POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Müşteri kaybı tahmininde sınıf dengesizliği problemi

Class imbalance problem in churn prediction

***Yazar(lar) (Author(s))**: M. Aslı AYDIN[1]*

ORCID[1]: 0000-0002-8905-7518

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Aydın M. A., "Müşteri kaybı tahmininde sınıf dengesizliği problemi", ***Politeknik Dergisi***, 25(1): 351-360, (2022).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.734916**

Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi

Class Imbalance Problem in Churn Prediction

Önemli noktalar (Highlights)

- *Telekomünikasyon sektöründe Müşteri Kaybı Tahmini problemi ele alınmıştır. / We consider Churn Prediction in Telecommunication Industry.*
- *Bu problem verisinde tipik olan Sınıf Dengesizliği Problemine değinilmiştir. / We refer the Class Imbalance Problem which is typical in data of this problem.*
- *Yeniden Örnekleme tekniklerinin sınıf dengesizliği problemini çözmekteki etkisi gözlenmiştir. / We observe the effect of Resampling Techniques on handling class imbalance.*
- *Makine Öğrenmesi tekniklerinin karşılaştırmalı çalışması sunulmaktadır. / This is a comparative study of Machine Learning Techniques.*
- *Yeniden Örnekleme ve Çapraz Geçerlemenin birlikte doğru kullanımına değinilmiştir. / We refer to correct way of using Resampling and Cross Validation jointly.*

Grafik Özet (Graphical Abstract)

Sınıf dengesizliği bulunan Telekom verisinde makine öğrenmesi yöntemleriyle müşteri kaybı tahmini/Churn prediction with machine learning techniques on imbalanced Telecom data

Şekil. Çalışmada Kullanılan Adımlar/ **Figure.** Steps used in this work

Amaç (Aim)

Sınıf dengesizliği bulunan Telekom verisi kullanılarak müşteri kaybı tahmininde yeniden örnekleme ve çapraz geçerlemenin birlikte doğru kullanımının makine öğrenmesi tekniklerinin performansı üzerindeki etkisinin incelenmesi / We investigate the effect of correct use of resampling and cross validation on the performance of machine learning algorithms in churn prediction on an imbalanced Telecom data.

Tasarım ve Yöntem (Design & Methodology)

Açık erişimli Telekom verisinde önce öznitelik seçimi yapılmış, yeniden örnekleme teknikleri ile sınıf dengesizliği problemi çözülmüştür. Makine öğrenmesi teknikleri ile sınıflandırma problemi çözülerek performans değerlendirmesi yapılmıştır. / First we apply feature selection on a publicly available Telecom data. Then we use resampling techniques to handle class imbalance. We use machine learning algorithms for binary classification and compare their performances.

Özgünlük (Originality)

Bu çalışma sınıf dengesizliği bulunan Telekom verisine uygulanan 7 farklı makine öğrenmesi ve 6 farklı yeniden örnekleme tekniğinden elde edilen sonuçların karşılaştırmalı bir analizini sunmaktadır. Yeniden örnekleme ve çapraz geçerlemenin birlikte doğru kullanımının önemini vurgulamaktadır. / We present a comparative analysis of seven machine learning algorithms and six resampling techniques for churn prediction on an imbalanced Telecom data. We emphasize the correct way of using resampling and cross validation jointly.

Bulgular (Findings)

Yeniden örnekleme ve çapraz geçerleme doğru kullanıldığında ROC Eğrisi ölçüsüne göre en fazla performans artışı ROSE-SMO birlikte kullanıldığı durumda elde edilmiştir. /Among the others, we observe the highest increase in ROC curve performance measure by resampling the data with ROSE and applying Support Vector Machines.

Sonuç (Conclusion)

Sınıf dengesizliği probleminde yeniden örneklemenin sınıflandırıcı performansını artırabileceği ancak çapraz geçerlemenin birlikte doğru kullanılmadığında sınıflandırıcı performansını olduğundan daha iyi gösterebileceği gözlenmiştir./ We observe an increase in the classifier's performance with a resampling method while handling imbalanced data, however it may cause overoptimism in case of incorrect use jointly with cross validation.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi

Araştırma Makalesi / Research Article

M. Aslı AYDIN*

Mühendislik Fakültesi, Endüstri Mühendisliği, Doğuş Üniversitesi, Türkiye

(Geliş/Received : 09.05.2020 ; Kabul/Accepted : 14.09.2020 ; Erken Görünüm/Early View : 28.09.2020)

ÖZ

Müşteri kaybı tahmini, müşteri verilerinin incelenerek ayrılması muhtemel müşterinin önceden tespit edilmesidir. Çözümünde makine öğrenmesi yöntemlerinden faydalanılmaktadır. Yapısı itibariyle veride *Ayrılan* sınıfındaki veri sayısının *Ayrılmayan* sınıfındakinden çok daha az olduğu gözlenir. Dengesiz sınıf dağılımı, makine öğrenmesi yöntemlerinin performansını olumsuz etkilediğinden verinin dengelenmesi önemlidir. Çalışmada telekomünikasyon sektöründeki müşteri kaybı tahminine odaklanılmıştır. Uygulama, açık erişimli veri tabanından elde edilen 7043 müşteriye ait 21 öznitelik içeren veri üzerinde gerçekleştirilmiştir. Öncelikle Minimum Fazlalık Maksimum Bağımlılık yöntemiyle öznitelik seçimi yapılmıştır. Yeniden örnekleme, Sentetik Azınlık Aşırı Örnekleme Yöntemi (SMOTE), Uyarlanabilir Sentetik Örnekleme (ADASYN), Çoğunluk Ağırlıklı Azınlık Örnekleme (MWMOTE), Hızlı Yakınsayan Gibbs Algoritması (RACOG), Rastgele Yürüyüş Aşırı Örnekleme (RWO) ve Rastgele Aşırı Örnekleme (ROSE) yöntemleriyle uygulanmıştır. Sınıflandırma problemi için Naïve Bayes, Karar Ağaçları, Rastgele Orman, Yapay Sinir Ağları, Lojistik Regresyon, Destek Vektör Makineleri ve K-En Yakın Komşuluk yöntemleri 10 kat Çapraz Geçerlemeyle uygulanmıştır. Deneyler yeniden örneklemenin çapraz geçerlemeden önce ve çapraz geçerleme sırasında uygulandığı iki farklı yaklaşımla gerçekleştirilmiştir. Yöntemlerin performansı Doğruluk, Kesinlik, Duyarlılık, F-Ölçütü, Alıcı İşletim Karakteristiği Eğrisiyle ölçülmüştür. Uygulanan yöntemlerin performansı orijinal verininkilerle kıyaslanmıştır. Destek Vektör Makinelerinin performansında ROSE'la çapraz geçerleme esnasında yeniden örneklenmiş veride orijinal veriye göre %5.7 iyileşme gözlenmiş, yeniden örnekleme çapraz geçerlemeden önce yapıldığında iyileşme miktarının gerçek değerinin üzerinde kaydedildiği sonucuna ulaşılmıştır.

Anahtar Kelimeler: Müşteri kaybı tahmini, makine öğrenmesi, sınıf dengesizliği problemi, telekomünikasyon.

