Aydın, M. A. (2025). Telekomünikasyon Sektöründe Müşteri Kaybı Tahmini İçin Açıklanabilir Makine Öğrenmesi. EMO Bilimsel Dergi, 15(2), 71-79.

Makale Türü: Araştırma Makalesi
Makale Gönderilme Tarihi: 08.03.2025
Kabul Tarihi: 27.05.2025

Telekomünikasyon Sektöründe Müşteri Kaybı Tahmini İçin Açıklanabilir Makine Öğrenmesi

Explainable Machine Learning for Customer Churn Prediction in Telecommunication Industry

M. Aslı Aydın[1]
0000-0002-8905-7518
[1]İşletme Mühendisliği Bölümü
Bahçeşehir Üniversitesi
asli.aydin@bau.edu.tr

Özet- ***Telekomünikasyon sektöründe ayrılacak müşterilerin önceden tahmin edilmesi, firmaların bu müşterilere özel elde tutma stratejileri geliştirerek müşteri kaybını önlemesinde oldukça önemlidir. Bu çalışmada müşteri kaybı tahmini için geleneksel makine öğrenmesi yöntemlerinden Lojistik Regresyon, Destek Vektör Makineleri ve Yapay Sinir Ağlarının performansı, topluluk öğrenme yöntemleri olan Rastgele Orman, Gradyan Güçlendirme Makineleri ve Aşırı Gradyan Güçlendirme yöntemlerinin performansıyla kıyaslanmıştır. Açık erişimli telekomünikasyon verisinde bulunan sınıf dengesizliği problemi, modellerin performansına olan olumsuz etkisinin önlenmesi amacıyla SMOTE yeniden örnekleme yöntemiyle giderilmiştir. Deney sonuçlarına göre Rastgele Orman modeli tahminlerin %85,9 doğruluk oranı ile diğer modellere üstünlük sağlamıştır. Model sonuçlarına en çok etki eden faktörlerin tespit edilip yorumlanabilmesi için Shapley Katkı Açıklamaları yöntemi kullanılmıştır. Bu yöntemle müşterilerin ayrılma kararına etki eden temel faktörler değerlendirilip ayrılacağı tahmin edilen müşterilere yönelik caydırıcı teşvik ve promosyonlar önerilmiştir.***

Anahtar Kelimeler- **Müşteri Kaybı Tahmini, Telekomünikasyon, Makine Öğrenmesi, Açıklanabilir Yapay Zeka, Shapley Değeri**

Abstract— In telecommunication industry, prediction of the customers who are the most likely to leave is very crucial for companies to prevent customer churn by developing special retention strategies for those. This study presents a comparison of the performances of traditional machine learning methods Logistic Regression, Support Vector Machines and Artificial Neural Networks with the performance of ensemble learning methods Random Forest, Gradient Boosting Machines and Extreme Gradient Boosting for customer churn prediction. We address class imbalance problem with SMOTE resampling method to reduce the adverse impact of imbalanced data on model performance. The results show Random Forest model outperforms the other models with an accuracy of 85,9%. We utilized Shapley Additive Explanation method to identify and interpret the features affecting the customer's decisions to leave most. This approach enables the subsequent recommendation of tailored incentives and promotions for those customers who are predicted to be at risk of leaving.

Keywords—Churn Prediction, Telecommunications, Machine Learning, Explainable AI, Shapley Value

I. GİRİŞ

Telekomünikasyon (telekom) sektörü, hızlı teknolojik gelişmeler ve değişen müşteri talepleri ile karakterize edilen son derece rekabetçi bir ortamda faaliyet göstermektedir. Mevcut müşterilerin rakip firmalara geçmesi olarak tanımlanabilen müşteri kaybı, telekom operatörleri için gelir, pazar payı ve marka itibarını etkileyen önemli bir tehdit oluşturmaktadır. Ayrıca sürdürülebilir büyüme ve kârlılık için müşteriyi elde tutmak çok önemlidir. Çünkü yeni müşteri edinmenin maliyeti, mevcut müşterileri elde tutmanın maliyetinden çok daha fazladır [1]. Burez ve Van den Poel [2] müşteri kaybını yönetmek için belirlenen yaklaşımların reaktif ve proaktif olmak üzere kategorize edilebileceğini söyler. Reaktif yaklaşımı benimseyen firmalar müşteriden ayrılma talebi gelene kadar bekler. Böyle bir durumla karşılaştıklarında aksiyon alıp ayrılacak müşterilere teşvik sunarlar. Ancak bu tedbirler sıklıkla çok geç uygulanmaktadır. Öte yandan, proaktif yaklaşımı benimseyen firmalar hangi müşterilerin ayrılma olasılığının yüksek olduğunu önceden tahmin etmeyi hedefler. Böylece bu müşteriler henüz ayrılmayı talep etmeden onlara özel ihtiyaçları karşılamak için hedefe yönelik kampanya ve promosyon gibi teşviklerle müşteri kaybını azaltmayı amaç edinir. Proaktif yaklaşımın reaktif yaklaşıma göre daha düşük teşvik maliyetlerine sahip olma gibi avantajları vardır. Ancak, müşteri kaybı tahminleri yanlışsa bu sistemler çok masraflı olabilir. Örneğin, firma ayrılmayacak müşteriyi ayrılacak diye tahmin ederse kaybetmeyeceği müşteri için gereksiz teşvik vermiş olur. Ya da ayrılacak müşteriyi doğru tahmin edemezse müşteriyi kaybedebilir. Bu nedenle, tahminlerin doğruluğu, sağlanacak kampanya veya promosyonların maliyeti ve ayrılacak müşterinin doğru tespit edilememesi durumunda yaşanacak müşteri kaybı düşünüldüğünde önem kazanmaktadır.

Müşteri kaybı tahmini probleminin temelde mevcut müşterilerin *Ayrılanlar* ve *Ayrılmayanlar* şeklinde iki sınıfa ayrılması (İkili Sınıflandırma) problemi olarak da tanımlanabiliyor olması, son yıllarda ön plana çıkan makine öğrenmesi yöntemlerinin de çözüm yaklaşımları arasında olmasını gündeme getirmiştir. Makine öğrenmesi yöntemleri, büyük boyutlardaki müşteri verisini analiz etme, müşteri kaybı eğilimini gösteren karmaşık kalıpları ve ilişkileri belirleme ve mevcut müşterilerin ayrılma davranışını yüksek doğrulukla tahmin etme potansiyeli sunmaktadır. Ancak üzerinde çalışılan verinin özellikleri uygulanan makine öğrenimi yönteminin tahmin başarısını etkileyebilmektedir. Örneğin denetimli öğrenme yöntemleriyle çözülecek ikili sınıflandırma problemlerinde bir sınıfa ait örneklerin diğerine göre önemli ölçüde fazla olması Veride Sınıf Dengesizliği olarak tanımlanır [3]. Telekom sektöründen elde edilen çok sayıda veri setinde *Ayrılan* sınıfındaki müşterilerin oranı *Ayrılmayan* sınıfındaki müşterilerin

oranına göre oldukça düşüktür. Bu da telekom verilerinde sınıf dengesizliği probleminin varlığına işaret eder [4]. Birçok makine öğrenmesi yöntemi verideki dengesiz yapıyı ayırt edebilecek şekilde tasarlanmamıştır. Bu sebeple, bu tip veriye uygulanan makine öğrenmesi yöntemlerinin performansı olumsuz etkilenebilmektedir [5]. Bu sebeple öncelikle verideki sınıf dengesizliğinin giderilmesi, sonra dengelenmiş veriye makine öğrenmesi yöntemlerinin uygulanması tahminlerin doğruluğunu artırmaktadır.

Telekom sektöründe müşteri kaybını önlemede belirli bir müşterinin ayrılıp ayrılamayacağının başarılı bir şekilde tahmin edilmesi kadar ayrılmaların ardında yatan nedenlerin tespit edilip açıklanabilmesi de önem arz etmektedir. Hızla gelişen makine öğrenmesi yöntemleri verideki dengesizliğin giderilmesi gibi desteklerle oldukça başarılı tahminlerde bulunsa da ayrılmaların arkasındaki sebepleri açıklamak gibi bir amaçları yoktur. Yapılan tahminlerin arkasındaki mantık insanlar tarafından bir bakışla anlaşılamadığı için bu yöntemler bir kara kutu olarak kabul edilir [6].

