data Charges	Belirtilen çeyreğin sonuna kadar müşterinin planında belirtilenin üzerindeki ekstra veri indirmeleri için ödeyeceği toplam ücreti belirtir.
Total Long Distance Charges	Belirtilen üç aylık dönem sonu itibarıyla müşterinin planında belirtilenin üzerindeki toplam uzak mesafe ücretini gösterir.
Satisfaction Score	Bir müşterinin şirkete ilişkin genel memnuniyet derecesi 1'den (Hiç Memnun Değil) 5'e (Çok Memnun).
Satisfaction Score Label	Puanın (1-5) metin sürümünü bir metin dizesi olarak gösterir.
Customer Status	Üç aylık dönem sonunda müşterinin durumunu belirtir: (Bıraktı, Kaldı veya Katıldı)
Churn Label	Doğrudan Kayıp Değeri ile ilgilidir. (Evet = müşteri bu çeyrekte şirketten ayrıldı, Hayır = müşteri şirkette kaldı)
Churn Value	Doğrudan Churn Label ile ilgilidir. (1 = müşteri bu çeyrekte şirketten ayrıldı, 0 = müşteri şirkette kaldı)
Churn Score	IBM SPSS Modeler tahmin aracı kullanılarak hesaplanan 0-100 arası bir değer. Model, kayıplara neden olduğu bilinen birçok faktörü içermektedir. Puan ne kadar yüksek olursa müşterinin ayrılma olasılığı da o kadar yüksek olur.
Churn Score Category	Kategorilerden birine bir Kayıp Puanı atayan bir hesaplama: (0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, ve 91-100)
CLTV	Müşteri Yaşam Boyu Değer. Tahmin edilen CLTV, kurumsal formüller ve mevcut veriler kullanılarak hesaplanır. Değer ne kadar yüksek olursa müşteri o kadar değerli olur. Yüksek değere sahip müşteriler kayıp açısından izlenmelidir.
CLTV Category	Kategorilerden birine CLTV değeri atayan bir hesaplama: (2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, ve 6501-7000)
Churn Category	Doğrudan Kayıp Nedeniyle ilgilidir. Müşterinin vazgeçme nedeni için üst düzey bir kategori: (Tutum, Rakip, Memnuniyetsizlik, Diğer, Fiyat). Şirketten ayrılırken tüm müşterilere ayrılma nedenleri soruluyor.
Churn Reason	Bir müşterinin şirketten ayrılmasının özel nedeni. Doğrudan Kayıp Kategorisi ile ilgilidir.

Sadece tek bir değer yazdığı için anlamlı bilgi elde edilemeyen kolonlar (“Quarter”, “Country”, “State”, “Count”) silinmiştir. 3 kolon eksik değer içermektedir. Bu kolonlardan 2 tanesi (“Churn Category”, “Churn Reason”) incelendiğinde eksik değer içermediği, ilgili gözlemde yer alan müşterilerin şirketten ayrılmadığı için o değerin boş olduğu analiz edilmiştir. Bu boşluk değerleri “No Churn” ile doldurulmuştur. 11 tane boş değerin yer aldığı “Total Charges” kolonunun müşterinin şirkette bulunduğu toplam ay olan “Tenure in Months” kolonu ile her ay ödediği toplam ücret olan “Monthly Charges” kolonunda yer alan verilerin çarpımından elde edildiği analiz edilmiştir. Bu sebeple boş veriler çarpım matematik işlemiyle doldurulmuştur. “Churn Label”, “Churn Reason”, “Churn Category”, “Customer Status” kolonları bağımlı değişken olan “Churn Value” kolonuyla aynı

bilgileri içerdiği için, veri sızıntısını engellemek amacıyla ve aynı zamanda değerli bir bilgi içermediği için veri setinden çıkarılmıştır.