Class Imbalance Problem In Churn Prediction

ABSTRACT

Customer churn prediction is determining the customers most likely to leave by examining customer data. Machine learning is one of the solution approaches. Class imbalance is typical for this problem since more customers are labeled as *Non-Churn* than *Churn.* Handling class imbalance is crucial since the classifier's performance is highly affected. This work focuses on churn prediction in telecommunications. We use publicly available churn data from 7043 customers having 21 features. We use Minimal Redundancy Maximal Relevance for feature selection. We handle class imbalance problem with resampling methods including Synthetic Minority Oversampling Technique, Adaptive Synthetic Sampling, Majority Weighted Minority Oversampling, Rapidly Converging Gibbs Algorithm, Random Walk Oversampling and Random Oversampling Examples. We employ classifiers including Naïve Bayes, Decision Trees, Random Forest, Artificial Neural Networks, Logistic Regression, Support Vector Machines and k-Nearest Neighbours with 10-fold cross validation (CV). We try two approaches in experiments: i) resampling during CV and ii) resampling before CV. We compare the results with original data using Accuracy, Precision, Recall, F-measure and ROC-Curve as performance measures. The results show 5.7% increase in model performance for Support Vector Machines with ROSE when we apply first approach. We observe that classifier's performance is overoptimized if second approach is applied.

Keywords: Churn prediction, machine learning, class imbalance problem, telecommunication.

1. GİRİŞ (INTRODUCTION)

Müşteri kaybı telekomünikasyon, bankacılık, sigortacılık, oyun ve eğlence gibi üyelik temeline dayanan sektörlerdeki firmalar için müşterinin üyesi olduğu bir firmadan diğer firmanın üyeliğine geçmesi olarak tanımlanabilir.

Yeni müşteri kazanmak, mevcut müşteriyi korumaktan daha maliyetli olduğu için firmalar müşterilerinin rakip firmalara geçişini en aza indirmek isterler [1]. Bu yüzden müşteri davranışlarını inceleyerek ayrılma olasılığı yüksek müşterileri önceden tespit edip düzenleyecekleri kampanya ve promosyon gibi teşviklerle müşterilerinin üyeliğinin devamını sağlamak isterler. Bu müşterilere sağlanacak teşviklerin de firma için bir maliyetinin olduğu düşünülünce firmadan ayrılmaya eğilimli müşterilerin doğru tahmin edilmesi önem kazanmaktadır.

Müşteri kaybı yaşanan sektörler içinde özellikle telekomünikasyonda, operatörler arası numara taşımanın kolaylaşması, gelişen teknoloji ile müşteri servis ve hizmetlerinde beklentinin artması, operatörler arasındaki yoğun rekabet gibi sebeplerle müşteri hareketliliği sıkça

*Sorumlu Yazar (Corresponding Author)
e-posta : aaydin@dogus.edu.tr

görülmektedir. Koçoğlu vd. tarafından yapılan çalışmada telekomünikasyon sektörünün müşteri kaybı tahmininde diğer sektörlere göre %60 oranla üzerinde en fazla çalışılan sektör olduğu belirlenmiştir [2]. Bu çalışmada da telekomünikasyon sektöründe müşteri kaybı tahminine odaklanılmıştır.

Müşteri kaybı tahmini problemi yapı olarak müşterilerin *Ayrılanlar* ve *Ayrılmayanlar* olarak iki farklı şekilde sınıflandırılması (İkili Sınıflandırma) problemi olarak da tanımlanabilir. Bu yüzden çözüm yolları arasında makine öğrenmesinin sınıflandırma yöntemleri de sıkça yer almaktadır. Bunlar arasında Naïve Bayes, Karar Ağaçları, Rastgele Orman, Yapay Sinir Ağları, Lojistik Regresyon, Destek Vektör Makineleri ve K-En Yakın Komşuluk en yaygın kullanılan sınıflandırıcılar olarak sayılabilir. Telekomünikasyon sektöründe müşteri kaybı tahmini için son yıllarda yapılan çalışmalar arasında Yapay Sinir Ağları yönteminin tahmin başarısının diğer yöntemlere göre daha yüksek olduğunu görülmektedir [3-5]. Ayrıca Günay ve Ensarı, Lojistik Regresyon ve Naïve Bayes tabanlı geliştirilen hibrit bir yöntemin tek başlarına Lojistik Regresyon ve Naïve Bayes yöntemlerine göre daha iyi sonuç vermesine rağmen Yapay Sinir Ağlarından daha iyi sonuç alınamadığı bildirmektedir [5]. Yıldız ve Albayrak, Aykırı Değer Analizi ve öznitelik seçiminin sınıflandırma performansına etkisini gözlemlemişlerdir [6]. Vafeiadis vd. müşteri kaybı tahmininde makine öğrenmesi yöntemlerinin karşılaştırmasını sunarken, kullanılan yöntemlerin tahmin başarısını artırmak amacıyla yükseltme yöntemlerini de denemiş ve başarılı olunduğunu gözlemlemiştir [7]. Ullah vd. ise önce öznitelik seçimi yaparak Rastgele Orman sınıflandırıcısını kullanmışlardır. Ayrıca yükseltme ve torbalama yöntemlerini de kullanıp bunların tahmin başarısı üzerindeki etkilerini incelemişlerdir [8]. Amin vd. çalışmasında öznitelik ağırlıklarının Genetik Algoritma kullanılarak belirlendiği Naïve Bayes yöntemi denemiştir [9]. Öznitelik ağırlıklarının rastgele atandığı duruma göre daha iyi tahmin oranı elde etmiştir.

Uygulanan makine öğrenmesi yönteminin başarısını verinin özellikleri de etkilemektedir. Gözetimli öğrenme yöntemleri ile çözülecek ikili sınıflandırma problemlerinde kullanılan verilerde tahmin edilecek bir sınıfa ait örnek sayısı diğer sınıfa ait örnek sayısından çok daha fazla ise bu veri Dengesiz Veri olarak bilinir [10]. Telekomünikasyon sektöründen elde edilen birçok veride *Ayrılanlar* oranı *Ayrılmayanlar* oranından oldukça az olduğundan müşteri kaybı tahmini problemi verisi dengesiz veri tanımına uygundur [11]. Makine öğrenmesi yöntemleri sınıflar arasındaki oran farklılıklarını göz önüne alacak şekilde tasarlanmadıklarından bu tip verilerle çalışılırken yanıltıcı sonuçlar üretebilirler.

Sınıf dengesizliği bulunan veriye makine öğrenmesi yöntemleri uygulandığında karşılaşılan problemler ile başa çıkmanın çeşitli yolları vardır. Bunlardan biri makine öğrenmesi yöntemini değerlendirmek için doğru performans ölçütünün seçilmesidir. Örneğin, *Ayrılanlar* sınıfındaki örnek oranı tüm verideki örnek sayısının sadece %5'i kadar olduğu dolayısıyla tahmin sınıfları arasında dengesiz dağılım olan bir veriyle çalışıldığını düşünelim. Böyle bir veri ile çalışılırken elde edilen yüksek doğru tahmin oranı, kullanılan makine öğrenmesi yönteminin gerçek performansını yansıtmayabilir. Çünkü tüm müşteriler *Ayrılmayanlar* şeklinde tahmin edildiğinde doğru tahmin oranı %95 olmasına rağmen ayrılacak olan müşterilerin hiçbiri doğru tahmin edilememiş olacaktır. Dolayısıyla dengesiz veri ile çalışılırken doğru tahmin oranını veren Doğruluk ölçütü dışında Kesinlik (Precision), Duyarlılık (Recall), Alıcı İşletim Karakteristiği (Receiver Operating Curve-ROC) Eğrisi, F-Ölçütü gibi başka performans ölçütlerini de kullanmak daha sağlıklı bir değerlendirme sunacaktır.