Açıklanabilir Yapay Zekanın (XAI) telekom sektöründeki müşteri kaybı tahminine entegrasyonu, müşteri memnuniyetsizliğini anlama ve ele alma konusunda önemli bir ilerlemeyi temsil etmektedir. XAI teknikleri, tahminler için şeffaf açıklamalar ve yorumlanabilirlik sağlayarak makine öğrenmesi yöntemlerini geliştirir ve firmaların müşteri kaybına neden olan temel faktörleri anlamalarına olanak tanır. Yönetimsel bakış açısıyla değerlendirildiğinde, XAI tüm paydaşların geliştirilmekte olan tahmin modelini daha iyi anlaması ve güven duymasını sağlayarak otomatik algoritmaları insani içgörüyle bütünleştirir [7]. Böylece firmaların hedeflenen müşteriler için daha etkin elde tutma stratejileri geliştirebilmeleri kolaylaşır.

Bu çalışma telekom sektöründe müşteri kaybını etkin bir biçimde tahmin edilirken aynı zamanda müşteri kaybının ardındaki nedenlerin de anlaşılmasını sağlamayı hedeflemektedir. Çalışmanın katkıları aşağıda özetlenmiştir:

- Sınıf dengesizliği bulunan telekom verisine öne çıkan makine öğrenmesi yöntemleri uygulanarak müşteri kaybı tahminindeki performansları analiz edilmiştir.
- Makine öğrenmesi yöntemlerinden elde edilen sonuçların nedenlerini yorumlayabilmek için açıklanabilir yapay zeka yöntemi kullanılmıştır.
- Özniteliklerin müşteri kaybı tahminine olumlu ya da olumsuz etkileri yönetim perspektifinden değerlendirilmiştir.

Bu çalışma aşağıdaki şekilde düzenlenmiştir. 2. bölümde ilgili çalışmalar özetlenerek bu çalışmayla olan ilişkileri incelenmiştir. 3. bölümde üzerinde çalışılan veri tanıtılıp veri ön işleme işlemlerinden bahsedilmiştir. Ayrıca sınıf dengesizliği problem için uygulanan yeniden örnekleme yöntemi ve makine öğrenmesi yöntemlerinden bahsedilmiştir. 4. bölümde test sonuçları verilmiş ve bunların XAI ile değerlendirilmesi yapılmış, elde edilen bulgular tartışılmıştır. 5. bölümde ise çalışmada elde edilen sonuçlar özetlenmiştir.

II. LİTERATÜR TARAMASI

Makine öğrenmesi yöntemleri müşteri kaybı tahminindeki başarıları dolayısıyla kısa bir süre içinde yaygın bir kullanım ve popülerlik kazanmıştır. Telekom sektöründeki müşteri kaybı tahmininde erken dönem çalışmalar arasında sıklıkla Naïve Bayes (NB) [8, 9], Karar Ağaçları (KA) [10, 11], Lojistik Regresyon (LR) [12, 13], Destek Vektör Makineleri (DVM) [14, 15], K-En Yakın Komşuluk (KNN) [16] ve Yapay Sinir Ağları (YSA) [17-19] yöntemlerinin kullanıldığı görülmektedir. Bu yöntemler birçok araştırmacı tarafından farklı Telekom veri setlerine uygulanmış ve performansları kıyaslanmıştır. Bunlar arasında Keramati vd. [20] KA, DVM, KNN ve YSA, Kaynar vd. [21] ise NB, DVM ve YSA yöntemlerinin performansını kıyaslamış, en başarılı tahminlere YSA ile ulaştıklarını açıklamışlardır. Aydın [22] çalışmasında telekom verisindeki sınıf dengesizliği problemine de değinmiş, uygulanan yöntemlerden DVM'nin diğerlerine göre daha iyi performans gösterdiği açıklamıştır.

Güncel çalışmalarda araştırmacıların temel makine öğrenme yöntemlerine ek olarak Rastgele Orman (RO), Gradyan Güçlendirme Makineleri (Gradient Boosting Machines-GBM), Aşırı Gradyan Güçlendirme (Extreme Gradient Boosting-XGBoost) gibi daha gelişmiş topluluk yöntemleri de kullanıldığı görülmektedir. Sikri vd. [23] çalışmalarında farklı yöntemlerle dengeledikleri telekom verisine NB, KA, LR, YSA, KNN, XGBoost yöntemlerini uygulamış, XGBoost yöntemiyle daha başarılı tahmin sonuçları elde etmişlerdir. Afzal vd. [24] ise KA, KNN, RO ve XGBoost yöntemlerini kıyasladıkları çalışmalarında RO yönteminin %98,25 doğruluk oranı ile diğer yöntemlerden daha başarılı sonuç verdiğini açıklamışlardır. Kök [25] çalışmasında öncelikle verideki sınıf dengesizliğine değinip sonra NB, KA, DVM, KNN ve RO yöntemlerini uygulamışlardır. RO yöntemi ile tahminlerde %81 doğruluk oranı elde etmiştir. Daha sonra bu modele ait tahminleri açıklanabilir yapay zeka yöntemleriyle yorumlamıştır. Boozary vd. [26] çalışmalarında sınıf dengesizliğini giderdikleri veriye LR, DVM, KNN, RO ve XGBoost yöntemlerini uygulamışlar, topluluk yöntemlerinin, spesifik olarak XGBoost yönteminin diğerlerinden daha yüksek performans gösterdiğini açıklamışlardır. Ayrıca tahminlere öznitelik katkılarını ölçmek, model yorumlanabilirliğini artırmak için açıklanabilir yapay zeka yöntemleri kullanmışlardır. Özkurt [27] çalışmasında geleneksel makine öğrenmesi yöntemlerine ek olarak RO, XGBoost, Gradyan Güçlendirme, LightGBM, CatBoost, AdaBoost gibi birçok güncel topluluk yöntemlerine yer vermiştir. Elde edilen sonuçlara göre LightGBM yöntemi %73,085 doğruluk oranı ile diğer yöntemlerden daha iyi performans sergilemiştir. Elde edilen makine öğrenmesi sonuçları iki farklı XAI yöntemi ile yorumlanmıştır. Asif vd. [28] çalışmalarında geleneksel makine öğrenmesi yöntemlerine yer vermezken, sınıf dengesizliği giderilmiş veriseti üzerinde XGBoost, CatBoost ve LightGBM yöntemlerinin bir kombinasyonu olan TriBoost yöntemini kullanmışlardır. Bu yöntemi bireysel temel yöntemlerle karşılaştırmış ve performans ölçütlerinde iyileşme gözlemlemişlerdir. Daha sonra elde edilen sonuçlar XAI yöntemleriyle ile açıklanmıştır.

Bu çalışmanın amacı telekom sektöründe müşteri kaybını etkin bir biçimde tahmin edebilmek ve sonrasında müşteri kaybına yol açan esas nedenlerin anlaşılmasını sağlamaktır. Böylece firmaların müşterileri elde tutma stratejileri geliştirmesi kolaylaşacaktır.

Çalışmada uygulanan yöntemlerle mevcut çalışmalarda kullanılan yöntemlerin kıyaslaması Tablo 1'de görülmektedir. Buna göre [20] ve [21]'de daha yüksek performans gösterme potansiyeli olan topluluk yöntemleri kullanılmamışken bu çalışmada bu yöntemler de yer verilmiştir. Ayrıca bu çalışmalarda uygulanan makine öğrenmesi yöntemlerinin performansını olumsuz etkileyecek sınıf dengesizliği problemine değinilmemişken bu çalışmada öncelikle veri dengeli hale getirilmektedir. [22], [23] ve [25]'de sadece bir topluluk yöntemi kullanılmıştır. Bu çalışmada geleneksel makine öğrenmesi yöntemlerine farklı topluluk yöntemleri de eklenerek daha kapsamlı bir değerlendirme hedeflenmiştir. [24] ve [27] 'de çeşitli yöntemler denenmiş ancak verideki sınıf dengesizliği problemi giderilmemiştir. Bu çalışmada [26]'te kullanılmayan ama başka çalışmalarda yüksek performans gösterdiği açıklanan YSA ve GBM yöntemleri de araştırmaya dahil edilmiştir. [28] sadece

topluluk yöntemleri arasında bir kıyaslama yaparken, bu çalışmada hem geleneksel makine öğrenme yöntemlerine hem de topluluk yöntemlerine yer verilmiştir. Ayrıca [20-24]'te açıklanabilir yapay zeka teknikleri kullanılmamışken bu çalışma tahminlerin yorumlanabilirliğine de değinmiştir.