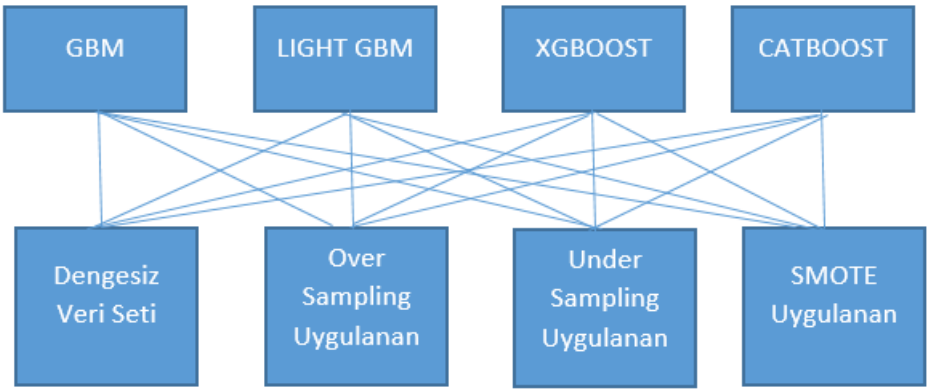
İkili (binary) kategorik kolonlar 1 ve 0 şeklinde değiştirilmiştir. Nominal kolon değerleri ise tek değer kodlama metoduyla kolondaki değerlerden yeni kolonlar türetilerek değiştirilmiştir. Tek değer kodlama metodunun asıl amacı aralarında büyüklük küçüklük ilişkisi olmayan kategorik değerlerden yeni kolonlar oluşturup bunları 1 ve 0 şeklinde ikili değerlere çevirmektir. Yaygın olarak kullanılan bu enkode yönteminin dezavantajlarından biri yüksek boyuta sebep olacağından dolayı hesaplama süresini ve maliyetini uzatmasıdır [22].

Makalede model performansları inceleneceği için fazla özellik mühendisliği yapılmadan farklı artırma algoritmalarının performansları kıyaslanmıştır.

Veri seti %80 eğitim seti, %20 test seti olmak üzere iki veri setine ayrılarak eğitim seti üzerinde Gradyan Artırma Algoritması, Hafif Gradyan Artırma Algoritması, Aşırı Gradyan Artırma Algoritması ve Catboost modelleri çalıştırılmıştır. Eğitim esnasında 5 katmanlı çapraz doğrulama metodu uygulanmıştır, bu sayede aşırı öğrenmenin önüne geçilmiştir. Çapraz doğrulama bir model doğrulama yöntemidir. Modelde kullanılan eğitim seti rastgele k farklı alt sete bölünür. Model bölünen alt kümelerden k-1 tanesiyle eğitilir ve geriye kalan diğer alt küme ile test edilerek performansı ölçülür. Bu işleme her bir farklı alt küme test seti olana kadar devam edilir. Her bir test setinden elde edilen hataların ortalaması alınarak çapraz doğrulamanın performansı bulunur. Bu yöntemin amacı hem modelin genellenmesini sağlamak hem de aşırı uyumun önüne geçmektir [23].

Hiperparametre optimizasyonu için Grid Araması (grid-search) tekniği kullanılmıştır. Bu teknikte makineye verilen hiperparametre olasılıklarının tümü deneme yanılma yöntemiyle tek tek kullanılıp en iyi performansı sağlayan parametreler seçilir.

Yukarıda bahsedildiği gibi veri seti dengesiz bir veri setidir. Bu sebeple modeller çoğunluk sınıfı öğrenmeye daha meyilli olmaktadırlar. Bu durumda dengesiz veri setlerinde sadece doğruluk (accuracy) değerlendirme metriğine bakmak araştırmacıyı yanıltabilir. Çünkü modelin doğruluk performansı ne kadar yüksek olsa da kaybedilen müşteriyi modeller öğrenememektedir. Doğruluk performansının yüksek olma sebebi kaybedilmeyen müşteri sayısının çok fazla olmasıdır. Bu durum gözlem sayısı az olan sınıf üzerinde büyük bir hakimiyet kurmaktadır. Bu sebeple veri seti dengesiz halden dengeli hale getirilmelidir. Bu çalışmada veri setine eksik örnekleme, aşırı örnekleme ve sentetik azınlık aşırı örnekleme metotları uygulanarak dengeli hale getirilmiştir. Aşırı örnekleme yönteminde eğitim setinde terk eden müşteri sayısı sentetik bir şekilde artırılarak terk etmeyen müşteri sayısına getirilmiştir. Eğitim setinde 1496 olan terk edenler sayısı 4138'e çıkarılmıştır. Eksik örnekleme yönteminde ise eğitim setindeki terk etmeyen müşteri sayısı azaltılarak terk eden müşteri sayısına eşitlenmiştir. Eğitim setinde 4138 olan terk etmeyenler sayısı 1496'ya düşürülmüştür. Burada dikkat edilmesi gereken nokta yeniden örnekleme yöntemlerinin sadece eğitim setine uygulanması gerekliliğidir. Bütün veri setine model eğitiminden önce uygulanır ise eğitim ve test seti ayrımında eğitim setindeki gözlemlere benzer üretilen sentetik veriler test setin içinde yer alabilir bu durum veri sızıntısına sebep olur ve gerçek hayatta şirketlerin kararlarını yanlış yönde etkiler. Sonuç olarak yukarıda bahsedilen tüm makine öğrenmesi algoritmaları toplamda 4 farklı veri setine uygulanmıştır (Dengesiz veri seti, aşırı örnekleme (over sampling), eksik örnekleme (under sampling), sentetik azınlık aşırı örnekleme (SMOTE)). Şekil 3'de veri setleri ve algoritmalar arasındaki ilişki açıkça görülebilir.



