Müşteri Metrikleri üzerinden Segmentasyon ve Kayıp Tahmini

Customer Segmentation and Churn Prediction via Customer Metrics

Tunahan BOZKAN
Information Technology Graduate Program
MEF University
Istanbul, TURKEY
bozkant@mef.edu.tr

Tuna ÇAKAR
Computer Engineering Department
MEF University
Istanbul, TURKEY
cakart@mef.edu.tr

Alperen SAYAR
Information Technology Graduate Program
MEF University
Istanbul, TURKEY
sayara@mef.edu.tr

Seyit ERTUĞRUL
Department of Information Technology
MEF University
Istanbul, TURKEY
ertugruls@mef.edu.tr

2022 30th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-5092-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SIU55565.2022.9864781

Özetçe

Bu çalışmada faktoring sektöründe faaliyet gösteren müşterilerin geçmişte yapmış oldukları işlem hareketleri ve sahip oldukları risk, limit ve şirket verileri üzerinden, son işlem tarihlerinden sonra gelecek üç ay içerisinde işlem yapmaya devam edip etmemelerini veri güdümlü makine öğrenimi modelleri kullanarak tahmin edilmesi amaçlandı. Kurulan modeller sonucunda iki farklı müşteri grubunun (Gerçek ve Tüzel şirket) Kayıp Analizi (Churn) gerçekleştirildi. XGBoost modeli ile %74 ve %77 oranında F1-Skoru ile tahmin edildi. Bu modelleme sayesinde ayrılacak olan müşterilerin tahminlemesi ile birlikte bu müşteri gruplarına yapılacak özel promosyonlar, kampanyalar sayesinde müşterileri elde tutma oranının artırılması amaçlandı. Elde tutma oranlarının artması sayesinde şirket bazında işlem hacmine doğrudan katkı yapılması sağlandı.

Anahtar Kelimeler: Faktoring, Kayıp Analizi, Makine Öğrenmesi

Abstract:

In this study, it is aimed to predict whether customers operating in the factoring sector will continue to trade in the next three months after the last transaction date, using data- driven machine learning models, based on their past transaction movements and their risk, limit and company data. As a result of the models established, Loss Analysis (Churn) of two different customer groups (Real and Legal factory) wascarried out. It was estimated by the XGBoost model with anF1 Score of 74% and 77%. Thanks to this modeling, it was aimed to increase the retention rate of customers through special promotions and campaigns to be made to these customer groups, together with the prediction of the customerswho will leave. Thanks to the increase in retention rates, a direct contribution to the transaction volume on a company basis was ensured.

Keywords: Factoring, Churn Analysis, Machine Learning

I. TEORİK ARKA PLAN

Hizmet sektöründe faaliyet gösteren işletmeler için veri tabanlı uygulamaların büyük çoğunluğu müşteri metriklerini kullanarak segmentasyon çalışması yapmayı hedeflemektedir. Doğru şekilde segmente edilen müşteriler için doğru aksiyonlar alınarak eldeki müşteriden sağlanan faydayı en üst düzeye çıkarmak ya da onu elde tutabilmeyi başarabilmek işletmeye doğrudan katkı sağlayacaktır. Ayrıca bilinmektedir ki var olan müşterileri elde tutabilmenin maliyeti, yeni müşteri kazanmaya nazaran çok daha az maliyetlidir [1]. Araştırmacıların pazarlama performansını ölçmesi ve ölçüm sonucunda gerekli önlemleri alması işletmelerin varlığını sürdürmesinde hayati bir öneme sahiptir. Literatürde pazarlama stratejisi olarak sıkça bahsedilen ve başarısı kanıtlanmış müşteri metrikleri ile, uzman görüşünden faydalanılarak oluşturulan müşteri metriklerinin birlikte değerlendirilerek segmentasyon çalışması yapmak ve çıkan sonuçlara göre aksiyonlar almak işletmelerin doğrudan büyümesine olanak tanıyacaktır [2]. Tek iterasyonlu müşteri metriklerinin, müşteri bağlılığına etkisi günümüzün öne çıkan ve araştırılan konuları arasında yerini aldı [3]. Müşteri metrikleri ile