TABLO 1. ÇALIŞMALARDA KULLANILAN TEKNİKLERİN ÖZETİ

	Geleneksel Makine Öğrenmesi Yöntemleri	***Topluluk Yöntemleri***	***Yeniden Örnekleme Yöntemleri***	***XAI Yöntemleri***
[20]	KA, DVM, KNN, YSA	X	X	X
[21]	NB, DVM, YSA	X	X	X
[22]	NB, KA, LR, DVM, KNN, YSA	RO	✓	X
[23]	NB, KA, LR, YSA, KNN	XGBoost	✓	X
[24]	KA, KNN	RO, XGBoost	X	X
[25]	NB, KA, DVM, KNN	RO	✓	✓
[26]	LR, DVM, KNN	RO, XGBoost	✓	✓
[27]	NB, KA, LR, KNN	RO,XGBoost, LightGBM,GB AdaBoost, CatBoost,	X	✓
[28]	--	TriBoost	✓	✓
Bu Çalışma	LR, DVM, YSA	RO, XGBoost, GBM	✓	✓

III. YÖNTEM

Bu çalışmada açık erişimli telekom verisi üzerinde müşteri kaybı tahmini farklı makine öğrenmesi yöntemleriyle yapılmış ve performans değerlendirilmesi sunulmuştur. Ayrıca en başarılı yöntemin sonuçlarının yorumlanabilmesini sağlamak için açıklanabilir yapay zeka yöntemi kullanılmıştır. Çalışmanın deney tasarımı Şekil 1'de şematize edilmiştir. Buna göre ham veriye öncelikle veri önişleme adımları uygulanmıştır. Daha sonra veri 80:20 oranında eğitim ve test verisi olmak üzere iki parçaya ayrılmıştır. Verideki sınıf dengesizliği problemini gidermek için sadece eğitim verisi olarak ayrılan kısım sentetik veri üretimi tekniği kullanılarak dengeli hale getirilmiştir. Bu işlem test için ayrılan kısma uygulanmamıştır. Aksi halde, test verisindeki üretilmiş veriler, makine öğrenmesi yönteminin eğitildiği veriyle benzerlik gösterecektir. Makine öğrenmesi yöntemleri dengeli hale getirilmiş veriye uygulanarak eğitilmiştir. Performans ölçümü için uygulanan testler sınıf dengesizliği bulunan test verisi üzerinde yapılmıştır. Böylece makine öğrenmesi yöntemlerinin performansının olduğundan daha iyi olduğu algısına yol açan aşırı uyum probleminin önüne geçilmiştir [22].

Makine öğrenmesi yöntemlerinden elde edilen sonuçlara göre en başarılı modele açıklanabilir yapay zeka yöntemi uygulanarak sonuçlar yorumlanmıştır.

A. Veri Seti

Bu çalışmada müşteri kaybı tahmini için açık erişimli, sınıf dengesizliği bulunan bir telekom verisi [29] kullanılmıştır. Veri setinde 7043 müşterinin *Cinsiyet (gender)* ve *Yaşlı/Emekli olma durumu (SeniorCitizen)* gibi demografik bilgilerini, *Abonelik Süresi (tenure)* ve *Kontrakt bilgisi (Contract)* gibi hesap bilgilerini, *Telefon hizmeti (PhoneService)* ve *İnternet Hizmeti (InternetService)* olup olmadığı gibi hizmet bilgilerini içeren öznitelikler yer almaktadır. Biri ikili sınıf değişkeni olmak üzere farklı türlerde toplam 21 öznitelik bulunmaktadır. Öznitelikler ve açıklamaları Tablo 2'de ayrıntılarıyla görülebilir. İkili sınıf değişkeninde 1849 müşteri *Ayrılan,* 5194 müşteri *Ayrılmayan* olarak etiketlendiğinden veride sınıf dengesizliği bulunduğu anlaşılmaktadır. Dengesizlik oranı, baskın sınıfın örneklem büyüklüğünün azınlık sınıfının örneklem büyüklüğünün oranı olarak hesaplanabilir. Çalışmada kullanılan veri seti için bu oran %35,5'tir.

B. Veri Ön İşleme

Veri ön işleme aşamasında öncelikle veri setinde tekrarlayan ya da eksik değerler bulunup bulunmadığı kontrol edilmiş, bu tip örneklere rastlanmamıştır. Sonrasında verideki kategorik değerler one-hot kodlama tekniğiyle nümerik değerlere dönüştürülmüştür. Artan sütun sayısının kontrol edilebilir boyutta kalmasını sağlamak için dönüştürülen her öznitelik için kategori sayısı 4'e indirilmiştir. Bütün özniteliklerin müşteri kaybına etkisinin ölçülebilmesi, açıklanabilir yapay zeka modelinde değerlendirilebilmesi için öznitelik seçme işlemi uygulanmamıştır. Ancak *Müşteri Numarası (customerID)* her müşteri için farklı bir değere sahip olduğundan ayrılacak müşteri tahmininde belirleyici olmadığı kanaatiyle silinmiştir. Bununla veri boyutunun makul seviyede tutulması amaçlanmıştır.

TABLO 2. ÖZNİTELİKLER

	Öznitelik	***Açıklama***	***Veri Tipi***
1	*customerID*	Müşteri Numarası	Nümerik
2	*gender*	Cinsiyeti	Kategorik
3	*SeniorCitizen*	Yaşlı/Emekli olup olmaması	Kategorik
4	*Partner*	Eş Durumu	Kategorik
5	*Dependents*	Bakmak zorunda olduğu kişi olup olmadığı	Kategorik
6	*tenure*	Abonelik 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 olup olmadığı	Kategorik
11	*OnlineBackup*	Çevrimiçi yedeği sahibi olup olmadığı	Kategorik
12	*DeviceProtection*	Cihaz korumasına sahip olup olmadığı	Kategorik
13	*TechSupport*	Teknik destek alıp almadığı	Kategorik
14	*StreamingTV*	Internet TV kullanıp kullanmadığı	Kategorik
15	*StreamingMovies*	Internet Film kullanıp kullanmadığı	Kategorik
16	*Contract*	Kontrat süresi	Kategorik
17	*PaperlessBilling*	Elektronik fatura kullanıp kullanmadığı	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

C. Verinin Dengeli Hale Getirilmesi

Sınıf dengesizliği bulunan verinin dengeli hale getirilmesi için literatürde farklı yöntemler mevcuttur. Bu yöntemler üç ana grupta incelenebilir:

1) Alt Örnekleme:

Bu yöntem, baskın sınıfa ait verilerden rastgele seçilen bir bölümün silinmesiyle gerçekleşir. Veri setini daha dengeli hale getirmenin yanı sıra, özellikle büyük veri kümelerinde daha küçük boyutlu veriyle çalışılmasını sağlar. Bu da makine öğrenmesi yöntemlerinin hesaplama süresini kısaltır. Ancak, silinen verilerdeki bilgiler kaybolduğundan yetersiz uyum (underfitting) problemine yol açabilir.

2) Aşırı Örnekleme:

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.

