Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Müşteri Kayıplarının Tahmini Üzerine Bir Veri Madenciliği Uygulaması

A Data Mining Application in Customer Churn Prediction

Mustafa Büyükkeçeci [1*], Mehmet Cudi Okur [2]

[1] Univerlist, İzmir, TÜRKİYE
[2] Yaşar Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği Bölümü, İzmir, TÜRKİYE
Sorumlu Yazar / Corresponding Author *: mustafa.buyukkececi@univerlist.com

Geliş Tarihi / Received: 12.11.2021 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 10.03.2022 *DOI:10.21205/deufmd.2022247218*
Atıf şekli/ How to cite: BÜYÜKKEÇECİ, M.,OKUR, M.C.(2022). Müşteri Kayıplarının Tahmini Üzerine Bir Veri Madenciliği Uygulaması. DEÜ FMD 24(72), 887-900.

Öz

Müşteri memnuniyeti ve sadakati uygun fiyat, ürün çeşitliliği, hızlı tedarik ve sevkiyat, ürün kalitesi, satış öncesi ve sonrası hizmetler ve müşteri davranışlarının analiz edilmesi ile sağlanır. Müşteri davranışlarını analiz eden işletmeler hem mevcut müşterilerini koruyabilir hem de yenilerini kazanabilir. Bu çalışmanın amacı işletmeleri terk etme ihtimali olan müşterileri tahmin edebilen gözetimli modeller üretmektir. Bu amaçla toplamda 21 sınıflandırma yöntemi ve telekomünikasyon, bankacılık ve e–ticaret sektörlerine ait veri kümeleri kullanılarak deney çalışmaları gerçekleştirilmiştir. Ayrıca işletmelerin harcama alışkanlıklarına göre müşterileri sıralamak ve sınıflandırmak için kullandıkları basit ama etkili bir pazarlama analiz aracı olan RFM (Recency, Frequency, Monetary Value) bölümlemesi, Ki-Kare Testi ile birlikte boyut indirgeme metodu olarak kullanılmıştır. Böylelikle optimal eleman sayısına sahip öznitelik altkümelerinin elde edilmesi ve öznitelik seçim öncesi ve sonrası model performanslarının kıyaslanması hedeflenmiştir.

Anahtar Kelimeler: *Veri Tabanlarında Bilgi Keşfi (VTBK), Veri Madenciliği, RFM Bölümlemesi, Öznitelik Seçimi, Boyut İndirgeme, Sınıflandırma*

Abstract

Customer satisfaction and loyalty can be achieved through reasonable prices, product variety, fast supply and delivery, product quality, pre and post-sales services and analysis of customer behaviors. Businesses that analyze customer behavior can both retain existing customers and gain new ones. This study aims to build supervised models that can predict customers who are likely to leave businesses. To this end, experiments were carried out using a total of 21 classification methods and datasets from the telecommunications, banking, and e-commerce industries. In addition, RFM (Recency, Frequency, Monetary Value) segmentation, a simple but effective marketing analysis tool used by businesses to rank and classify customers according to their spending habits, was used as a dimension reduction method together with the Chi-Square Test. Thus, it is aimed to obtain feature subsets with optimal number of elements and to compare model performances before and after feature selection.

Keywords: *Knowledge Discovery in Databases (KDD), Data Mining, RFM Segmentation, Feature Selection, Dimension Reduction, Classification*

1. Giriş

İşletmeler varlıklarını geçmişe göre çok daha rekabetçi bir ortamda sürdürmektedirler. Rekabet gücünü korumak için değişen koşulların, müşteri tutum ve davranışlarının iyi analiz edilip, yeni stratejilerin geliştirilmesi gerekmektedir. Bu sebeple işletmelerin çeşitli kaynaklardan elde ettikleri veriler her zamankinden daha değerli hale gelmiştir. Ancak ham ve değerli (anlamlı) olmayan bu veri yığını, Veri Tabanlarında Bilgi Keşfi (VTBK) sürecinden geçirilerek değerli hale getirilebilir. Değerli hale gelen veriler işletmelere büyük avantajlar sağlar. Bu avantajların başında müşterilerini daha iyi anlama fırsatı gelir. Müşterilerini daha iyi anlayan işletmeler de müşteri memnuniyeti ve sadakatini sağlamada üstünlük elde eder.

İşletmeler için yeni müşteriler elde etmek, mevcut olanları korumaktan daha maliyetlidir. Özellikle ekonominin sıkıntılı olduğu dönemler dışında müşterilerin işletmeyi terk etme eğilimleri bir iş analitiği aracı olan veri madenciliği ile tespit edilip, gerekli önlemler alınarak engellenebilir. İş zekâsı tanımlayıcı, iş analitiği ise tahmin edici çözümlemeye (analize) odaklanır. Bu sebeple iş zekâsı geçmişte veya günümüzde ne olduğu sorusuna cevap verirken, iş analitiği gelecekte ne olacağı sorusunu cevaplar [1].

Günümüzde işletmeler rekabet baskısını her zamankinden daha çok hissetmektedirler. Özellikle çok sayıda işletmenin faaliyet gösterdiği sektörlerde müşteriler hakkında detaylı bilgi sahibi olmak işletmelere ciddi bir rekabet avantajı sağlar. Veri madenciliği uygulamaları kullanılarak müşteri davranışları ve gereksinimleri tespit edilip, mevcut müşterilerin korunması ve yeni müşterilerin kazanılması için yeni satış ve pazarlama stratejileri geliştirilebilir. Bunlara ek olarak işletmeler satış tahmini, pazar araştırması ve risk analizi gibi amaçlar için de veri madenciliği uygulamalarından yararlanmaktadırlar.

Bu çalışmada müşteri kaybını tahmin etmeye yönelik yüksek doğruluğa sahip gözetimli modellerin oluşturulması hedeflenmiştir. Elde edilen modeller sayesinde işletmeler işletmeyi terk etme riski taşıyan müşterileri tespit edebilir, müşteri memnuniyetini ve sadakatini sağlamak için çeşitli pazarlama eylemlerini uygulayabilirler. Gerçekleştirilen deneylerde telekomünikasyon, bankacılık ve e–ticaret sektörlerine ait veri kümeleri ve modellerin oluşturulması için 6 farklı, toplamda ise 21 sınıflandırıcı kullanılmıştır. Ayrıca analiz kalitesini arttırmak için Bayes Hiper Parametre Optimizasyonu, Ki-Kare Testi ve RFM skoru boyut indirgeme yöntemi olarak kullanılmıştır. Model kalitesi yani sınıflandırma performansı doğruluk ve Eğri Altında Kalan Alan (AUC) metrikleri ile ölçülmüştür. Sonuçlar öznitelik seçiminden önce ve sonra olmak üzere iki ayrı tabloya aktarılmış ve oluşan performans farklılıkları kıyaslanmıştır.

2. Literatür Taraması

Bu başlık altında müşteri kayıp analizi için gözetimli makine öğrenmesi yöntemlerini kullanmış çalışmaların özetleri sunulmuştur. Patricia ve ark. [2] Yapay Sinir Ağları ve istatistiksel metotlar kullanarak müşteri tercihlerini tespit etmeye çalışmışlardır. Eiben ve ark. [3] 1998 yılında yaptıkları çalışmada Lojistik Regresyon, Genetik Programlama, Kaba Veri Analizi ve CHAID veri analizi tekniklerini pazarlama alanındaki üç veri kümesine uygulayıp kıyaslamışlardır. Madden ve ark. [4] Avustralya ISP pazarındaki abone kayıplarını tespit etmeye yönelik gerçekleştirdikleri çalışmada bir Regresyon Analiz yöntemi olan İki Terimli Probit Modelini kullanmışlardır. Datta ve ark. [5] cep telefonu abonelerinin davranış özelliklerini modellemek için CHAMP adını verdikleri bir sistem geliştirmişlerdir. Koçoğlu ve ark. [6] çalışmalarında müşteri kayıp analizi ile ilgili çalışmaların ve bu çalışmalarda kullanılmış olan veri madenciliği yöntemlerinin bir listesini sunmuşlardır.

Huang ve ark. [7] telekomünikasyon sektöründeki müşteri kayıplarının tahmini için Lojistik Regresyon, Doğrusal Sınıflandırıcı, Naïve Bayes, Karar Ağaçları, Çok Katmanlı Algılayıcı Sinir Ağları, Destek Vektör Makinesi ve Evrimsel Öğrenme sınıflandırıcılarını kullanmışlardır. Çalışmada ayrıca öznitelik çıkarımı yapılarak bir dizi yeni öznitelik tanımlanmış, orijinal ve yeni öznitelik kümesi ve 7 sınıflandırıcı çeşitli deneyler ile kıyaslamıştır. Xie ve ark. [8] çalışmalarında bankacılık sektöründeki müşteri kayıplarını tahmin etmek için IBRF (Improved Balanced Random Forests) adını verdikleri yeni bir yöntem sunmuş ve bu yöntemi Yapay Sinir Ağları, Karar Ağaçları ve Sınıf Ağırlıklı Çekirdek Destek Vektör Makinesi yöntemleri ile kıyaslamışlardır. Tsai ve Lu [9] çalışmalarında Geri Yayılımlı Yapay Sinir Ağları ve Kendi

Kendini Organize Eden Haritalar yöntemlerini kullanarak iki yeni melez model geliştirmişler ve modellerin sınıflandırma performansını değişik deney kurguları ile telekomünikasyon sektörüne ait üç test kümesi üzerinde değerlendirmişlerdir.

