Bankacılık Sektöründe Müşteri Segmentasyonu ile Müşteri Kaybı (Churn) Tahmini: Makine Öğrenmesi Yaklaşımı

Dilek Altaş KARACA[1] **Hülya ERDOĞDU**[2]

[1]Prof.Dr., Marmara Üniversitesi, İktisat Fakültesi, Ekonometri Bölümü, dilekaltas@marmara.edu.tr, ORCID: 0000-0001-5103-9018

[2]Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstatistik Bilim Dalı, hulyaerdogdu@marun.edu.tr, ORCID:0009-0009-0418-9801

Özet: Bu çalışmada, bankacılık sektöründe müşteri kaybı (churn) tahmini, müşteri segmentasyonu temelinde makine öğrenmesi algoritmalarıyla analiz edilmiştir. Araştırmanın amacı, müşterilerin davranışsal özelliklerine dayalı olarak churn risklerini önceden öngörmek ve segmentasyon bilgisinin model performansına katkısını değerlendirmektir. Çalışmada, açık kaynaklı bir bankacılık veri seti kullanılmış ve veri ön işleme aşamalarında aykırı değer analizi, değişken dönüşümleri ve ölçekleme işlemleri gerçekleştirilmiştir. Ardından, müşteri segmentasyonu amacıyla RFM (Recency, Frequency, Monetary) analizi yapılmış ve K-Means algoritması ile müşteriler benzer özelliklerine göre kümelere ayrılmıştır. Segment bilgisi daha sonra tahmin modellerine bağımsız değişken olarak eklenmiştir. Churn değişkeninin dengesiz dağılımı nedeniyle SMOTE yöntemi uygulanarak eğitim verisinde sınıf dengesi sağlanmıştır. Müşteri kaybı tahmini için Logistic Regression, Random Forest, XGBoost, LightGBM ve CatBoost algoritmaları uygulanmıştır. Modellerin performansı doğruluk, duyarlılık, kesinlik, F1-Skoru ve ROC-AUC metrikleri ile değerlendirilmiştir. Analiz bulgularına göre, CatBoost modeli %85 ROC-AUC değeri ile churn sınıfını en iyi ayırt eden algoritma olarak öne çıkarken, LightGBM modeli %83.33 doğruluk ve %0.59 F1-Skoru ile en dengeli performansı göstermiştir. Logistic Regression modeli yüksek duyarlılığa rağmen düşük kesinlik nedeniyle sınırlı başarı sunmuştur. Elde edilen sonuçlar doğrultusunda, müşteri segmentasyonu bilgisinin churn tahmin modellerinin başarımını anlamlı düzeyde artırdığı tespit edilmiştir. Bu kapsamda, alternatif hipotez kabul edilmiş ve segmentasyonun makine öğrenmesi uygulamalarında stratejik bir değişken olarak değerlendirilmesi önerilmiştir.

Anahtar Kelimeler: Müşteri Kaybı(Churn), Müşteri Segmentasyonu, RFM, K-Means, SMOTE, Makine Öğrenmesi, ROC-AU

Customer Churn Prediction Through Customer Segmentation In The Banking Sector: A Machine Learning Approach

Abstract: This study analyzes customer churn prediction in the banking sector using machine learning algorithms based on customer segmentation. The main objective of the research is to predict churn risks in advance based on customers' behavioral characteristics and to evaluate the contribution of segmentation information to model performance. An open-source banking dataset was used, and during the data preprocessing phase, outlier analysis, variable transformation, and scaling procedures were conducted. Subsequently, an RFM (Recency, Frequency, Monetary) analysis was performed for customer segmentation, and customers were clustered using the K-Means algorithm based on their similar characteristics. The segment information was then included as an independent variable in the prediction models. Due to the imbalanced distribution of the churn variable, the SMOTE method was applied to ensure class balance in the training data. Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost algorithms were implemented for churn prediction. Model performance was evaluated using accuracy, recall, precision, F1-score, and ROC-AUC metrics. According to the analysis results, the CatBoost model stood out as the best-performing algorithm in distinguishing the churn class with an ROC-AUC score of 85%, while the LightGBM model exhibited the most balanced performance with 83.33% accuracy and an F1-score of 0.59. Although the Logistic Regression model had high recall, it demonstrated limited success due to its low precision. The findings indicate that customer segmentation information significantly enhances the performance of churn prediction models. Accordingly, the alternative hypothesis was accepted, and it is recommended to consider segmentation as a strategic variable in machine learning applications.

Keywords: Customer Churn, Customer Segmentation, RFM, K-Means, SMOTE, Machine Learning, ROC-AUC

1. GİRİŞ

Dijitalleşmenin yaygınlaşması ve rekabetin artması ile birlikte müşteri bağlılığını sürdürmek, bankacılık sektöründe hayati öneme sahip bir konu haline gelmiştir. Yeni müşteri kazanımı yüksek maliyetli bir süreç olup, mevcut müşterilerin elde tutulması kurumların kârlılığı açısından çok daha önemlidir.

Bu bağlamda, müşteri kaybı (churn) bankalar için yalnızca bir müşteri kaybı değil, aynı zamanda gelecekteki potansiyel gelirin azalması anlamına da gelmektedir.

Literatürde müşteri kaybının tahmini üzerine yapılan birçok çalışmada, çeşitli demografik ve davranışsal özelliklerin churn ile olan ilişkisi incelenmiş ve farklı tahmin modelleri geliştirilmiştir (Verbeke et al., 2012; Larivière & Van den Poel, 2005). Ancak bu çalışmaların birçoğunda, tüm müşteriler homojen bir yapıdaymış gibi ele alınmakta, müşteriler arasındaki farklılıklar göz ardı edilmektedir.

Oysa müşteri davranışları birbirinden oldukça farklılık göstermekte ve bu farklılıkların dikkate alınması churn tahmininde model başarısını artırabilmektedir. Bu noktada müşteri segmentasyonu, yani benzer özelliklere sahip müşterilerin gruplandırılması, tahmin modellerinin doğruluğunu artıran önemli bir yaklaşımdır. Segmentasyon sayesinde her müşteri grubuna özel tahminleme yapılabilmekte, böylece daha hedefli stratejiler geliştirilebilmektedir.

Bu çalışmanın özgün yönü, churn tahmininde müşteri segmentasyonunun modelleme sürecine doğrudan entegre edilmesidir. RFM (Recency, Frequency, Monetary) analizi ile elde edilen davranışsal veriler üzerinden K-Means algoritması ile yapılan segmentasyon, makine öğrenmesi modellerine bağımsız değişken olarak dahil edilmiştir. Bu sayede geleneksel yaklaşımlarda eksik kalan "müşteri çeşitliliğini dikkate alma" problemi giderilmeye çalışılmıştır.

Çalışma, hem farklı makine öğrenmesi algoritmalarının performanslarını karşılaştırmakta hem de segmentasyon temelli yaklaşımların churn tahmininde nasıl bir katkı sunduğunu analiz etmektedir. Bu yönüyle araştırma, bankacılık sektöründe veri odaklı müşteri yönetimi stratejilerinin geliştirilmesine bilimsel katkı sağlamayı amaçlamaktadır.

Çalışmada, RFM (Recency, Frequency, Monetary) analizi ile elde edilen müşteri davranış skorları üzerinden K-Means algoritması kullanılarak segmentasyon gerçekleştirilmiş ve bu segment bilgisi makine öğrenmesi modellerine entegre edilmiştir.

Araştırma kapsamında Logistic Regression, Random Forest, XGBoost, LightGBM ve CatBoost gibi farklı makine öğrenmesi algoritmaları kullanılarak churn tahmini yapılmış ve segment bilgisinin modele dahil edilmesinin, tahmin başarısına katkısı analiz edilmiştir.

Müşteri bağlılığını sürdürmek, çağdaş bankacılık uygulamalarında yalnızca rekabet avantajı değil, aynı zamanda sürdürülebilir kârlılık için de temel bir strateji haline gelmiştir. Ancak müşteri davranışları oldukça çeşitlidir ve bu durum, müşteri kaybı (churn) tahmininde kullanılacak modellerin karmaşık hale gelmesine neden olmaktadır. Bu araştırma, müşteri segmentasyonu gibi davranışsal ayrımları göz önünde bulundurarak, churn tahmini sürecine yeni bir boyut kazandırmayı amaçlamaktadır. Bu çalışmanın teorik önemi, müşteri segmentasyonu ile tahmin modelleri arasındaki etkileşimi analiz etmesidir. RFM analizi ve K-Means kümeleme algoritması ile oluşturulan segment bilgisi, geleneksel churn tahmin modellerine entegre edilmiştir. Böylece, müşteri heterojenliği veri düzeyinde dikkate alınarak tahmin performansının iyileştirilip iyileştirilemeyeceği değerlendirilmiştir. Bu durum, veri temelli karar verme süreçlerinin daha hassas ve kişiselleştirilmiş hale getirilmesine katkı sağlamaktadır. Uygulama açısından bu araştırma, bankaların müşteri kaybını önlemeye yönelik geliştirecekleri stratejilere doğrudan katkı sunmaktadır. Elde edilen bulgular, müşteri segmentlerine özel risk analizlerinin yapılabilmesini sağlayarak daha etkili kampanya yönetimi, sadakat programları ve erken müdahale mekanizmaları geliştirilmesine olanak tanımaktadır.

Çalışma ayrıca, farklı makine öğrenmesi algoritmalarının churn tahminindeki performanslarını karşılaştırarak, bu modellerin bankacılık sektöründe ne ölçüde başarılı olduklarına dair ampirik bir çerçeve sunmaktadır. Böylece araştırma, ileride yapılacak benzer tahminleme çalışmalarına referans oluşturabilecek nitelikte bilimsel bir temel sağlamaktadır.

Bu araştırmanın alana katkısı, müşteri segmentasyonunun yalnızca pazarlama stratejileri için değil, aynı zamanda kayıp riskinin öngörülmesi gibi kritik bir problemde nasıl işlevsel kullanılabileceğini göstermesidir. Bu yönüyle çalışma, literatürdeki müşteri segmentasyonunu ve churn tahminini ayrı ayrı ele alan birçok çalışmadan farklılaşmakta ve bu iki yapıyı bir araya getirerek literatürdeki bir boşluğu doldurmaktadır.

Çalışma ayrıca, farklı makine öğrenmesi algoritmalarının churn tahminindeki performanslarını karşılaştırarak, bu modellerin bankacılık sektöründe ne ölçüde başarılı olduklarına dair ampirik bir çerçeve sunmaktadır. Böylece araştırma, ileride yapılacak benzer tahminleme çalışmalarına referans oluşturabilecek nitelikte bilimsel bir temel sağlamaktadır.

Bu araştırmanın alana katkısı, müşteri segmentasyonunun yalnızca pazarlama stratejileri için değil, aynı zamanda kayıp riskinin öngörülmesi gibi kritik bir problemde nasıl işlevsel kullanılabileceğini göstermesidir. Bu yönüyle çalışma, literatürdeki müşteri segmentasyonunu ve churn tahminini ayrı ayrı ele alan birçok çalışmadan

farklılaşmakta ve bu iki yapıyı bir araya getirerek literatürdeki bir boşluğu doldurmaktadır.

