Müşteri Adayının İşlem Yapma İhtimalinin Tahminlenmesi Modeli

Model for Estimating the Probability of a Customer to Have a Transaction

Alperen Sayar
Information Technology Graduate Program
MEF University
Istanbul, Türkiye
sayara@mef.edu.tr

Tunahan Bozkan
Information Technology Graduate Program
MEF University
Istanbul, Türkiye
bozkant@mef.edu.tr

Tuna Çakar
Computer Engineering Department
MEF University
Istanbul, Türkiye
cakart@mef.edu.tr

Seyit Ertuğrul
Department of Information Technology
MEF University
Istanbul, Türkiye
ertugruls@mef.edu.tr

2022 7th International Conference on Computer Science and Engineering (UBMK) | 978-1-6654-7010-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/UBMK55850.2022.9919439

***Öz*—Bu çalışmada aktif olarak faktöring sektöründe faaliyet yürüten bir şirkette mal ve hizmet satışından doğan alacakların temlik alınarak satıcı firmaya finansman sağlanması amacıyla kuruma ilk defa gelen müşteri adayının gelecek 3 ay içerisindeki işlem yapma ihtimalinin veri güdümlü makine öğrenmesi modelleri kullanılarak tahminlenmesi hedeflendi. Buna bağlı olarak da yüksek potansiyelli ve düşük potansiyelli müşterilerin bulunmasıyla daha etkin, verimli ve doğru yaklaşımlar içerisinde hareket edip aksiyon alarak, işletme bazında işlem hacmine doğrudankatkı sağlanması amaçlandı. Bu kapsamda KKB (Kredi Kayıt Bürosu) tarafından sağlanan; müşteri adaylarının Risk, Mersis, GİB bilgilerinden yararlanılarak ve veri tabanında tutulan müşterilerin, çek keşidecilerinin, müşteri temsilcilerinin ve şubelerin geçmiş bilgilerinden yola çıkarak öznitelik mühendisliği ve keşifsel veri analizleriyle makine öğrenmesi modellerinde kullanılacak veri seti oluşturuldu. Kuruma gelen müşteri adayları iki farklı kuruluş türünde olduğundan dolayı (Şahıs ve Tüzel) iki farklı tahminlenme modeli uygulandı. Birden çok sınıflandırma modelleri denenmiş olup, Şahıs şirketler için en yüksek F1-Skoru %86 ile Rastgele Orman Ağaçlar (Random Forest) modeli, Ticari şirketler için de en yüksek F1-Skoru %82 ile Rastgele Orman Ağaçlar (Random Forest) modeli ile sağlandı.**

Anahtar Sözcükler — Faktoring, İşlem Tahmini, Makine Öğrenmesi

***Abstract*—In this study, it is aimed to estimate the probability of a customer who comes to the institution for the first time to make a transaction in the next 3 months, using data-driven machine learning models, in order to provide financing to the seller company by assigning the receivables arising from the sale of goods and services in a company actively operating in the factoring sector. Accordingly, it was aimed to directly contribute to the transaction volume on a business basis by acting and taking action with more effective, efficient and correct approaches by finding high-potential and low-potential customers. In this context, provided by KKB (Credit Registration Bureau); The data set to be used in machine learning models was created with feature engineering and exploratory data analysis, using the Risk, Mersis, GİB information of the prospective customers and the historical information of the customers, check issuers, customer representatives and branches kept in the database. Since the leads coming to the institution are in two different types of organizations (Individual and Legal), two different forecasting models were applied. Multiple classification models were tried, and the highest F1-Score of 86% for private companies was obtained with the Random Forest model, and the highest F1-Score for commercial companies was obtained with the Random Forest model with 82%.**

Keywords — Factoring, Transaction Forecast , Machine Learning

I. GİRİŞ

Faktöring işletmeleri, varlığını ve devamlılığını büyük ölçüde çek işlemleriyle sürdürmektedir. İlk çek işlemleri söz konusu olduğunda ise bazı müşteri adayları işlem gerçekleştirmekte, bazıları ise işlem gerçekleştirmemektedir. Bundan ötürü işlem gerçekleştirecek kişileri ve gerçekleştirmeyecek kişileri tespit edebilmek, işletmelerin farklı aksiyonlar alabilmesine imkân tanıyacaktır. Bu işletmelerin gelir kaynağı olan çek konusunda ileri istatistiksel metotlar ve yapay zekâ kullanılarak müşteri adaylarını sınıflandırabilmek, işletmenin cirosuna ve dolayısıyla kârına doğrudan katkı yapılmasına imkân tanıyacaktır.