978-1-6654-5092-8/22/$31.00 ©2022 IEEE

kârlılık arasındaki ilişkileri anlama ihtiyacının önemi günümüzün son yıllarında artarak, firmaların pazarlama harcamalarını takip etmek ve doğrulamak için önem arz etmektedir [4]. Müşteri segmentasyonun nihai hedefinde, müşteri grupları arasındaki en iyi ayrıştırmayı sağlayan kümeleri bulmaktır [5]. Kendi bulunduğu grup içerisinde homojen özellikleri barındıran, kendi dışındaki gruplardan ise olabildiğince heterojen özellikler içeren kümelere bölmek şirketlere "doğru kişiye, doğru aksiyon" metodolojisini uygulayabilme yetisini kazandıracaktır [6]. Ayrıca segmentasyon yapılırken kullanılacak olan 'katkı' değişkenin etkisinin yüksek düzeyde olması beklenmektedir. Müşteri karlılığını segmentasyona dahil etmek, stratejileri, bir kuruluşun pazarlama programlarının etkinliğini ve verimliliğini geliştirmesini sağlar [7]. Çalışmanın katkısı göz önünde bulundurulduğunda yapılan literatür araştırmaları neticesinde görülmektedir ki müşteri metrikleri baz alınarak, yapay zeka temelli incelemeler içerisinde öncü rolüne sahiptir. Bu doğrultuda işletmelere farklı bakış açısı kazandırılması hedeflendi. Faktoring sektörü açısından bakıldığında ise uzman görüşlerinden faydalanılarak üretilen müşteri metriklerinin kendi içinde özgün ve eşsiz bir model girdisi olduğu görülmektedir.

Şekil 1. Akış Diyagramı

II. YÖNTEM

A. *Veri Setinin Hazırlanması*

Veri setinin hazırlanması aşamasında Tam Finans şirketinde 2019 yılının Mart ayından 2021 yılının Aralık ayına kadar olan veriler kullanıldı. Bu veri seti bütün işlem hareketleri dahil olmak üzere 3.009.763 adet gözlem ve ham olarak 173 adet öznitelik içermektedir. Proje çıktılarına uygunluk açısından elde edilen verilerin tümünü kullanmak için müşterilerin farklı tarihlerdeki tüm işlemleri alınarak tekilleştirme müşteri bazında yapılmadı. Kayıp analizi (Churn) gerçekleştirmek için farklı iş birimleri ile görüşerek müşterinin ne kadar sürede işlem yapılmaması sonrasında kaybedildiği tartışılarak geçerli süre 3 ay olarak belirlendi. Yani son işleminden itibaren 3 ay içinde işlem gerçekleştirmeyen müşteri kayıp olarak saptandı. Alınan karar doğrultusunda son işleminden sonra 3 ay içinde işlem yapmayan müşterilerin hedef değişkeni 1, en az 1 işlem yapan müşterilerin hedef değişkeni ise 0 olmak üzere iki ayrı sınıf etiketi atıldı.