Dengesiz veri ile başa çıkmanın bir diğer yolu ise verinin yeniden örnekleme yöntemleriyle dengeli hale getirilmesidir. Sınıf dengesizliği öğrenme yöntemleri i) Alt Örnekleme (undersampling), ii) Aşırı Örnekleme (oversampling) ve iii) Sentetik Veri Üretilmesi şeklinde üç başlıkta toplanabilir. Yeniden örnekleme yöntemleriyle veriyi dengelemek farklı performans ölçütlerinin sonuçlarının birbirleriyle daha uyumlu olmasını sağlayabilmektedir. Durahim çalışmasında farklı örnekleme yöntemlerinin çalışma süresini, yeniden örnekleme yapılmış verinin üzerinde makine öğrenmesi yöntemlerinin doğruluklarını kıyaslamıştır [12].

Telekomünikasyon sektöründe müşteri kaybı tahmini için yapılan çalışmalar arasında da dengesiz veriye değinen çalışmalar mevcuttur. Bunlar arasında, Effendy vd. %0.7 *Ayrılanlar* oranına sahip verideki dengesizlik etkisini azaltmak için Alt Örnekleme ve Sentetik Azınlık Aşırı Örnekleme Yönteminin (Synthetic Minority Oversampling Technique- SMOTE) kombinasyonu bir örnekleme yöntemi geliştirmişlerdir [13]. Model performans ölçütü olarak F-Ölçüsünü kullanmışlardır. Geliştirdikleri örnekleme yönteminin sonucunun örnekleme yöntemi kullanılmadığı ve varolan örnekleme yöntemlerinin kullanıldığı durumlardan daha iyi sonuç verdiğini gözlemlemişlerdir. Amin vd. ise dört farklı veri seti üzerinde Mega-Trend Dağılım Fonksiyonu (Mega-Trend Diffusion Function-MTDF), SMOTE, Uyarlanabilir Sentetik Örnekleme (Adaptive Synthetic Sampling-ADASYN), En Üst-N Ters K-En Yakın Komşuluk (Top-N Reverse k-Nearest Neighbours-TRkNN), Çoğunluk Ağırlıklı Azınlık Örnekleme (Majority Weighted Minority Oversampling-MWMOTE) ve Bağışıklık Merkezli Aşırı Örnekleme (Immune Centroids Oversampling Technique-ICOTE) yöntemlerini kıyaslamışlar, MTDF yönteminin tahmin başarısına etkisinin en fazla olduğunu söylemişlerdir [14]. Aditsania vd. %4 *Ayrılanlar* oranına sahip veri seti üzerinde ADASYN yeniden örnekleme yöntemini kullanarak F-Ölçüsünde %80 civarında iyileşme sağlamışlardır [15]. Koçoğlu ve Özcan %14.14 *Ayrılanlar* oranına sahip veri üzerinde Aşırı Örnekleme, Alt Örnekleme, SMOTE ve Rastgele Aşırı Örnekleme (Random Oversampling Examples-ROSE) yöntemlerinin makine öğrenmesi modelinin performansına etkisini kıyaslamışlardır [16]. Buna göre dengesiz veri probleminin çözümünde Aşırı Örnekleme

veya SMOTE yöntemlerinin öncelikle tercih edilebilecek yöntemler olduğu sonucuna ulaşmışlardır.

Makine öğrenmesi yöntemlerinin performansı değerlendirilirken bağımsız bir test verisi bulunmuyorsa kullanılan yöntemlerden biri Çapraz Geçerleme (Cross Validation) 'dir. *k*-kat çapraz geçerleme yönteminde eldeki veri *k* parçaya bölünür. Parçalardan biri test için ayrılırken geriye kalan *k-1* parça makine öğrenmesi yönteminin eğitiminde kullanılır. Bu işlem *k* parçadan her biri test verisi olarak kullanılacak şekilde tekrarlanır. Makine öğrenme yönteminin performansı elde edilen sonuçların ortalaması alınarak bulunur.

Sınıf dengesizliği bulunan veriyle çalışırken çapraz geçerleme ve yeniden örnekleme yöntemlerini birlikte kullanmak gerekebilir. Böyle bir durumda bu yöntemlerin uygulanma sırası, makine öğrenmesi yönteminin performansının doğru değerlendirilebilmesi için önemlidir. Literatürde tüm veri yeniden örnekleme ile dengeli hale getirildikten sonra çapraz geçerleme uygulayan çalışmalar görmek mümkündür [13,15]. Ancak doğru yaklaşım yeniden örneklemenin çapraz geçerleme esnasında uygulanması olmalıdır. Buna göre, önce verinin test ve eğitim verisi olarak parçalanmalı, ardından yeniden örnekleme tekniği sadece eğitim verisine uygulanarak makine öğrenmesi yöntemi eğitilmelidir. Performans ölçümü yeniden örnekleme uygulanmayan test verisi üzerinde yapılmalıdır. Aksi halde, test verisinde de yeniden örnekleme ile üretilmiş, makine öğrenmesi yönteminin eğitildiği veriyle aynı ya da çok benzer veriler bulunur. Bu da makine öğrenmesi yönteminin performansının olduğundan daha iyi olduğu algısına yol açacaktır [17].

Bu çalışmada müşteri kaybı tahmini problemi için açık erişimli, dengesiz sınıf dağılımı gösteren bir telekomünikasyon verisi [18] çeşitli yeniden örnekleme teknikleri ile dengeli hale dönüştürülmüş ve farklı makine öğrenmesi yöntemleri ile elde edilen modellerin performansları kıyaslanmıştır. Çalışmanın katkısı aşağıdaki gibi özetlenebilir:

i) Sınıf dengesizliği bulunan Telekom verisine uygulanan 7 farklı makine öğrenmesi ve 6 farklı yeniden örnekleme tekniğinden elde edilen sonuçların karşılaştırmalı bir analizini sunmak,

ii) Yeniden örnekleme ve çapraz geçerleme yöntemlerinin birlikte kullanıldığı durumda işlem sırasının model performansına etkisini sergilemek.

Bu çalışma aşağıdaki şekilde düzenlenmiştir. 2. Bölümde üzerinde çalışılan veri, veri ön işleme işlemlerinden bahsedilmiştir. Ayrıca kullanılan yeniden örnekleme yöntemleri sunulmuştur. 3. Bölümde elde edilen test sonuçları verilmiş ve bunların değerlendirilmesi yapılmıştır. 4. Bölümde ise çalışmada elde edilen sonuçlar özetlenmiştir.