Klasik koşullu yaklaşımlardan ziyade yapay zeka metotları kullanılarak yapılan bu çalışma sayesinde faktöring ve finans sektöründe yapay zeka yaklaşımlarının artırılması amaçlanmaktadır. Ayrıca, araştırmada finans sektöründe kullanılabilecek istatistiksel ve makine öğrenmesi tabanlı metotların birçoğu kullanılmıştır. Böylece, akademik dünyada oldukça kısır bir yer kaplayan finans ve faktöring alanındaki çalışmalara bir örnek niteliği taşıyabilecek yapay zeka tabanlı yaklaşımlar ile problem çözüme kavuşturulmuştur. Bilim

978-1-6654-7010-0/22/$31.00 ©2022 IEEE

dünyasına katkısı açısından değerlendirildiğinde faktöring sektörleri içerisinde yenilikçi yaklaşım amacı taşımaktadır. Bu doğrultuda projenin nihai çıktısı, sektör bazında göz önünde bulundurulduğunda, yapay zekâ metodolojisinin kullanıldığı nadir proje araştırmalarından birini kapsamaktadır.

II. LİTERATÜR ÖZETİ

Günümüzdeki işletmelerin en önemli yapı taşlarının başında gelen ve yoğun rekabet koşullarında varlıklarını sürdürmesini sağlayabilecek temel etmen veri odaklı yaklaşımlardır. Birçok farklı alanda kitleler halinde elde edilen veriler işlenerek bilgiye dönüştürülmekte ve ortaya çıkarılan bu bilgiler üzerinden karar verme şansı sağlamaktadır [1]. Bu bilgiler ışığında mekanizmayı oluşturup, hammaddeden ürüne kadar olan süreci doğru yapılandırarak, katma değer sağlayan süreçlerin oluşması muhtemeldir. Bu çerçevede yeni nesil analitik yöntemlerin en temelinde veriye olan yaklaşımlar açısından değerlendirildiğinde sistemler arasında farklılaşma görülmektedir. Dolayısıyla, yeni girilen veriye dayalı yaklaşım büyük bir paradigma ve algı değişikliğine neden olmuştur [2]. Bu noktada özellikle ön plana çıkan yöntemlerin başında büyük veri analitiğinin dışında makine öğrenimi ve yapay zekâ bazlı metotlar gelmektedir. Bu çerçevede, işletmeden işletmeye (B2B) pazarlama süreçlerinde yapılan hesaplamalar ve değerlendirmeler konusunda farklı yöntemler gündeme gelmektedir. B2B süreçlerini de kapsayacak şekilde yapılacak bir hizmet yapısı birçok farklı alanı kapsayacağından dolayı çok eksenli bir yaklaşımı gerektirmektedir [3]. Büyük veri analitiği çerçevesindeki bu çok disiplinli bakış metodolojisi içinde gerçekleşmekte olan bu dönüşüm süreci araştırma çerçevesi ve yaklaşımlarla birlikte pratikte de birçok yeniliği ve fırsatı beraberinde getirmiştir [4]. Diğer bir açıdan değerlendirildiğinde B2B süreçlerde işletmelere daha uygun pazarlama süreçlerini yapay zekâ teknolojilerini büyük veri analitiği, veri madenciliği, veri bilimi çerçevesinde sağlayarak daha uygun pazarlama stratejilerinin belirlenmesi ve daha doğru kararların verilebilmesi öncelikli hedeftir [5].

Bu yaklaşım, işletmelerin çabalarını gerçekten kaybetme riski altında olan müşterilere odaklamasına olanak tanır ve potansiyel olarak onlara ihtiyacı olmayan müşterilere teşvik sağlamak için harcanacak paradan tasarruf sağlar [6]. Ayrıca ekonomi perspektifinden bakıldığında da bu konuda yapılan tüm bilimsel çalışmalar gösteriyor ki, var olan müşterileri elde tutabilmenin maliyeti, yeni müşteri kazanmaya nazaran çok daha az maliyetlidir [7]. Bu noktada projenin çıktısıyla birlikte maliyet azaltımına olanak tanınacaktır. Bir diğer nokta ise müşteri adayının, müşteriye dönüşme serüveninin başarıyla tamamlanmasıyla birlikte, daha sonraki işlemler için; o kişilerin işlem yapma ihtimalleri daha da artırılmış olacaktır. Bu yaklaşım, işletmelerin çabalarını gerçekten kaybetme riski altında olan müşterilere odaklanmasına olanak tanır ve potansiyel olarak onlara ihtiyacı olmayan müşterilere teşvik sağlamak için harcanacak paradan tasarruf sağlar [8].