Şekil 3. Modellerin 4 farklı veri setine uygulanma şeması

3. Sonuçlar

Çalışma sonucunda elde edilen bulgular Tablo 2'de verilmiştir.

Tablo 2. Veri setlerine uygulanan modellerin performans sonuçları

Model	Metrik	Dengesiz veri seti	Under Sampling	Over Sampling	SMOTE
Light GBM	Accuracy	0,96	0,95	0,95	0,96
	F1 Score (0-1)	0,97-0,91	0,96-0,90	0,97-0,91	0,97-0,91
XGBoost	Accuracy	0,96	0,95	0,96	0,96
	F1 Score (0-1)	0,97-0,92	0,96-0,90	0,97-0,92	0,97-0,92
CatBoost	Accuracy	0,96	0,94	0,96	0,96
	F1 Score (0-1)	0,97-0,92	0,96-0,90	0,97-0,92	0,97-0,92
Gradient Boost	Accuracy	0,96	0,94	0,96	0,96
	F1 Score (0-1)	0,97-0,92	0,96-0,90	0,97-0,92	0,97-0,92

Model performanslarını ölçmek için Tablo 2'de yer alan doğruluk metriği sonuçlarına bakıldığında, dengesiz veri seti kullanılarak yapılan tüm tahmin modellerinin aynı performansı gösterdiği görülmektedir. Tabloda F1-Skoru için yer alan ikili sonuç değerleri sırasıyla terk etmeyen müşteriler ve terk eden müşterilerin değerleri şeklindedir. Bu metrik incelendiğinde dengesiz veri seti için Hafif Gradyan Artırma Algoritması dışındaki tüm algoritmalar aynı performansı göstermişlerdir. Yeniden örnekleme tekniklerine bakıldığında doğruluk metriği açısından herhangi bir performans artışı göze çarpmamaktadır. Aynı durum F1-Skoru için de geçerlidir.