2. YÖNTEM (METHODOLOGY)

2.1. Veri Kümesi (Input Data)

Çizelge 1. Öznitelik İsimleri ve Türleri (Features' Name and Type)

No	Öznitelik	Açıklama	Türü
1	*customerID*	Müşteri Numarası	Nümerik
2	*gender*	Cinsiyet	Kategorik
3	*SeniorCitizen*	Yaşlı/Emekli olma durumu	Kategorik
4	*Partner*	Eş Durumu	Kategorik
5	*Dependents*	Bağlılık Durumu	Kategorik
6	*tenure*	Abone olma süresi (ay)	Nümerik
7	*PhoneService*	Telefon servisi alma durumu	Kategorik
8	*MultipleLines*	Birden fazla hatta sahip olma durumu	Kategorik
9	*InternetService*	İnternet servis sağlayıcısı	Kategorik
10	*OnlineSecurity*	Çevrimiçi güvenliğinin olma durumu	Kategorik
11	*OnlineBackup*	Çevrimiçi yedeği sahip olma durumu	Kategorik
12	*DeviceProtection*	Cihaz korumasına sahip olma durumu	Kategorik
13	*TechSupport*	Teknik destek alma durumu	Kategorik
14	*StreamingTV*	Internet TV kullanma durumu	Kategorik
15	*StreamingMovies*	Internet Film kullanma durumu	Kategorik
16	*Contract*	Kontrat tipi	Kategorik
17	*PaperlessBilling*	Elektronik fatura kullanma durumu	Kategorik
18	*PaymentMethod*	Ödeme şekli	Kategorik
19	*MonthlyCharges*	Aylık ödeme miktarı	Nümerik
20	*TotalCharges*	Toplam ödeme miktarı	Nümerik
21	*Churn (Sınıf Değişkeni)*	Ayrılma durumu	Kategorik

Bu çalışmada açık erişimli bir telekomünikasyon verisi kullanılmıştır [18]. Veride 7043 müşteriye ait biri sınıf değişkeni olmak üzere farklı türlerde 21 öznitelik vardır. Eksik veri bulunmamaktadır. Öznitelikler ve ilgili bilgiler Çizelge 1'de görülebilir. İkili sınıf değişkeni *Ayrılan / Ayrılmayan* olarak tanımlanmış, 1849 müşteri *Ayrılan* olarak etiketlenmiştir. *Ayrılan* etiketine sahip müşteri sayısı verideki tüm müşterilerin yaklaşık %26'sını oluşturduğundan sınıf niteliğine göre dengesiz bir dağılım olduğu anlaşılmaktadır.

2.2 Veri Ön İşleme (Data Preprocessing)

Veri ön işleme aşamasında öncelikle deneylerde kullanılacak öznitelikler seçilmiştir. Öznitelik seçimi ile veriyi açıklayıcı özelliği daha fazla olan özniteliklerin belirlenmesi amaçlanır. Örneğin ayrılacak müşteri tahmininde Abone olma Süresi (*tenure*) 'nin Müşteri Numarası (*customerID*) 'ndan daha belirleyici olduğu açıktır. Bunun gibi veride bulunan gereksiz özniteliklerin belirlenerek çıkarılması veri boyutunu küçülterek hesaplama yükünün azalmasına, makine öğrenmesi yönteminin daha anlamlı öznitelik kümesi ile çalışmasını sağlayarak daha iyi performans sergilemesine yarar.

Öznitelik seçme işlemi filtre yöntemleri (filter methods), zarflayıcı yöntemler (wrapper methods) ve gömülü yöntemler (embedded methods) olarak üç grupta incelenebilir. Bu çalışmada öznitelik seçimi için zarflayıcı yöntem kategorisindeki Minimum Fazlalık Maksimum Bağımlılık (Minimum Redundancy Maximum Relevance – mRMR) yöntemi kullanılmıştır. Bu yöntemin tercih edilme nedeni, yöntemin bir özniteliği sadece tahmin için uygunluğuna göre seçmeyi değil aynı zamanda seçilen diğer özniteliklerle korelasyonunu değerlendirerek veri fazlalığını azaltmayı amaçlamasıdır. Yöntem aynı veri için Amin vd. [14] tarafından da kullanılmıştır. Seçilen 10 öznitelik, sınıf değişkeniyle birlikte Çizelge 2' de görülebilir.

Çizelge 2. mRMR ile Seçilen Öznitelikler (Features Subset After mRMR)

Öznitelik	**Öznitelik**
gender	*PhoneService*
SeniorCitizen	*DeviceProtection*
Partner	*Contract*
Dependents	*PaperlessBilling*
tenure	*PaymentMethod*
Churn (Sınıf Değişkeni)	

2.3 Yeniden Örnekleme Yöntemleri (Resampling Techniques)

Verinin dengeli hale getirilmesi için yeniden örnekleme yöntemleri uygulanmıştır. Yeniden örnekleme yöntemleri Alt Örnekleme (undersampling), Aşırı Örnekleme (oversampling) ve Sentetik Veri Üretilmesi şeklinde üç grupta incelenebilir. Bunlardan Alt Örnekleme, baskın sınıfa ait verilerden rastgele seçilen bir bölümün silinmesiyle gerçekleşir. Bu yöntem veriyi daha dengeli getirmenin yanında, özellikle veri boyutunun çok büyük olduğu durumlarda daha küçük boyutlu veriyle çalışılmasını sağladığından sınıflandırma yönteminin çalışma süresini kısaltabilir. Diğer yandan, silinen verilerdeki bilgiler kaybolduğundan yetersiz uyum (underfitting) problemine yol açabilir.

Aşırı Örnekleme ise azınlık sınıfına ait verilerden rastgele seçilen bir bölümünün tekrarlanarak bu sınıfa ait veri sayısının artırılması olarak tanımlanabilir. Bu yöntemde bilgi kaybı olmadığından alt örneklemeye göre daha üstün bir yöntemdir. Ancak, verilerin bir kısmı aynen tekrarlandığı için aşırı uyum (overfitting) problemine yol açabilir.

Sentetik Veri Üretilmesi ileri aşırı örnekleme yöntemi olarak tanımlanabilir. Azınlık sınıfına ait verilerin bir bölümünün rastgele seçilerek tekrarlanması yerine bu sınıfa ait yeni verilerin belli bir algoritma ile üretilmesi yöntemidir. Bu çalışma Sentetik Azınlık Aşırı Örnekleme Yöntemi (SMOTE), Uyarlanabilir Sentetik Örnekleme (ADASYN), Çoğunluk Ağırlıklı Azınlık Örnekleme (MWMOTE), Hızlı Yakınsayan Gibbs Algoritması (Rapidly Converging Gibbs Algorithm-RACOG), Rastgele Yürüyüş Aşırı Örnekleme (Random Walk Oversampling-RWO) ve Rastgele Aşırı Örnekleme (ROSE) sentetik veri üretme tekniklerinin karşılaştırmasını sunmaktadır. Bu yöntemlere ait kısa bilgiler aşağıda yer almaktadır.

2.3.1 Sentetik azınlık aşırı örnekleme yöntemi (Synthetic minority oversampling technique-SMOTE)

En yaygın kullanılan sentetik veri üretme tekniklerinden biri Sentetik Azınlık Aşırı Örnekleme yöntemidir. Chawla vd. tarafından geliştirilen bu yöntem birçok sentetik veri üretme algoritması için de temel oluşturur [19]. Ana fikri şöyle özetlenebilir. Azınlık sınıfından rastgele bir örnek, *i*, seçilir. Daha sonra bu örneğin yine azınlık sınıfına ait *k* en yakın komşusu belirlenir. *k* en yakın komşu içerisinden bir tanesi, *j*, seçilerek *i* ile *j* bir doğru ile birleştirilir. Bu doğru üzerinde seçilecek nokta üretilen yeni sentetik örnek olur. Bu yöntemle oluşturulan sentetik örnek azınlık sınıfından seçilen rastgele örneklem olan *i* ve *j*'nin bir doğrusal kombinasyonu olarak hesaplanmış olur.