PROBLEMİN TANIMLANMASI

Projenin uygulanacağı finans şirketine ayda ortalama binin üzerinde tekil şirket, çek bozdurma işlemi yaptırmak için girişimde bulunmaktadır. Bu sayı senelik bilançoda 400 bin seviyesindedir. Bu şirket ile ilk kez temas eden müşteri adaylarının üçte biri farklı sebeplerle işlem onayı almasına karşın işlem yapmamaktadır. Bu tercihin fiyat, hizmet, süre vb. gibi farklı gerekçeleri olabilmekle birlikte yapılan saha çalışmaları neticesinde fiyat ve hizmetin önemli göstergeler olduğu görüldü. Bu çalışma kapsamında işletmeyle ilk kez temasta bulunan müşteriler arasında potansiyeli yüksek olanların veri güdümlü modelleme yaklaşımlarıyla saptanarak daha uygun bir teklif sunulması, ilgili müşteri adaylarına özel arama yapılması, bu aramanın daha kıdemli bir müşteri temsilcisi tarafından yapılması gibi farklı pazarlama stratejileri kullanılarak işlem oranının doğrudan arttırılması hedeflendi. Dolayısıyla, başvuruları içeriye alma oranını artırarak şirketin hem kısa vadeli kârlılığına doğrudan katkı sağlanması hem de orta vadede potansiyeli yüksek olduğu saptanan müşterilerle iş yapma olasılığını artırmaktır.

Eski olarak nitelendirdiğimiz, daha önce işlem yapmış olan müşterilerin senelik ortalamada %47,2'si onaylanan başvurularını işleme çevirirken daha önce işlem yapmamış müşteri adaylarından onay almış olmalarına yine senelik ortalamada %28,9'u karşın bu işlemleri gerçekleştirmektedir. İşlem yapmayı tercih etmeyen müşteri adaylarının fiyat, süre, hizmet, sistemsel sorun, başka faktöring firmasını tercih etme ve diğer nedenlerle şirket ile işlem yapmama sebepleri olarak kaydedilmiştir. Dolayısıyla, şirketin hedeflemesi gereken en yüksek potansiyele sahip müşteri adaylarını tespit edilmesidir. Bu temasta bulunan müşteri adaylarının başvurularının işleme dönüştürülmesi neticesinde gerçekçi senaryoda bile (iyimser ve kötümser senaryoların ortalaması olarak düşünüldüğünde) günlük 15 bin TL'den fazla net kâr, senelik bilançoda da yaklaşık 4 milyon TL net kar artış potansiyeli bulunmaktadır. Bu net kâr artışının sağlanması, Faktöring sektörü özelinde düşünüldüğünde yüksek hacimli cirolara karşılık gelmektedir. Dolayısıyla, şirketin senelik ciro ve kâr marjını artırmak için kullanması uygun olan stratejilerden biri bu ilk kez temas eden müşteri adaylarına odaklanarak ilk kez işlem yapmalarını sağlamaktır. Bundan ötürü, bu proje hedefinin başarıyla tamamlanmasının sağlanması neticesinde ölçülebilir bir sonuç ortaya koyması beklenmektedir.

İlgili akademik literatürde de sıklıkla bahsedildiği üzere müşterilerin sınıflandırılabilmesi için gözlemlenen en büyük sorun hangi müşteri özelliklerinin kullanılması ve ölçümlenmesinin en doğru ve en etkili sonuçları ortaya çıkaracağıdır. Bundan dolayı bu çalışma kapsamında faktöring bazında değerlendirilecek ve kullanılacak tüm müşteri verilerinin incelenmesi yapılacaktır. Bu projenin hedefi olan yüksek potansiyelli kişileri bulmak ve hangi özellikleri taşıdığını, hangi müşteri profili özellikleri dağılımları gösterdiklerinin yanı sıra düşük potansiyelli

müşteriler için de bu etmenler ölçülebilecektir. Bu sınıflandırmaların başarıyla tamamlanması neticesinde bu proje "Doğru kişiye, doğru aksiyon" metodolojisini uygulayabilme yetisini kazandıracaktır [9]. Christopher ve arkadaşlarının yaptıkları bir diğer bilimsel araştırmalar da gösteriyor ki "Herkese uyan tek beden!" yaklaşımı günümüzdeki koşullar göz önünde bulundurulduğunda geçerliliğini yitirmiştir [10].