3) Sentetik Veri Üretilmesi:

Bu yöntemde, azınlık sınıfına ait yeni sentetik veriler oluşturularak veri setine eklenir. Bu şekilde dengeli hale getirilen veride hem aşırı uyum hem de yetersiz uyum sorunları daha az görülür. Literatürde sentetik veri üretimi için farklı stratejilere sahip birçok yöntem vardır. Bunlar arasında Sentetik Azınlık Aşırı Örnekleme Yöntemi (Synthetic Minority Over-sampling Technique - SMOTE) en yaygın kullanılan sentetik veri üretme tekniklerinden biridir. Chawla vd. [30] tarafından geliştirilen bu yöntem birçok sentetik veri üretme algoritması için de temel oluşturur. Bu yöntemin çalışma prensibi şu şekildedir: Öncelikle azınlık sınıfından rastgele bir örnek seçilir. Bu örneğin azınlık sınıfına ait k en yakın komşusu belirlenir. Bu komşular arasından bir tanesi rastgele seçilir ve seçilen örnekler bir doğru üzerinde birleştirilir. Bu doğru üzerindeki seçilen rastgele bir nokta, yeni sentetik veri olarak veri setine eklenir. Bu şekilde oluşturulan sentetik veri, seçilen iki örneğin bir doğrusal kombinasyonu olarak hesaplanmış olur.

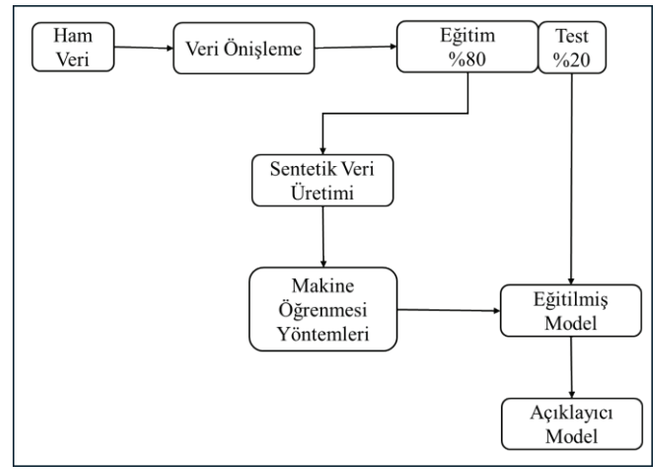
Bu çalışmada verideki sınıf dengesizliği problem SMOTE yöntemi ile sentetik veri üretilerek giderilmiştir. Yeniden örneklenmiş veri setinde dengesizlik oranı 1'dir. Yani *Ayrılan* ve *Ayrılmayan* müşteri sayıları birbirine eşitlenmiştir.

D. Makine Öğrenmesi Yöntemleri

Bu çalışmada telekom verisinde müşteri kaybı tahmini için verimliliği ve güvenirliliği dolayısıyla sıklıkla kullanılan makine öğrenmesi yöntemlerinden Lojistik Regresyon, Destek Vektör Makineleri, Yapay Sinir Ağları, Rastgele Orman, Gradyan Güçlendirme Makineleri ve Aşırı Gradyan Güçlendirme yöntemlerinin performanslarının karşılaştırılması sunulmaktadır. Kullanılan yöntemlere ait kısa bilgiler aşağıda yer almaktadır.

1) Lojistik Regresyon (LR)

Lojistik regresyon, özellikle ikili sınıflandırma problemleri için kullanılan, amacın bağımsız değişkenlere (özniteliklere) göre bir olayın meydana gelme olasılığını tahmin etmek olduğu istatistiksel bir yöntemdir. Oluşturulan model, bağımsız değişkenlerin doğrusal kombinasyonunu olasılıklara dönüştürmek için logit fonksiyonunu kullanır ve tahmin edilen olasılıkların 0 ile 1 arasında olmasını sağlar. Lojistik regresyondaki katsayılar, diğer tüm değişkenleri sabit tutarken bağımsız değişkendeki bir birimlik değişiklik için olasılık oranındaki değişikliği temsil eder. Bu, sonuçların olayın



Şekil 1. Deney Tasarımı.

gerçekleşme olasılığı açısından yorumlanmasına olanak tanır.

2) Destek Vektör Makineleri (DVM)

Destek Vektör Makineleri hem sınıflandırma hem de regresyon için kullanılan bir denetimli öğrenme yöntemidir [31]. DVM ana fikri öznitelik uzayında sınıfları aralarındaki uzaklığı maksimize edecek şekilde ayıran optimum hiper düzlemi belirlemektir. Bu hiper düzlem, hiper düzlem ile destek vektörleri olarak bilinen her bir sınıfın en yakın veri noktaları arasındaki mesafe olan en büyük marja sahip olacak şekilde seçilir. DVM'ler özellikle verilerin doğrusal olarak ayrılamadığı durumlarda polinom, sigmoid gibi çekirdek fonksiyonları kullanarak verilerin doğrusal olarak ayrılabilir hale geldiği daha büyük boyutlu bir uzaya yansıtır ve etkili sınıflandırmaya izin verir. DVM'lerin amacı, marjı en üst düzeye çıkararak yanlış sınıflandırma riskini en aza indirmektir. Bu, DVM'leri özellikle öznitelik sayısının veri sayısına kıyasla büyük olduğu durumlarda etkili kılar ve güçlü sınıflandırma performansı sunar.

3) Yapay Sinir Ağları (YSA)

Yapay sinir ağları, insan beyninin yapısı ve işlevinden esinlenen temel bir makine öğrenmesi yöntemidir. Birbirine bağlı düğümlerden veya bir giriş katmanı, bir veya daha fazla gizli katman ve bir çıkış katmanı olmak üzere katmanlar halinde düzenlenmiş nöronlardan oluşur. Giriş katmanı, karmaşık, doğrusal olmayan ilişkileri modellemek için ağırlıklar ve aktivasyon fonksiyonları kullanılarak gizli katmanlar aracılığıyla işlenen ham verileri alır. Her nöron girdilerinin ağırlıklı bir toplamını hesaplar, doğrusal olmayan bir aktivasyon fonksiyonu uygular ve sonucu bir sonraki katmana aktarır. Eğitim sırasında ağ, tahmin edilen ve gerçek çıktılar arasındaki hatayı en aza indirmek için geriye yayılma gibi teknikler kullanarak bu ağırlıkları hesaplar. Bu yinelemeli öğrenme süreci, yapay sinir ağlarının verilerden otomatik olarak örüntüler ve özellikler çıkarmasını sağlar. YSA verilerden genelleme yapma ve karmaşık sorunları yüksek doğrulukla ele alma yetenekleri nedeniyle görüntü tanıma, doğal dil işleme, tahmine dayalı modelleme ve karar verme gibi uygulamalarda yaygın olarak kullanılmaktadır.

4) Rastgele Orman (RO)

Rastgele Orman, tahminlerin doğruluğunu artırmak için birden fazla karar ağacını birleştiren bir topluluk öğrenme yöntemidir. Rastgele Ormanın ana fikri, her biri verinin rastgele bir alt kümesi üzerinde

eğitilmiş çok sayıda karar ağacı oluşturmaktır. Bu rastgelelik, aşırı uyumu azaltmaya yardımcı olur ve modelin genelleme yeteneğini artırır. Ormandaki her ağaç hedef değişkeni tahmin eder ve nihai tahmin, sınıflandırma problemlerinde oylama veya regresyon problemlerinde ortalama alma yoluyla tüm ağaçların tahminlerinin değerlendirilmesiyle yapılır. RO birden fazla karar ağacının güçlü yönlerinden yararlanarak hem kategorik hem de sayısal verileri işleyebilir. Bundan dolayı araştırmacılar tarafından büyük ve karmaşık veri setlerinde sıklıkla tercih edilir.