2.3.2 Uyarlanabilir sentetik örnekleme (Adaptive synthetic sampling-ADASYN)

Uyarlanabilir Sentetik Örnekleme, SMOTE yönteminin geliştirilmiş bir versiyonudur [20]. SMOTE ile benzer bir algoritma ile sentetik veri üretir. Ancak, SMOTE azınlık sınıfından seçilen her rastgele örneklem için aynı sayıda sentetik veri üretirken, ADASYN hangi sayıda sentetik veri üretileceğine bir olasılık dağılım fonksiyonu kullanarak karar verir. Bu dağılım fonksiyonu azınlık sınıfından alınan farklı örneklemlere öğrenilme zorluğuna göre bir ağırlık verir. Böylece ağırlığı fazla olan örneklemden daha fazla sayıda sentetik veri üretilir.

2.3.3 Çoğunluk ağırlıklı azınlık örnekleme (Majority weighted minority oversampling-MWMOTE)

SMOTE yönteminin bir versiyonu olan Çoğunluk Ağırlıklı Azınlık Örnekleme yöntemi Barua vd. tarafından sunulmuştur [21]. Çalışma prensibi üç adımda özetlenebilir. Önce öğrenilmesi zor olan azınlık sınıfı örneklemleri belirlenir. Bu örneklemlerin çoğunlukla karar sınırına daha yakın, yoğun kümeler yerine daha seyrek kümeler içindeki, çoğunluk sınıfının yoğun kümelerine daha yakın örneklerden oluştuğu gözlemlenmiştir. İkinci adımda, belirlenen örneklemlere en yakın çoğunluk sınıfı örneklerine olan uzaklıklarına göre ağırlıklar atanır. Son adımda ise SMOTE ile benzer bir stratejiyle sentetik veri üretilir.

2.3.4 Hızlı yakınsayan gibbs algoritması (Rapidly converging Gibbs algorithm-RACOG)

Hızlı Yakınsayan Gibbs Algoritması ayrık nümerik veriler için kullanılabilen bir yöntemdir. Gibbs Örnekleyicisi kullanılarak dağılımı azınlık sınıfının olasılık dağılımına yakınsayan sentetik örnekler üretir. Bu yöntemin detayları Das vd.'nin çalışmasında anlatılmaktadır [22].

2.3.5 Rastgele yürüyüş aşırı örnekleme (Random walk oversampling-RWO)

Rastgele Yürüyüş Aşırı Örnekleme yöntemi Zhang ve Li tarafından sunulmuştur [23]. Azınlık sınıfının ortalama ve standart sapmasını mümkün olduğunca koruyacak şekilde sentetik örnekler üretmeyi hedefleyen bir yöntemdir.

2.3.6 Rastgele Aşırı Örnekleme (Random oversampling examples-ROSE)

Rastgele Aşırı Örnekleme verinin özelliklerini değiştirmeden sentetik veri üretmek yoluyla sınıf dengesizliği problemini çözmeyi amaçlar. Seçilen örneklemde her iki sınıfın da koşullu olasılık dağılımı tahmini için düzleştirilmiş özyükleme (smoothed bootstrap) yaklaşımını kullanan bir yöntemdir. Dolayısıyla bu yöntemde alt örnekleme ve aşırı örnekleme yöntemlerinin birleştirdiği söylenebilir. Yöntemin ayrıntıları Menardi ve Torelli'nin çalışmasında anlatılmaktadır [24].

3. DENEYSEL BULGULAR (EXPERIMENTAL RESULTS)

Bu bölümde çalışmada kullanılan açık erişimli telekomünikasyon verisine uygulanan yeniden örnekleme ve makine öğrenmesi yöntemlerinin performans değerlendirmeleri sunulmuştur. Daha sonra elde edilen sonuçlar yorumlanmıştır. Sonuç bölümünde yapılan çalışma özetlenmiş, elde edilen sonuçlardan gelecek çalışmalar için araştırma fikirleri geliştirilmiştir.

3.1. Bulgular (Results)

Çalışmada kullanılan telekomünikasyon verisinde öncelikle öznitelik seçimi uygulanmıştır. Dengesizlik probleminin giderilmesi için SMOTE, ADASYN, MWMOTE, RACOG, RWO, ve ROSE yeniden örnekleme yöntemleri uygulanmıştır. *R* programlama dilinde [25] *imbalance* [26] ve *ROSE* [27] paketleri kullanılarak her bir yeniden örnekleme yöntemine göre yeni veri oluşturulmuştur. *WEKA* 'da [28] bulunan *Naive Bayes (NB), Karar Ağaçları (J48), Rastgele Orman (Random Forest-RF), Yapay Sinir Ağları (MultiLayerPerceptron-MLP), Lojistik Regresyon (Logistic-LR), Destek Vektör Makineleri (SMO) ve K-En Yakın Komşuluk (IBk)* algoritmaları 10 kat çapraz geçerleme ile uygulanarak ikili sınıflandırma problemi çözülmüştür. Makine öğrenmesi yöntemlerinin başarısı Doğruluk, Duyarlılık, Kesinlik, F-Ölçütü ve ROC Eğrisi performans değerlendirme ölçütlerine göre değerlendirilmiştir.

Çapraz geçerleme ve yeniden örnekleme yöntemlerinin birlikte kullanımı için iki farklı yaklaşım kullanılmıştır: i) çapraz geçerleme esnasında yeniden örnekleme, ii) çapraz geçerlemeden önce yeniden örnekleme. İlkinde veri 10 eşit parçaya bölünmüş, bir parça test verisi olarak ayrılmıştır. Geriye kalan 9 parça yeniden örnekleme yöntemleri ile dengeli hale getirilip bu veri üzerinde makine öğrenmesi yöntemleri eğitilmiştir. Oluşturulan modellerin performansı test verisi üzerinde ölçülmüştür. Bu işlem iteratif şekilde her parça için tekrarlanarak sonuçların ortalaması Çizelge 3'te sunulmuştur. Bu uygulamadan 1. *Yaklaşım* olarak bahsedilecektir.

İkincisinde ise veri yeniden örnekleme yöntemleri ile dengeli hale getirildikten sonra makine öğrenmesi yöntemleri 10 kat çapraz geçerleme ile uygulanmıştır. Bu yaklaşımın ilkinden en önemli farkı test verisinin de yeniden örnekleme yöntemleri ile dengeli hale getirilmiş olmasıdır. Bu yaklaşımla elde edilen sonuçlar Çizelge 4'te sunulmuştur. Bu uygulamadan *2. Yaklaşım* olarak bahsedilecektir.