Bu çalışmanın çıktısı olarak geliştirilecek modelin özellikle yüksek potansiyelli müşterileri (işlem kapasitesi ve sıklığı kapsamında) doğrudan ilk dokunuşlarındaki ("first touch") özelliklerinden yola çıkılarak müşterinin potansiyeli hakkında kestirimde bulunacak bir tahmin modeli geliştirilecektir. İkinci aşamada bu modelin adaptif hale getirilerek yeni veri kayıt ve işlem süreciyle birlikte bu süreçlerin otomatik şekilde işleme alınması ve sonuçlandırılarak ilgili birimlere gerekli bilgilendirmenin ulaşması sağlanmaktadır. Yukarıda da belirtildiği gibi eski müşterilerin %50 kadarı, yeni müşterilerin

%30 kadarı işlem aşamasına geçmektedir. Bu Ar-Ge projesinin beklenen temel faydası ilk defa işlem yapan müşterilerdeki geliştirilecek modelin yönlendirdiği müşterilerdeki işlem sayı ve oranının artırılmasıdır.

III. YÖNTEM

A. *Veri setinin hazırlanması*

2019 yılının Mart ayından 2022 yılının Temmuz ayına kadar olan sadece ilk işlemini gerçekleştiren müşterilerin verileri kullanıldı. Faktöring işlemi gerçekleştiren müşteriler iki farklı grup üzerinde toplanmaktadır. Bunlar: Tüzel (Ticari) şirketler ve Şahıs (Gerçek) şirketlerdir. Veri seti Şahıs şirketleri ve Tüzel şirketlere ait özelliklerin bazılarının ortak olmasına karşın, çoğu değişken farklılık gösterdiğinden iki ayrı gruba bölünerek farklı stratejik yaklaşımlarda bulunuldu. Bu bölümlemeden sonra ana veri seti ortaya çıkarıldı. Şahıs şirketleri için 36.973 adet gözlem ve 71 adet öznitelik bulunmaktadır. Tüzel şirketler için 26.093 adet gözlem ve 60 adet öznitelik bulunmaktadır. Son olarak hedef değişken için belirtilen tarihler arasında ilk sorgulamasını gerçekleştiren müşterilerin 3 ay içerisinde işlem yapmalarına göre 1 ve 0 olmak üzere 2 aynı sınıf belirlenmiştir.

B. *Veri Analizi ve Önişleme*

Daha sonra bu iki veri seti üzerinde öznitelik analizleri gerçekleştirildi ve öznitelik tipleri tanımlandı. Bunlar sayısal değerler, kategorik değerler, tarih hem kategorik hem sayısal bilgiler taşıyan karışık değerler olarak ifade edilebilir. Sayısal değerler içerenler ayrık ve sürekli değerlere sahiptirler. Kategorik değişkenler ordinal (sıralı) ve nominal (sırasız) öznitelikler içermektedir. Bazı özniteliklerin veri tipleri gözlenenden farklı olduğu tespit edildi. Bu özniteliklerin veri tipleri uygun şekilde dönüştürülmesi gerçekleştirildi. Eksik veriler, bir çalışmanın istatistiksel gücünü azaltabilir ve geçersiz sonuçlara yol açan önyargılı tahminler üretebilir [11]. Betimleyici istatistikler ile verilerdeki boş değerler tespit edildi ve ön analiz sonrasında verilerin kaybolması mı yoksa o durum için mi bulunmadığı tespit edilerek, durumuna göre farklı atamalar yapıldı [12].