Vafeiadis ve ark. [10] deney çalışmalarında AdaBoost.M1 güçlendirme[1] yönteminin sunduğu performans iyileşmesini tek gizli katmana sahip Geri Beslemeli Yapay Sinir Ağları, Karar Ağaçları, Destek Vektör Makinesi, Naïve Bayes ve Lojistik Regresyon sınıflandırıcılar ve doğruluk, duyarlılık, keskinlik ve F–Skoru performans ölçüm metrikleri kullanarak analiz etmişlerdir. Burez ve Van Del Poel [11] çalışmalarında veri kümelerindeki sınıf dağılımı dengesizliği ile nasıl daha iyi başa çıkılabileceğini 6 adet gerçek dünya veri kümesi, 4 sınıflandırıcı ve 3 performans ölçüm metriği kullanarak incelemişlerdir. 1998 ve 2008 yılları arasında gerçekleştirilen çalışmaların detaylı özetinin bulunduğu Verbeke ve ark. [12] ait deney çalışmasında Karınca Kolonisi Optimizasyonu kullanan AntMiner+ ve ALBA kural çıkarım yöntemlerinin yanı sıra daha geleneksel yöntemlerden olan C4.5 ve RIPPER ile kural kümeleri oluşturmuşlardır. Xia ve Jin [13] Destek Vektör Makinesi, Geri Yayılımlı Yapay Sinir Ağları, Karar Ağacı C4.5, Lojistik Regresyon ve Naïve Bayes sınıflandırıcıları telekomünikasyon sektörüne ait veri kümeleri kullanarak karşılaştırmıştır. Verbeke ve ark. [14] 2014 yılında gerçekleştirdikleri çalışmada müşterileri kayıplarını tahmin etmek için sosyal ağ bilgilerini ve ilişkisel (ağ, çizge) sınıflandırıcıları kullanmışlardır.

Lu ve ark. [15] ait gerçek hayat veri kümelerinin kullanıldığı çalışmada Gentle AdaBoost algoritması müşterileri iki kümeye ayırmak, Lojistik Regresyon sınıflandırıcı ise her bir küme üzerinde bir müşteri kayıp tahmin modeli oluşturmak için kullanılmıştır. Caigny ve ark. [16] deney çalışmalarında LLM adını verdikleri ve Karar Ağacı ve Lojistik Regresyon kullanılarak oluşturulmuş melez bir sınıflandırıcı kullanmışlardır. Geliştirdikleri yöntemi Karar Ağaçları, Lojistik Regresyon, Rastgele Ormanlar ve Lojistik Model Ağaçları ile kıyaslayan yazarlar ayrıca 2011 ile 2017 yılları arasında gerçekleştirilmiş müşteri kayıp tahmini modellemesi ile ilgili çalışmaların bir listesini de sunmuşlardır. Khan ve ark. [17] ise İran'da hizmet veren bir internet servis sağlayıcı tarafından sağlanan veri kümesini kullanarak müşteri kayıp analizi gerçekleştirmişlerdir. De Bock ve Van den Poel [18] müşteri kayıp tahmini için 4 gerçek hayat veri kümesi ve iki farklı Rotasyon Tabanlı Sınıflandırıcı Topluluğu (Rotasyon Tabanlı Kolektif Öğrenme) yöntemi kullanmışlardır. Çalışmalarında ayrıca 2000 ile 2009 yılları arasında Sınıflandırıcı Topluluğu yöntemi kullanılarak gerçekleştirilmiş çalışmaları da listelemişlerdir.

Mishra ve Reddy [19] çalışmalarında Evrişimsel Sinir Ağı ile Derin Öğrenme yöntemini telekomünikasyon sektörüne ait veri kümesine uygulamış ve elde edilen modelleri doğruluk, hata oranı, duyarlılık, keskinlik ve F–Skoru gibi farklı performans ölçüm metrikleri ile değerlendirmişlerdir. Çalışmada ayrıca Derin Öğrenme yöntemi kullanılarak gerçekleştirilmiş sekiz çalışmanın özetleri de sunulmuştur. Kim ve ark. [20] mobil ve çevrimiçi oyunlarda yaşanan oyuncu kayıp tahminine odaklanmışlar, 3 farklı oyun verisini Lojistik Regresyon, Eğim Güçlendirme, Rastgele Ormanlar gibi klasik yöntemlerin yanında CNN ve LSTM Derin Öğrenme yöntemleri ile de analiz etmişlerdir. Spanoudes ve Nguyen [21] çalışmalarında müşteri kayıp tespiti için Derin Öğrenme yöntemini kullanmışlar ve kullanıcı eylem günlüğü kaydedebilen abonelik tabanlı herhangi bir şirkete uygulanabilecek bir temsil mimarisi sunmuşlardır.

Bahsedilen yöntemlerin dışında literatürde Tehlike Modeli [22], Kaplan-Meier Sağkalım Analizi [23], Ampirik Bayes Metodu [24] ve Cox Regresyonu [25] gibi metotlar kullanılarak gerçekleştirilmiş müşteri kayıp analiz çalışmaları da mevcuttur.

3. Veri Tabanlarında Bilgi Keşfi (VTBK) ve Veri Madenciliği

Veri tabanlarında bilgi keşfi ham verinin değerli hale (bilgiye) dönüştürülme sürecine verilen isimdir. Literatürde farklı örnekleri [26, 27] olsa da bu çalışmada veri tabanlarında bilgi keşfi süreci 6 aşamaya bölünmüştür.

[1] Zayıf (rastgele tahminden biraz daha iyi performans gösteren) sınıflandırıcıları güçlü hale çevirmek için kullanılan topluluk (kolektif) meta algoritmalara verilen isimdir.

1. **Ön hazırlık (Araştırma):** Uygulama alanının anlaşılması, gerekli ön bilginin toplanması ve VTBK süreç amacının belirlenmesi bu adımda gerçekleştirilir.
2. **Amaca uygun verilerin seçilmesi:** Veriler birincil ve/veya ikincil olmak üzere iki kaynaktan elde edilebilir. Birincil veri kaynakları, araştırmacının deney, gözlem ve anketler yoluyla kişisel olarak elde ettiği verilerden oluşur. İkincil veri kaynakları ise başkaları tarafından çeşitli ortam ve formatlarda toplanan ve derlenen verilerdir [28]. Bu adımda analiz için kullanılacak uygun veriler seçilir ve VTBK amacına uygun veri kümesi oluşturulur.
3. **Veri önişleme ve temizleme:** Bu adım VTBK sürecinin niteliğini arttırmaya yöneliktir. Veri kümeleri eski (güncel olmayan), eksik, fazla, tutarsız ve hatalı (gürültülü) veriler içerebilir. Yanıltıcı ve hatalı analiz sonuçlarının engellenmesi için veri kümeleri veri önişleme yöntemleri kullanılarak kusurlu verilerden arındırılır. Farklı kaynaklardan gelen verilerin birleştirilmesi, veri dönüştürme, veri azaltma ve veri ayrıklaştırma işlemleri de bu aşamada gerçekleştirilir.
4. **Verilerin veri ambarlarına aktarılması:** Bu adımda veriler veri ambarı adı verilen ilişkisel yapıda özelleşmiş veri tabanına aktarılır. Analiz, raporlama, sorgulama ve benzeri amaçlarla kullanılacak veriler burada saklanır ve periyodik olarak güncellenir.
5. **Veri madenciliği:** Veri madenciliği VTBK sürecinin analiz adımıdır ve betimleyici ve tahmin edici olmak üzere iki sınıfa ayrılır. Betimleyici veri madenciliğinde amaç veri kümesi içerisindeki verilerin iç yapısını, birbirleri ile olan ilişkilerini ve genel özelliklerini nitelemektir. Bunun için veri madenciliğinin Kümeleme, Birliktelik Kuralı Tespiti, Özetleme işlevleri kullanılır. Tahmin edici veri madenciliğinde ise amaç mevcut veriyi kullanarak tahmin yapabilen modeller oluşturmaktır. Bunun için veri madenciliğinin Sınıflandırma, Regresyon ve Zaman Serileri Analizi işlevleri kullanılır. Veri madenciliği işlevleri ve uygulama alanları hakkında detaylı bilgi [29–32] numaralı referanslardan elde edilebilir.
6. **Değerlendirme:** Son aşamada, sonuçlar çeşitli sunum teknikleri ile son kullanıcıya sunulur ve geçerlilik, güncellik ve kullanışlılık gibi kriterlere göre yorumlanır ve değerlendirilir.