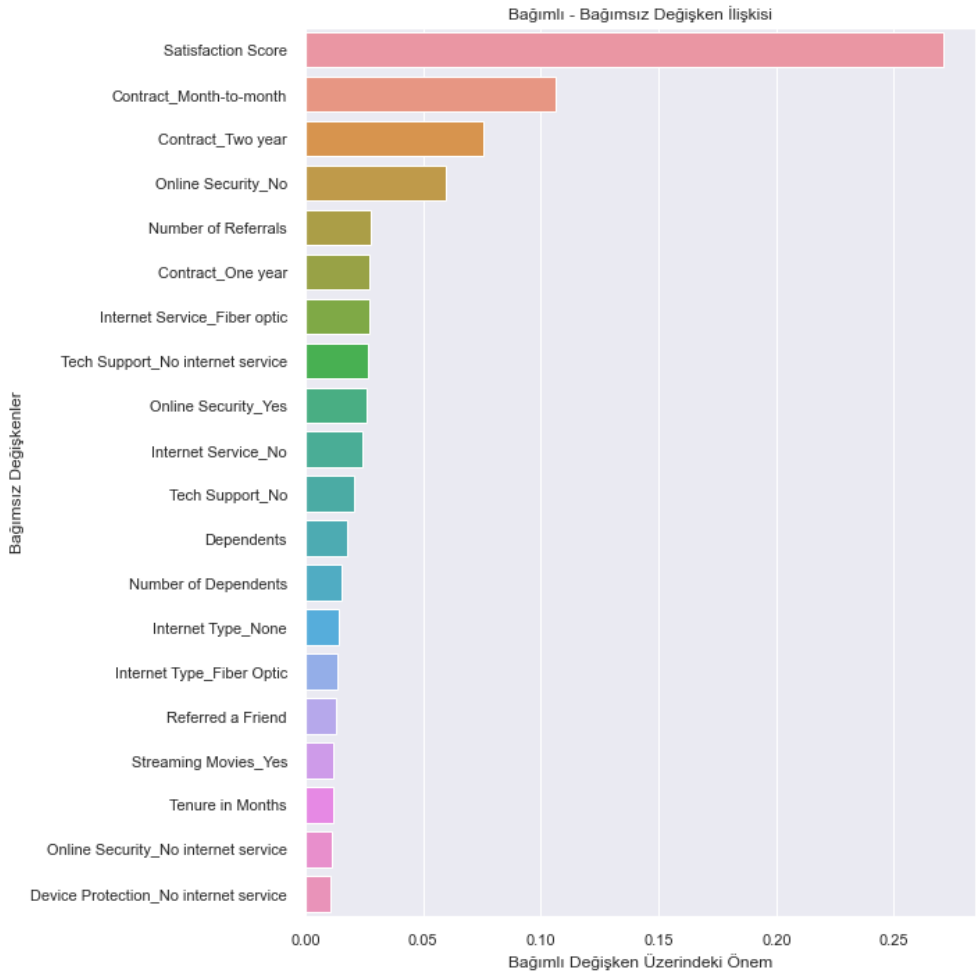
Genel anlamda performans yükselmemiş gibi görünse de bir şirket için önemli olan terk edecek olan müşteriyi doğru belirleyip bazı aksiyonlar alınmasıdır. Terk etmeyecek müşteriyi terk edecekmiş gibi tahmin edip bu duruma aksiyon alınması bir şirket için büyük bir problem olmazken asıl problem terk

edecek müşteriyi analiz edememektir. Bu sebeple bir diğer değerlendirme metriği olan duyarlılık metriğinin nasıl değiştiğini ölçmek önemlidir.

Tablo 3. Duyarlılık metriği sonuçları

Model	Dengesiz Veri Seti	Under Sampling	Over Sampling	SMOTE
Light GBM	0,887	0,946	0,884	0,898
XGBoost	0,892	0,949	0,906	0,895
Catboost	0,892	0,941	0,906	0,898
Gradient Boost	0,892	0,946	0,922	0,898

Yukarıda yer alan Tablo 3 incelendiğinde, dengesiz veri seti için yapılan modellerin duyarlılık sonuçlarına bakıldığında en yüksek duyarlılık değerini Hafif Gradyan Artırma Algoritması dışındaki diğer modeller 0,892 başarı oranıyla vermişlerdir. Yeniden örnekleme metotlarının kullanıldığı veri setleri incelendiğinde ise eksik örnekleme metodunun kullanıldığı veri setinde Aşırı Gradyan Artırma Algoritması 0,949 oranla en yüksek duyarlılık metriğine sahiptir. Başarı performansında %6,39 artış görülmektedir. Dolayısıyla sadece şirketi terk edecek müşteriyi tahmin etmede büyük bir performans artışı görülmüştür denebilir. Büyük müşteri kitlesine sahip şirketler için bu orandaki performans artışı, yüksek sayıdaki müşteriyi elde tutabilmek anlamına gelebilir.



Şekil 4. Bağımsız değişkenlerin önem sırası

Şekil 4 incelendiğinde Mutluluk skoru (Satisfaction Score), Aylık sözleşme (Contract_Month-to-month) ve 2 yıllık sözleşme (Contract Two year) kolonlarının bağımlı değişkene etki eden en önemli 3 özellik olduğu görülmektedir. Dolayısıyla bir müşterinin şirketi terk edip etmemesinde mutluluk skoru ve sözleşme tipleri önemli rol oynamaktadır.