ÇİZELGE I. EKSİK VERİ DURUM ANALİZ YÖNTEMLERİ

Eksik Veri Durum Analiz Yöntemleri	
Tamamen Rastsal Kayıp (MCAR)	Eksik verilerin tamamen şans eseri oluşması
Rastsal Olmayan Kayıp (MNAR)	Eksik verilerin bir başka değişkene bağlı olarak oluşması
Rastsal Kayıp (MAR)	Eksik verilerin rastgele oluşması

XGBoost ve LightGBM modelleri boş değerleri ele alabilirken diğer modeller için bu durum geçerli değildir. Boş verilerin birden çok farklı bakış açısına ve temeline dayanan yöntemlerle inceleme ve uygulaması yapıldı. %80'nin üstünde boş veri içeren değişkenler veri setinden çıkarıldı. Aynı şekilde kişi bazında %60'ın üstünde boş verisi olan satırlar da veri setinden çıkarıldı. Kalan diğer değişkenlerin boş verileri için çok değerli (multivariate) ve tek değerli (univariate) doldurma yöntemleri kullanıldı. Elde edilen sonuçlar doğrultusunda farklı öznitelikler için birden çok yönteme başvuruldu. Bu amaç doğrultusunda; ortalama atama, k-En Yakın Komşular yöntemi ile doldurma ve kategorik değişkenler için sabit değer atama yöntemleri kullanıldı.

ÇİZELGE II. EKSİK VERİ DOLDURMA YÖNTEMLERİ

Eksik Veri Doldurma Yöntemleri	
Ortalama Değerler ile Doldurma Yöntemi	Eksik gözlemlerin bulunduğu özniteliğin ortalama ile doldurulması
k-EnYakın Komşular ile Doldurma Yöntemi (kNN)	Eksik gözlemlerin öklid mesafesine göre seçilecek komşu sayısına göre doldurulması
Sabit Değer ile Doldurma Yöntemi	Eksik gözlemlerin sabit bir değer ile doldurulması

Eksik veri analizi sonrasında kategorik değişkenlerin analizi gerçekleştirildi. Yüksek kardinaliteye sahip olan kategorik değişkenler içerisindeki az sayıda veriye sahip olan sınıflar hedef değişkene göre analiz edilip uygun sınıflar ile birleştirildi. Yüksek kardinalite durumunu analiz edip, çözümlenmesi model performansı açısından olumlu etki yaratacaktır [13]. Çoğu makine öğrenmesi algoritmaları kategorik değişkenleri ele alamamasından dolayı kategorik değişkenler çözümlenerek (encode) sayısal verilere dönüştürüldü. İlk olarak ikili (binary) hedef değişkene sahip olduğumuzdan dolayı genel olarak finans sektöründe yaygın kullanılan kanıtların ağırlığı yöntemi (WOE) uygulandı [14]. Bu yönteme ek olarak hedef değişkene göre kodlama (Target Encoding), İkili kodlama (Binary Encoding) ve One-Hot kodlaması yöntemlerine başvuruldu. Kodlamalar sonucunda elde edilen değerlerin, hedef değişken ile olan ilişkisi analiz edildiğinde en iyi sonucu hedef değişkene göre kodlama (Target Encoding) yöntemi verdiği görüldü.

Makine öğrenmesi modellerinin genel olarak kabul ettiği varsayımlar doğrusallık (lineerlik), çoklu bağlantı (multicollinearity), normal dağılım ve homoskedastisitedir.

Bu durumların asgari düzeyde karşılanması durumu kontrol edildi. Verilerin normal dağılıma yakınlığını kontrol etmek için histogram grafiği ve Kolmogorov-Smirnov testi kullanıldı [15]. Homoskedastisite için artıkların (residual) görselleştirilmesi yöntemi uygulandı. Bunlara ek olarak Levene's testi, Bartlett's testi ve Goldfeld-Quandt testi kullanıldı. Çoklu bağlantı durumu için ise korelasyon matrisi kullanılarak değişkenler arasındaki ikili ilişkiler kontrol edildi. Değişken dağılımlarının analiz edilmesi için olasılıksal dağılım yöntemlerine başvuruldu. Ayrık ve sürekli değişkenler için farklı metodlar kullanıldı. Ayrık değişkenler için Binom ve Poisson dağılımları ile test edilirken sürekli değişkenler için Gauss dağılımı ve çarpıklık testi gerçekleştirildi. Gerçekleştirilen testler sonucunda değişkenleri normalizasyonları için logaritmik dönüşüm, reciprocal dönüşüm, karekök dönüşümü ve yeo-johnson dönüşümü yöntemleri kullanıldı. Normalleştirmeler sonrasında verilerin ölçeklendirilmesi gerçekleştirildi. Ölçeklendirme için standartlaştırma, en büyük-en küçük ölçeklendirme (min-max scaling), ve güçlü (robust) ölçeklendirme yöntemleri denendi. Analizler sonucunda en büyük-en küçük ölçeklendirme yöntemine karar verildi ve uygulandı. Veri dağılımları analizi sonrasında değişkenlerin dağılımını normalliğe yaklaştırmak için aykırı gözlemler analizi edildi. Aykırı gözlemlerin ele alınması model performansı açısından doğrudan pozitif yönde etki yaratacaktır [16].