Bu makalenin deney aşamasında işletmeleri terk etme ihtimali olan müşterilerin tahmini için çeşitli sınıflandırma algoritmaları kullanılmıştır. Gözetimli öğrenme şekli olan sınıflandırma amaç sınıfı bilinmeyen verinin (gözlemin) önceden belirlenmiş olan bir dizi sınıftan hangisine ait olduğunu belirlemektir. İki sınıflı, örneğin fotoğrafları renkli veya siyah beyaz olarak sınıflandırmak, çok sınıflı, örneğin kredi risk değerlendirmesi ve çok etiketli, örneğin filmleri türlerine göre sınıflandırmak, olmak üzere 3 sınıflandırma problemi vardır.

Sınıflandırma problemlerinde veri kümesi eğitim ve test (sınama) olmak üzere iki parçaya ayrılır. Daha sonra kullanılacak sınıflandırma yöntemi önce eğitim kümesi ile eğitilir daha sonra test kümesi ile test edilir. Eğer sınıflandırma algoritmasının hiper parametreleri ayarlanacaksa, örneğin Yapay Sinir Ağları katmanlarındaki gizli birimlerin sayısının belirlenmesi, veri kümesi eğitim, test ve doğrulama kümeleri olmak üzere üç parçaya bölünür ve doğrulama kümesi ile hiper parametre ayarlanması yapılır. Oluşturulan modelin başarısı ise doğruluk, yanlış sınıflandırma oranı, AUC, hassaslık, belirginlik, kaldıraç, F–Skoru gibi model değerlendirme metrikleri [33] yardımı ile ölçülür.

4. Deney Aşamasında Kullanılan Yöntemler

4.1. Öznitelik Seçimi

Veri kümelerindeki gürültülü, gereksiz ve konu dışı özniteliklerin seçilip, veri kümesinden çıkartılmasına veri boyutu indirgemesi, öznitelik seçimi, değişken seçimi veya değişken altküme seçimi denir. Öznitelik seçimi [34] veri kümelerinin boyutunu küçültmek, veri analiz maliyetini ve süresini düşürmek, ölçeklenebilirliği arttırmak, basit, anlaşılabilir, genelleme kabiliyeti yüksek ve kolay güncellenebilen modeller üretmek için kullanılır. Öznitelik seçim yöntemleri filtreleyen, sarıcı, gömülü, melez ve topluluk (kolektif) olmak üzere 5 sınıfa ayrılır. Bu çalışmada hem kategorik hem de sürekli değişkenler ile kullanılabilen ve filtreleyen yöntemlerden biri olan Ki-Kare Testi [35] (Formül 1) kullanılmıştır. Filtreleyen seçim yöntemleri, öznitelikler

(bağımsız değişkenler) ile hedef değişken (sınıf değişkeni veya bağımlı değişken) arasındaki korelasyonu ölçer.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Formülde i gözlem sayısı, O gözlemlenen ve E beklenen değeri (değerleri) ifade etmektedir. Tek değişkenli bir test olan Ki-Kare testi esasen iki değişkenin birbirinden ne kadar bağımsız (serbest) olduğunu test eder. Öznitelik seçiminde ise tam tersi durum test edilir çünkü bağımsız değişken ve bağımlı değişken (sınıf, hedef) arasındaki ilişki ne kadar fazla (kuvvetli) ise ilgili bağımsız değişken (öznitelik) sınıflandırma işlevi için o denli gereklidir. Bu sebeple Ki-Kare testi sonucunda elde edilen öznitelik önem skorları p değerinin[2] negatif logaritması ($-log(p)$) alınarak hesaplanır.

4.2. RFM Bölümlemesi (Sınıflandırması)

RFM (**R**ecency, **F**requency, **M**onetary Value) bölümlemesi müşterileri harcama alışkanlıklarını dikkate alarak sıralayan ve bölümleyen pazarlama analiz araçlarından bir tanesidir. RFM bölümlemesi sadece en iyi müşterileri belirlemek için kullanılmaz. Alışveriş yapma potansiyelinin yüksek olduğu müşterileri ve müşteri gruplarını, yeni ve mevcut müşterilerden ne kadar gelir elde edilebileceğini ve sürekli müşteri olma potansiyeli olan müşterileri tahmin etmeye de yardımcı olur [36].

RFM bölümlemesi müşteri ölçeklendirmesi ve müşteri sadakatinin ölçülmesi için kullanılan basit bir yöntemdir. Yöntem sadece R, F ve M değerlerini üretebilecek özniteliklere ihtiyaç duyduğu için hem az hem de çok sayıda kayıt ve özniteliğe sahip uygun veri kümelerine kolayca uygulanabilir. Bu çalışmada RFM bölümlemesi normal işlevine ek, öznitelik çıkarımı (boyut indirgeme) metodu olarak kullanılmıştır. Öznitelik çıkarımı aynı öznitelik seçimi gibi bir boyut indirgeme tekniğidir ancak öznitelik çıkarımında amaç mevcut öznitelik kümesini kullanarak yeni öznitelik veya öznitelikler üretmektir.

RFM bölümlemesi üç nicel faktöre dayanmaktadır: R değeri alışveriş tarihinin günümüze ne kadar yakın olduğunu veya en son alışveriş tarihini, F değeri yapılan alışverişlerin sıklığını ve M değeri ise yapılan alışverişin nakdi kıymetini, yani toplam harcamayı, ifade eder. Müşterilere her bir kategori için 1 ile 5 veya 1 ile 9 arasında puan atanır ve puanların yan yana yazılmasıyla üç haneli RFM skoru elde edilir. RFM skoru yüksek olan müşteriler işletme için önemli müşterilerdir. RFM skorunun RFD, RFE, RFM–I, RFMTC gibi çeşitli varyasyonları vardır [37]. RFM değerlerinin gruplaması iki türlü yapılabilir [38].

1. **İç içe gruplama:** Bu metotta öncelikle müşterilere en son alışveriş tarihi dikkate alınarak R (güncellik) skoru atanır. Her bir R sırası içinde, müşterilere bir F (alışveriş sıklığı) sırası ve her bir F sırası içinde müşterilere bir M (mali) sırası atanır. Bu metotta F değeri R'ye, M değeri de F'ye bağımlı olduğu için F ve M değerlerinin yorumlanması zordur.
2. **Bağımsız gruplama:** R, F ve M değerleri birbirinden bağımsız şekilde atanır.

RFM bölümlemesinde ideal olan grupların eşit sayıda müşteri içermesidir. Aynı skoru alan müşteriler gruplara iki farklı şekilde atanabilir. İlk yaklaşımda aynı skora sahip müşteriler aynı gruba atanır. Bu yaklaşım da gruplardaki müşteri adeti dikkate alınmadığı için grup dağılımları homojen değildir. İkinci yaklaşımda ise aynı skora sahip müşteriler rastgele gruplara atanır. Bu yaklaşımda da amaç her bir gruba eşit sayıda müşteri atamaktır. Örneğin, 5 gruplu bir bölümleme yapılıyorsa her grup toplam müşteri sayısının %20'si kadar müşteri içermelidir. Bu çalışmada RFM skorları iç içe gruplama ile oluşturulmuş ve aynı skoru alan müşterilerin grupları ilk yaklaşıma göre belirlenmiştir.

4.3. Sınıflandırma Teknikleri

Bu bölümde deney aşamasında kullanılan sınıflandırma algoritmaları genel hatları ile açıklanmıştır. Asıl versiyonları iki sınıflı problemlerde kullanılabilen algoritmalar "Bire–Karşı–Bir" veya "Bire–Karşı–Hepsi" yöntemleri kullanılarak çok sınıflı problemlere uyarlanabilir [39].

[2] Serbestlik derecesi ve Ki-Kare skoruna göre hesaplanan bir değerdir.