2. Müşteri Kaybı (Churn) Kavramı

Müşteri kaybı (churn), bir müşterinin belirli bir dönemde hizmet aldığı işletmeyle ilişkisini sonlandırması durumunu ifade etmektedir. Bankacılık, telekomünikasyon, sigortacılık ve perakende gibi rekabetin yoğun olduğu sektörlerde müşteri kaybı, işletmelerin gelir kaybı yaşamasına ve yeniden müşteri kazanımı için ek maliyetler oluşmasına neden olmaktadır. Özellikle bankacılık sektöründe, bir müşterinin kaybedilmesi sadece mevcut gelirin değil, aynı zamanda uzun vadeli müşteri değeri ve çapraz satış fırsatlarının da kaybı anlamına gelmektedir (Hwang et al., 2004).

Literatürde churn kavramı farklı yaklaşımlarla ele alınmıştır. Ngai et al. (2009), müşteri kaybını geçmiş davranışlara dayalı olarak gelecekte hizmet alımını durdurma olasılığı olarak değerlendirmiştir. Bu tanımlar, churn'ün hem geçmiş veriye dayalı bir tahmin süreci olduğunu hem de pazarlama stratejileri açısından önleyici eylemlerle ilişkilendirilmesi gerektiğini ortaya koymaktadır.

Churn analizi, şirketlerin müşteri bağlılığı stratejileri geliştirmelerine olanak tanımakta ve müşteri ilişkileri yönetiminin (CRM) temel yapı taşlarından biri olarak değerlendirilmektedir. Bu kapsamda churn tahmini, veri madenciliği ve makine öğrenmesi teknikleri ile desteklenmekte ve müşterilerin hizmeti ne zaman ve neden bırakabileceğine dair erken uyarı mekanizmaları geliştirilmesine katkı sağlamaktadır (Coussement & Van den Poel, 2008).

Bu çalışmada churn, banka müşterisinin mevcut hizmet sağlayıcısı ile olan ilişkisinin sonlandığı durumu ifade etmektedir. Araştırma kapsamında kullanılan veri seti, müşteri kaybını ikili sınıf (0: churn değil, 1: churn) şeklinde etiketlemekte; bu durum, sınıflandırma modelleri ile tahminleme yapılmasını mümkün kılmaktadır. Böylece bankalar churn riski yüksek müşterileri önceden tespit edebilir ve uygun aksiyonlar geliştirerek müşteri bağlılığını artırabilir.

2.1. Müşteri Segmentasyonu ve RFM Yaklaşımı

Müşteri segmentasyonu, işletmelerin geniş müşteri kitlesini benzer özellikler gösteren alt gruplara ayırarak bu gruplara yönelik daha etkili ve hedeflenmiş stratejiler geliştirmesini sağlayan bir pazarlama tekniğidir. Segmentasyon, müşteri ihtiyaçlarının, davranışlarının ve potansiyel değerinin farklılaşması nedeniyle önemli hale gelmiştir. Özellikle müşteri ilişkileri yönetimi (CRM) uygulamalarında, doğru segmentasyon; çapraz satış, müşteri tutundurma ve kişiselleştirilmiş hizmet sunma gibi stratejilerin başarısını doğrudan etkilemektedir (Kotler & Keller, 2012).

Segmentasyon sürecinde kullanılan yöntemlerden biri olan RFM analizi (Recency, Frequency, Monetary), müşteri davranışlarının üç temel boyutta ölçülmesine olanak tanır:

Recency (Yenilik): Müşterinin en son ne zaman işlem yaptığı,

Frequency (Sıklık): Belirli bir dönemde kaç kez işlem yaptığı,

Monetary (Harcama): Toplam işlem tutarıdır.

RFM analizi, müşterilerin hizmetle ne kadar aktif olduklarını ve şirkete sağladıkları değeri kolayca ölçebilmek açısından son derece pratik ve güçlü bir yöntemdir. Özellikle büyük veri ortamlarında hızlı segmentasyon yapılmasını sağladığı için pazarlama analitiğinde ve veri bilimi uygulamalarında yaygın şekilde kullanılmaktadır (Hughes, 1994).

Bu çalışmada RFM analizi, banka müşterilerinin davranışsal segmentasyonu amacıyla kullanılmıştır. RFM skorları her müşteri için ayrı ayrı hesaplanmış ve ardından K-Means kümeleme algoritması ile benzer davranışlara sahip müşteriler aynı segment altında gruplanmıştır. Böylece churn tahminine yalnızca bireysel demografik değişkenlerle değil, aynı zamanda müşterinin bankayla olan davranışsal ilişkisini yansıtan segment bilgisi de dâhil edilmiştir.

RFM tabanlı segmentasyon, churn tahmininde yalnızca bireysel bazda değil, grup davranışı bazında da öngörüler yapılmasını mümkün kılmakta; bu sayede bankalar, yüksek riskli segmentlere özel stratejiler geliştirebilmektedir.

2.2. Makine Öğrenmesi Yaklaşımı

Makine öğrenmesi (machine learning), verilerden örüntüler ve ilişkiler öğrenerek tahminleme ya da sınıflandırma gibi görevleri otomatik biçimde gerçekleştiren algoritmalar bütünüdür. Özellikle büyük veri çağında, geleneksel istatistiksel modellerin ötesine geçerek daha esnek, güçlü ve yüksek doğruluk oranlarına sahip çözümler sunmaktadır. Bankacılık, sağlık, e-ticaret, telekomünikasyon gibi birçok sektörde yaygın olarak kullanılmakta; müşteri davranışı tahmini, dolandırıcılık tespiti ve kredi risk skorlama gibi alanlarda etkin sonuçlar sağlamaktadır (Kelleher et al., 2015).

Makine öğrenmesi algoritmaları temelde üç ana gruba ayrılmaktadır :

Denetimli öğrenme (supervised learning) – Giriş ve çıkış verileri bilinir. Örnek: Sınıflandırma (classification) ve regresyon (regression).

Denetimsiz öğrenme (unsupervised learning) – Çıkış bilgisi yoktur, örüntüler ya da gruplar bulunmaya çalışılır. Örnek: Kümeleme (clustering).

Bu çalışmada hem denetimli hem de denetimsiz öğrenme teknikleri kullanılmıştır: K-Means algoritması ile segmentasyon (denetimsiz), Churn tahmini için ise sınıflandırmaya dayalı denetimli algoritmalar uygulanmıştır.

Makine öğrenmesi, geleneksel regresyon analizlerinden farklı olarak çok sayıda değişkenle çalışabilmekte, karmaşık doğrusal olmayan ilişkileri modelleyebilmekte ve büyük veri setlerinde yüksek doğrulukla sonuç üretebilmektedir. Ayrıca farklı algoritmaların performansları karşılaştırılarak en başarılı modelin seçilmesine de olanak sağlar.

Bu çalışmada kullanılan makine öğrenmesi yaklaşımı, yalnızca churn tahmini yapmakla kalmamakta; aynı zamanda müşterilerin segment bazında değerlendirilmesini sağlayarak daha derinlikli öngörüler sunmaktadır. Böylece veri temelli karar alma süreçleri desteklenmekte, müşteri yönetimi stratejileri bilimsel temellere dayandırılmaktadır.

2.3. Kullanılan Sınıflandırma Algoritmaları

Bu araştırmada, müşteri kaybını tahmin etmek amacıyla farklı sınıflandırma algoritmaları uygulanmıştır. Sınıflandırma algoritmaları, makine öğrenmesinde bağımlı değişkenin

iki veya daha fazla sınıfa ait olup olmadığını belirlemeye yönelik yöntemlerdir. Her algoritma farklı matematiksel temellere dayansa da amaç müşterilerin churn olup olmayacağını yüksek doğrulukla tahmin etmektir. Aşağıda bu çalışmada kullanılan algoritmalar açıklanmıştır

Logistic Regression (LR) : İkili sınıflandırma problemleri için kullanılan istatistiksel bir yöntemdir. Bağımlı değişkenin iki değerli (örneğin churn: evet/hayır) olması durumunda, giriş değişkenleri ile sonuç arasındaki ilişkiyi modellemek için kullanılır. Parametrik bir modeldir ve sonuçlar yorumlanabilirlik açısından avantaj sağlar (Hosmer et al., 2013).

Random Forest (RF) : Birden fazla karar ağacının rastgele alt örneklem ve değişken setleri ile oluşturularak topluca karar vermesine dayanan ansambl (ensemble) bir algoritmadır. Aşırı öğrenmeyi (overfitting) azaltması ve yüksek doğruluk oranları sağlaması nedeniyle tercih edilir. Özellikle çok sayıda bağımsız değişkenin olduğu veri setlerinde etkili sonuçlar verir (Breiman, 2001).

XGBoost (Extreme Gradient Boosting):Boosting algoritmalarından biri olan XGBoost, model hatalarını adım adım azaltmaya yönelik çalışan güçlü bir sınıflandırma yöntemidir. Hızlı çalışması ve paralel işlem kabiliyeti sayesinde büyük veri setlerinde tercih edilmektedir. Kaggle yarışmalarında sıkça birinciliği kazanan modellerin temelini oluşturur (Chen & Guestrin, 2016).

LightGBM (Light Gradient Boosting Machine) : Microsoft tarafından geliştirilen ve XGBoost'un daha hafif, hızlı ve bellek dostu alternatifi olarak tasarlanmış bir boosting algoritmasıdır. Özellikle büyük boyutlu veri setlerinde düşük gecikmeli işlem gerektiren uygulamalarda tercih edilir. Leaf-wise büyüme stratejisi sayesinde daha hassas ayrımlar yapabilir (Ke et al., 2017).

CatBoost : Yandex tarafından geliştirilen bu algoritma, kategorik değişkenleri otomatik olarak işleyebilme kabiliyetiyle öne çıkar. Ön işleme gereksinimini azaltması ve aşırı öğrenmeye karşı dayanıklı olması sayesinde kullanıcı dostu bir sınıflandırma yöntemidir. Özellikle finans ve e-ticaret verileri üzerinde etkili performans göstermektedir (Dorogush et al., 2018).

Bu algoritmalar, veri setine uygulanarak churn tahmini için modeller oluşturulmuş, performansları çeşitli başarı metrikleri (accuracy, recall, precision, f1-score, ROC-AUC) ile karşılaştırılmıştır. Böylece yalnızca tek bir model üzerinden değil, çoklu modeller üzerinden değerlendirme yapılmış ve en uygun algoritma belirlenmiştir.

2.4. İlgili Araştırmalar/Literatür Taraması

Müşteri kaybı (churn) tahmini ve müşteri segmentasyonu, veri bilimi ve pazarlama analitiği alanlarında sıklıkla ele alınan konular arasında yer almaktadır. Son yıllarda yapılan çalışmalar, müşteri davranışlarının modellenmesi, veriye dayalı stratejik karar alma süreçlerinin geliştirilmesi ve makine öğrenmesi algoritmalarının etkinliğinin test edilmesi üzerine yoğunlaşmaktadır. Bu bölümde churn tahmini ile segmentasyon arasındaki ilişkiyi ele alan bazı öncü çalışmalara yer verilmiştir.