Çizelge 3. Test Sonuçları - Çapraz Geçerleme Esnasında Yeniden Örnekleme (Experimental Results-Resampling During Cross Validation)

Metod	*Algoritma*	*Doğruluk*	*Duyarlılık*	*Kesinlik*	*F-Ölçütü*	*ROC Eğrisi*
SMOTE	*NB*	%73.72	0.737	**0.793**	0.750	0.815
	J48	%77.63	0.776	0.780	0.778	0.805
	RF	%74.00	0.740	0.739	0.740	0.771
	MLP	%76.78	0.768	0.768	0.768	0.810
	LR	%77.97	0.780	0.785	0.782	**0.822**
	SMO	**%78.32**	**0.783**	0.792	**0.787**	0.748
	IBk	*%71.73*	*0.717*	*0.721*	*0.719*	*0.676*
ADASYN	*NB*	%72.41	0.724	**0.794**	0.738	0.814
	J48	%75.48	0.755	0.774	0.762	0.790
	RF	%73.66	0.737	0.740	0.738	0.769
	MLP	**%76.90**	**0.769**	0.787	**0.775**	0.810
	LR	%76.84	0.768	0.786	**0.775**	**0.817**
	SMO	%76.16	0.762	0.788	0.770	0.747
	IBk	*%71.22*	*0.712*	*0.723*	*0.717*	*0.678*
MWMOTE	*NB*	%74.97	0.750	**0.790**	0.760	0.811
	J48	%77.46	0.775	0.765	0.768	0.801
	RF	%74.91	0.749	0.743	0.746	0.775
	MLP	%76.50	0.765	0.760	0.762	0.803
	LR	**%78.09**	0.781	0.778	0.779	**0.821**
	SMO	%78.03	**0.780**	0.775	**0.777**	0.712
	IBk	*%71.33*	*0.713*	*0.710*	*0.712*	*0.679*
RACOG	*NB*	%72.53	0.725	0.794	0.739	0.817
	J48	%75.36	0.754	0.775	0.761	0.790
	RF	%73.43	0.734	0.755	0.742	0.784
	MLP	**%76.04**	**0.760**	0.772	**0.765**	0.804
	LR	%74.57	0.746	**0.795**	0.758	**0.825**
	SMO	%70.77	*0.708*	0.788	*0.723*	0.742
	IBk	*%71.67*	0.717	*0.738*	0.725	*0.709*
RWO	*NB*	%74.40	0.744	**0.790**	0.756	0.806
	J48	%77.46	0.775	0.773	0.774	0.810
	RF	%73.89	0.739	0.732	0.735	0.760
	MLP	%77.12	0.771	0.760	0.763	0.804
	LR	**%78.43**	**0.784**	0.778	**0.780**	**0.819**
	SMO	%78.14	0.781	0.774	0.777	0.708
	IBk	*%70.09*	*0.701*	*0.698*	*0.699*	*0.663*
ROSE	*NB*	%70.26	0.703	0.792	0.718	0.815
	J48	%52.89	0.529	**0.801**	0.528	0.756
	RF	%56.12	0.561	**0.801**	0.567	0.813
	MLP	%61.52	0.615	0.786	0.630	0.806
	LR	**%73.49**	**0.735**	0.799	**0.748**	**0.826**
	SMO	%70.60	0.706	0.796	0.722	0.748
	IBk	*%62.20*	*0.622*	*0.758*	*0.640*	*0.686*

Yeniden örneklemenin makine öğrenme yöntemlerinin başarısına etkisinin daha iyi gözlemlenebilmesi için sınıf dengesizliği bulunan orijinal veriye de aynı makine öğrenmesi yöntemleri aynı şekilde uygulanmıştır. Elde edilen araştırma sonuçları Çizelge 5'te sunulmuştur.

Çizelge 4. Test Sonuçları-Yeniden Örnekleme Ardından Çapraz Geçerleme (Experimental Results-First Resampling, Then Cross Validation)

Metod	*Algoritma*	*Doğruluk*	*Duyarlılık*	*Kesinlik*	*F-Ölçütü*	*ROC Eğrisi*
SMOTE	*NB*	*%73.41*	*0.734*	0.772	0.741	0.826
	J48	%78.42	0.784	0.782	0.783	0.829
	RF	%76.94	0.769	0.768	0.769	0.826
	MLP	%78.00	0.780	0.782	0.781	0.839
	LR	**%79.17**	**0.792**	**0.790**	**0.791**	**0.860**
	SMO	%78.95	0.79	0.788	0.789	0.759
	IBk	%73.93	0.739	*0.738*	*0.739*	*0.739*
ADASYN	*NB*	*%73.78*	*0.738*	*0.743*	*0.737*	0.816
	J48	%81.52	0.815	0.817	0.815	0.869
	RF	**%82.43**	**0.824**	**0.826**	**0.824**	**0.89**
	MLP	%80.80	0.808	0.809	0.808	0.89
	LR	%81.08	0.811	0.812	0.811	0.89
	SMO	%80.53	0.805	0.805	0.805	*0.805*
	IBk	%80.05	0.801	0.802	0.800	0.821
MWMOTE	*NB*	%74.32	0.743	0.780	0.750	0.833
	J48	%79.04	0.790	0.787	0.788	0.838
	RF	%76.00	0.760	0.756	0.758	0.826
	MLP	%78.23	0.782	0.780	0.781	0.846
	LR	**%80.30**	**0.803**	**0.800**	**0.800**	**0.871**
	SMO	%79.88	0.799	0.795	0.796	0.762
	IBk	*%73.89*	*0.739*	*0.735*	*0.737*	*0.737*
RACOG	*NB*	%72.64	0.726	0.766	0.734	0.815
	J48	%75.96	0.760	0.756	0.758	0.795
	RF	%73.60	0.736	0.733	0.734	0.798
	MLP	%75.72	0.757	0.758	0.757	0.815
	LR	%76.10	0.761	0.759	0.760	**0.823**
	SMO	**%76.15**	**0.762**	**0.760**	**0.761**	0.728
	IBk	*%71.00*	*0.710*	*0.705*	*0.707*	*0.702*
RWO	*NB*	%73.81	0.738	0.771	0.745	0.821
	J48	%77.89	0.777	0.772	0.774	0.830
	RF	%75.40	0.754	0.750	0.751	0.819
	MLP	%78.13	0.781	0.779	0.780	0.847
	LR	**%80.27**	**0.803**	**0.798**	**0.799**	**0.872**
	SMO	%79.87	0.799	0.795	0.795	0.759
	IBk	*%73.11*	*0.731*	*0.726*	*0.728*	*0.725*
ROSE	*NB*	%76.33	0.763	0.768	0.762	0.856
	J48	%82.20	0.822	0.823	0.822	0.860
	RF	**%84.93**	**0.849**	**0.850**	**0.849**	**0.920**
	MLP	%82.50	0.825	0.826	0.825	0.899
	LR	%74.81	0.748	0.752	0.747	0.811
	SMO	%74.71	0.747	0.756	0.745	0.747
	IBk	*%70.25*	*0.703*	*0.715*	*0.698*	*0.653*

Çizelge 5. Test Sonuçları-Sınıf Dengesizliği Bulunan Veri (Experimental Results-Imbalanced Data)

	Algoritma	***Doğruluk***	***Duyarlılık***	***Kesinlik***	***F-Ölçütü***	***ROC Eğrisi***
Orijinal Veri	*NB*	%76.06	0.785	0.761	0.769	0.817
	J48	%77.66	0.777	0.765	0.769	0.798
	RF	%74.91	0.749	0.742	0.745	0.769
	MLP	%76.64	0.766	0.768	0.767	0.804
	LR	**%78.67**	**0.787**	**0.776**	**0.779**	**0.827**
	SMO	%78.33	0.783	0.775	0.778	0.701
	IBk	*%72.66*	*0.727*	*0.720*	*0.723*	*0.681*