TABLO 3. MODEL HİPERPARAMETRELERİ

	Hiperparametre	***Değer***
DVM	Kernel	Linear
	C	10
YSA	Gizli katman boyutu	(64,32)
	Aktivasyon	Relu
	Çözümleyici	Adam
	Max-iterasyon	200
RO	n-tahminci	100
GGM	n-tahminci	100
	Öğrenme-hızı	0.3
	Max- derinlik	3
XGB	Yineleme sayısı	100
	Öğrenme-hızı	0.1
	Max- derinlik	3

5) Gradyan Güçlendirme Makineleri (GGM)

Gradyan Güçlendirme Makineleri hem regresyon hem de sınıflandırma problemleri için kullanılan güçlü bir topluluk öğrenme yöntemidir. GGM'lerin çalışma prensibi karar ağaçları gibi birden fazla zayıf öğreniciyi sırayla birleştirerek güçlü bir tahmin modeli oluşturmaktır. Sıradaki her ağaç, tahmin edilen ve gerçek değer arasındaki farklara odaklanarak kendinden önceki ağacın hatalarını düzeltmek üzere eğitilir. Bu iteratif süreç, gradyan iniş optimizasyonunu kullanarak ortalama karesel hata veya log kaybı gibi belirli bir kayıp fonksiyonunu en aza indirir. GGM'ler, tüm ağaçların tahminlerini bir araya getirerek verideki karmaşık ilişkileri yakalayan sağlam bir tahmin modeli oluşturur. GGM farklı kayıp fonksiyonlarına göre uyarlanabilir. Ayrıca düzenlileştirme ve öğrenme hızı ayarı gibi teknikler aracılığıyla aşırı uyumu azaltmada etkilidir. Bu özellikler GGM'leri çeşitli uygulamalarda yüksek performanslı modelleme için popüler bir seçenek haline getirmektedir.

6) Aşırı Gradyan Güçlendirme (XGBoost)

Aşırı Gradyan Güçlendirme hızı, ölçeklenebilirliği ve tahmin performansı ile öne çıkan gradyan artırma algoritmasının optimize edilmiş ve gelişmiş bir uygulamasıdır. GGM'inkine benzer şekilde karar ağaçları gibi zayıf öğrenicileri yinelemeli olarak birleştirerek tahmin modelleri oluşturur. Ancak XGBoost'da aşırı uyumu kontrol etmek için bulunan L1 (Lasso) ve L2 (Ridge) düzenlemeleri GGM'de bulunmamaktadır. Ayrıca XGBoost, paralel işleme, etkin bellek yönetimi ve önbelleğe duyarlı algoritmalar gibi iyileştirmeler sayesinde GGM'den daha hızlı sonuç üretebilmektedir. Dağıtık hesaplamayı desteklemesi büyük ölçekli veri kümelerinde çalışmasını kolaylaştırmaktadır. XGBoost, seyrekliğe duyarlı algoritmalar kullanarak verideki eksik değerleri kendi içindeki yapılarla işlerken, GGM genellikle bunun için veri ön işleme gerektirir. XGBoost, aşırı uyumla daha iyi mücadele edebilmek için ağaçlar için bırakma düzenlemesi getiren DART (Bırakma Katkılı Regresyon Ağaçları) gibi ek teknikler içerir. Bu tür özellikler GGM uygulamalarında mevcut değildir. XGBoost'un sahip olduğu bu özellikler araştırmacılar tarafından sıklıkla tercih edilmesini sağlamaktadır.

IV. DENEYSEL BULGULAR VE TARTIŞMALAR

Bu bölümde deneylerden elde edilen sonuçlar, bu sonuçlara göre özniteliklerin değerlendirilmesi ve bunların tartışılması yer almaktadır.

A. Makine Öğrenmesi Modellerinin Performansı ve Tartışma

Makine öğrenmesi modelleri Scikit-learn paketi [32] kullanılarak Python programla diliyle uygulanmıştır. Modellerin performansını optimize etmek için ızgara arama (grid search) yöntemiyle hiperparametre ayarlaması yapılmıştır. Modeller için seçilen bazı hiperparametreler Tablo 3'te görülmektedir. Bunlar dışındaki parametreler için varsayılan değerler kabul edilmiştir.

Modellerin performansı Doğruluk, Kesinlik, Duyarlılık, F1-Skoru ve ROC-Eğrisinin Altında Kalan Alan (ROC-AUC) ölçütlerine göre değerlendirilmiştir. Elde edilen sonuçlar Tablo 4'te sunulmuştur. Her bir ölçüt için en iyi performans değeri koyu, en düşük performans değeri ise italik ile gösterilmiştir. Çalışmada kullanılan spesifik veri setinden elde edilen sonuçlara göre Rastgele Orman bütün ölçütlere göre en iyi performansı sergileyerek diğerlerine üstünlük sağlayan yöntem olmuştur. Destek Vektör Makineleri Doğruluk ve Keskinlik ölçütlerine göre, Yapay Sinir Ağları ise Duyarlık, F1-skoru ve ROC-AUC ölçütlerine göre en düşük performansı sergileyen yöntemler olmuşlardır.

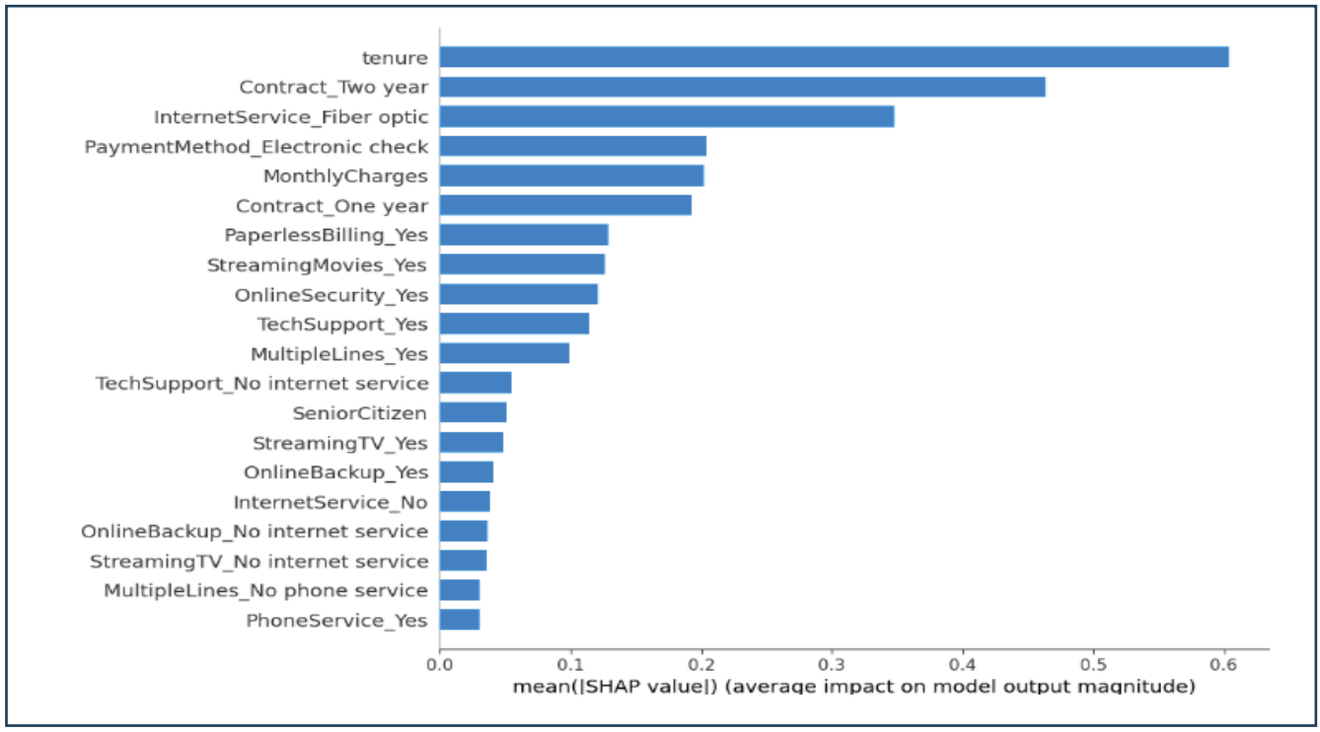
TABLO 4. MAKİNE ÖĞRENMESİ MODELLERİNİN PERFORMANSI