Verbeke, Martens, Mues ve Baesens (2012) tarafından gerçekleştirilen çalışmada, telekomünikasyon sektöründe müşteri kaybını tahmin etmeye yönelik kâr odaklı bir veri madenciliği yaklaşımı önerilmiştir. Çalışmada lojistik regresyon, Random Forest ve diğer sınıflandırma

algoritmaları kullanılarak farklı modeller karşılaştırılmış ve pazarlama kararları için önerilerde bulunulmuştur. Model başarıları yalnızca istatistiksel doğrulukla değil, aynı zamanda finansal etkilerle birlikte değerlendirilmiştir (Verbeke et al., 2012).

Larivière ve Van den Poel (2005), banka müşterilerinin kârlılık ve bağlılık düzeylerini Random Forest ve regresyon ormanları aracılığıyla modellemişlerdir. Bu kapsamda, müşteri segmentasyonu ile tahmin modelleri entegre edilerek yüksek doğruluk oranları elde edilmiş; müşteri değeri ile kayıp riski birlikte analiz edilmiştir (Larivière & Van den Poel, 2005).

Coussement ve Van den Poel (2008) ise churn tahmini sürecinde metin madenciliği ve veri madenciliği tekniklerini bir arada kullanarak, müşteri e-posta içerikleri ile davranışsal verileri bütüncül şekilde analiz etmişlerdir. Bu yaklaşım, CRM sistemlerinin entegre veri analitiği potansiyelini ortaya koyması bakımından önem arz etmektedir (Coussement & Van den Poel, 2008).

3. Araştırma Modeli

Bu araştırma, nicel araştırma desenlerinden betimsel modele dayalı olarak yapılandırılmıştır. Mevcut bir veri seti üzerinden, müşteri segmentasyonu ve churn tahmini süreçleri analiz edilmiş; değişkenler arasındaki ilişkiler değerlendirilmiş ve sınıflandırma algoritmaları kullanılarak tahmin modelleri geliştirilmiştir.

Araştırma kapsamında veri madenciliği ve makine öğrenmesi tabanlı bir modelleme yaklaşımı benimsenmiştir. Bu kapsamda, müşterilerin davranışsal özelliklerine göre segmentlere ayrılması ve bu segment bilgilerinin churn tahmini sürecine entegre edilmesi hedeflenmiştir. Segmentasyon için RFM (Recency, Frequency, Monetary) analizine dayalı K-Means kümeleme algoritması kullanılmış; churn tahmini için ise Logistic Regression, Random Forest, XGBoost, LightGBM ve CatBoost algoritmaları uygulanmıştır.

Araştırma modeli, kestirimsel analiz ve model karşılaştırması içeren bir yapıya sahiptir. Her bir algoritmanın churn tahminindeki başarısı accuracy, recall, f1-score ve ROC-AUC gibi performans metrikleri ile değerlendirilmiş; ayrıca segmentasyon değişkeninin modele etkisi analiz edilmiştir.

Bu yönüyle çalışma, veri temelli tahminleme ve sınıflandırma yaklaşımının müşteri yönetimi alanında nasıl uygulanabileceğini ortaya koyan uygulamalı bir araştırma niteliği taşımaktadır.

3.1. Evren ve Örneklem / Veri Seti

Bu araştırmada kullanılan veri seti, veri bilimi projeleri için açık kaynaklı platformlardan biri olan Kaggle üzerinden temin edilmiştir. “Bank Customer Churn Prediction” adlı veri seti, bankacılık sektöründe bireysel müşteri davranışlarını temel alan yapısıyla, müşteri kaybı tahmini amacıyla hazırlanmıştır. Veri seti toplam 10.000 müşteri gözleminden ve 12 değişkenden oluşmaktadır. Araştırmanın evrenini, bir bankanın bireysel müşteri portföyü oluşturmaktadır. Ancak doğrudan saha verisi toplanmadığı için, örneklem bu hazır veri seti ile sınırlıdır. Çalışma grubu, farklı ülkelerden (Fransa, Almanya ve İspanya) müşterileri temsil eden anonimleştirilmiş müşteri kayıtlarından oluşmaktadır.

Bu nedenle, demografik ve davranışsal özellikler üzerinden genel değerlendirme yapılmış, kişisel tanımlayıcı bilgilere yer verilmemiştir.

Bu çalışmada kullanılan değişkenler, veri türleri ve ölçme düzeylerine ilişkin bilgiler Tablo 3.1’de verilmiştir.

Tablo 3. 1 Veri Setine Ait Değişken Tanımları ve Ölçüm Düzeyleri

Değişken Adı	Açıklama	Veri Türü
customer_id	Her müşteriye ait benzersiz kimlik numarası	Metrik (int)
credit_score	Müşterinin kredi puanı	Metrik (int)
country	Müşterinin ülkesi (Fransa, Almanya, İspanya)	Kategorik
gender	Müşterinin cinsiyeti (Kadın/Erkek)	Kategorik
age	Müşterinin yaşı	Metrik (int)
tenure	Bankada geçirilen yıl sayısı	Metrik (int)
balance	Müşterinin banka hesabındaki bakiye	Metrik (float)
products_number	Sahip olunan banka ürün sayısı	Metrik (int)
credit_card	Kredi kartı kullanma durumu (0: Hayır, 1: Evet)	Kategorik(binary)

active_member	Aktif müşteri olma durumu (0: Hayır, 1: Evet)	Kategorik(binary)
estimated_salary	Müşterinin tahmini maaşı	Metrik (float)
churn	Müşteri kaybı durumu (0: Hayır, 1: Evet)	Kategorik(binary)

3.2. Veri Toplama Araçları

Bu araştırmada birincil veri toplanmamış, mevcut bir veri seti kullanılmıştır. Veri, Kaggle platformu üzerinden indirilen ve anonimleştirilmiş bir banka müşterisi veri setidir. Veriler, Python programlama dili ve Pandas kütüphanesi ile işlenmiş, Scikit-learn ve ilgili makine öğrenmesi kütüphaneleri analiz aracı olarak kullanılmıştır

3.3. Verilerin Toplanması

Veri seti, Kaggle üzerinden .csv formatında indirilmiş ve Python ile işlenmiştir. Veriler üzerinde öncelikle eksik değer kontrolü yapılmış, veri tipi dönüşümleri gerçekleştirilmiş, kategorik değişkenler uygun şekilde kodlanmıştır. Özellikle gender, country gibi değişkenlerde Label Encoding ve One-Hot Encoding yöntemleri uygulanmıştır. Veriler daha sonra eğitim ve test kümelerine ayrılarak analiz süreci başlatılmıştır.

3.4. Verilerin Analizi

Bu bölümde, araştırmada kullanılan veri seti üzerinde gerçekleştirilen analiz adımları detaylı biçimde sunulmaktadır. Öncelikle veri setinin genel yapısı, betimsel istatistikler ve eksik/aykırı değer durumları incelenmiştir. Ardından RFM analizi gerçekleştirilmiş ve müşteriler K-Means algoritması yardımıyla segmentlere ayrılmıştır. Son aşamada ise müşteri kaybı (churn) tahmini için çeşitli makine öğrenmesi algoritmaları uygulanmış ve model performansları değerlendirilmiştir. Her analiz adımı ilgili alt başlıklar altında açıklanmıştır.

3.4.1. Betimsel İstatistikler

Veri analiz sürecinin ilk aşamasında, veri setinde yer alan değişkenlerin genel yapısı incelenmiş ve temel betimsel istatistikler değerlendirilmiştir. İnceleme sonucunda, veri setinin 10.000 gözlem ve 12 değişkenden oluştuğu, herhangi bir eksik verinin bulunmadığı görülmüştür. Sayısal değişkenlerin dağılımı, merkezi eğilim (ortalama, medyan) ve yayılım (standart sapma, minimum–maksimum) ölçütleriyle özetlenmiştir.

Veri setindeki yaş, kredi skoru, bakiye ve maaş gibi metrik değişkenlerde geniş bir dağılım gözlenmiştir. Örneğin, yaş değişkeni 18 ile 92 arasında değişirken, ortalama yaklaşık 39'dur. Benzer şekilde, hesap bakiyesi değişkeninde 0 TL ile 250.000 TL arasında değişen değerlere rastlanmış, bu durum bazı müşterilerin hiç bakiye bulundurmadığını, bazılarının ise oldukça yüksek tutarlar taşıdığını göstermektedir. Ürün sayısı, kredi kartı kullanımı ve aktif müşteri durumu gibi kategorik değişkenler incelendiğinde ise müşterilerin çoğunlukla 1 veya 2 ürün kullandığı, kredi kartı ve aktif üyelik oranlarının da dengeli dağıldığı gözlemlenmektedir.

Bağımlı değişken olan müşteri kaybı (churn) ikili (0/1) bir yapıya sahiptir ve dağılım incelendiğinde müşterilerin yaklaşık %20'sinin bankadan ayrıldığı, %80'inin ise mevcutta kaldığı tespit edilmiştir.

Sonuç olarak, veri seti analiz öncesi herhangi bir eksik veya ciddi tutarsızlık içermemekte olup, modellemeye uygun ve istatistiksel olarak değerlendirilebilir durumdadır.

Tablo 3. 2 Sayısal Değişkenlere Ait Betimsel İstatistikler

```
>>> df.shape
(10000, 12)
>>> df.describe().T
                   count          mean           std          min  \
customer_id      10000.0  1.569094e+07  71936.186123  15565701.00
credit_score     10000.0  6.505288e+02     96.653299       350.00
age              10000.0  3.892180e+01     10.487806        18.00
tenure           10000.0  5.012800e+00      2.892174         0.00
balance          10000.0  7.648589e+04  62397.405202         0.00
products_number  10000.0  1.530200e+00      0.581654         1.00
credit_card      10000.0  7.055000e-01      0.455840         0.00
active_member    10000.0  5.151000e-01      0.499797         0.00
estimated_salary 10000.0  1.000902e+05  57510.492818        11.58
churn            10000.0  2.037000e-01      0.402769         0.00

                         25%           50%           75%           max
customer_id      15628528.25  1.569074e+07  1.575323e+07  15815690.00
credit_score         584.00  6.520000e+02  7.180000e+02       850.00
age                   32.00  3.700000e+01  4.400000e+01        92.00
tenure                 3.00  5.000000e+00  7.000000e+00        10.00
balance                0.00  9.719854e+04  1.276442e+05    250898.09
products_number        1.00  1.000000e+00  2.000000e+00         4.00
credit_card            0.00  1.000000e+00  1.000000e+00         1.00
active_member          0.00  1.000000e+00  1.000000e+00         1.00
estimated_salary   51002.11  1.001939e+05  1.493882e+05    199992.48
churn                  0.00  0.000000e+00  0.000000e+00         1.00
```