B. *Veri Analizi ve Önişleme*

Faktoring işlemlerinde, işlem öncesi sorgulama ve fiyat araştırması yaygın olduğundan dolayı aynı müşteriler aynı çekleri sorgulattıkları günden bir süre sonra getirip işleme dönüştürmektedir. Bu sebepten ötürü veri setinde işlem bazında olmasa da sorgulama bazında gözlem çoklaması durumu mevcuttur. Bu veri çoklaması durumu için veri setinde filtrelemeler yapıldı. Yapılan filtreleme ise aynı müşteri eğer ki aynı çeki getiriyorsa en yakın tarihli işlem ele alındı. Yapılan bu filtreleme sonrasında gözlem sayısı 2.409.980'e düşürüldü. Bu duruma ek olarak genel analizlerin öncesinde faktoring işlemlerindeki müşteriler 2 farklı gruba ayrılmaktadır. Bunlar: Gerçek (Şahıs) şirketleri ve Ticari (Tüzel) şirketlerdir. Bu iki farklı müşteri grubuna ait öznitelikler kendi aralarında kıyaslama ve hesaplama ölçütleri açısından farklılık göstermekle beraber Gerçek şirketine sahip müşterilerin sahip olduğu bazı öznitelikler Ticari şirket müşterilerini kapsamamaktadır. Bu sebeplerden ötürü veri seti iki farklı gruba bölündü. Gerçek şirkete sahip müşteriler için gözlem sayısı 1.128.474 olurken Tüzel şirkete sahip müşterilerin gözlem sayısı 1.281.506 adettir. Ticari müşterilerin öznitelik sayısı ise 140'a düşürüldü. Uygun veri setlerinin elde edilmesinden sonra öznitelik tiplerinin kontrolü sağlandı. Veri setleri 4 farklı veri tipi içermekteydi. Makine öğrenimi modellerinin çoğunluğu eksik veri kabul etmemekte ve boş verilere karşı güçlü (robust) davranamamaktadır [8]. Boş verilerin üstesinden gelmek için birden çok yöntem bulunmakla beraber, bu yöntemlerden bazıları veri setleri üzerinde uygulandı. Öznitelik bazında %80 ve gözlem bazında %60 üzerinde boş veri var olması durumda bu öznitelikler veya gözlemler veri setinden çıkarıldı. Diğer boş değerler için ortalama atama, k-En Yakın Komşular (k-NN) ve sabit değer atama gibi birden çok yöntem kullanıldı. Kategorik veriye sahip öznitelikler sayısallaştırıldı. Sayısallaştırma aşamasında farklı yöntemler kullanıldı. Veri setlerimizdeki hedef değişkenin ikili (binary) olması sebebiyle finans alanında genel olarak kullanılan kanıtların ağırlığı yöntemi (WOE) uygulandı. Sonrasında hedef değişkene göre (Target Encoding), James-Stein yöntemi ve Hash yöntemi uygulandı. Kategorik değişkenlerin dönüştürülmesinde kullanılan yöntemlerin sonuçlarının analiz edilmesinden sonra hedef değişkene ait en yüksek ilişki ağını temsil eden yönteme başvuruldu. Çoklu bağlantı durumu için korelasyon matrisi kullanılarak birbirleri arasında %90'dan daha fazla ilişki bulunan özniteliklerden bir tanesi, ya çıkarıldı ya da birleştirme mümkün ise özniteliklerden yeni bir öznitelik elde edilerek kullanıldı [9]. Veri setlerindeki değerlerin normalliğe yakınlığını kontrol etmek için histogram grafiği ve Kolmogorov-Smirnov testi kullanıldı. Ayrık değişkenler için Binom dağılımı kullanılırken sürekli değişkenler için Gauss dağılımı ve Çarpıklık analizi gerçekleştirildi. Verilerin normalize etmek için logaritmik dönüşüm [10], karekök dönüşümü ve yeo-johnson dönüşümü yöntemleri kullanıldı [11]. Uygulanan normalizasyon sonrasında veri setlerinde bulunan farklı özniteliklerin aynı değer aralıklarına getirilmesi gerekmektedir. Bu ölçeklendirme hem kümeleme hem de akabinde gerçekleştirilecek modelleme için oldukça önemlidir. Ölçeklendirme işlemi için Standartlaştırma, Min-Max ölçeklendirme ve güçlü (robust ölçeklendirme) yöntemleri uygulandı [12]. Uygulanan yöntemlerin analizi sonrasında Min-Max ölçeklendirme yöntemi kullanıldı.

C. *Öznitelik Mühendisliği*

Veri ön işleme süreci sonrasında iş bilgisi, verilerin analizi ve literatür taramasıyla birlikte elimizde hazır bulunan veri

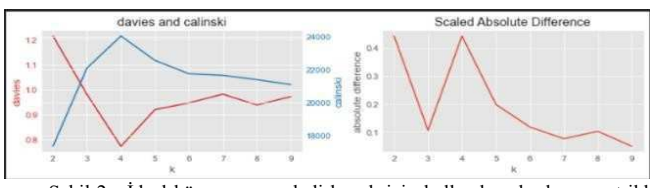
setine yeni öznitelikler eklendi. İlk olarak Pazarlama-Satış ve Analitik birim ile yapılan görüşmeler sonucunda ortak bir konsensüs kararı ile 10 farklı metriğin eklenmesi kararlaştırıldı. Müşterilerin genel olarak son işlem tarihinden önce geçmiş tarihli 1, 2, 3, 6, 9 ve 12 ay içerisindeki toplam sorgulama sayıları ve toplam işlem sayıları öznitelik olarak eklendi. Bu eklenen geçmiş tarihlerdeki işlem ve sorgulama sayıları için ağırlıklandırma yöntemine başvuruldu. En yüksek ağırlık katsayısına sahip öznitelik 1 ay olurken en azı 12 ay olarak baz alındı. Eklenen müşteri öznitelikleri sonrasında aynı öznitelik ekleme işlemleri, işlemleri gerçekleştiren müşteri temsilcileri ve çeklerin sahibi (keşideci) içinde uygulandı. İş bilgisi dahilinde çek ile işlem yapan müşterilerin bir çoğunun şirket yaşlarının 2'den küçük olduğu görüldü. Bunun için şirket yaşları 2'den büyük veya küçük olma durumu da değişken olarak eklendi. Son olarak modelleme aşaması öncesinde elde edilen metriklerin kümeleme sonuçları da veri setine yeni değişken olarak eklendi. Kümeleme sonucu eklenen veriler ile model kurulup daha sonrasında ise kümelemenin etkisinin görülmesi amacıyla kümeleme sonucu olmadan model kuruldu.