4. Değerlendirme

Önemli bir konumda olan telekomünikasyon şirketleri, değişken ve hızlı büyüme göstermektedir. Analitik yöntemleri kullanarak yeni teknolojilerle birlikte ihtiyaçlara değer sağlayabilmeleri ve pazar dinamiğinde mücadele edebilmeleri önemlidir. Telekomünikasyon şirketlerinin sorunlarından biri olan müşteri kaybı analizi tahmine dayalı ve ileri analitik yöntemlerle çözüme ulaştırılabilir. Sonuç olarak bu çalışmada açık erişime sunulan telekomünikasyon şirketi verileri üzerinde makine öğrenmesi teknikleriyle modeller geliştirilerek performansları detaylı bir şekilde değerlendirilmiştir.

Bulgular, doğru başarı metriğini tespit ederek buna göre aksiyon almanın önemli olduğunu göstermiştir. Ayrıca, sadece doğruluk veya F1-Skoru metriğine odaklanmak performansta herhangi bir değişiklik göstermezken terk edecek olan müşteriyi gerçekten tespit etmek için duyarlılık metriğinin önemi ve bu metriğin nasıl geliştiği kanıtlanmıştır.

Çalışmada verinin dengesiz dağılması sınırlardan birini oluştursa da yeniden örnekleme teknikleri kullanılarak dengeli hale getirilen veri setiyle modellerin %6,39 oranında bir başarı artışı gösterdiği görülmüştür. Bu performans artışı küçük şirketler için çok fark yaratmayabilir ancak orta ve büyük ölçekli şirketler için çok fazla müşteri tespiti anlamına gelmektedir.

Çalışmada, literatürde halihazırda yer alan ve farklı sektör verileriyle yapılmış müşteri kaybı analizi çalışmalarında kullanılan makine öğrenmesi modelleri ve yeniden örnekleme yöntemlerinin ışığında, bu çalışmalarla ilişkili olarak telekomünikasyon şirketlerinin müşteri kaybı analizini ölçmesinin önemi vurgulanmıştır. Literatürde yer alan çalışmalara ek olarak doğruluk metriğinin de müşteri kaybı analizinde önemli bir rol oynadığı ve şirketlerin elde tutmaları gereken müşteriye odaklanmanın dışında aynı zamanda gidecek olan müşteriye de dikkat etmesinin gerekli olduğu ve bunun için de Aşırı Gradyan Artırma Algoritması'nın diğer artırma metotlarına göre daha iyi performans sağladığı gösterilmiştir.

Şirketi terk eden müşterilerin tespitinde mutluluk skorunun önemli olduğu vurgulanmıştır. Müşterilerin kaybını azaltmak için daha sıklıkla mutluluk anketleri yapılarak müşteri kaybının yaşanmaması için önlemler alınarak bu durum azaltılabilir.

İleriki araştırmalarda daha fazla geliştirme için farklı özellik mühendisliği teknikleri uygulanabilir. Literatürde yer alan diğer yeniden örnekleme teknikleri kullanılarak oluşturulan farklı veri setleri arasında performans kıyaslaması yapılabilir. Veri setinin doğal yollardan dengeli hale getirilebilmesi için veri setine dengeli hale getirecek yeni müşteri verileri eklenebilir.

Deklarasyon ve Etik Standartlar

Yazarlar bu makalenin araştırılması, yazarlığı ve/veya yayınlanmasıyla ilgili olarak herhangi bir potansiyel çıkar çatışması beyan etmemiştir. Bu makalenin yazarları, bu çalışmada kullanılan materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel izin gerektirmediğini beyan eder.

Yazar Katkısı

Bütün yazarlar birlikte sunulan fikri tasarladı, teoriyi geliştirdi, hesaplamaları yaptı ve deneyleri gerçekleştirdi ve bu çalışmanın bulgularını denetledi. Bütün yazarlar birlikte sonuçları tartıştı ve makaleyi son haline getirdi.

Kaynaklar

[1] Çiçek, A., & Arslan, Y., "Müşteri Kayıp Analizi İçin Sınıflandırma Algoritmalarının Karşılaştırılması," İleri Mühendislik Çalışmaları ve Teknolojileri Dergisi, 1(1), 13-19, (2020).

[2] Gold, C, "Fighting churn with data. the science and strategy of customer retention," Manning Pubn, (2020).

[3] Almana, A. M., Aksoy, M. S., & Alzahrani, R., "A survey on data mining techniques in customer churn analysis for telecom industry," International Journal of Engineering Research and Applications, 4(5), 165-171, (2014).