C. *Öznitelik Mühendisliği*

Veri setindeki analizler ve ön işleme aşamaları sonrasında elde edilen verilere ek olarak eldeki veriler ile yeni veriler üretildi. İlk olarak işlem gerçekleştirmek isteyen müşterilerin bazıları birden fazla çek getirebilmektedir. Bir kişiye ait birden çok gözlemin bulunması yanlılığa sebep olacağından dolayı veriler işlem bazında tekilleştirildi. Hedef değişken üzerinden toplulaştırma ("aggregation") yapıldı ve sayısal değerler en büyük, en küçük, ortalama ve standart sapma değerleri kullanıldı. Kategorik değişkenler için ise frekans analizi yapıldı. Toplulaştırma sonrasında özellikle tarih değişkenleri kullanılarak müşterilerin çek getirme tarihleri için analizler yapıldı ve tarih değerleri yıl, ay, gün, çeyrek, hafta sayıları gibi değerlere dönüştürüldü ve kuruluş tarihleri üzerinden kuruluş yaşları hesaplandı.

D. *Model Geliştirme*

Model geliştirme aşamasına geçildiğinde öncelikle hedef değişkenin kategorik olması sebebiyle gözetimli makine öğrenmesi yöntemleri seçildi. Modelleme için literatürde genel olarak kullanılan birden çok sınıflandırma yöntemi kullanılarak model sonuçları karşılaştırıldı. Lojistik Regresyon, Karar Ağaçları, Rastgele Ormanlar, XGBoost, LightGBM, Ekstra ağaçlar ve Destek Vektör Makineleri sınıflandırma modelleri kuruldu. Veri setindeki hedef değişken sınıf sayıları arasında eşitlik olmaması sebebiyle dengesiz ("imbalanced") bir veri setinin mevcut olduğu görüldü. Veri setinin %61'ini işlem yapmayan müşteriler oluştururken %39'unu ise işlem yapan müşteriler oluşmuştur. Bu oran her iki veri seti hemen hemen eşittir. Birkaç örnek kullanan önceki çalışmada, dengesizliğin doğruluğun değeri ve anlamı ile diğer bazı iyi bilinen performans ölçütleri üzerinde büyük bir etki gösterebileceği gösterilmiştir [17]. Dengesiz veri seti problemini ortadan kaldırmak için birden fazla yol denendi. Bu yollar, örnek azaltımı (undersampling), örnek artırımı (oversampling) ve sınıf ağırlığı ayarlama yöntemleridir. Denenen yöntemler sonucunda örnek azaltımı yöntemi iyi sonuçlar vermezken sınıf ağırlığı ayarlama ve örnek artırımı yöntemlerinin ise doğruluk üzerinde büyük bir etkisi olmamıştır. Bundan dolayı modeller ana veri seti üzerinden geliştirilip ilgili çıktılar elde edilmiştir.

ÇİZELGE III. DENGESİZ VERİ SETİ ÇÖZÜM YÖNTEMLERİ

Yöntem	Açıklama
Örnek Azaltma (Undersampling)	Sınıf sayısı fazla olan değerlerden rastgele değerler çıkararak dengesizliği ortadan kaldırma
Örnek Artırma (Oversampling)	Az olan sınıfa ait yeni değerler üreterek hedef değişkendeki dengesizliği ortadan kaldırma
Sınıf Ağırlığı Artırma	Makine öğrenmesi modellerinin bir çoğunda 'class_weight' parametresi ile azınlık sınıfa verilen ağırlığı dengesizlik oranında artırma