Bu bölümde sınıf dengesizliği bulunan orijinal veri ile çapraz geçerleme esnasında ve çapraz geçerleme öncesinde SMOTE, ADASYN, MWMOTE, RACOG,

RWO ve ROSE yeniden örnekleme tekniklerinin kullanıldığı veri üzerinde NB, J48, RF, YSA, LR, DVM ve IBk yöntemleriyle yapılan sınıf tahmininden elde

edilen sonuçlar yorumlanmıştır. Kullanılan yöntemlerin performans ölçümleri için Doğruluk, Duyarlılık, Kesinlik, F- Ölçütü ve ROC Eğrisi metrikleri kullanılmıştır. Sonuçlar Çizelge 3, Çizelge 4 ve Çizelge 5'te sunulmuş, elde edilen en iyi sonuçlar koyu renk ile en kötü sonuçlar ise italik ile gösterilmiştir. Tüm yöntemler göz önüne alındığında elde edilen en iyi ve en kötü sonuçlar ise bu işaretlemelere ek olarak alt çizgi ile belirtilmiştir.

Elde edilen sonuçların yorumlanmasında, dengesiz verilerle yapılan çalışmalarda daha sık tercih edilen bir ölçüt olduğundan [24], ROC Eğrisinden elde edilen sonuçlar esas alınmıştır. Sonuçlar bir yeniden örnekleme ve bir makine öğrenmesi yöntemi uygulanarak elde edildiğinden yorumlanırken *Yeniden Örnekleme Yöntemi- Makine Öğrenmesi Yöntemi* kombinasyonu (SMOTE-NB gibi) kullanılmıştır.

Çizelge 4'te sunulan 2. Yaklaşımla elde edilen sonuçlara göre en iyi tahmin başarısının 0.920 ROC Eğrisi değeri ile ROSE-RF ile elde edildiği görülmektedir. Bu değer Orijinal Veri-RF değeri olan 0.769 ile kıyaslandığında verinin daha dengeli hale getirilmesiyle tahmin başarısında %19.6 oranında iyileşme sağlandığı görülmektedir. Ancak bu değer Çizelge 3'te sunulan 1. Yaklaşımla elde edilen ROSE-RF değeri olan 0,813 ile kıyaslandığında iyileşmenin sadece %5.7 olduğu görülmektedir. Sonuçlar arasındaki farkın sebebi 2. Yaklaşımda kullanılan test verisinin de dengeli hale getirilmiş olmasıdır. Bu durumda makine öğrenmesi yöntemi eğitildiği veriye benzer veriyle test edilmiş olduğundan tahmin başarısı gerçek değerinden daha iyi görünerek yanıltıcı bir sonuç doğurmaktadır. Benzer farkın diğer yöntem ikilileri için de görülmesi beklenen bir durumdur. Doğru değerlendirme için yeniden örneklenmemiş test verisi kullanılan 1. Yaklaşım doğru yaklaşım olarak kabul edilmelidir.

1. Yaklaşımla elde edilen sonuçlar Çizelge 5'te sunulan yeniden örnekleme kullanılmayan durumla kıyaslandığında en fazla iyileşme SMOTE-SMO ve ROSE-SMO ile elde edilmiştir. Buna göre ROC Eğrisi değeri 0.701'ten 0.748'e çıkarak tahmin başarısında %6.7 iyileşme kaydedilmiştir. ADASYN-SMO da %6.5 oranında iyileşmeyle bunlara yakın bir sonuç sergilemiştir. En az iyileşme %1.5 oranıyla MWMOTE-SMO'dan elde edilmiştir. Makine öğrenmesi yöntemlerinden SMO hemen her yeniden örnekleme yöntemi ile en fazla iyileşmenin sağlandığı yöntem olmuştur.

Makine öğrenmesi yöntemlerinin tahmin başarısı kıyaslandığında her iki yaklaşım için de en küçük ROC Eğrisi değerinin IBk yönteminde gözlenmektedir. Bunun sebebi olarak bu yöntemin çalışma prensibi gösterilebilir. Bu yöntemde sınıfı bilinmeyen veri noktasının sınıf tahmini ona en yakın *k* komşusunun sınıfına bakılarak yapılmaktadır. Çalışılan verideki sınıf dengesizliği sebebiyle seçilen *k* en yakın komşuda çoğunluk sınıfına ait daha çok veri olacağından, tahminin çoğunluk sınıfı lehine olması daha muhtemeldir. Yeniden örnekleme yöntemleri ile verinin daha dengeli hale getirilmesiyle bile IBk yönteminin en düşük tahmin başarısına sahip olduğu görülmektedir. Bunun sebebi yeniden örnekleme yöntemlerinin yeni veri üretirken çoğunluk ve azınlık sınıflarının dağılımlarını gözetmesi olarak düşünülebilir.

4. SONUÇ (CONCLUSION)

Müşteri kaybı tahmini müşteri davranışlarından elde edilen verinin incelenerek ayrılması muhtemel müşterinin önceden tespit edilmesi olarak tanımlanabilir. Ayrılacak müşteri doğru tahmin edilebilirse ayrılmaması için çeşitli promosyon ve kampanyalar düzenlenerek müşteri kaybıyla oluşacak zarar en aza indirilebilir.

Müşteri kaybı tahmini için kullanılan çözüm yaklaşımları arasında makine öğrenmesi yöntemlerinin de sıklıkla yer aldığı görülmektedir. Bu yöntemlerin performansları çalışılan verideki sınıf dengesizliğinden olumsuz etkilenmektedir. Verideki sınıf dağılımını daha dengeli hale getirebilmek için yeniden örnekleme yöntemlerinden faydalanılabilir.

Bu çalışmada rekabetin yoğun olduğu telekomünikasyon sektöründeki açık erişimli veri üzerinde müşteri kaybı tahmini problemi ele alınmıştır. 7043 müşteriden toplanan veride öncelikle Minimum Fazlalık Maksimum Bağımlılık öznitelik seçme yöntemi ile 21 olan öznitelik sayısı 10'a düşürülmüştür. SMOTE, ADASYN, MWMOTE, RACOG, RWO ve ROSE yeniden

örnekleme yöntemleri ile Naive Bayes, Karar Ağaçları, Rastgele Orman, Yapay Sinir Ağları, Lojistik Regresyon, Destek Vektör Makineleri ve K-En Yakın Komşuluk makine öğrenmesi yöntemleri 10 kat çapraz geçerlemeyle uygulanmıştır. Bu uygulamada iki farklı yaklaşım denenmiştir. Yeniden örnekleme 1.yaklaşımda çapraz geçerleme esnasında, 2. yaklaşımda ise çapraz geçerlemeden önce uygulanmıştır. Makine öğrenmesi yöntemlerinin başarısı Doğruluk, Duyarlılık, Kesinlik, F-Ölçütü ve ROC Eğrisi performans değerlendirme ölçütleriyle kıyaslanmıştır.