Metrikler
Müşterilerin en son ne zaman sorgulama işlemi yaptığı
Farklı günlerde toplam kaç kere sorgulatma ve işlem yaptığı
Son 1 yıl için kaç farklı günde sorgulatma ve işlem yaptığı
Yapılan sorgulamaların toplam kaç adedinin işleme döndüğü
Müşterilerin günlük 1000 TL başına katkısı
Müşterilerin onaylanan işlemlerinin kaç adedinin işleme döndüğü
Son 6 ay içerisindeki maksimum faktoring risk bilgisi
Müşterilerin toplam nakdi riskinin nakdi limite oranı
Müşterilerin son sorgulamadaki veya işlemindeki nakdi risk büyüklüğü
Müşterilerin getirdiği kaliteli çeklerin toplam çek sayısına oranı

Tablo 1. Modelleme sürecinden önce oluşturulan müşteri metrikleri

D. Model Geliştirme

Öznitelik mühendisliği sonucunda elde edilen ve bazı var olan değişkenlerin kümelenmesi için segmentasyon yapıldı. Kümeleme yöntemleri arasından öncelikle K-ortalama (K-means) yöntemi ile kümeleme yapıldı. K-ortalama yöntemi ölçeklendirmeye karşı duyarlıdır ve bu sebepten dolayı veriler kümele işlemi öncesinde farklı yöntemler ile ölçeklendirildi. Uygulanan yöntemler ise sırasıyla: Unscaled, Standard Scale, Min-Max Scaling, Max Abs, Robust Scaling, Yeo-Johnson, Gaussian-PDF, Uniform-PDF ve L2 normalization yöntemleridir. Yeo-Johnson uygulanırken lambda'nın ön tanımlı değeri olan 'None' olarak bırakıldı. Çünkü bu yöntemde en iyi lambda değerini bulmaktadır. Kümeleme sonuçlarını karşılaştırmak için 4 farklı ölçüm metriği belirlendi. Bunlar, Silhouette Score, Davies Bouldin Score, Calinski Harabasz Score ve Dirsek (Elbow) yöntemleridir.Yapılan segmentasyon çalışmaları sonrasında ideal küme sayısı 4 olarak seçildi ve en iyi ayrışma ise Min-Max ölçeklendirme yöntemi ile elde edildi.



Şekil 2. İdeal küme sayısını belirlemek için kullanılan skorlama metrikleri

Hedef değişkeninin sürekli olmaması sebebiyle gözetimli makine öğrenimi modelleri kullanıldı. Modelleme aşamasında birden çok model denendi. Kullanılan modeller: Destek Vektör Makineleri, XGBoost, Lojistik Regresyon, Rassal Ormanlar, Karar Ağaçları ve k-En Yakın Komşular'dır. Veri setleri üzerinde yapılan analizler sonucunda hedef değişkene bağlı olarak veri setinin dengesiz olduğu tespit edildi. Şahıs şirketine sahip müşteriler %41 oranında işlem yapmazken, %59 oranında işlem yapmaya devam edildiği tespit edildi. Ticari şirketlere sahip müşterilerde ise oran %45'e %55'tir. Yüksek oranda bir dengesizlik olmamasına rağmen dengesiz veri seti problemi çözümü için birkaç yol denendi. Bunlar örnek azaltımı, örnek artırımı ve sınıf ağırlığı dengelemesidir. Örnek azaltma yöntemi sonucunda model başarısı azalırken, örnek artırımı konusunda model başarısı %1 oranında arttırıldı. Sınıfların ağırlığı yöntemi ise bazı modellerde geçerliliği olmadığı için sadece test amaçlı uygulandı. Modellerin sonuçlarının güvenilir olması ve daha test edilebilir ve kullanılabilir olması için çapraz doğrulama (cross validation) yöntemine başvuruldu. Bu aşamada tüm gözlemlerin %70'i eğitim seti, %30'u ise test seti olarak kullanıldı. Çapraz doğrulama aşamasında çapraz doğrulama kat sayısı veri setinin büyüklüğüne bağlı olarak 5 seçildi.

III. SONUÇ VE TARTIŞMA

Hem Gerçek (Şahıs) şirketler için hem de Ticari (Tüzel) şirketleri için en iyi model sonucu XGBoost algoritmasının verdiği görüldü. Öznitelik sayısının çok olmasından ötürü Temel Bileşen Analizi (PCA) uygulanarak tüm modeller tekrar çalıştırıldı. Çalıştırma sonucunda F1-Skoru'na etkisi yok denecek kadar az olduğu görüldü. Kullanılan modellerin doğruluk metrikleri açısından şahıs şirketleri için ortalama F1-Skoru değeri %74 iken ticari müşterilere ait kurulan modelde ortalama F1-Skoru değeri %77 oldu. Şahıs şirketler için eşik değeri %45,7 iken ticari şirketler için eşik değeri %48,1 oldu.