- **Karar Ağaçları:** Karar Ağaçları sıklıkla kullanılan tahmin modeli oluşturma metotlarından bir tanesidir [40]. Kategorik hedefleri (örneğin, kredi talebinin banka tarafından kabulü veya reddi) tahmin etmek için kullanılan Sınıflandırma Karar Ağaçları ve nümerik hedefleri (örneğin, döviz kurundaki yıllık artış miktarı) tahmin etmekte kullanılan Regresyon Karar Ağaçları olmak üzere iki sınıfa ayrılırlar. Ağaç yapısı kök ile başlar ve yukarıdan aşağıya doğru dallar, iç düğümler (karar düğümü) ve yaprak (uç) düğümlerinden oluşur. Ağacın tek bir kökü olur ve yaprak düğümden sonra başka bir düğüm gelemez. Kök ve iç düğümlere ise iç veya yaprak düğüm eklenebilir. Her iç düğüm bir öznitelik üzerinde gerçekleştirilen testi, her dal test sonucunu ve yaprak düğümler de sınıf etiketini temsil eder. Sınıflandırma Karar Ağacı oluşturuluyorsa kök düğümün ve iç düğümlerin belirlenme işleminde Gini İndeksi veya Bilgi Kazancı, Regresyon Karar Ağacı oluşturuluyorsa Ortalama Karekök Hatası (MSE) veya Artık Hatası (Residual Error) gibi yöntemler kullanılır.
- **Lojistik Regresyon:** Lojistik Regresyon iki sınıflı gözetimli öğrenme problemlerinde kullanılır [41]. Doğrusal regresyondan farklı olarak tahmin için bir doğru yerine biçimsel fonksiyon adı verilen, tekdüze (monoton), 0 ve 1 aralığında sürekli ve S şekline benzer bir eğri olan sigmoid fonksiyonu kullanılır. Naïve Bayes gibi olasılık temelli bir yaklaşımdır. Bu yöntem için bağımsız değişkenler sürekli veya kategorik özellikte, bağımlı değişken ise kategorik ve iki değerli olmalıdır.
- **Naïve Bayes:** Bayes sınıflandırıcılar [42] sınıflandırma işlemini olasılık tabanlı olan Bayes teoremine göre gerçekleştirirler. Bayes teoremi örneklerin belirli bir sınıfa dahil olma olasılığını hesaplamak için kullanılır. Genel olarak algoritma sınıfı bilinmeyen bir örnek ile karşılaştığında tüm sınıflar için bu değeri hesaplar ve veriyi olasılığı en yüksek olan sınıfa atar. Koşullu olasılığa dayalı sınıflandırıcılara örnek olan Bayes sınıflandırıcılar hızlı, kolay uygulanabilir ve çoğu zaman yüksek sınıflandırma performansına sahip olmalarından ötürü sıklıkla kullanılırlar. Buna karşılık Bayes sınıflandırıcılar öznitelikler arasındaki ilişkileri dikkate almazlar ve bütün özniteliklere aynı önem derecesine (ağırlığa) sahipmiş gibi davranırlar.
- **Destek Vektör Makinesi:** Destek Vektör Makinesi [43] sınıflandırma ve regresyon problemlerinde kullanılır. Algoritma sınıfları her iki sınıfa da eşit uzaklıkta olacak doğrusal bir hiper düzlem yardımıyla ayırmaya çalışır. Hiper düzleme en yakın olan ve her iki sınıfa ait örneklere destek vektörleri denir. Destek vektörleri hiper düzlem ortada kalacak şekilde birbirine paralel iki sınır çizgisi oluşturur. Sınır çizgileri içerisinde herhangi bir sınıfa ait örnek bulunmaz ve bu alana sınır denir. Algoritma hem sınır genişliği (daraltıp genişletebilir) hem de sınırın konumuyla oynayarak (açısı) sınıfları birbirinden ayırmaya çalışır. Ancak örnekleri her zaman bir doğru yardımıyla ayırmak mümkün değildir. Doğrusal şekilde ayrılamayan örnek kümelerinde doğrusal olmayan hiper düzlemleri kullanan destek vektör makinesi kullanılır.
- **Topluluk Metotları:** Topluluk Öğrenmesi [44] zayıf sınıflandırıcıların bir araya getirilmesiyle oluşturulur. Zayıf sınıflandırıcı rastgele yani şansa dayalı sınıflandırmadan biraz daha iyi performansa sahip sınıflandırıcılara verilen isimdir. Her bir sınıflandırıcının sınıflandırma performansı tek başına zayıf olsa da çeşitli yöntemlerle topluluk haline getirildiklerinde yüksek sınıflandırma performansı ve kolay ölçeklenebilirlik elde edilmektedir. Ayrıca elde edilen sonucun tek bir modelin (sınıflandırıcının) performansına bağlı kalmasının da önüne geçilmiş olur. Topluluk metotlarında elde edilen sonuçlar sınıflandırma problemleri için oy çokluğu veya ağırlıklı oylama, regresyon problemleri için ise ortalama veya ağırlıklı ortalama gibi tekniklerle birleştirilirler.
- **Yapay Sinir Ağları:** Yapay Sinir Ağları [45] gerçek biyolojik sinir hücrelerini taklit eden ve düğüm (node) veya nöron adı verilen yapay sinir hücrelerinin genellikle katmanlar halinde bir araya gelmesiyle oluşan sisteme verilen isimdir. Yapay Sinir Ağlarını oluşturan ve bilginin işlendiği yer olan yapay sinir hücreleri girdi (veriler),

girdiye ait ağırlık değeri, transfer fonksiyonu, aktivasyon fonksiyonu ve çıktı kısımlarından oluşur. Girdiler sırası ile bu katmanlardan geçerek işlenirler. Yapay sinir hücreleri aynı biyolojik sinir hücreleri gibi birbirlerine bağlı ve etkileşim içindedirler. Böylelikle eş zamanlı (paralel) olarak çalışan her bir düğüm aldığı girdiyi işledikten sonra bağlı olduğu diğer düğüm veya düğümlere girdi olarak iletir. Bu sebeple Yapay Sinir Ağları katmanlar halinde tasarlanır. Bu katmanlara sırası ile girdi, gizli ve çıktı adı verilmiştir. Sadece girdi ve çıktı katmanlarından oluşan ağlara tek, girdi, gizli ve çıktı katmanlarından oluşan ağlara ise çok katmanlı sinir ağı denir. İleri besleme ve geri yayılım olmak üzere iki çeşit Yapay Sinir Ağı vardır.

4.4. Çapraz Doğrulama

Üretilen modellerin daha önce karşılaşmadığı veriler üzerinde test edilmesi ve sınıflandırma performans ölçüm metrikleri tarafından değerlendirilmesi gereklidir. Böylelikle tahmin için kullanılacak modelin pratikte ne kadar hatasız performans göstereceğini kestirebiliriz.

K–Katlamalı Çapraz Doğrulamada veri seti k adet eşit miktarda örnek içeren, eşit bölünmezse bir küme diğerlerinden daha fazla olacak şekilde, alt kümeye ayrılarak eğitim ve test süreçleri k kere tekrar edilir. Her seferde bir küme test, diğer kümeler ise eğitim amaçlı kullanılırlar. Bu yöntemde her örnek en az 1 kere test, $(k-1)$ kere de eğitim kümesinde bulunur. Genellemek gerekirse k değeri ile eğitim kümesi arasında doğru, test kümesi ile ters orantı vardır. Bu sebeple k değeri ne kadar büyürse veri o kadar büyük parçalara ayrılacağından eğitim kümesinin içerdiği örnek sayısı artar. Ancak test kümesinin boyutu da küçülür. Bu durumda varyans artar ve eğitim için ihtiyaç duyulan süre uzar. Tersi durumda ise varyans azalır ve eğitim için ihtiyaç duyulan süre kısalır. K–Katlamalı Çapraz Doğrulama yönteminde başarı oranı her yinelemedeki başarı oranının ortalamasına eşittir. Bu çalışmada k değeri genelde tercih edilen değer olan 10 olarak alınmıştır.

Tabakalı K–Katlamalı Çapraz Doğrulama [46] sınıf başına düşen örnek sayısının dengesiz diğer bir deyişle asimetrik dağılıma sahip olduğu durumlarda varyans farkını azaltmak ve çeşitliliği sağlamak için kullanılır. Tabakalı K–Katlamalı Çapraz Doğrulama iki farklı şekilde uygulanabilir. İlk şekilde her örnek grubundan (sınıftan) eşit adette örnek alınır. İkinci şekilde ise her örnek grubundan küme içerisindeki oranı dikkate alınarak örnek alınır. Geri kalan süreç her iki uygulama şekli için de K–Katlamalı Çapraz Doğrulama ile aynıdır.

5. Bulgular

Deney çalışması Tablo 1'de bazı özellikleri paylaşılan farklı iş kollarına ait, genel kullanıma açık, eksik ve mükerrer veri içermeyen 3 veri kümesi ile gerçekleştirilmiştir. Veri kümeleri seçilirken RFM bölümlemesine uygun olmalarına (RFM skoru elde edilecek öznitelikler içermesine), müşterilerin işletmeyi terk edip etmediğini gösteren sınıf değişkeninin olmasına ve mümkün oldukça fazla öznitelik (boyut) içermesine dikkat edilmiştir.