Elde edilen sonuçlara göre, benzer özelliklere sahip eğitim ve test verisi kullanılan 2. yaklaşımdan elde edilen makine öğrenmesi performans değerlerinin, orijinal verinin bir parçası üzerinde test edildiği 1. yaklaşımdan elde edilen performans değerlerinden daha iyi olabildiği gözlenmiştir. 2. yaklaşımdaki eğitim ve test verisi benzerliğinin performans değerlerindeki yapay artışa neden olduğundan doğru yaklaşımın yeniden örnekleme ile dengeli hale getirilmiş eğitim verisi ile oluşturulan modellerin orijinal veriden ayrılan test verisi üzerinde sınandığı 1. yaklaşım olduğu sonucuna varılmıştır. Buna göre, ROSE ve Destek Vektör Makineleri doğru yaklaşımla uygulandığında elde edilen ROC Eğrisi performansının aynı yöntemin orijinal verideki performansından %5.7 daha iyi olduğu gözlenmiştir. Buradan sınıf dengesizliği bulunan verinin yeniden örnekleme yöntemleri ile dengeli hale getirilmesinin makine öğrenmesi yöntemlerinin performansına olumlu etki ettiği sonucu çıkarılabilir. Bu etkinin doğru ölçülebilmesi için yeniden örnekleme ve çapraz geçerleme yöntemlerinin doğru uygulanması gerektiği anlaşılmıştır. Uygulanan yöntemlerin performansının farklı özelliklere sahip veriler üzerindeki etkisi ayrı bir çalışma konusu olabilir.

ETİK STANDARTLARIN BEYANI
(DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

M. Aslı AYDIN: Deneyleri yapmış ve sonuçlarını analiz etmiştir. Makalenin yazım işlemini gerçekleştirmiştir.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur.

KAYNAKLAR (REFERENCES)

[1] Cao, J., Yu, X. & Zhang, Z., "Integrating OWA and data mining for analyzing customers churn in E-commerce.", ***J Syst Sci Complex,*** 28**:** 381–392, (2015).

[2] Koçoğlu, F.Ö., Özcan, T., Baray, Ş.A., "Veri madenciliğinde ayrılan müşteri analizi problemi üzerine bir literatür araştırması", ***Uluslararası katılımlı 16. Üretim Araştırmaları Sempozyumu***, 868-874, (2016).

[3] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., Abbasi, U., "Improved churn prediction in telecommunication industry using data mining techniques", ***Applied Soft Computing***, 24: 994-1012, (2014).

[4] Kaynar, O. , Tuna, M. , Görmez, Y. , Deveci, M., "Makine öğrenmesi yöntemleriyle müşteri kaybı analizi", ***Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi***, 18:1 , 1-14, (2017).

[5] Günay, M. and Ensarı,T., "Predictive churn analysis with machine learning methods.**" *26th Signal Processing and Communications Applications Conference (SIU)***,** Izmir, 1-4, (2018).

[6] Yıldız, M. and Albayrak, S., "Customer churn prediction in telecommunication", **23nd Signal Processing and Communications Applications Conference (SIU)**, Malatya, 256-259, (2015).

[7] Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.Ch., "A comparison of machine learning techniques for customer churn prediction", ***Simulation Modelling Practice and Theory***, 55: 1-9, (2015).

[8] Ullah,I., Raza,B., Malik, A. K. , Imran, M., Islam, S. U. and Kim, S. W., "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", **IEEE Access**, 7: 60134-60149, (2019).

[9] Amin A., Shah B., Abbas A., Anwar S., Alfandi O., Moreira F., "Features Weight Estimation Using a Genetic Algorithm for Customer Churn Prediction in the Telecom Sector", ***In: Rocha Á., Adeli H., Reis L., Costanzo S. (eds) New Knowledge in Information Systems and Technologies. WorldCIST'19 2019. Advances in Intelligent Systems and Computing***, 931: 483-491, (2019).

[10] Kartal, E., Özen, Z., "Dengesiz Veri Setlerinde Sınıflandırma", ***Mühendislikte Yapay Zekâ Uygulamaları***, Sakarya, 109-131, (2017).

[11] Gui, C., "Analysis of imbalanced data set problem: The case of churn prediction for telecommunication", ***Artif. Intell. Research,*** 6:2, 93, (2017).

[12] Durahim, A., "Comparison Of Sampling Techniques For Imbalanced Learning". ***Yönetim Bilişim Sistemleri Dergisi*** , 2:2, 181-191, (2016).

[13] Effendy, V., Adiwijaya and Baizal, Z. K. A., "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest", ***2nd International Conference on Information and Communication Technology (ICoICT)***, Bandung, 2014, 325-330, (2014).

[14] Amin, A. et al., "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study**,** **IEEE Access**, 4: 7940-7957, (2016).

[15] Aditsania, A., Adiwijaya and Saonard, A. L., "Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm", ***3rd International Conference on Science in Information Technology (ICSITech)***, Bandung, 2017, 533-536, (2017).

[16] Koçoğlu, F. ve Ozcan, T., "Dengeli-Dengesiz Veri Seti Dağılımının Aşırı Öğrenme Makinesi Yöntemi Performansına Etkisi", ***Mühendislik ve Teknoloji Yönetimi Zirvesi-ETMS2018***, İstanbul, 201-209, (2018).

[17] Blagus, R. and Lusa,L. "Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models," ***BMC bioinformatics***, 16:1, 1–10, (2015).

[18] https://www.kaggle.com/blastchar/telco-customer-churn/version/1 (Son erişim tarihi: 05/06/2019)

[19] Chawla, N. V. et.al., "SMOTE: Synthetic Minority Over-Sampling Technique", ***Journal of Artificial Intelligence Research,*** 16, 321–357 (2002).

[20] He, H., Bai, Y., Garcia, E. A. and Li, S., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning**",** ***IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)***, Hong Kong, 1322-1328, (2008).

[21] Barua, S., Islam, M.M., Yao, X., Murase, K., "MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning", ***IEEE Trans. Knowl. Data Eng.*** 26:2, 405–425, (2014).

[22] Das, B., Krishnan, N.C., Cook, D.J., "RACOG and wRACOG: two probabilistic over- sampling techniques", ***IEEE Trans. Knowl. Data Eng.*** 27:1, 222–234, (2015).

[23] Zhang, H., Li, M., "RWO-sampling: a random walk oversampling approach to imbalanced data classification", ***Inf. Fusion*** 20: 99–116, (2014).

[24] Menardi, G. and Torelli, N., "Training and assessing classification rules with imbalanced data", ***Data Mining and Knowledge Discovery***, 28: 92–122, (2014).

[25] R Development Core Team, "R: A language and environment for statistical computing", ***R Foundation for Statistical Computing***, Vienna, Austria. ISBN 3-900051-07-0, (2008). URL http://www.R-project.org. (Son erişim tarihi: 07/01/2020)

[26] S, Fernández A, Herrera F., "Imbalance: Oversampling algorithms for imbalanced classification in R", ***Knowledge-Based Systems***, 161: 329-341, (2018).

[27] Lunardon, N., Menardi, G., and Torelli, N., "ROSE: a Package for Binary Imbalanced Learning", ***R Journal***, 6:1, 82-92, (2014).

[28] Weka. https://www.cs.waikato.ac.nz/ml/weka/index.html. (Son erişim tarihi: 07/01/2020