Makine Öğrenmesi Modelleri	Sınıflar	Hassaslık(Precision)	Duyarlılık(Recall)	F1-Skoru
ExtraTrees Sınıflandırıcı	0	0.75	0.80	0.77
	1	0.67	0.61	0.64
Lojistik Regresyon	0	0.79	0.60	0.68
	1	0.57	0.76	0.65
Rassal Ormanlar	0	0.76	0.79	0.78
	1	0.68	0.65	0.66
XGBoost	0	0.80	0.76	**0.78**
	1	0.68	0.72	**0.70**

Tablo 2. Şahıs şirketlerine ait verilerin model skorları

Makine Öğrenmesi Modelleri	Sınıflar	Hassaslık(Precision)	Duyarlılık(Recall)	F1-Skoru
ExtraTrees Sınıflandırıcı	0	0.75	0.83	0.79
	1	0.76	0.66	0.71
K En Yakın Komşu Sınıflandırıcı	0	0.74	0.76	0.75
	1	0.70	0.69	0.69
Rassal Ormanlar	0	0.75	0.82	0.79
	1	0.76	0.67	0.71
XGBoost	0	0.78	0.81	**0.80**
	1	0.76	0.73	**0.74**

Tablo 3. Tüzel şirketlere ait verilerin model skorları

Model sonuçlarının karşılaştırılması ve doğruluklarının ölçümlenmesini analiz etmek etmek için farklı doğruluk metrikleri kullanıldı. Kullanılan metrikler: Karmaşıklık matrisi, Sınıflandırma raporudur. Karmaşıklık matrisi ve Sınıflandırma

raporu modellerin kesinlik, hassaslık ve F1 skorlarını karşılaştırmak ve analiz etmek için kullanıldı. Eşik değeri ölçümleme ile farklı sınıfların doğruluk oranları birbirlerine en yakın hale getirildi. Modelleme ölçümlerinin sonrasında en iyi sonucu veren model için ızgara araması yöntemi (grid search) ile hiper parametre optimizasyonu yapıldı. Izgara arama yönteminde parametreler belirlenirken her bir parametreye liste şeklinde birçok değer verilerek model başarısını en yüksekte tutan parametre değerleri kullanıldı.

Bu çalışmanın müşterilerin ve müşteri adaylarının daha doğru şekilde tanımlanması, bu kitleye daha doğru stratejilerle yaklaşım sağlanması hedeflendi. Aynı zamanda faktoring sektöründe özellikle kişilerin aldıkları hizmetlerden azami düzeyde faydalanmasının sağlanması, finansal verileri yeterli olmayan kişilerin alternatif yöntemlerle değerlendirilerek (makine öğrenim teknikleri gibi) şirketin sunduğu imkanlardan faydalanmalarının önünün açılması ile katma değeri yüksek bir çalışma olması ve dolayısıyla işletme hacmine doğrudan katkı yapması amaçlandı. Elde edilen bulgular ve oluşturulan modeller birkaç iterasyon sonrasında canlıya alınması planlanan modeller olması nedeniyle önemlidir.

IV. KAYNAKLAR

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

[2] Wiesel, T., Verhoef, P., & De Haan, E. (2012). There is no single best measure of your customers.

[3] Kahreh, M. S., Tive, M., Babania, A., & Hesan, M.3 (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.

[4] Gupta, S., & Zeithaml, V. (2006). Customer metrics and their impact on financial performance. *Marketing science*, *25*(6), 718-739.

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

[6] Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological bulletin*, *81*(6), 358.

[7] Niraj, R., Gupta, M., & Narasimhan, C. (2001). Customer profitability in a supply chain. Journal of marketing, 65(3), 1-16.

[8] Royston, P. (2004). Multiple imputation of missing values. Stata Journal, 4 (3), 227-241.

[9] Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, Sriram S. Modeling customer lifetime value. Journal of service research. 2006 Nov;9(2):139-55.

[10] Changyong, F. E. N. G., Hongyue, W. A. N. G., Naiji, L. U., Tian, C. H. *E.* N., Hua, H. E., & Ying, L. U. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, *26*(2), 105.

[11] Weisberg, S. (2001). Yeo-Johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, *1*, 2003.

[12] Martin, R. D., & Zamar, R. H. (1989). Asymptotically min—maX bias robust M-estimates of scale for positive random variables. Journal of the American Statistical Association, 84(406), 494-501.