Deney çalışmasında kullanılan veri kümelerinden ilki müşterilerine ev telefonu ve internet hizmetleri sağlayan hayali bir telekomünikasyon şirketine aittir. İlgili veri kümesi müşterilerin demografik özellikleri, ikamet yerleri, kullandıkları hizmetler ve mevcut durumları ile ilgili bilgiler içermektedir. Bankacılık sektörüne ait diğer veri kümesi müşterilerin demografik bilgileri, kullandıkları kredi kartı türü (kategorisi), kredi kartı limiti, banka ve kredi kartı kullanım sıklığı gibi bilgilerden oluşmaktadır. E-ticaret ile ilgili veri kümesi ise müşteriler ile ilgili çeşitli kişisel bilgilerin yanı sıra en son sisteme ne zaman giriş yaptıkları (gün olarak), en son ne zaman satış yaptıkları (gün olarak), satış miktarı ve kazançları (dolar cinsinden) gibi bilgileri de içermektedir.

Veri kümeleri karakter dizisi, kategorik ve nümerik değerler içerdiği için Diskriminant Analizi ve K–En Yakın Komşu sınıflandırıcılar kullanılamamıştır. Bu sebeple deneyler Karar Ağaçları, Lojistik Regresyon, Naïve Bayes, Destek Vektör Makinesi, Topluluk Yöntemleri ve Yapay Sinir Ağları sınıflandırıcıları ile gerçekleştirilmiştir. Sınıflandırıcılar öncelikle varsayılan değerlerle daha sonra da Bayes Hiper Parametre Optimizasyonu sonucunda elde edilen parametrelerle çalıştırılmıştır. Veri kümelerindeki sınıf dağılımları dengesiz (asimetrik dağılıma sahip) olduğundan tüm model eğitim ve test aşamalarında Tabakalı 10–Katlamalı Çapraz Doğrulama kullanılmıştır.

Tablo 1. Deney aşamasında kullanılan veri kümelerinin bazı özellikleri

Veri Kümesi	İlgili Sektör	Örnek Sayısı	Öznitelik Sayısı	Eksik Veri	Veri Tipi	Sınıflar ve Sınıf Dağılımı		Kaynak
IBM Telco	Telekom.	7043	20	Yok	Karakter Dizisi Kategorik Nümerik	"No"	5174	IBM[3]
						"Yes"	1869	
Bank Churners	Bankacılık	10127	20	Yok		"No"	8500	Kaggle[4]
						"Yes"	1627	
IBM Customer Churn	E–Ticaret	2066	16	Yok		"High"	983	GitHub[5]
						"Low"	699	
						"Medium"	384	

Ayrıca dengesiz veri dağılımlarında kullanılabilen ve Topluluk Yöntemlerinden biri olan RUS (Rastgele Seyrek Örnekleme) Güçlendirme sınıflandırıcı da deney çalışmasına dahil edilmiştir. Model performansları doğruluk ve Eğri Altında Kalan Alan (AUC) [47] yardımı ile değerlendirilmiştir. Üç sınıfa sahip olan e–ticaret veri kümesinde AUC değerleri "Bire–Karşı–Hepsi" yöntemiyle hesaplanmıştır. Deneyler MATLAB® 2021b ve 8 çekirdekli Intel Core i9 CPU (3.6 GHz ve 128 GB DDR4 RAM) bilgisayar kullanılarak gerçekleştirilmiştir.

RFM bölümlemesi ve öznitelik seçimi (boyut indirgemesi) yapılmadan önce 6 temel toplamda ise 21 adet sınıflandırıcı ile oluşturulan modellerin performansları doğruluk ve AUC metrikleri kullanılarak ölçülmüş ve elde edilen değerler Tablo 2'de sunulmuştur. Tablo 2'de ve deney çalışmasının devamında veri kümelerinin asıl isimleri yerine ait oldukları sektör isimleri (Tablo 1'de ikinci sütun) kullanılmıştır. Tablo 2 için kullanılan dipnotlar Tablo 4 için de geçerlidir. Elde edilen en yüksek değerler kalın, en düşük değerler ise yatık şekilde yazılmıştır.

Tablo 2'deki model başarım sonuçlarına göre telekomünikasyon veri kümesinde Lojistik Regresyon, bankacılık veri kümesinde optimize edilmiş Topluluk Yöntemi ve e–ticaret veri kümesinde optimize edilmiş Destek Vektör Makinesi sınıflandırıcıları doğruluk ve AUC metrikleri cinsinden diğer sınıflandırıcılara göre daha başarılı modeller oluşturmuşlardır.

[3] https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113
[4] https://kaggle.com/sakshigoyal7/credit-card-customers/tasks?taskId=2729
[5] https://github.com/IBM/icp4d-customer-churn-classifier

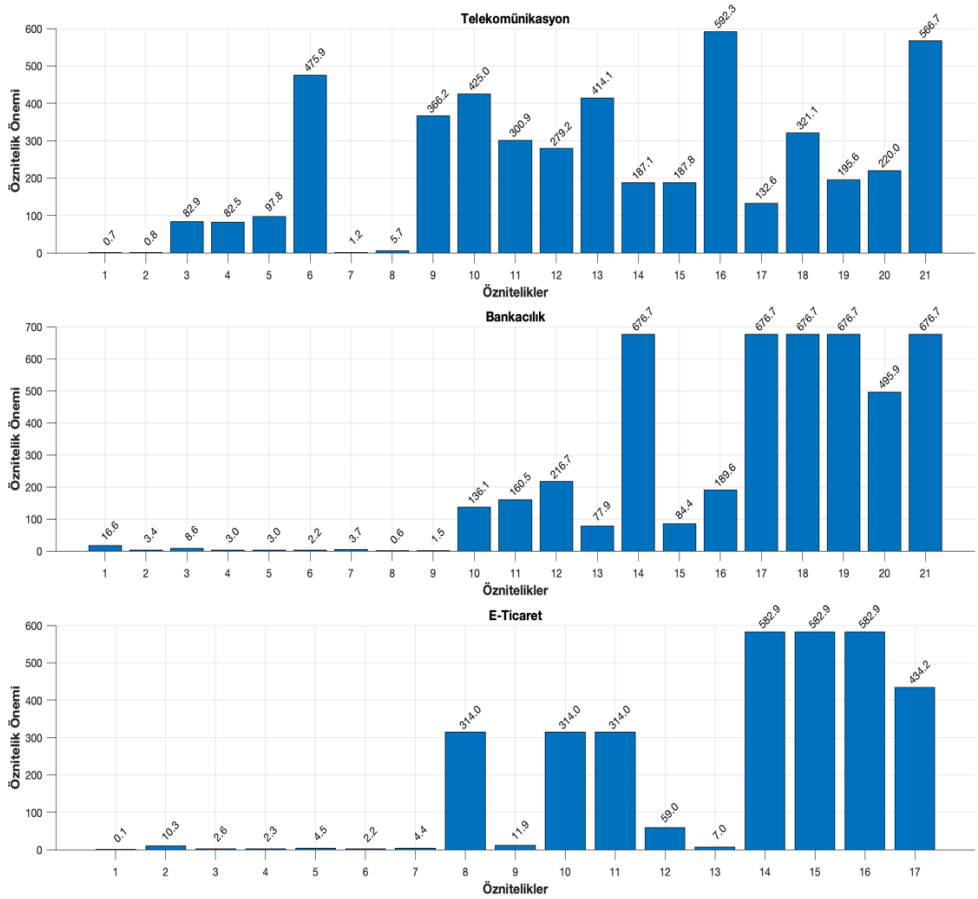
Tablo 2. Oluşturulan modellerin başarım değerleri

Sınıflandırıcı	Tip	Veri Kümesi					
		Telekom.		Bankacılık		E–Ticaret	
		Doğruluk	AUC	Doğruluk	AUC	Doğruluk	AUC[6]
Karar Ağaçları	Çok Ypr.[7]	78.9	0.83	94.6	0.94	90.9	0.94–0.98–0.87
	Ortalama Ypr.	79.0	0.82	93.2	0.94	93.1	0.96–0.99–0.89
	Az Ypr.	78.8	0.79	90.3	0.89	91.1	0.94–0.99–0.87
	Opt.[8]	79.3	0.81	94.4	0.92	94.2	0.96–0.99–0.90
Lojistik Regresyon	—	**80.6**	**0.85**	90.3	0.92	—[9]	—
Naïve Bayes	Gaussian	*72.8*	0.82	*88.8*	*0.87*	—[10]	—
	Kernel	73.6	0.82	89.7	0.95	*87.9*	0.96–0.99–0.87
	Opt.	73.6	0.82	90.2	0.94	89.2	0.96–0.99–0.87
Destek Vektör Makinesi	Doğrusal	80.0	0.83	90.5	0.92	88.8	0.96–0.99–0.89
	Karesel	79.9	0.81	93.4	0.96	94.2	0.97–0.99–0.89
	Kübik	77.6	0.78	93.7	0.96	92.4	0.96–0.99–0.88
	Opt.	80.3	0.81	93.4	0.96	**95.1**	**0.97–0.99–0.91**
Topluluk Yöntemleri	Uyar. Güç.[11,12]	80.1	0.84	96.0	**0.99**	94.7	0.97–0.99–0.90
	Torbalama	79.6	0.83	96.3	**0.99**	94.2	0.97–0.99–0.89
	RUS Güç.	76.0	0.84	92.9	0.98	93.8	0.97–0.99–0.90
	Opt.	79.8	0.84	**97.4**	**0.99**	93.9	0.96–0.99–0.88
Yapay Sinir Ağları	Dar	78.8	0.82	93.9	0.97	92.4	0.96–0.99–0.89
	Ortalama	77.0	0.79	93.4	0.92	88.7	*0.93–0.98–0.86*
	Geniş	74.1	*0.75*	93.6	0.96	89.7	0.94–0.99–0.86
	2 Katmanlı	78.4	0.81	94.4	0.97	92.2	0.95–0.99–0.86
	3 Katmanlı	78.3	0.81	94.0	0.97	91.4	0.95–0.99–0.88