	Doğruluk	***Kesinlik***	***Duyarlılık***	***F1-Skoru***	***ROC-AUC***
LR	0.759	0.543	0.796	0.645	0.770
DVM	*0.748*	*0.528*	0.817	0.641	0.770
YSA	0.774	0.599	*0.543*	*0.570*	*0.702*
RO	**0.859**	**0.874**	**0.837**	**0.855**	**0.859**
GBM	0.796	0.647	0.572	0.607	0.726
XGB	0.777	0.576	0.719	0.639	0.759

Deney sonuçlarına göre Rastgele Orman modeli müşteri kaybını %85,9 gibi yüksek bir doğruluk oranıyla tahmin edebilmektedir. Rastgele Orman birden çok karar ağacının bir araya gelerek oluşturduğu bir topluluk öğrenme modeli olduğu için topluluk öğrenme modeli olmayan LR, DVM ve YSA'dan daha iyi performans göstermesi şaşırtıcı olmamıştır. Çalışmada yer verilen diğer topluluk öğrenme yöntemleri olan GGM ve XGBoost modelleri doğruluk ölçütüne göre topluluk öğrenme modeli olmayan LR, DVM ve YSA'dan daha iyi performans sergilemiş olmalarına rağmen özellikle dengesiz veri setlerindeki performans değerlendirmelerinde sıklıkla tercih edilen F1-skoru ve ROC-AUC ölçütlerine göre bu modellerden daha düşük başarı değerlerine sahip olmuşlardır. Bu sonucun kullanılan veri setinin özelliklerinden kaynaklandığı düşünülmektedir. Çünkü GGM ve XGBoost modellerinin daha çok büyük ve karmaşık bir yapıya sahip veri setlerinde daha başarılı sonuçlar verdikleri bilinmektedir [33].

İleriki bir çalışma konusu olarak GGM ve XGBoost yöntemlerinin daha büyük boyutlu ve öznitelik ilişkilerinin daha karmaşık olduğu veri setlerindeki performanslarının gözlemlenmesi planlanabilir.

setini kullanan [25]'in Rastgele Orman modelinden elde ettiği sonuç %77,9 iken bu çalışmada aynı yöntemle %85,9 başarı elde edilmiştir. Bu çalışmada üretilen model mevcut çalışma sonuçlarından daha başarılı sonuçlar üreterek telekom sektöründe



Şekil 2. RO modeline göre özniteliklerin katkılarını gösteren SHAP değerleri

Topluluk öğrenme modelleri kendi aralarında değerlendirildiğinde de Rastgele Orman modelinin GGM ve XGBoost'tan daha iyi performans sergilemesinin sebebi olarak Rastgele Orman modelinin küçük ve dengesiz veri setlerinde diğerlerinden daha başarılı tahminlerde bulunması gösterilebilir.

Çalışmanın bulguları yeniden örnekleme ile dengelenmiş veri setinde topluluk öğrenme yöntemlerini de içeren farklı modellerin performansını değerlendiren mevcut çalışmalarla da kıyaslanmıştır. Tüm çalışmalar modellerin F1-skorunu raporladığından kıyaslama bu ölçüte göre yapılmıştır. Buna göre [22]'de SMOTE ile dengelenmiş veri setinden iyi performansı %78,9 ile DVM göstermiştir. Topluluk öğrenme modeli olarak sadece RO kullanılmış, performansı %74 olmuştur. [23] topluluk öğrenme yöntemleri olarak XGboosting ve Gradyan Artışı yöntemleri kullanılmış, XGboosting %68,8 ile en iyi performansı sergilerken buna en yakın geleneksel makine öğrenmesi modeli %46,72 başarı ile Karar ağaçları modeli olmuştur. [25]'te ise SMOTE ile dengelenmiş veri setinde topluluk öğrenme yöntemi olan Rastgele Orman %77,9 başarı oranı göstermiştir. [26]'da ise XGBoost algoritması %99 başarı ile yüksek bir performans sergilemiştir. [27]'de LightGBM yöntemi %73,085 doğruluk oranı ile en iyi performansa sahip yöntem olmuştur. [28]'de ise topluluk öğrenme yöntemlerinin kendi aralarında yapılan değerlendirmede kombine yöntemin bireysel yöntemlere göre kayda değer bir iyileştirme sağlamadığı gözlemlenmiştir.

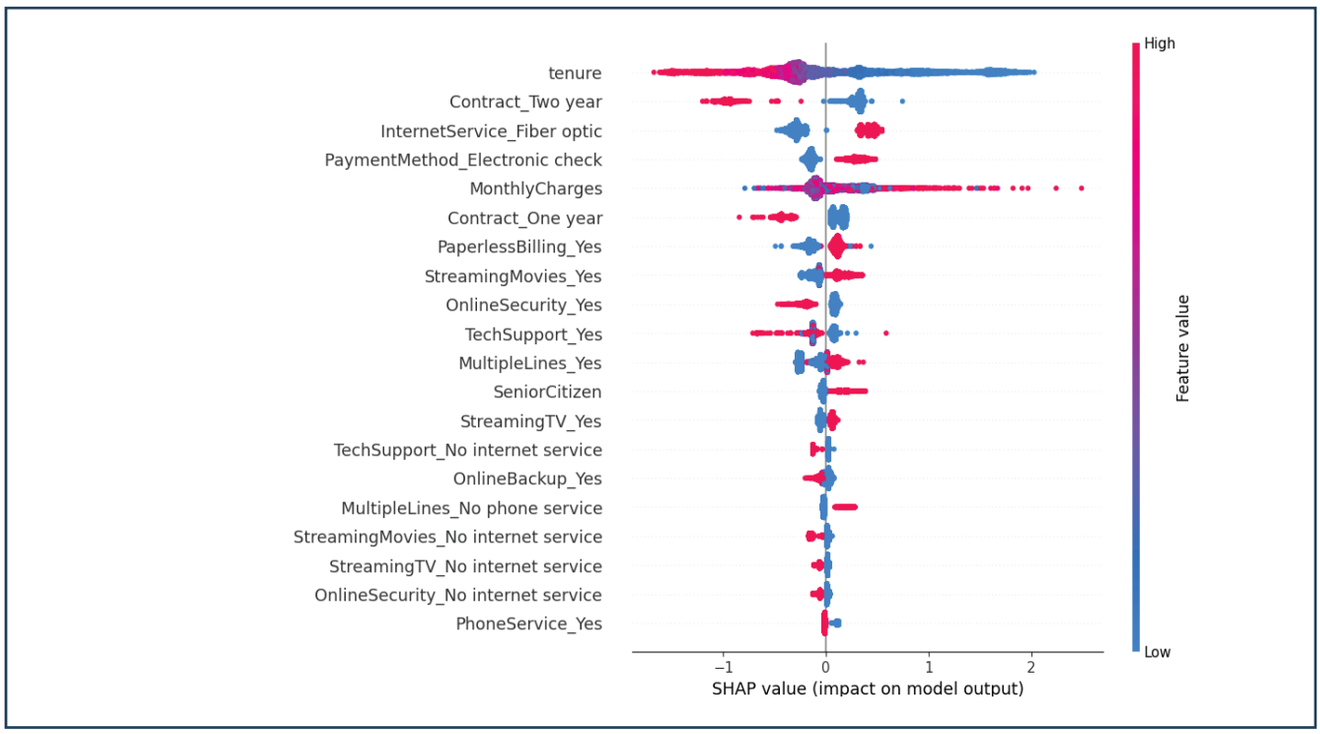
Mevcut çalışmalardan elde edilen sonuçlar, bu çalışmadan elde edilen topluluk yöntemlerinin geleneksel makine öğrenmesi yöntemlerinden daha iyi performans sergilediği sonucunu desteklemektedir. Bu çalışmadaki yönteme en yakın yöntem ve veri

müşteri kaybı tahminine katkı sağlamıştır.

B. Özniteliklerin Değerlendirilmesi

Telekom sektöründe ayrılacak müşterilerin önceden doğru tahmin edilmesi müşteri kaybını önlemede tek başına yeterli olmamaktadır. Ayrılacak müşteri tahmin edildikten sonra bu müşterilere sağlanacak teşvik ve promosyonların doğru belirlenerek müşteri kaybının en aza indirilmesi hedeflenmektedir. Makine öğrenmesi modelleri yüksek tahmin performansı göstermelerine rağmen teşviklerin planlanması için bilgi sunmazlar. Ancak açıklanabilir yapay zeka yöntemleri kullanılarak makine öğrenmesi modelinin tahminlerinin arkasındaki sebeplerin şeffaf bir şekilde yorumlanması sağlanabilir.