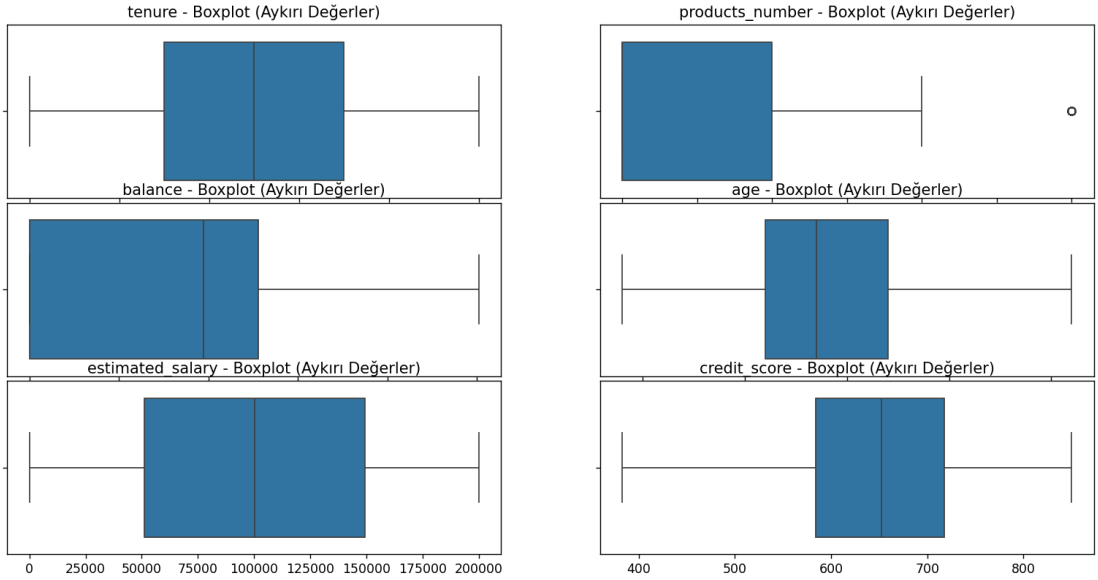
3.4.2. Aykırı Değer ve Normal Dağılım Analizi

Modelleme öncesi yapılan ön analizlerde, bazı sayısal değişkenlerde istatistiksel olarak uçta yer alan değerler tespit edilmiştir. Özellikle credit_score ve age değişkenlerinde aykırı değerler bulunduğu belirlenmiş ve bu gözlemler analiz sonuçlarını bozabilecek potansiyele sahip olduğu için değerlendirmeye alınmıştır.

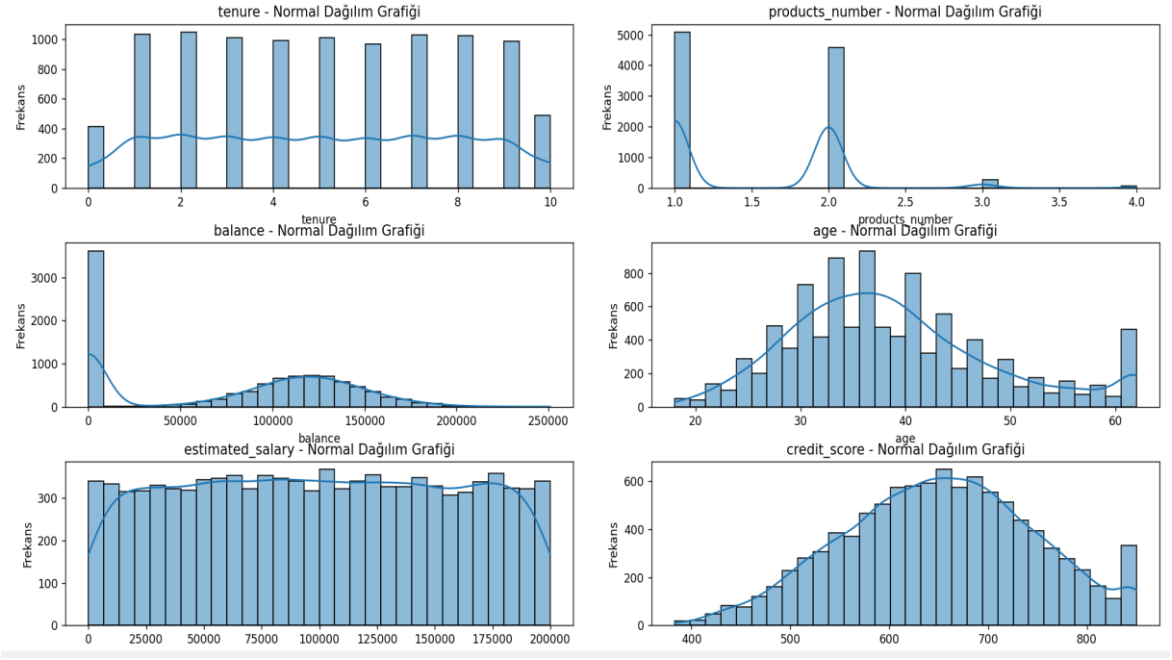
Bu kapsamda, credit_score değişkeninde minimum değerin 383'ün altında olması nedeniyle, daha düşük değerlerin veri bütünlüğünü bozabileceği düşünülmüş ve bu değerin altındaki tüm gözlemler 383 ile sınırlandırılmıştır. Benzer şekilde, age değişkeninde maksimum değerin 62'nin üzerinde olduğu gözlenmiş ve bu durum, yaş dağılımında dengesizlik yaratabileceği için üst sınır 62 olarak belirlenmiştir. Bu işlem, veri setindeki uç değerlerin etkisini azaltmak amacıyla uygulanmış basit bir winsorize yaklaşımıdır.

Uygulanan bu sınırlandırma işlemleri sayesinde, credit_score ve age değişkenlerinde analizde istenmeyen sapmaların önüne geçilmiş ve verilerin daha sağlıklı bir dağılım göstermesi sağlanmıştır. Diğer sayısal değişkenlerde yapılan incelemelerde aykırı değerlerin sınırlı sayıda ve sistematik bir bozulmaya yol açmadığı görülmüş, bu nedenle müdahale edilmeden analize dahil edilmiştir.

Şekil 3. 1 Sayısal Değişkenlere Ait Aykırı Değer Analizi (Boxplot)



Şekil 3. 2 Sayısal Değişkenlere Ait Histogram ve Normal Dağılım Eğrileri



Veri setindeki metrik değişkenlerin dağılımı ve normal eğrileri birlikte gösterilmiştir. Age ve credit_score değişkenleri normal dağılıma yakınken, diğer değişkenlerde çarpıklık gözlemlenmiştir.

3.4.3. Değişken Dönüşümü ve Ön İşleme

Veri setinde yer alan gender ve country değişkenleri, makine öğrenmesi algoritmalarına uygun hale getirilmek amacıyla sayısal forma dönüştürülmüştür. gender değişkeni ikili bir yapıya sahip olduğundan, Female değeri 0, Male değeri ise 1 olarak kodlanmıştır. country değişkeni üç kategoriden oluştuğu için one-hot encoding yöntemi uygulanmış, referans kategori olarak France çıkarılmış ve country_Germany, country_Spain adında iki yeni değişken oluşturulmuştur. Oluşan bu sütunlar, int veri tipine dönüştürülerek modele hazır hale getirilmiştir.

3.5. Müşteri Segmentasyonu

Müşteri segmentasyonu, bir kurumun müşteri kitlesini belirli benzerliklere göre alt gruplara ayırma sürecidir. Bu işlem, farklı müşteri gruplarının davranışsal özelliklerini analiz ederek hedefe yönelik stratejiler geliştirmek amacıyla uygulanır. Pazarlama, müşteri ilişkileri yönetimi ve müşteri yaşam boyu değeri analizi gibi birçok alanda müşteri segmentasyonu kritik bir rol oynamaktadır.

Bu çalışmada müşteri segmentasyonu, iki farklı yöntem kullanılarak gerçekleştirilmiştir. İlk olarak, müşteri davranışlarını tanımlamada yaygın olarak kullanılan RFM (Recency, Frequency, Monetary) analizi uygulanmış ve önceden tanımlı kurallara dayalı olarak müşteriler farklı segmentlere ayrılmıştır. RFM yöntemi, geçmiş müşteri davranışlarına dayalı olarak kullanıcıları belirli davranış kategorilerine yerleştirerek müşteri değerini anlamaya yardımcı olur.

İkinci aşamada ise daha dinamik ve veri odaklı bir yaklaşım olan K-Means kümeleme algoritması kullanılmıştır. Bu algoritma, benzer özelliklere sahip müşteri gruplarını otomatik olarak belirleyerek daha esnek bir segmentasyon sağlar. RFM skorları üzerinden standartlaştırma işlemi gerçekleştirilmiş ve benzer müşteri profilleri kümelenmiştir.

Bu iki yöntem bir arada kullanılarak hem kural tabanlı hem de veri odaklı segmentasyon perspektifleri değerlendirilmiş; müşteri davranışlarının farklı açılardan analiz edilmesi sağlanmıştır. Böylece hem yorumlanabilir hem de model odaklı güçlü bir segmentasyon yapısı oluşturulmuştur**.**

3.5.1. RFM Analizi ve Skor Hesaplaması

Müşteri segmentasyonu amacıyla bu çalışmada RFM (Recency, Frequency, Monetary) analizi kullanılmıştır. RFM, müşterilerin zamansal ve finansal davranışlarını değerlendirerek gruplandırma yapmaya yarayan yaygın bir yöntemdir. Bu çalışmada her müşteri için üç temel boyutta puanlama yapılmıştır:

Recency (R): Müşterinin bankada ne kadar süredir aktif olduğunu gösterir. Tenure değişkeni bu amaçla kullanılmış ve en yeni müşterilere en yüksek puan (5), en eski müşterilere en düşük puan (1) verilmiştir.

Frequency (F): Müşterinin bankada sahip olduğu ürün sayısını ifade eder. Products_number değişkeni kullanılarak, daha fazla ürün kullanan müşterilere daha yüksek puanlar atanmıştır. Veride aynı ürün sayısına sahip müşterilerin sıralanması için rank() yöntemi kullanılmıştır.

Monetary (M): Müşterinin hesap bakiyesi ile ilişkilidir. Balance değişkeni kullanılarak değerlendirme yapılmıştır. Bakiyesi 0 olan müşterilere otomatik olarak 1 puan verilmiş, geri kalan müşteriler bakiye düzeylerine göre 4 eşit parçaya ayrılarak 2 ile 5 arasında puanlandırılmıştır.

Her müşteriye ait R_score, F_score ve M_score değerleri birleştirilerek üç basamaklı bir RFM_score oluşturulmuştur. Örneğin R=5, F=4, M=3 olan bir müşteri 543 skorunu almıştır.

Oluşturulan RFM_score değerleri yardımıyla müşteriler, davranışsal benzerliklerine göre önceden tanımlanmış segmentlere atanmıştır. Segment atama işlemi, özellikle R_score ve F_score kombinasyonlarına dayalı olarak yapılmış ve aşağıdaki segment tanımları kullanılmıştır:

- Best_Customers: R=5 ve F≥4
- Potential_Loyalists: R=5 ve F=2–3
- Lost_Customers: R=1 ve F≥3
- Loyal_Customers, Moderate_Customers, Low_Engagement gibi çeşitli davranış grupları...

Benzer müşteriye ayrıca, monetary score üzerinden bir değer katmanı atanmıştır (Platinum, Gold, Silver, Bronze, LowValue). Böylece her müşteri için Segment ve Monetary_Tier değerleri birleştirilerek bir Final_Segment oluşturulmuştur.(örn.Loyal_Customer_Gold).