Telekomünikasyon veri kümesinde RFM skorları sırası ile ay cinsinden sözleşme süresi (sözleşme süresi uzun olanın R değeri büyük), müşterinin sahip olduğu toplam abonelik (kullandığı toplam servis) ve operatöre ödediği aylık ücret değerlerini içeren öznitelikler dikkate alınarak hesaplanmıştır. Bankacılık ve e–ticaret ile ilgili veri kümelerinde RFM skorları ise müşterinin/kullanıcının aktif olmadığı süre, gerçekleştirilen toplam aylık işlem adeti ve aylık toplam işlem tutarı dikkate alınarak hesaplanmıştır. RFM skorları hesaplandıktan ve her bir veri kümesinin sonuna (sınıf değişkeninden önce) yeni bir öznitelik olarak eklendikten sonra, veri kümelerindeki öznitelikler tek değişkenli ve filtreleyen tipte olan Ki-Kare Testi ile sıralanmıştır (Şekil 1). Sıralama işlemi özniteliklerin sınıflandırmaya olan katkılarına göre yapılır. Önem puanı arttıkça özniteliklerin sınıflandırmaya katkısı (bağımlı değişken ile olan ilişkisi) artar.

Ki-Kare Testi artı sonsuz (+∞) puanlar üretebilmektedir. Çubuk grafikte bu puanların gösterilebilmesi mümkün olmadığı için artı sonsuz puanlar en büyük sonlu puana eşitlenmiştir. Şekil 1'de paylaşılan çubuk grafiklerdeki son çubuklar RFM skorunun önem derecesini ifade eder.

[6] AUC değerleri sırasıyla "High", "Low" ve "Medium" sınıflarına ait değerlerdir.
[7] Ypr. = Yapraklı.
[8] Opt. = Optimize edilmiş (Bayes Hiper Parametre Optimizasyonu uygulanmış).
[9] Algoritma iki sınıflı sınıflandırma problemlerinde kullanılabilir.
[10] Algoritma normal (Gauss) dağılıma sahip veri kümelerinde kullanılabilir.
[11] Uyar. Güç. = Uyarlamalı Güçlendirme.
[12] İki sınıflı problemler için AdaBoost.M1, çok sınıflı problemler için AdaBoost.M2 metodu kullanılmıştır.



Şekil 1. Öznitelik önemlerinin 2B çubuk grafiği ile ifade edilmesi

Ki-Kare Testi gibi filtreleyen tipteki öznitelik seçim algoritmaları öznitelik altkümesi oluşturmaz. Model oluşturma sürecine katılacak özniteliklerin tespiti, yani öznitelik altkümesinin oluşturulması, özniteliklerin önem sırası ve puanı dikkate alınarak, kullanıcı tarafından belirlenen (öznel, nispi) bir eşik değere göre yapılır. Bu sebeple aynı veri kümesi için oluşturulan öznitelik altkümeleri eşik değer seçimine göre farklılık gösterebilir. Tablo 3'te her bir veri kümesi için öznitelik önem sırası (öznitelik önem puanı yüksekten düşüğe doğru olacak şekilde) paylaşılmıştır. Ki-Kare Testi sonucunda elde öznitelik önem puanlarındaki düşüşler dikkate alınarak seçilen ve RFM bölümlemesi için kullanılan öznitelikler sırasıyla Tablo 3'ün üçüncü ve dördüncü sütununda sunulmuştur. Deney çalışmasında RFM bölümlemesi boyut indirgeme yöntemi olarak kullanılmıştır. Bu sebeple RFM skorunun tespiti için kullanılan öznitelikler model oluşturma aşamasına dahil edilmemiş, ilgili öznitelikleri temsilen RFM skorlarına ait öznitelikler kullanılmıştır. Bu özniteliklerin Ki-Kare Testi sonucunda yüksek öneme sahip olduğu tespit edilmiştir. Daha önce bahsedildiği gibi telekomünikasyon veri kümesi için F değeri müşterinin abonesi olduğu toplam servis adeti dikkate alınarak hesaplanmıştır. Bu sebeple Tablo 3'te telekomünikasyon veri kümesinin RFM bölümlemesi için kullanılan öznitelikleri kısmına üçüncü öznitelik yazılmamıştır. Boyut indirgeme metotlarının uygulanmasından sonra elde edilen nihai öznitelik altkümeleri Tablo 3'ün son sütununda, bu altkümeler kullanılarak oluşturulan modellerin doğruluk ve AUC metrikleri cinsinden sınıflandırma performansları ise Tablo 4'te yer almaktadır.

Tablo 3. Ki-Kare Testine göre öznitelik önem sırası, Ki-Kare Testi sonuçlarına göre seçilen öznitelikler, RFM bölümlemesi için kullanılan öznitelikler ve boyut indirgemeden sonra elde edilen nihai öznitelik altkümelerinin elemanları

Veri Kümesi	Öznitelik Önem Sırası	Seçilen Öznitelikler	RFM Bölümlemesi için Kullanılan Öznitelikler	Nihai Öznitelik Altkümeleri
Telekom.	16, 21, 6, 10, 13, 9, 18, 11, 12, 20, 19, 15, 14, 17, 5, 3, 4, 8, 7, 2, 1	İlk 9	16, 19	6, 9, 10, 11, 12, 13, 18, 21
Bankacılık	17, 18, 19, 21, 14, 20, 12, 16, 11, 10, 15, 13, 1, 3, 7, 2, 4, 5, 6, 9, 8	İlk 6	11, 17, 18	14, 19, 20, 21
E–Ticaret	15, 16, 14, 17, 8, 10, 11, 12, 9, 2, 13, 5, 7, 3, 4, 6, 1	İlk 7	8, 9, 14	10, 11, 15, 16, 17

Tablo 4. Nihai öznitelik altkümeleri ile oluşturulan modellerin başarım değerleri

Sınıflandırıcı	Tip	Veri Kümesi					
		Telekom.		Bankacılık		E–Ticaret	
		Doğruluk	AUC	Doğruluk	AUC	Doğruluk	AUC
Karar Ağaçları	Çok Ypr.	79.0	0.83	90.3	0.89	92.5	0.95–0.98–0.89
	Ortalama Ypr.	79.4	0.81	89.9	0.85	94.1	0.96–0.99–0.90
	Az Ypr.	79.0	0.79	88.0	0.81	91.1	0.95–0.99–0.87
	Opt.	79.5	0.81	90.7	0.89	94.5	0.96–0.99–0.90
Lojistik Regresyon	—	**79.8**	**0.84**	87.5	0.82	—	—
Naïve Bayes	Gaussian	73.8	0.83	86.9	0.81	—	—
	Kernel	*71.4*	0.82	87.5	0.87	81.4	0.95–0.99–0.86
	Opt.	73.8	0.83	87.6	0.86	82.5	0.95–0.99–0.86
Destek Vektör Makinesi	Doğrusal	79.1	0.83	83.9	0.72	93.0	0.96–0.99–0.88
	Karesel	79.2	0.81	85.3	0.80	93.7	0.97–0.99–0.88
	Kübik	78.3	*0.77*	*72.3*	*0.70*	*75.4*	*0.85–0.99–0.61*
	Opt.	79.2	0.80	90.4	0.79	94.2	0.97–0.99–0.87
Topluluk Yöntemleri	Uyar. Güç.	**79.8**	**0.84**	90.8	0.92	94.2	0.97–0.99–0.90
	Torbalama	78.6	0.82	91.1	0.93	93.1	0.96–0.99–0.89
	RUS Güç.	75.9	**0.84**	84.6	0.90	93.8	0.97–0.99–0.90
	Opt.	79.7	**0.84**	**92.5**	**0.94**	94.5	0.97–0.99–0.90
Yapay Sinir Ağları	Dar	79.5	0.83	89.1	0.88	**95.0**	**0.97–0.99–0.91**
	Ortalama	79.1	0.82	90.1	0.90	94.4	0.97–0.99–0.89
	Geniş	76.6	0.79	90.2	0.90	93.5	0.96–0.99–0.87
	2 Katmanlı	79.4	0.83	89.3	0.89	94.7	0.97–0.99–0.90
	3 Katmanlı	78.7	0.82	89.6	0.89	**95.0**	0.97–0.99–0.90

Tablo 4 incelendiğinde, telekomünikasyon veri kümesinde Lojistik Regresyon ve Uyarlamalı Güçlendirme, bankacılık veri kümesinde optimize edilmiş Topluluk Yöntemi ve e–ticaret veri kümesinde ise Dar ve 3 Katmanlı Yapay Sinir Ağları sınıflandırıcıları diğer sınıflandırıcılara göre doğruluk ve AUC değeri daha yüksek modeller oluşturmuşlardır. Tablo 2 ve Tablo 4'teki model başarım değerleri kıyaslandığında aşağıdaki sonuçlara ulaşılmıştır.