Bu çalışmada açıklanabilir yapay zeka yöntemlerinden Shapley Katkı Açıklamaları (SHAP-Shapley Additive Explanations) kullanılmıştır. SHAP yöntemi ana fikri Lloyd Shapley'nin oyun teorisinde oyuncuların oyunun skoruna bireysel katkılarını bulmak için hesapladığı Shapley değerlerine dayanmaktadır [34]. Benzer mantıkla verideki her bir özniteliğin Shapley değeri hesaplanarak makine öğrenmesi modelinin performansına olan katkısı açıklanmaya çalışılır. Bu yöntem telekom sektöründe müşteri kaybı tahmini probleminde müşterilerin ayrılacak diye tahmin edilmesinde hangi özniteliklerin katkısının daha fazla olduğunun belirlenmesini kolaylaştırır. Makine öğrenmesi modelinin tahminin ardındaki sebebin açıklanabilmesi belirlenen özelliklere sahip müşterilerin daha iyi analiz edilmesine olanak sağlar. Böylece bu müşterilerin ayrılmalarını önleyecek teşvik ve promosyonlar daha kolay ve doğru bir şekilde belirlenebilir



Şekil 3. Rastgele Orman modeline göre Arı Sürüsü Grafiği

Bu çalışmada kullanılan veri setinde %85,9 doğruluk oranıyla en iyi performans gösteren Rastgele Orman modelinin sonuçlarını yorumlayabilmek amacıyla SHAP yöntemi kullanılmıştır. Bu yöntemle hesaplanan ağırlıklara göre sıralanmış en önemli on öznitelik Tablo 5 'te sunulmuştur. Şekil 2'de ise tüm özniteliklerin modele katkısı görülebilir. Bu sonuçlara göre Abonelik Süresi *(tenure)*, Kontrat süresi *(Contract)* ve İnternet Servis Sağlayıcısı *(InternetService)* türü ayrılıp ayrılmama kararına en çok etki eden öznitelikler olarak görülmektedir. Müşterinin Internet TV kullanıp kullanmadığı *(StreamingTV),* Birden fazla hatta sahip olup olmadığı *(MultipleLines)* ve Telefon servisi alma durumunun *(PhoneService)* modele katkısı en az olmuştur.

TABLO 5. RO MODELİNE GÖRE EN ÖNEMLİ 10 ÖZNİTELİK

	Öznitelik	Ağırlık
1	tenure	0.604147
2	Contract_Two year	0.463502
3	InternetService_Fiber optic	0.347486
4	PaymentMethod_Electronic check	0.203429
5	MonthlyCharges	0.202202
6	Contract_One year	0.192431
7	PaperlessBilling_Yes	0.128885
8	StreamingMovies_Yes	0.126132
9	OnlineSecuriy_Yes	0.121431
10	TechSupport_Yes	0.113812

Öznitelik değerlerinin RO modeliyle müşteri kaybı tahminine etkisinin anlaşılabilmesi için SHAP yöntemiyle çizilen Arı Sürüsü Grafiği (BeeSwarm plot) incelenmiştir. Şekil 3'te görülen grafikte öznitelikler önem sıralarına göre yukarıdan aşağıya sıralanmışlardır. Her bir veri noktasının özniteliklere göre dağılımını kırmızı veya mavi renkteki noktalarla gösterilmiştir. Her noktanın rengi, ilgili özniteliğin orijinal değerini gösterir. Genellikle, yüksek öznitelik değerleri kırmızı, düşük değerler ise mavi renkte gösterilir. Her bir noktanın yataydaki yeri, o özniteliğin SHAP değerine göre belirlenir. Pozitif SHAP değerleri, yani merkezdeki dikey çizginin sağındaki noktalar, özniteliğin bu değerlerinin modelin sonucunu artırdığını, negatif SHAP değerleri, merkezdeki dikey çizginin solundaki noktalar ise modelin sonucunu azalttığını gösterir. Bu, öznitelik değerinin modelin çıktısına nasıl etkide bulunduğunu anlamaya yardımcı olur.

Şekil 3 bu bilgiler ışığında incelendiğinde müşteri kaybı tahminine etkisi en fazla olan Abonelik süresi *(tenure)*'nin değeri büyüdükçe müşterinin ayrılma olasılığının azaldığı anlaşılmaktadır. Buna göre müşteri kaybını en aza indirmek isteyen bir telekom firması uygulayacağı promosyon ve teşvikleri daha çok Abonelik süresi daha kısa olan müşterilerine yönelik planlayabilir. Benzer şekilde Kontrat süresi *(Contract)* arttıkça da ayrılma olasılığı azalmaktadır. Özellikle Kontrat süresi iki yıl olan müşterilerde bu olasılığın oldukça düşük olduğu görülmektedir. Buna göre firmanın müşterilerinin Kontrat Süresini uzattığında müşteri kaybının azalacağı beklenebilir. Ayrıca İnternet Servis Sağlayıcısı Fiber Optik olan müşterilerin ayrılma olasılığının daha fazla olduğu görülmektedir. Bu telekom firmasının fiber optik alt yapısının yeterli olmadığını düşündürmektedir. Buna göre firmanın fiber optik internet servis sağlayıcısına yatırım yaparak şartlarını geliştirmesi müşteri kaybını azaltmaya yardımcı olabilir. Bunun yanında Ödeme Şekli (PaymentMethod) incelendiğinde Elektronik Çek (Electronic Check) olan müşterilerin de ayrılma olasılığının daha fazla olduğu ama diğer ödeme türlerinin (Kredi kartı ve Posta Çeki) müşterilerin ayrılma kararında önemli bir etkisinin olmadığı görülmektedir. Bu da firmada Elektronik Çek ödeme türüyle ilgili bir problem olabileceğini düşündürmektedir. Firmanın bunu fark ederek problemi çözmesi ya da müşterileri diğer ödeme türlerine yönlendirmesi müşteri kaybını azaltmaya yönelik önlemler arasında

olabilir. Aylık ödeme miktarı (MonthlyCharges) bilgileri incelendiğinde ise ödeme miktarı çok fazla olan müşterilerin ayrılma olasılığının yüksek olduğu anlaşılırken ortalama bir miktar ödeyen müşterilerin ayrılıp ayrılmayacağına dair net bir çıkarım yapılamamaktadır.

V. SONUÇ

Telekomünikasyon sektöründe müşteri kaybı sebeplerini anlamak ve azaltmak büyüme ve karlılığı sürdürebilmek için oldukça önemlidir. Bu yüzden firmalar ayrılma riski yüksek müşterilerin önceden tahmin ederek bu müşterileri elde tutmaya yönelik geliştirecekleri teşvik ve promosyonlarla müşteri kaybını en aza indirmek isterler. Bunu gerçekleştirmede müşteri verilerini işleyip analiz eden makine öğrenmesi yöntemleri hangi müşterilerin ayrılabileceğini etkili bir şekilde tahmin etme yeteneği ile öne çıkmaktadır.

Bu çalışmada müşteri kaybı tahmini için başarılı bir makine öğrenmesi yöntemi belirlerken aynı zamanda müşteri kaybının arkasındaki nedenlere ilişkin öngörüler de sunmak hedeflenmiştir. Bunun için geleneksel makine öğrenmesi yöntemlerinden Lojistik Regresyon, Destek Vektör Makineleri ve Yapay Sinir Ağlarının performansı daha gelişmiş topluluk öğrenme yöntemleri olan Rastgele Orman, Gradyan Güçlendirme Makineleri ve Aşırı Gradyan Güçlendirme yöntemlerinin performansıyla kıyaslanmıştır. Telekom verilerinde sıklıkla karşılaşılan sınıf dengesizliği problemi model performansına olumsuz etkileri sebebiyle ele alınıp SMOTE yeniden örnekleme yöntemiyle giderilmiştir. Testlerde Rastgele Orman modeli %85,9 doğruluk oranı ile diğer modellere göre üstünlük sağlamıştır.