Model kurulumu sonucunda doğru ve güvenilir çıktıların sağlanabilmesi adına çapraz doğrulama (cross- validation) yapıldı. Çapraz doğrulama, tahmine dayalı modellerin genelleme yeteneğini değerlendirmek ve fazla uydurmayı önlemek için bir veri yeniden örnekleme yöntemidir [18,19].Çapraz doğrulama yöntemi için verilerin %80'i eğitim ve %20'si test olarak bölündü. Veri seti dengesiz olması dolayısıyla tabakalı örnekleme (stratified sampling) kullanılmış olup, çapraz doğrulama değeri 5 olarak seçildi. Veri setinin dengesiz olmasından ötürü model sonuçlarını analiz etmek için direk doğruluk oranı kullanılmadı. Model doğruluklarının analizi için Karmaşıklık Matrisi, Dengelenmiş Doğruluk, Geometrik Ortalama, Dominans, Dengesizlik Indeksi, ROC-AUC Eğrisi, Kesinlik-Hassaslık (Precision-Recall) eğrisi kullanıldı. Özellikle ROC-AUC eğrisi ve Kesinlik-Hassaslık eğrileri eşik değerini ("Threshold") ayarlamak için kullanıldı. Eşik değeri ayarlanarak sınıflar arasındaki doğruluk farkı azaltıldı. Modelleme ölçümleri sonrasında en iyi sonucu veren modellerin için yarılanma ızgarası araması ("Halving Grid Search") ile hiper parametre ayarlaması (optimizasyonu) Destek Vektör Makineleri, Lojistik Regresyon, Rastgele Ormanlar ve XGBoost için yapıldı.

IV. SONUÇ VE TARTIŞMA

Şahıs şirket müşterileri için en iyi sonucu Destek Vektör Makineleri (SVM) verirken Ticari şirket müşterileri için en iyi sonucu XGBoost sınıflandırıcısı vermiştir. Elde edilen skorların gerçek müşteriler için ortalama F1-Skoru değerleri %79 olurken, ticari müşteriler için elde edilen ortalama F1-Skoru %75 seviyesi üzerinde olmuştur. Baskın hedef değişkeninin skor değerleri yüksek olduğundan dolayı ROC-AUC eğrisi kullanılarak Şahıs şirketler için eşik değeri 0.46 belirlenmiştir. Buna karşın ticari şirketlerin eşik değeri ise 0,38 olarak belirlendi. Çalışma sonucunda geliştirilen modeller, şirket bünyesinde günümüzden itibaren

kullanılmakta olup önümüzdeki 3 ay içinde sonuçlar kontrol edilip duruma göre tekrardan gerekli ayarlamalar yapılarak, yaşam döngüsü içinde devam edecektir. Bu proje kapsamında nihai hedef olarak yüksek potansiyelli müşteri adaylarının bulunabilmesinden ötürü duyarlılık skorlarının daha çok önem arz ettiği söylenebilir.

ÇİZELGE IV. ŞAHIS ŞİRKETLERİ İÇİN MODEL BAŞARI SKORLARI

MODEL	precision	recall	f1-score
LogisticRegression: 0	0,8023	0,7064	0,7513
LogisticRegression: 1	0,7217	0,8155	0,7658
DecisionTreeClassifier: 0	0,8287	0,7562	0,7908
DecisionTreeClassifier: 1	0,7579	0,8301	0,7924
RandomForestClassifier: 0	**0,88**	**0,8461**	**0,8627**
RandomForestClassifier: 1	**0,8332**	**0,8684**	**0,8505**
ExtraTreesClassifier: 0	0,8375	0,7380	0,7846
ExtraTreesClassifier: 1	0,7515	0,8486	0,7971
XGBClassifier: 0	**0,8628**	**0,8446**	**0,8536**
XGBClassifier: 1	**0,8263**	**0,8454**	**0,8357**
KNeighborsClassifier: 0	0,742118912	0,69736377	0,71905
KNeighborsClassifier: 1	0,687524155	0,73274115	0,70941
SVC: 0	0,8045	0,7707	0,7872
SVC: 1	0,7554	0,790247	0,77244