RFM analizinde "Monetary" değişkeni müşterinin bankaya sağladığı finansal katkıyı yansıtmakla birlikte, segmentasyon sürecinde ikincil önceliğe sahiptir. Müşteri bağlılığı doğrudan bakiyeye göre değerlendirildiğinde yanıltıcı sonuçlar ortaya çıkabilir. Örneğin, yalnızca yüksek bakiye tutan ancak aktif olmayan bir müşteri sadık kabul edilemez. Bu nedenle, ilk aşamada müşteri bağlılığına dair sinyaller veren Recency ve Frequency skorlarına göre davranışsal segmentler oluşturulur; ardından Monetary değeri kullanılarak bu segmentlerin finansal önemi değerlendirilir. Bu yaklaşım, RFM analizinin özgün mantığına da uygundur: R ve F skorları müşterinin davranışsal sınıfını belirlerken, M skoru bu sınıfın işletme açısından taşıdığı ekonomik değeri ortaya koyar.

Final Segmentlerin Açıklanması

Aşağıda her bir segmentin genel anlamı açıklanmıştır:

Best_Customers_[Platinum/Gold/Silver/Bronze/LowValue]:

İşlem sıklığı ve yenilik düzeyi yüksek olan, bankaya en sık ve yakın zamanda etkileşimde bulunan müşteriler. Finansal değer düzeyine göre farklılaştırılmıştır. Platinum olanlar en değerli gruptur.

Loyal_Customers_[Platinum/Gold/Silver/Bronze/LowValue]:

Sadakat göstergesi taşıyan, belirli sıklıkla işlem yapan ve genellikle bankayla sürekli ilişkide kalan müşteriler. Finansal katkıları seviyelerine göre değişmektedir.

Moderate_Customers_[Platinum/Gold/Silver/Bronze/LowValue]:

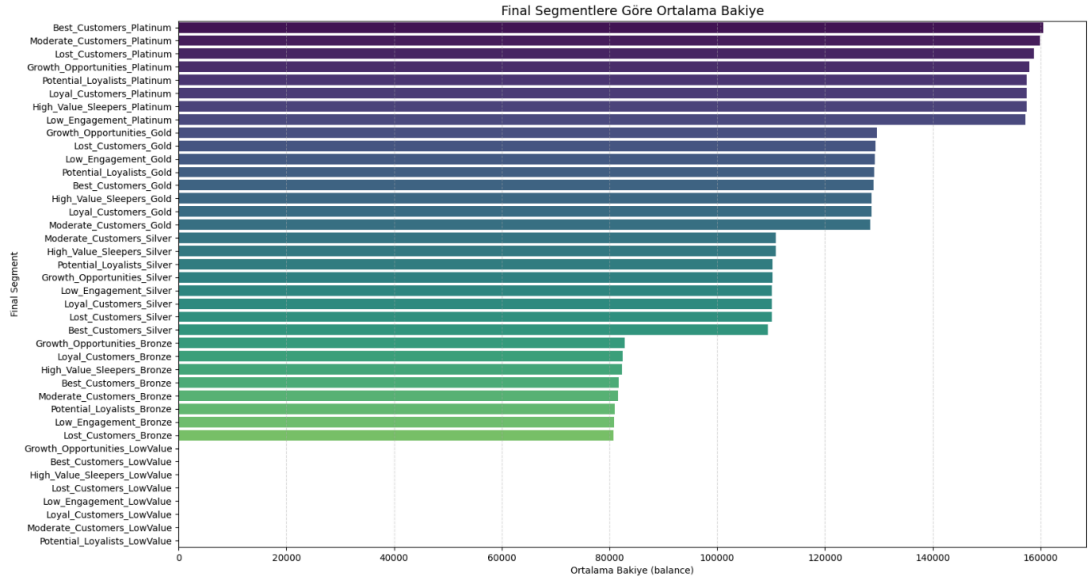
Orta düzeyde işlem sıklığına sahip, ne çok aktif ne de pasif olan dengeli müşteri grubudur. Müşteri potansiyeli segment içinde farklılık gösterir.

High_Value_Sleepers_[Platinum/Gold/Silver/Bronze/LowValue]:

Çok yüksek bakiye tutan ancak işlem sıklığı düşük olan müşterilerdir. Banka için finansal açıdan önemli olmakla birlikte, sadakatleri zayıf olabilir.

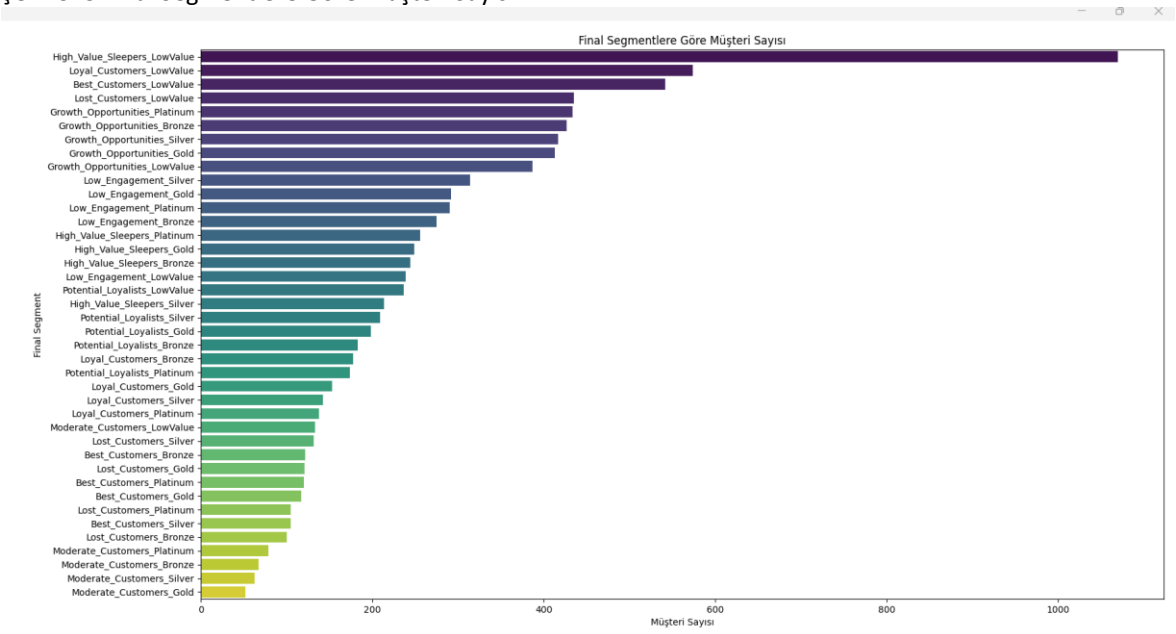
Growth_Opportunities_[Platinum/Gold/Silver/Bronze/LowValue]:
Davranışsal olarak orta düzeyde aktif olan ancak potansiyel barındıran müşteri grubudur. Doğru stratejilerle sadakat artırılabilir.
Potential_Loyalists_[Platinum/Gold/Silver/Bronze/LowValue]:
Son dönemde etkileşimde bulunmuş, işlem sıklığı sınırlı fakat sadakat kazanılabilecek müşteri segmentidir. Finansal değeri seviyelere göre ayrılır.
Lost_Customers_[Platinum/Gold/Silver/Bronze/LowValue]:
Bankayla ilişkisini büyük ölçüde sonlandırmış ya da uzun süredir pasif durumda olan müşterilerdir. Sadakat ve katkı seviyesi segment ismine göre değişmektedir.
Low_Engagement_[Platinum/Gold/Silver/Bronze/LowValue]:
Hem işlem sıklığı hem de son işlem zamanı açısından düşük performans gösteren, bankayla zayıf ilişkili müşteri grubudur.

Şekil 3. 3 Final Segmentlere Göre Ortalama Bakiye



RFM analizine göre oluşturulan segmentlerin her biri için ortalama bakiye değerleri görselleştirilmiştir. En yüksek bakiyeye sahip segmentler, genellikle "Platinum" düzeyindeki müşterilerden oluşmaktadır. Özellikle Best_Customers_Platinum segmenti, en yüksek ortalama bakiyeye sahip gruptur. Tersine, LowValue grubundaki müşteriler, daha düşük bakiye seviyeleri ile dikkat çekmektedir.

Şekil 3. 3 Final Segmentlere Göre Müşteri Sayısı



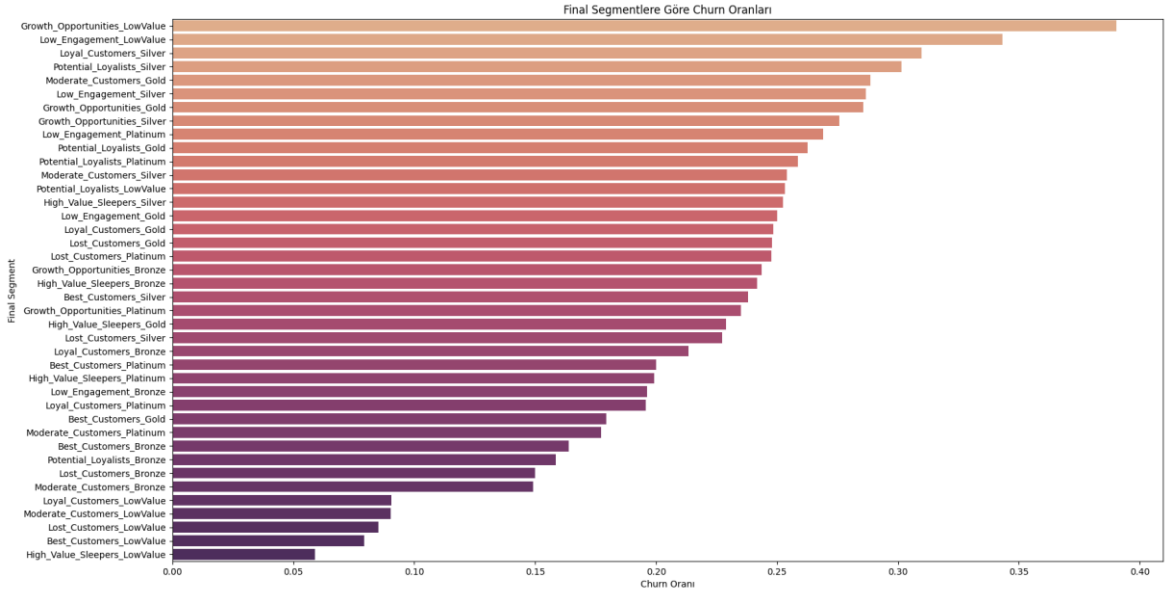
RFM skorları ve segment etiketleri kullanılarak oluşturulan Final_Segment değişkenine göre her segmentte kaç müşteri bulunduğu görselleştirilmiştir. Grafik, müşteri kitlesinin davranış ve değer düzeylerine göre nasıl dağıldığını özetlemektedir.