Makine öğrenmesi yöntemleri müşteri kaybı tahmininde yüksek başarı göstermesine rağmen özniteliklerin tahminleri nasıl etkilediğine dair bir bilgi sunmamaktadır. Müşterilerin ayrılma kararlarının ardındaki sebebi anlayabilmek, özniteliklerin önemi hakkında fikir edinmek için Rastgele Orman modelinin tahminleri için SHAP değerlerinden yararlanılmıştır. Bu bilgiler, müşteri kaybını en çok etkileyen faktörleri belirlemek isteyen telekom şirketleri için yol gösterici olmaktadır. Bu çalışmada kullanılan müşteri verisine göre SHAP değerleri kullanılarak, abonelik süresi, kontrat süresi ve internet servis sağlayıcısı türü özniteliklerinin müşteri kayıp tahmini üzerinde önemli etkiye sahip olduğunu belirlenmiştir. Böylece müşteri kaybına neden olan temel faktörlerin şeffaf bir şekilde yorumlanabilmesini sağlanmış, telekomünikasyon sektöründe müşteri kaybını en aza indirmek için geliştirilecek stratejilerin belirlenmesi kolaylaşmıştır.

KAYNAKLAR

[1] J. Cao, X. Yu & Z. Zhang, “Integrating OWA and data mining for analyzing customers churn in E-commerce.”, J Syst Sci Complex, vol.28, pp. 381–392, (2015) https://doi.org/10.1007/s11424-015-3268-0

[2] J. Burez, & D. Van den Poel. “Crm at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services”. Expert Systems with Applications, 32, 277–288. (2007)

[3] E. Kartal, Z. Özen, “Dengesiz Veri Setlerinde Sınıflandırma”, Mühendislikte Yapay Zekâ Uygulamaları, Sakarya, 109-131, (2017).

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

[5] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” Expert Syst. Appl., vol. 36, no. 3 PART 1, pp. 4626–4636, (2009) doi: 10.1016/j.eswa.2008.05.027.

[6] F. Emmert-Streib, O. Yli-Harja, M. Dehmer. “Explainable artificial intelligence and machine learning: A reality rooted perspective”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), Article e1368. (2020).

[7] A. Barredo Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Inform Fusion, vol. 58, pp. 82-115, (2020).

[8] C. Kirui, et al. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining." International Journal of Computer Science Issues (IJCSI) 10.2 Part 1: 165 (2013).

[9] P. Kisioglu and Y.I. Topcu, “Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey,” Expert Syst Appl, vol. 38, no. 6, pp. 7151 7157, (2011).

[10] I. Brandusoiu, G. Toderean, “Churn prediction modeling in mobile telecommunications industry using decision trees”. Journal of Computer Science and Control Systems, *6*(1), 14, (2013).

[11] S. Nisha, G. K. Monika, and K. Garg. "Churn prediction in telecommunication industry using decision tree." International Journal of Engineering Research & Technology *(IJERT)* 6, no. 4: 439-433 (2017).

[12] K. Preeti, et al. "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression." 2016 symposium on colossal data analysis and networking (CDAN). IEEE,(2016).

[13] J. Manďák, and J. Hančlová. "Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications." Statistika: Statistics & Economy Journal 99.2 (2019).

[14] Z. Yu, et al. "Customer churn prediction using improved one-class support vector machine." Advanced Data Mining and Applications: First International Conference, ADMA 2005, Wuhan, China, July 22-24, 2005. Proceedings 1. Springer Berlin Heidelberg, (2005).

[15] G. Xia, and W. Jin. "Model of customer churn prediction on support vector machine." Systems Engineering-Theory & Practice 28.1): 71-77. (2008).

[16] N. Sjarif, et al. "A customer Churn prediction using Pearson correlation function and K nearest neighbor algorithm for telecommunication industry." Int. J. Advance Soft Compu. Appl 11.2 (2019).

[17] C. F. Tsai and Y. H. Lu, “Customer churn prediction by hybrid neural networks,” Expert Syst. Appl., vol. 36, no. 10, pp. 12547–12553, (2009). doi: 10.1016/j.eswa.2009.05.032.

[18] S. and P. "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services" International Journal of Computer Applications (2011) doi:10.5120/3344-4605.

[19] O. Adwan, et al. "Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis." Life Science Journal 11.3): 75-81. (2014).

[20] A., Keramati et. Al. “Improved churn prediction in telecommunication industry using data mining techniques”, Applied Soft Computing, Volume 24, 994-1012, (2014). https://doi.org/10.1016/j.asoc.2014.08.041.

[21] O. Kaynar v.d.“Makine öğrenmesi yöntemleriyle müşteri kaybı analizi”, Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 18 (1) , 1-14, (2017)

[22] M. A. Aydın, “Müşteri kaybı tahmininde sınıf dengesizliği problemi”, Politeknik Dergisi, 25(1): 351-360, (2022).

[23] A. Sikri, R. Jameel, S.M. Idrees, *et al.* “Enhancing customer retention in telecom industry with machine learning driven churn prediction”. Sci Rep 14, 13097 (2024). https://doi.org/10.1038/s41598-024-63750-0

[24] M. Afzal, S. Rahman, D. Singh and A. Imran, "Cross-Sector Application of Machine Learning in Telecommunications: Enhancing Customer Retention Through Comparative

Analysis of Ensemble Methods," in *IEEE Access*, vol. 12, pp. 115256-115267, (2024) doi: 10.1109/ACCESS.2024.3445281.

[25] İ. Kök, "Açıklanabilir Yapay Zekaya Dayalı Müşteri Kaybı Analizi ve Elde Tutma Önerisi", Müh.Bil.ve Araş.Dergisi, c. 6, sy. 1, ss. 13–23, (2024) doi: 10.46387/bjesr.1344414.

[26] P. Boozary, S. Sheykhan, H. GhorbanTanhaei, C. Magazzino,"Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction", International Journal of Information Management Data Insights,Volume 5, Issue 1,100331, (2025)

[27] C. Özkurt. "Transparency in Decision-Making: The Role of Explainable AI (XAI) in Customer Churn Analysis", *ITEB*, vol. 2, no. 1, pp. 1–11, Jan. 2025,

[28] Malhotra, R., Shah, S., Bansal, S. "Customer Churn Analysis Using Machine Learning in Telecom Industry". Proceedings of Data Analytics and Management. ICDAM 2024. Lecture Notes in Networks and Systems, vol 1301. Springer, Singapore. (2025).

[29] https://www.kaggle.com/blastchar/telco-customer-churn/version/1 (Son erişim tarihi: 05/02/2025)

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

[31] B. E. Bernhard,I. M. Guyon, and V. N. Vapnik. "A training algorithm for optimal margin classifiers." Proceedings of the fifth annual workshop on Computational learning theory. (1992).

[32] F. Pedregosa, et al. "Scikit-learn: Machine Learning in Python." ArXiv abs/1201.0490 (2011)

[33] H. Alshari, A. Saleh, ve A. Odabas, "CPU Performansı için Gradyan Artırıcı Karar Ağacı Algoritmalarının Karşılaştırılması", Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, c. 37, sy. 1, ss. 157–168, (2021).

[34] S. M., Lundberg &, S. I. Lee, "A unified approach to interpreting model predictions". Advances in Neural Information Processing Systems, 30, 4765-4774, (2017).

Özgeçmişler

M. Aslı Aydın, lisans derecesini Boğaziçi Üniversitesi Matematik Bölümü'nden, yüksek lisans ve doktora derecelerini sırasıyla Bahçeşehir Üniversitesi ve Boğaziçi Üniversitesi Endüstri Mühendisliği Bölümü'nden almıştır. Araştırma ilgi alanları arasında yöneylem araştırması ve makine öğrenmesi uygulamaları yer almaktadır. Güncel araştırmaları kombinatoryal optimizasyon problemleri için takviyeli öğrenme yöntemleri üzerinde yoğunlaşmaktadır.