ÇİZELGE V. TİCARİ ŞİRKETLER İÇİN MODEL BAŞARI SKORLARI

MODEL	precision	recall	f1-score
LogisticRegression: 0	0,7703	0,6692	0,7162
LogisticRegression: 1	0,6897	0,7783	0,7313
DecisionTreeClassifier: 0	0,7967	0,719	0,7559
DecisionTreeClassifier: 1	0,7259	0,7929	0,7579
RandomForestClassifier: 0	**0,848**	**0,80**	**0,8280**
RandomForestClassifier: 1	**0,8012**	**0,8312**	**0,8159**
ExtraTreesClassifier: 0	0,8055	0,70	0,7495
ExtraTreesClassifier: 1	0,7195	0,8114	0,7627
XGBClassifier: 0	**0,8308**	**0,8074**	**0,8189**
XGBClassifier: 1	**0,7943**	**0,8082**	**0,80**
KNeighborsClassifier: 0	0,7101	0,6601	0,6842
KNeighborsClassifier: 1	0,6555	0,6955	0,6749
SVC: 0	0,7725	0,7334	0,7525
SVC: 1	0,7234	0,7530	0,7379

Duyarlılık skorları için, yani gerçekte yüksek potansiyelli olarak etiketlenenlerin ne kadarı için doğru tahminleme gerçekleştirildi. Bir diğeri Kesinlik skorları, yani yüksek potansiyelli müşteri adayları olarak tahminlemesi yapılanların, ne kadarı gerçekten yüksek potansiyelli olduğudur. Bu iki skorlara farklı bakış açılarıyla yaklaşmak, farklı aksiyonların alınabileceğini göstermektedir. Bu çalışma kapsamında nihai hedef olarak yüksek potansiyelli müşteri adaylarının bulunabilmesinden ötürü duyarlılık skorlarının daha ön plana çıktığı ifade edilebilir.

Çizelge IV'te görüldüğü üzere en yüksek başarı skorlarını Random Forest ve XGBClassifier vermiştir.

Kuruluş içinde aktif olarak kullanılan benzer bir model bulunmadığından şirket açısından değerlendirildiğinde oldukça yenilikçi bir etkiye ve geliştirilen model canlıya alındığından katma değeri yüksek bir çalışmaya sahip olmuştur. Henüz ilk temasta bulunmuş olan yüksek potansiyelli müşteri adayları doğru ve başarılı bir şekilde saptanabildiği takdirde pazarlama ekibini daha verimli kanallara yönlendirilmesi planlandı. Dolayısıyla, bu çalışma kapsamında gerekli geliştirmelerin yapılmasıyla birlikte şirket kârına doğrudan katkı sağlaması ön görülmektedir.

KAYNAKÇA

[1] A. Kusiak, (2009). Innovation: A data-driven approach. International Journal of Production Economics, 122(1), 440-448.

[2] S. Moro, P. Cortez,, & P. Rita,. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] A.J. Christy, A. Umamakeswari, L. Priyatharsini, & A. Neyaa (2021). RFM ranking–An effective approach to customer segmentation. Journal of King Saud University-Computer and Information Sciences, 33(10), 1251-1257.

[8] A. Lemmens, & C. Croux. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276-286.

[9] M.S. Kahreh, M. Tive, A. Babania & M. Hesan,(2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. Procedia-Social and Behavioral Sciences, 109, 590-594.

[10] M. Christopher, H. Peck,& D. Towill,(2006). A taxonomy for selecting global supply chain strategies. The International Journal of Logistics Management.

[11] J. Miao, & L. Niu, (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919–926. https://doi.org/10.1016/j.procs.2016.07.111

[12] S. Fielding, P.P. Fayers, A. McDonald, G. McPherson, & M. K. Campbell,(2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health and Quality of Life Outcomes, 6(1), 1-9.

[13] K. Potdar, T.S. Pardawala, & C.D. Pai,(2017). A comparative study of categorical variable encoding techniques for neural network classifiers. International journal of computer applications, 175(4), 7-9.

[14] D. L. Weed, (2005). Weight of evidence: a review of concept and methods. Risk Analysis: An International Journal, 25(6), 1545-1557.

[15] F.J Massey Jr, (1951). The Kolmogorov-Smirnov test for goodness offit. Journal of the American statistical Association, 46(253), 68-78.

[16] C.C. Aggarwal, (2017). An introduction to outlier analysis. In *Outlier analysis* (pp. 1-34). Springer, Cham.

[17] A. Luque, A. Carrasco, A. Martín, & A. de las Heras, (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231.

[18] T. Hastie, R. Tibshirani, & J. Friedman,(n.d.). Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition.

[19] R.O. Duda, P.E. Hart,& D. G.Stork, D. (2001). Pattern Classification