Segment dağılımına ilişkin grafik incelendiğinde, en kalabalık müşteri grubunun High_Value_Sleepers_LowValue segmenti olduğu görülmektedir. Bu segmentte yer alan müşteriler yüksek işlem sıklığına sahip olmalarına rağmen parasal değer açısından düşük profildedir.

Onu izleyen segmentler olan Loyal_Customers_LowValue,Best_Customers_Low Value ve Lost_Customers_LowValue gibi gruplar da benzer şekilde hem davranışsal hem finansal özellikler bakımından farklılıklara işaret etmektedir. Grafikte dikkati çeken bir başka unsur ise Platinum değer grubuna sahip segmentlerin sayıca daha az olmasıdır. Bu durum, yüksek değerli müşterilerin nicel olarak sınırlı ancak nitelik olarak stratejik öneme sahip olduğunu göstermektedir. Örneğin; Best_Customers_Platinum ve Growth_Opportunities_Platinum segmentleri az sayıda müşteri barındırsa da, bu müşterilerin elde tutulması banka açısından yüksek öneme sahiptir.

Bu analiz, müşteri tabanının hangi davranış ve değer segmentlerinde yoğunlaştığını görselleştirerek, ileride yapılacak pazarlama, çapraz satış ve churn stratejileri açısından yol gösterici bir temel sunmaktadır.

Şekil 3. 4 Final Segmentlere Göre Churn Oranları.



Her müşteri segmentinde yer alan müşterilerin bankadan ayrılma (churn) oranları görselleştirilmiştir. Segmentler arasında churn oranları açısından belirgin farklar olduğu gözlemlenmiştir. Final segmentlere göre churn oranlarının dağılımı incelendiğinde, müşteri kaybı açısından bazı segmentlerin diğerlerine kıyasla daha riskli olduğu gözlemlenmiştir.

Özellikle Growth_Opportunities_LowValue, Low_Engagement_LowValue ve Loyal_Customers_Silver segmentlerinde churn oranlarının %25'in üzerinde olduğu dikkat çekmektedir. Bu segmentlerdeki müşteriler ya düşük parasal değere sahip olsalar da davranışsal potansiyel barındırmakta, ya da sadakat göstergesi taşımasına rağmen elde tutulamama riski taşımaktadır.

Öte yandan High_Value_Sleepers_LowValue, Best_Customers_LowValue ve Lost_Customers_LowValue segmentlerinde churn oranlarının oldukça düşük olduğu görülmektedir. Bu durum, davranışsal olarak aktif olmayan ya da değer düzeyi düşük olan müşteriler arasında sadık bir kitle bulunduğuna işaret etmektedir. Özellikle Best_Customers_Platinum gibi yüksek değerli segmentlerde churn oranının düşük olması, bankanın bu segmentleri başarıyla koruduğunu göstermektedir.

Bu analiz, müşteri segmentlerinin sadece değer değil aynı zamanda risk düzeyi açısından da değerlendirilmesine olanak sağlamış ve ileride uygulanabilecek segment bazlı churn önleme stratejilerine temel oluşturmuştur.

Yapılan analizler sonucunda, segmentlerin sadece parasal değerlerine göre değil, aynı zamanda müşteri kaybı riski açısından da anlamlı farklılıklar gösterdiği görülmüştür. Özellikle Growth_Opportunities_LowValue ve

Low_Engagement_LowValue gibi segmentlerde churn oranları dikkat çekici seviyededir ve bu gruplar müdahale gerektiren kritik müşteri gruplarıdır. Buna karşılık, Best_Customers_Platinum ve Loyal_Customers_Gold gibi yüksek değerli segmentlerde churn oranlarının düşük olması, bu müşterilerin elde tutulmasında başarılı stratejiler uygulandığını göstermektedir.

Segment bazlı analiz, sadece müşteri gruplarını tanımayı değil, aynı zamanda risk ve fırsatları doğru bir şekilde haritalandırmayı da mümkün kılmıştır. Bu bulgular, sonraki aşamada uygulanacak churn tahminleme modellerinde segment bilgisinin kullanılmasının önemini pekiştirmektedir.

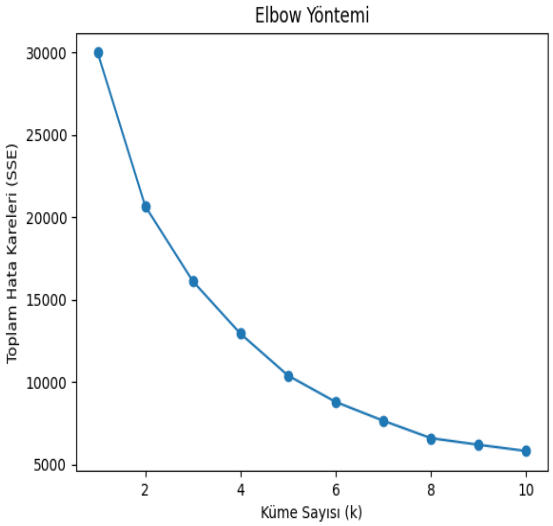
3.5.2. K-Means Algoritması ile Müşteri Kümeleme Analizi

Bu bölümde, müşteriler RFM skorlarına göre gruplandırılarak benzer davranış profiline sahip segmentlerin otomatik olarak belirlenmesi amaçlanmıştır. Bu amaç doğrultusunda, denetimsiz öğrenme algoritmalarından biri olan K-Means kümeleme yöntemi kullanılmıştır. Modelleme sürecine başlamadan önce, RFM skorları standardize edilmiş, ardından optimal küme sayısı Elbow (dirsek) yöntemi ile belirlenmiştir. Elde edilen kümeler, müşterilerin davranışsal benzerlikleri üzerinden anlamlandırılarak etiketlenmiştir.

RFM skorları R_score, F_score ve M_score değişkenlerinden oluşmaktadır. K-Means algoritması, mesafeye dayalı bir yöntem olduğundan değişkenlerin aynı ölçekte olması gerekmektedir. Bu nedenle değişkenler StandardScaler kullanılarak normalize edilmiştir.

Optimal küme sayısını belirlemek amacıyla 1'den 10'a kadar farklı küme sayıları için toplam hata kareleri (SSE - Sum of Squared Errors) hesaplanmıştır. Bu değerler görselleştirilerek analiz edilmiştir. Elbow grafiğinde k = 4 noktasında belirgin bir kırılma (dirsek) gözlemlenmiş ve bu değer optimal küme sayısı olarak kabul edilmiştir.

Şekil 3. 5 Elbow Yöntemi ile Optimal Küme Sayısının Belirlenmesi



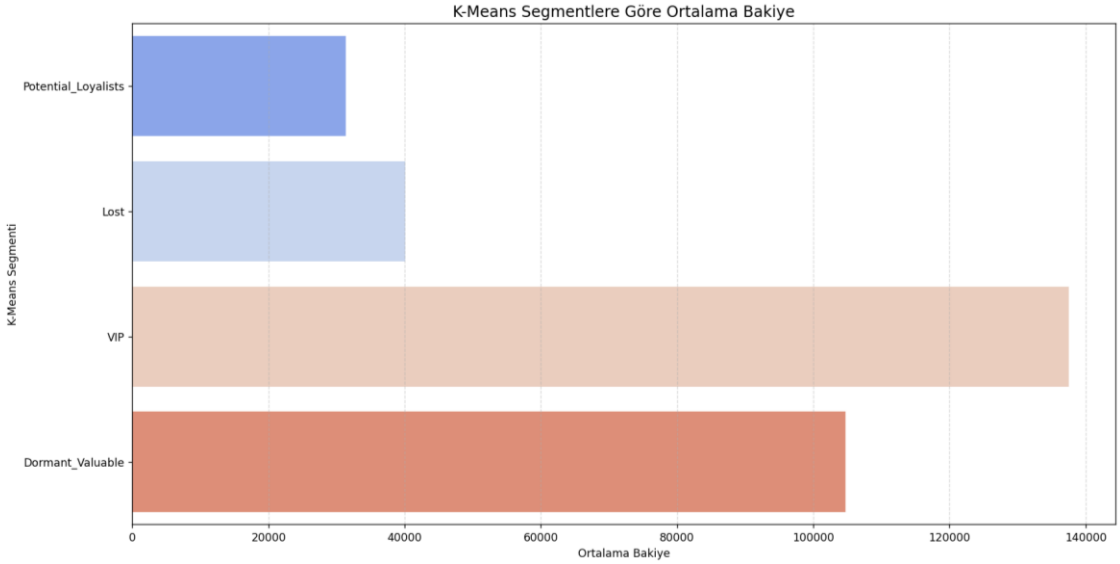
Bu grafikte x ekseninde küme sayısı, y ekseninde ise her bir modelin toplam hata kareleri (SSE) yer almaktadır. 4. kümede belirgin bir kırılma noktası gözlenmiştir.

Elbow yöntemi sonucunda k = 4 olarak belirlenen optimal küme sayısı ile K-Means modeli eğitilmiş ve her müşteri gözlemine ait segment belirlenmiştir. Segmentler, ortalama RFM skorlarına göre yorumlanarak etiketlenmiştir:

Tablo 3. 3 K-Means Algoritması ile Oluşturulan Müşteri Segmentleri ve Açıklamaları

K_Means_ Segment	Segment Etiketi	Açıklama
0	Dormant_Valuable	Düşük etkileşimli fakat finansal değeri yüksek
1	Potential_Loyalists	Sadakat potansiyeli taşıyan, gelişime açık
2	Lost	Etkileşimi ve katkısı düşük, terk riski yüksek
3	VIP	Yüksek sadakat ve finansal değere sahip en önemli grup

Şekil 3. 6 K-Means Segmentlerine Göre Ortalama Müşteri Bakiyesi



Grafikte, her bir müşteri segmentine ait ortalama banka bakiyesi yatay çubuk grafik ile gösterilmiştir. VIP ve Dormant_Valuable segmentleri en yüksek finansal değeri taşıyan gruplardır. Lost ve Potential_Loyalists segmentleri ise daha düşük ortalama bakiye düzeyine sahiptir.