[4] Arnold, T. J., (Er) Fang, E., & Palmatier, R. W., "The effects of customer acquisition and retention orientations on a firm's radical and incremental innovation performance," Journal of the Academy of Marketing Science, 39, 234-251, (2011).

[5] Gallo, A., "The value of keeping the right customers.Harvard Business Review," https://hbr.org/2014/10/the-value-of-keeping-the-right-customers, (2014, Ekim 29)

[6] Huang, B. Q., Kechadi, T. M., Buckley, B., Kiernan, G., Keogh, E., & Rashid, T., "A new feature set with new window techniques for customer churn prediction in land-line telecommunications," Expert Systems with Applications, 37(5), 3657-3665, (2010).

[7] Ahmad, A. K., Jafar, A., & Aljoumaa, K., "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, 6(1), 1-24, (2019). https://doi.org/10.1186/s40537-019-0191-6

[8] Brandusoiu, I., Toderean G., & Beleiu, H., "Methods for churn prediction in the prepaid mobile telecommunications industry," Konferans Sunumu, 2016 International Conference on Communications (COMM), Bükreş, Romanya, (2016, Haziran 09-10) https://doi.org/10.1109/ICComm.2016.7528311

[9] Öztürk, M.E., Tunç, A.A., & Akay, M.F., "Machine learning based churn analysis for sellers on the e-commerce Marketplace," International Journal of Mathematics and Computer in Engineering 1(2), 171–176, (2023). https://doi.org/10.2478/ijmce-2023-0013

[10] An, Z., Song, Z., & Wang, X., "Bank Customer Churn Based on Different Models, Oversampling, and Encoding Methods," BCP Business & Management 26, 703-713, (2022). https://doi.org/10.54691/bcpbm.v26i.2030

[11] Sayed, H., Abdel-Fattah, M. A., & Kholief, S., "Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study," International Journal of Advanced Computer Science and Applications, 9(11), (2018).

[12] Saleh, S., Saha, S., "Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university," SN Applied Sciences, 5(7), 173, (2023).

[13] Çelik, S., Tayalı, S.T., "Resampling and Ensemble Strategies for Churn Prediction," Bilişim Teknolojileri Dergisi, 16(4), 263-273, (2023).

[14] Verma, P., "Churn prediction for savings bank customers: A machine learning approach," Journal of Statistics Applications & Probability, 9(3), 535-547, (2020).

[15] Friedman, J. H., "Greedy function approximation: a gradient boosting machine," Annals of statistics, 1189-1232, (2001).

[16] Chen, T., & Guestrin, C., "Xgboost: A scalable tree boosting system," Konferans Sunumu, 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, CA, USA, (2016, Ağustos 13-17) https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

[17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y., "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, 30, (2017).

[18] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A., "CatBoost: unbiased boosting with categorical features," Advances in neural information processing systems, 31, (2018).

[19] Dorogush, A. V., Ershov, V., & Gulin, A., "CatBoost: gradient boosting with categorical features support," arXiv preprint arXiv:1810.11363, (2018).

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

[21] Microsoft., "Makine öğrenimi modellerinin sonuçları," https://learn.microsoft.com/tr-tr/dynamics365/finance/finance-insights/confusion-matrix, (2023, Mart 08)

[22] Udila, A., "Encoding methods for categorical data: A comparative analysis for linear models, decision trees, and support vector machines," Lisans Tezi, TU Delft Electrical Engineering, Mathematics and Computer Science, (2023).

[23] Berrar, D., "Cross-Validation," In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier Science: Amsterdam, The Netherlands , 542-545, (2019).

Yazar Biyografileri

Yazar 1	Başak Ceren SEÇİK GÖÇER Lisans: Yıldız Teknik Üniversitesi, Kimya-Metalurji Fakültesi, Matematik Mühendisliği Yüksek Lisans: Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Mühendisliği
Yazar 2	İbrahim EMİROĞLU Lisans: Yıldız Teknik Üniversitesi, Mühendislik Fakültesi, Matematik Mühendisliği Yüksek Lisans: Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Mühendisliği Doktora: University of Hertfordshire, Elektrik Elektronik Fakültesi, Elektronik Mühendisliği, Birleşik Krallık