- RFM bölümlemesi ve öznitelik seçiminden önce ve sonra oluşturulan gözetimli öğrenme modellerinin sınıflandırma performansları kıyaslandığında her iki sonuç kümesinin oldukça yakın olduğu görülmektedir.

- Optimize edilmiş sınıflandırıcılar genelde iyi performans göstermişlerdir. Farklı optimizasyon metotlarının, örneğin, rastgele arama veya evrimsel algoritmalar, kullanılan sınıflandırıcılar üzerindeki etkisi bu çalışmada incelenmemiş ancak gelecek çalışması olarak hedeflenmiştir.
- Telekomünikasyon veri kümesindeki tüm öznitelikler kullanıldığında hem doğruluk (%80,6) hem de AUC değeri (0,85) en yüksek model Lojistik Regresyon sınıflandırıcı ile elde edilmiştir. Boyut indirgemesi gerçekleştirildikten sonra 8 öznitelik kullanılarak Uyarlamalı Güçlendirme ve Lojistik Regresyon ile %79,8 doğruluk ve 0,84 AUC değerlerine sahip modeller elde edilmiştir.
- Bankacılık sektörüne ait veri kümesinde gerçekleştirilen boyut indirgemesinden sonra asıl veri kümesi kullanıldığında elde edilen performans değerlerine yaklaşılmış ancak doğruluk değerinde %5'lik, AUC değerinde ise 0.05'lik kayıplar yaşanmıştır.
- Son olarak, e–ticaret veri kümesinde boyut indirgemesinden önce optimize edilmiş Destek Vektör Makinesi sınıflandırıcı %95,1 doğruluk oranına ve 0,97–0,99–0,91 AUC değerlerine sahip model üretmiştir. Boyut indirgemesi yapıldıktan sonra Dar ve 3 Katmanlı Yapay Sinir Ağları sınıflandırıcı ile üretilen modeller %95,0 doğruluk değerini yakalamıştır. Ayrıca Dar Yapay Sinir Ağları Sınıflandırıcı ile 0.97–0.99–0.91 AUC değerleri elde edilmiştir.

6. Tartışma ve Sonuç

İşletmelerce toplanan ve saklanan veriler doğru kullanıldığında ticari üstünlük sağlayabilecek önemli bir faktördür. Veriyi üreten, saklayan ve iyi analiz edip, kullanabilen işletmelerin kârlılıkları, gelişme hızları ve rekabet güçleri artmaktadır. Günümüz teknoloji ve bilgi çağında değer yaratmak artık fiziksel imkân ve varlıklardan çok, mevcut ve elde edilebilir bilgi kaynaklarını etkin kullanmaktan geçmektedir. Bu çalışmada işletmelerin en büyük problemlerinden biri olan müşteri kayıplarının tespit edilmesi üzerine odaklanılmıştır. Çalışmada temel olarak kullanılan 6 adet sınıflandırıcının ve bu sınıflandırıcıların farklı türlerinin tahmin performansı öznitelik seçiminden önce ve sonra olmak üzere doğruluk ve AUC metrikleri cinsiden kıyaslanmıştır.

Çalışmada esasen bir pazarlama yöntemi olan RFM bölümlemesi boyut indirgeme yöntemi olarak kullanılmıştır. Böylelikle asıl veri kümesinden daha az elemanlı ancak asıl veri kümesine oldukça yakın performansa sahip öznitelik altkümelerinin üretilmesi hedeflenmiştir.

Gözetimli öznitelik seçiminde temel amaç tahmine dayalı performanstan ödün vermeden tahmin edicilerin (öznitelik) sayısını mümkün olduğunca azaltmaktır. Bu çalışmada öznitelik seçiminden sonra oluşturulmuş modellerin tahmin performansında ciddi bir artış veya azalış gözlemlenmemiştir. Böylelikle performanstan ödün vermeden öznitelik sayıları azaltılmış ve hedeflenen amaca ulaşılmıştır.

İşletmelerden kaçma eğilimleri olan müşteri gruplarının tespit edilip önlem alınması işletmelerin ticari rekabetini koruması ve güçlendirmesi için gereklidir. Schmitt'e [44] göre müşteri kayıpları istek dışı ve finansal sebeplerden ötürü ayrılanlar, istek dışı ve finansal olmayan sebeplerden ötürü ayrılanlar, isteyerek ve finansal nedenlerden ötürü ayrılanlar ve isteyerek ve finansal olmayan sebeplerden ötürü ayrılanlar olmak üzere 4 sınıfa ayrılır.

İstek dışı işletmeyi terk eden, yani birinci ve ikinci gruptaki müşterilerin hem tahmini hem de geri kazanılması oldukça zordur. Finansal sebeplerden işletmeyi terk eden müşterilerin finansal durumları düzeldiğinde işletmeye geri dönme ihtimalleri yüksektir. Bu sebeple isteyerek ve finansal olmayan sebeplerden ötürü ayrılan müşterilerin analizi yapılmalı, elde edilen çıkarımlar hem kaybedilen müşterilerin geri kazanılmasında hem de işletmeyi terk etme riski taşıyan müşterilerin tespitinde kullanılmalıdır. Buradan yola çıkarak gelecek çalışması olarak son gruba odaklanmış ve RFM bölümlemesinden farklı olarak finansal olmayan nedenleri de dikkate alan bir deney çalışması hedeflenmiştir.

Kaynakça

[1] Harvard Business Analytics Program Blog. 2021. Business Intelligence vs. Business Analytics. https://analytics.hbs.edu/blog/business-intelligence-vs-business-analytics (Erişim Tarihi: 26.09.2021).

[2] Patricia, M.W., Brockett, P.L., Golden, L.L. 1997. A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice, Marketing Science, Cilt. 16(4), s. 370–391. DOI: 10.1287/mksc.16.4.370

[3] Eiben, A.E., Koudijs, A.E., Slisser, F. 1998. Genetic Modelling of Customer Retention, EuroGP 1998: Genetic Programming, Cilt. 1391, s. 178–186. DOI: 10.1007/BFb0055937

[4] Madden, G.G., Savage, S.J., Coble-Neal, G. 1999. Subscriber Churn in the Australian ISP Market, Information Economics and Policy, Cilt. 11, s. 195–207. DOI: 10.1016/S0167-6245(99)00015-3

[5] Datta, P., Masand, B., Mani, D.R., Li, B. 2000. Automated Cellular Modeling and Prediction on a Large Scale, Artificial Intelligence Review, Cilt. 14, s. 485–502. DOI: 10.1023/A:1006643109702

[6] Koçoğlu, F.Ö., Özcan, T., Baray, Ş.A. 2016. Veri Madenciliğinde Ayrılan Müşteri Analizi Problemi Üzerine Bir Literatür Araştırması. Üretim Araştırmaları Sempozyumu (ÜAS 2016), 12-14 Ekim, İstanbul, Türkiye, 868–874.