Tablo 3. 4 Segmentlere Göre RFM Skorları, Ürün Sayısı, Bakiye ve Müşteri Dağılımı

K_Means_Label	R_score	F_score	M_score	Products Number	Balance	Müşteri Sayısı
Dormant_Valuable	1,93	1,8	3,19	1,08	104.694	2415
Lost	1,98	4,29	1,76	2,02	40.112	2388
Potential_Loyalists	4,39	3,34	1,43	1,67	31.368	2812
VIP	4,29	2,51	4,21	1,32	137.539	2385

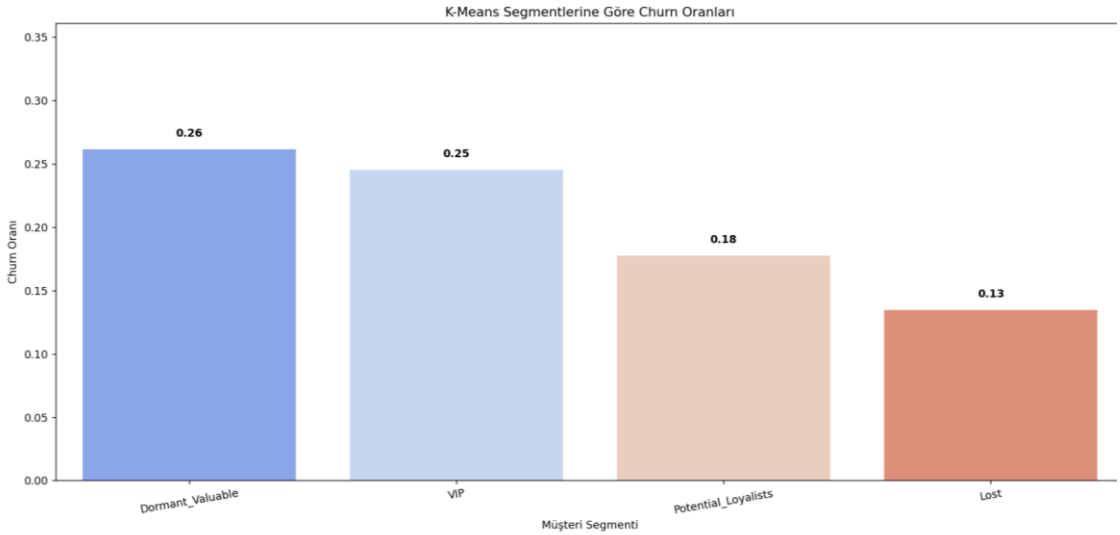
Dormant_Valuable segmenti, düşük etkileşim düzeyine (R=1.93, F=1.80) rağmen yüksek parasal katkı (M=3.19) ve ortalama 104.694 TL bakiye ile dikkat çekmektedir. Bu müşteriler finansal açıdan değerli olup, pasif davranış profilleri nedeniyle yeniden kazanım stratejileri kapsamında önceliklendirilmelidir.

Lost segmentinde yer alan müşteriler geçmişte yüksek sıklıkla işlem yapmış (F=4.29) ancak güncel etkileşimleri oldukça düşüktür (R=1.98). Parasal katkılarının (M=1.76) azalması, bu grubun hizmetten uzaklaştığını göstermektedir. Sadakat geçmişi göz önünde bulundurularak yeniden etkileşim sağlanabilir.

Potential_Loyalists segmenti, bankayla son dönemde sıkça etkileşimde bulunan (R=4.39, F=3.34) ancak düşük bakiye seviyesine (M=1.43, ortalama 31.368 TL) sahip müşterileri kapsamaktadır. Bu grup, doğru ürün ve kampanya eşleştirmeleriyle yüksek değerli müşterilere dönüştürülebilecek potansiyel taşımaktadır.

VIP segmenti, hem yakın zamanda etkileşimde bulunan (R=4.29) hem de yüksek parasal değer sunan (M=4.21, ortalama bakiye 137.539 TL) müşterilerden oluşmaktadır. Banka açısından en stratejik grubu oluşturan bu müşterilerin elde tutulması kritik öneme sahiptir.

Şekil 3. 7 K-Means Segmentlerine Göre Churn Oranları



Grafikte her bir müşteri segmentine ait müşteri kaybı (churn) oranı görselleştirilmiştir. En yüksek churn oranı %26 ile Dormant_Valuable segmentinde gözlemlenmiştir. Bu segmentin düşük etkileşime rağmen yüksek bakiyeye sahip olması, müşteri kaybının finansal etkisini artırmaktadır. VIP segmenti ise %25 churn oranı ile ikinci sırada yer almakta, sadık ve değerli müşterilerin dahi risk altında olabileceğini göstermektedir. Potential_Loyalists segmenti %18 churn oranıyla gelişime açık grubu temsil ederken, Lost segmenti %13 ile en düşük churn oranına sahip gruptur.

3.5.3 Segmentasyonun Derinleştirilmesi ve Modelleme Sürecine Entegrasyonu

Bu çalışmada müşteri segmentasyonu iki farklı yöntemle yapılmıştır: İlki klasik RFM analizine dayalı segmentler, ikincisi ise K-Means algoritmasıyla oluşturulan kümelerdir. RFM yöntemiyle müşteriler davranışsal skorlarına (Recency, Frequency, Monetary) göre oldukça detaylı şekilde sınıflandırılmış ve yaklaşık 40'tan fazla farklı segment oluşmuştur. Bu detaylı yapı, müşterileri tanımak açısından faydalı olsa da, çok sayıda küçük gruba ayrılması nedeniyle analizlerin ve özellikle modellemelerin içinde kullanılması zor hale gelmiştir.

Bu nedenle ikinci aşamada, aynı RFM skorları kullanılarak K-Means algoritmasıyla daha sade bir kümeleme yapılmıştır. Bu yöntemle müşteriler benzer davranış kalıplarına göre sadece 4 kümeye ayrılmış ve bu gruplara anlamlı etiketler verilmiştir (örneğin: VIP, Lost gibi). Böylece hem segment yapısı sadeleşmiş hem de bu bilgiler modellerde kullanılabilir hale gelmiştir.

Özetle, RFM ile müşterileri detaylı analiz edilebilir halde geldi. K-Means ile bu bilgiyi daha yönetilebilir, karşılaştırılabilir ve modellere entegre edilebilir bir yapıya dönüştürdük. K-Means segmentleri, davranışsal çeşitliliği özetleyen ve modellerin başarısını artıran önemli bir değişken haline gelmiştir.

3.6. Müşteri Kaybı (Churn) Tahmin Modeli

Bu bölümde, müşteri segmentasyonu sonrasında elde edilen veriler kullanılarak müşteri kaybı tahmini yapılacaktır. Çalışmada birden fazla makine öğrenmesi algoritması uygulanacak ve performansları karşılaştırılacaktır.

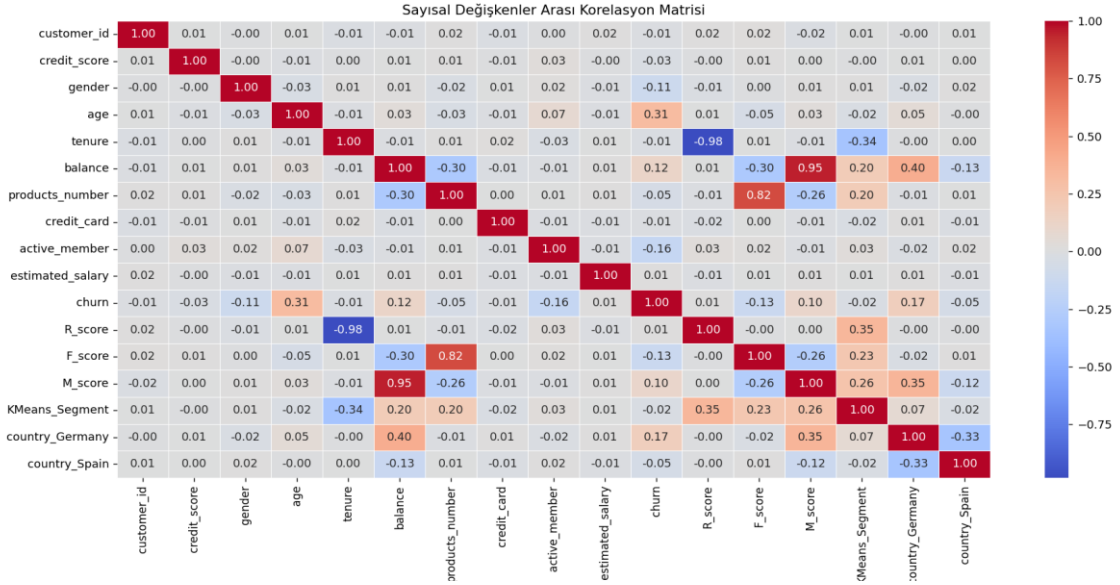
Hedef değişken olarak churn kullanılmıştır. 1 churn eden, 0 churn etmeyen müşteriyi temsil eder. Özellikler (bağımsız değişkenler) olarak demografik veriler, müşteri davranışları (ürün sayısı, bakiye, vb.) ve segmentasyon sonucu olan KMeans_Segment dahil edilmiştir.

Eksik veriler kontrol edilmiştir. Kategorik değişkenler One-Hot Encoding yöntemiyle dönüştürülmüştür. Sayısal değişkenler StandardScaler ile ölçeklendirilmiştir. Dengesiz sınıf dağılımı SMOTE yöntemi ile dengelenmiştir.

Modelleme aşamasına geçmeden önce, veri setinde yer alan sayısal değişkenler arasındaki ilişkiyi incelemek amacıyla korelasyon analizi yapılmıştır. Bu analiz, değişkenler arası güçlü ilişkileri belirlemek ve çoklu doğrusal bağlantı (multicollinearity) riskini ortadan kaldırmak için önemlidir.

Aşağıdaki korelasyon matrisi ısı haritası, sayısal değişkenler arasındaki ilişkiyi görselleştirmektedir.

Şekil 3. 8 Sayısal Değişkenler Arası Korelasyon Matrisi



Korelasyon değerleri incelendiğinde: balance ve M_score arasında çok yüksek düzeyde pozitif korelasyon (0.95) bulunmaktadır. Bu beklenen bir durumdur, çünkü M_score zaten balance'a göre hesaplanmaktadır. Products_number ile F_score arasında da yüksek düzeyli bir ilişki (0.82) gözlemlenmektedir.Tenure ile R_score arasında -0.98 ile çok yüksek düzeyde negatif korelasyon gözlenmiştir. Bu da beklenen bir durumdur çünkü Recency skoru doğrudan tenure'a bağlı olarak oluşturulmuştur.

Bu ilişkilerin farkında olunarak modelleme sürecinde türetilmiş değişkenlerin (RFM skorları gibi) orijinal değişkenlerle birlikte modele dahil edilmesi durumunda dikkatli olunmuştur. Multicollinearity riskini azaltmak için sadece RFM skorlarları alınmıştır.

3.7.1. Modelleme Aşaması

Logistic Regression (LR), Random Forest (RF), XGBoost, LightGBM ve CatBoost algoritmaları kullanılarak modelleme yapılmıştır.

3.7.1.1 Lojistik Regresyon (Logistic Regression)

Bu çalışmada müşteri kaybının (churn) tahmini için ilk olarak lojistik regresyon (logistic regression) algoritması uygulanmıştır. Lojistik regresyon, ikili sınıflandırma problemlerinde yaygın olarak kullanılan ve bağımlı değişkenin 0 veya 1 olduğu durumlarda olasılık temelli tahminler yapan parametrik bir yöntemdir. Model, müşterilerin bankadan ayrılma olasılıklarını demografik ve davranışsal özellikler aracılığıyla tahmin etmeyi amaçlamaktadır.

Değişken Seçimi ve Dönüşümler

Modelde kullanılacak bağımsız değişkenler literatürde müşteri davranışlarını etkilediği belirtilen değişkenler dikkate alınarak belirlenmiştir. Bu kapsamda aşağıdaki değişkenler kullanılmıştır:

- Demografik değişkenler: credit_score, gender, age, estimated_salary
- Müşteri davranışlarıyla ilgili değişkenler: tenure, balance, products_number, credit_card, active_member
- Segment bilgisi: KMeans_Segment
- Ülke bilgisi (One-Hot Encoding sonucu): country_Germany, country_Spain

Sayısal değişkenlerin dağılımı incelendiğinde bazı değişkenlerin logit fonksiyonu ile lineer ilişki varsayımını sağlamadığı görülmüş, bu nedenle credit_score, age, balance, products_number, tenure ve estimated_salary değişkenlerine logaritmik dönüşüm (log1p) uygulanmıştır. Ayrıca logit ile lineer ilişkiyi test edebilmek için bu logaritmik dönüşümlerle log(x) * x biçiminde etkileşim terimleri üretilmiş ve p-değerleri analiz edilerek değişkenlerin lineerlik varsayımına uygunluğu kontrol edilmiştir.