[7] Huang, B., Kechadi, M.T., Buckley B. 2012. Customer Churn Prediction in Telecommunications, Expert Systems with Applications, Cilt. 39(1), s. 1414–1425. DOI: 10.1016/j.eswa.2011.08.024

[8] Xie, Y., Li, X., Ngai, E.W.T., Ying, W. 2009. Customer Churn Prediction Using Improved Balanced Random Forests, Expert Systems with Applications, Cilt. 36(3-Part 1), s. 5445–5449. DOI: 10.1016/j.eswa.2008.06.121

[9] Tsai, C.–F., Lu, Y.–H. 2009. Customer Churn Prediction by Hybrid Neural Networks, Expert Systems with Applications, Cilt. 36(10), s. 12547–12553. DOI: 10.1016/j.eswa.2009.05.032

[10] Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.Ch. 2015. A Comparison of Machine Learning Techniques for Customer Churn Prediction, Simulation Modelling Practice and Theory, Cilt. 55, s. 1–9. DOI: 10.1016/j.simpat.2015.03.003

[11] Burez, J., Van den Poel, D. 2009. Handling Class Imbalance in Customer Churn Prediction, Expert Systems with Applications, Cilt. 36(3-Part 1), s. 4626–4636. DOI: 10.1016/j.eswa.2008.05.027

[12] Verbeke, W., Martens, D., Mues, C., Baesens, B. 2011. Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques, Expert Systems Applications, Cilt. 38(3), s. 2354–2364. DOI: 10.1016/j.eswa.2010.08.023

[13] Xia, G.-E., Jin, W.-D. 2008. Model of Customer Churn Prediction on Support Vector Machine, Systems Engineering – Theory & Practice, Cilt. 28(1), s. 71–77. DOI: 10.1016/S1874-8651(09)60003-X

[14] Verbeke, W., Martens, D., Baesens, B. 2014. Social Network Analysis for Customer Churn Prediction, Applied Soft Computing, Cilt. 14(Part C), s. 431–446. DOI: 10.1016/j.asoc.2013.09.017

[15] Lu, N., Lin, H., Lu, J., Zhang, G. 2014. A Customer Churn Prediction Model in Telecom Industry Using Boosting, IEEE Transactions on Industrial Informatics, Cilt. 10(2), s. 1659–1665. DOI: 10.1109/TII.2012.2224355

[16] Caigny, A.D., Coussement, K., De Bock, K.W. 2018. A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees, European Journal of Operational Research, Cilt. 269(2), s. 760–772. DOI: 10.1016/j.ejor.2018.02.009

[17] Khan, A.A., Jamwal, S., Sepehri, M.M. 2010. Applying Data Mining to Customer Churn Prediction in an Internet Service Provider, International Journal of Computer Applications, Cilt. 9(7), s. 8–14. DOI: 10.5120/1400-1889

[18] De Bock, K.W., Van den Poel, D. 2011. An Empirical Evaluation of Rotation-Based Ensemble Classifiers for Customer Churn Prediction, Expert Systems with Applications, Cilt. 38(10), s. 12293–12301. DOI: 10.1016/j.eswa.2011.04.007

[19] Mishra, A., Reddy, U.S. 2017. A Novel Approach for Churn Prediction Using Deep Learning, 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 14-16 December, Coimbatore, India, 1–4. DOI: 10.1109/ICCIC.2017.8524551

[20] Kim, S., Choi, D., Lee, E., Rhee, W. 2017. Churn Prediction of Mobile and Online Casual Games Using Play Log Data, PLoS ONE, Cilt. 12(7):e0180735, s. 1-19. DOI: 10.1371/journal.pone.0180735

[21] Spanoudes, P., Nguyen, T. 2017. Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors, Machine Learning (cs.LG), s. 1–22. arXiv:1703.03869

[22] Bhattacharya, C.B. 1998. When Customers are Members: Customer Retention in Paid Membership Contexts, Journal of the Academy of Marketing Science, Cilt. 26(1), s. 31–44. DOI: 10.1177/0092070398261004

[23] Lariviere, B., Van den Poel, D. 2004. Investigating the Role of Product Features in Preventing Customer Churn, By Using Survival Analysis and Choice Modeling: The Case of Financial Services, Expert Systems with Applications, Cilt. 27, s. 277–285. DOI: 10.1016/j.eswa.2004.02.002

[24] Greis, N.P., Gilstein, C.Z. 1991. Empirical Bayes Methods for Telecommunications Forecasting, International Journal of Forecasting, Cilt. 7(2), s. 183–197. DOI: 10.1016/0169-2070(91)90053-X

[25] Wong, K.K.-K. 2011. Using Cox Regression to Model Customer Time to Churn in The Wireless Telecommunications Industry, Journal of Targeting, Measurement and Analysis for Marketing, Cilt. 19(1), s. 37–43. DOI: 10.1057/jt.2011.1

[26] Fayyad, U. 1997. Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases, 9th International Conference on Scientific and Statistical Database Management, 11-13 August, Olympia, WA, USA, 2–11. DOI: 10.1109/SSDM.1997.621141

[27] Maimon, O., Rokach, L. 2005. Introduction to Knowledge Discovery in Databases. ss 1–17. Maimon, O., Rokach, L., eds. 2005. Data Mining and

Knowledge Discovery Handbook, Springer, Boston, MA, USA, 1383s.

[28] Hox, J., Boeije, H.R. 2005. Data Collection, Primary versus Secondary, Encyclopedia of Social Measurement, s. 593–599. DOI: 10.1016/B0-12-369398-5/00041-4

[29] Han, J., Kamber, M., Pei, J. 2011. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, USA, 744s.

[30] Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V. 2018. Introduction to Data Mining. 2nd Edition. Pearson, USA, 864s.

[31] Liao, S.-H., Chu, P.-H., Hsiao, P.-Y. 2012. Data Mining Techniques and Applications – A Decade Review from 2000 to 2011, Expert Systems with Applications, Cilt. 39(12), s. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063

[32] Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., Sun, X. 2011. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature, Decision Support Systems, Cilt. 50(3), s. 559–569. DOI: 10.1016/j.dss.2010.08.006

[33] Hossin, M., Sulaiman, M.N. 2015. A Review on Evaluation Metrics for Data Classification Evaluations, International Journal of Data Mining & Knowledge Management Process, Cilt. 5(2), s. 1–11. DOI: 10.5121/ijdkp.2015.5201

[34] Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A. 2015. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Cilt. 13(5), s. 971–989. DOI: 10.1109/TCBB.2015.2478454

[35] Jin, X., Xu, A., Bie, R., Guo, P. 2006. Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles, BioDM 2006: Data Mining for Biomedical Applications, Cilt. 3916, s. 106–115. DOI: 10.1007/11691730_11

[36] Segal, T. 2021. Recency, Frequency, Monetary Value. https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp (Erişim Tarihi: 22.09.2021).

[37] Ibitoye, A., Onime, C., Zaki, N.D., Onifade, O.F.W. 2022. Socio-Transactional Impact of Recency, Frequency, and Monetary Features on Customers' Behaviour in Telecoms' Churn Prediction, Iraqi Journal for Computer Science and Mathematics, Cilt. 3(2), s. 101-110. DOI: 10.52866/ijcsm.2022.02.01.011

[38] IBM SPSS Statistics Documentation. 2021. RFM Binning. https://www.ibm.com/docs/en/spss-statistics/28.0.0?topic=marketing-rfm-analysis (Erişim Tarihi: 24.09.2021).

[39] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F. 2011. An Overview of Ensemble Methods for Binary Classifiers in Multi-Class Problems: Experimental Study on One-Vs-One and One-Vs-All Schemes, Pattern Recognition, Cilt. 44(8), s. 1761–1776. DOI: 10.1016/j.patcog.2011.01.017

[40] Rokach, L., Maimon, O. 2005. Decision Trees. ss 165–192. Maimon, O., Rokach, L., eds. 2005. Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA, USA, 1383s.

[41] Press, S.J., Wilson, S. 1978. Choosing Between Logistic Regression and Discriminant Analysis, Journal of the American Statistical Association, Cilt. 73(364), s. 699–705. DOI: 10.1080/01621459.1978.10480080

[42] Hoare, Z. 2008. Landscapes of Naïve Bayes Classifiers, Pattern Analysis & Applications, Cilt. 11(1), s. 59–72. DOI: 10.1007/s10044-007-0079-5

[43] Cortes, C., Vapnik, V.N. 1995. Support-vector Networks, Machine Learning, Cilt. 20(3), s. 273–297. DOI: 10.1007/BF00994018

[44] Rokach, L. 2010. Ensemble-based Classifiers, Artificial Intelligence Review, Cilt. 33, s. 1–39. DOI: 10.1007/s10462-009-9124-7

[45] Jain, A.K., Mao, J., Mohiuddin, K.M. 1996. Artificial Neural Networks: A Tutorial, Computer, Cilt. 29(3), s. 31–44. DOI: 10.1109/2.485891

[46] Taherdoost, H. 2016. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research, International Journal of Academic Research in Management (IJARM), Cilt. 5(2), s. 18–27. DOI: 10.2139/ssrn.3205035

[47] Tharwat, A. 2021. Classification Assessment Methods, Applied Computing and Informatics, Cilt. 17(1), s. 168–192. DOI: 10.1016/j.aci.2018.08.003

[48] Schmitt, J. 1999. Churn: Can Carriers Cope? Skyrocketing Subscriber Defections Have Carriers Worldwide Seeking New Churn Solutions, Telecommunication North American Edition, s. 32–33.