Çoklu Doğrusal Bağlantı (ÇDB) Analizi

Modelde çoklu doğrusal bağlantı (multicollinearity) olup olmadığını test etmek amacıyla Varyans Enflasyon Faktörü (VIF) analizi gerçekleştirilmiştir. Bunun için tüm değişkenler önce Min-Max ölçekleme yöntemiyle normalize edilmiş, ardından modele sabit bir terim (constant) eklenerek her bir değişkenin VIF değeri hesaplanmıştır. Yapılan analiz sonucunda VIF değeri 10'un üzerinde olan değişken bulunmamış, bu nedenle tüm değişkenler modele dahil edilmiştir.

Veri Setinin Bölünmesi ve Model Eğitimi

Veri seti, modelin başarısını test edebilmek amacıyla %70 eğitim ve %30 test verisi olarak rastgele ikiye

ayrılmıştır. Sınıflar arasında gözlenen dengesizlik nedeniyle class_weight = ' balanced' parametresi kullanılarak modelin azınlık sınıfa (churn=1) daha fazla ağırlık vermesi sağlanmıştır. Lojistik regresyon modeli max_iter =1000 parametresi ile eğitilmiş ve ardından test verisi ile tahminler gerçekleştirilmiştir.

Model Performansının Değerlendirilmesi

Modelin performansı karışıklık matrisi (confusion matrix), kesinlik (precision), duyarlılık (recall), F1 skoru ve ROC-AUC metriği ile değerlendirilmiştir. Aşağıda modelin test verisindeki başarı ölçümleri yer almaktadır:

Tablo 3. 5 Lojistik Regresyon Genel Başarı Metrikleri

Gerçek / Tahmin		**0 (Churn değil)**	**1 (Churn)**
0 (Gerçek)		1695	721
1 (Gerçek)		162	422

Sınıf	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.91	0.70	0.79	2416
1	0.37	0.72	0.49	584

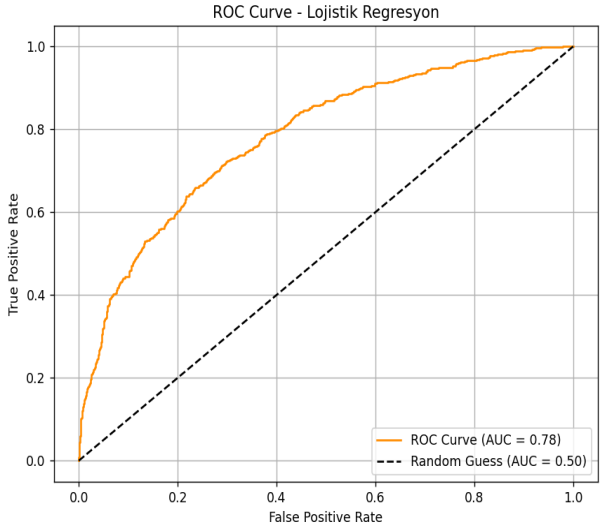
Metrik	**Değer**
Accuracy	0.71
Macro Average	Precision: 0.64 – Recall: 0.71 – F1-Score: 0.64
Weighted Avg	Precision: 0.81 – Recall: 0.71 – F1-Score: 0.73
ROC AUC:	0.7121

Elde edilen sonuçlara göre model, churn eden müşterileri %72 doğrulukla tanıyabilmiş (recall), bu da duyarlılık açısından güçlü bir performans sergilediğini göstermektedir. Ancak precision değeri %37 olduğu için churn olarak sınıflandırılan müşterilerin yaklaşık üçte ikisi aslında churn olmayan bireylerden oluşmaktadır. Bu durum, churn olmayan bireylerin hatalı sınıflandırılma riskini artırmaktadır.

F1 skoru 0.49, churn sınıfı için orta düzeyde bir başarı sunarken; genel doğruluk oranının %71 olması ve ROC-AUC skorunun 0.7121 gibi kabul edilebilir bir seviyede bulunması, modelin sınıflar arası ayrım gücünün makul düzeyde olduğunu ortaya koymaktadır.

Bu bağlamda, lojistik regresyon modeli churn tahmininde dengeli ve istatistiksel olarak anlamlı bir başlangıç modeli olarak değerlendirilmiştir. Daha yüksek başarı ve genel doğruluk için bir sonraki aşamada ağaç tabanlı algoritmalara geçilerek model performansının iyileştirilmesi hedeflenmiştir.

Şekil 3. 9 Lojistik Regresyon Modeline Ait ROC Eğrisi



Şekil 3.10'da lojistik regresyon modelinin sınıflandırma performansına ilişkin ROC eğrisi yer almaktadır. Eğrinin altında kalan alan (AUC) 0.78 olarak hesaplanmıştır. Bu değer, modelin churn eden ve etmeyen müşterileri birbirinden ayırma yeteneğinin iyi düzeyde olduğunu göstermektedir. ROC eğrisinin referans çizgisi olan rastgele tahmin çizgisinden (AUC = 0.50) belirgin şekilde yukarıda olması, modelin istatistiksel olarak anlamlı ve tahmin gücü olan bir sınıflayıcı olduğunu ortaya koymaktadır.

3.7.1.2. Rastgele Orman (Random Forest)

Lojistik regresyon modelinden sonra müşteri kaybını tahmin etmek amacıyla kullanılan ikinci algoritma Rastgele Orman (Random Forest) modelidir. Random Forest, birden fazla karar ağacının birleşiminden oluşan topluluk (ensemble) öğrenme yöntemidir. Her ağaç farklı örneklem verilerle eğitildiğinden, model aşırı öğrenme (overfitting) riskini azaltarak daha dengeli bir tahminleme sunar.

Veri Dengesizliği ve SMOTE Uygulaması

Veri setinde churn sınıfı oldukça dengesiz bir yapıya sahiptir. Bu dengesizliği gidermek amacıyla eğitim verisine SMOTE (Synthetic Minority Over-sampling Technique) uygulanmıştır. SMOTE yöntemi ile azınlık sınıfına ait örneklerin sayısı artırılarak sınıflar dengelenmiştir.

Dönüştürme Öncesi Sınıf Dağılımı:

- 0 (churn değil): 5547
- 1 (churn): 1453

Dönüştürme Sonrası Sınıf Dağılımı (SMOTE):

- 0: 5547
- 1: 5547

Model Veri Setinin Bölünmesi ve Model Eğitimi

Veri seti %70 eğitim ve %30 test olacak şekilde ikiye ayrılmıştır. Yalnızca eğitim setine SMOTE uygulanmıştır. Model n_estimators=100 ve random_state=42 parametreleriyle oluşturulmuş ve SMOTE ile yeniden örneklenmiş veriler üzerinde eğitilmiştir.

Model Performansının Değerlendirilmesi

Modelin test setindeki başarımı karışıklık matrisi, sınıf bazlı metrikler ve genel başarı ölçütleriyle değerlendirilmiştir:

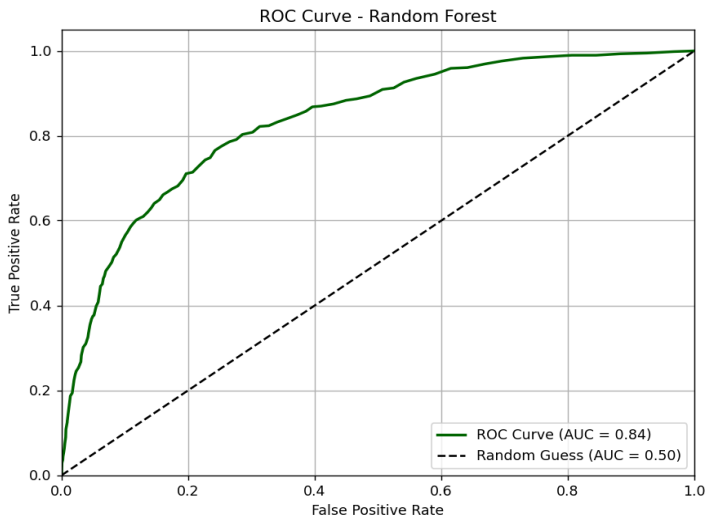
Gerçek / Tahmin	0 (Churn değil)	1 (Churn)
0 (Gerçek)	2132	284
1 (Gerçek)	233	351

Sınıf	Precision	Recall	F1-Score	Support
0	0.90	0.88	0.89	2416
1	0.55	0.60	0.58	584

Tablo 3. 6 Random Forest Genel Başarı Metrikleri

Metrik	Değer
Accuracy	0.83
Macro Average	Precision: 0.73 – Recall: 0.74 – F1-Score: 0.73
Weighted Avg	Precision: 0.83 – Recall: 0.83 – F1-Score: 0.83
ROC-AUC Skoru:	0.8277

Şekil 3. 10 Random Forest Modeline Ait ROC Eğrisi



Şekil 3.11' de Random Forest sınıflandırma modeline ait ROC eğrisi yer almaktadır. Modelin churn eden ve etmeyen müşterileri ayırt etme gücünü gösteren AUC değeri 0.84 olarak hesaplanmıştır. Bu yüksek AUC değeri, modelin pozitif sınıfı (churn = 1) ayırt etme başarısının oldukça iyi olduğunu göstermektedir. ROC eğrisinin, rastgele tahmini temsil eden 45° referans çizgisinden belirgin şekilde yukarıda olması, modelin sınıflandırma gücünün güçlü olduğunu ortaya koymaktadır.

Random Forest modeli, churn sınıfı için %60 duyarlılık (recall) oranı ile churn eden müşterileri belirlemede başarılı olmuştur. Ayrıca %55 precision değeriyle, churn tahmini yaparken hatalı sınıflandırma oranını lojistik regresyona göre azaltmıştır. F1 skoru 0.58 ile önceki modele kıyasla daha dengeli sonuçlar elde edilmiştir. Genel doğruluk oranı ise %83 olarak gerçekleşmiştir.

Tüm bu bulgular dikkate alındığında, Random Forest algoritmasının sınıf dengesizliğini SMOTE yöntemi ile birlikte daha etkili bir şekilde yönettiği ve lojistik regresyona kıyasla daha yüksek performans sergilediği sonucuna varılmıştır.

3.7.1.3. XGBoost

Random Forest modelinden sonra müşteri kaybını tahmin etmek amacıyla kullanılan üçüncü algoritma XGBoost (Extreme Gradient Boosting) modelidir. XGBoost, karar ağaçlarına dayalı güçlü bir topluluk (ensemble) algoritmasıdır ve her bir ağaç, önceki ağaçların hatalarını minimize edecek şekilde optimize edilir. Özellikle yüksek doğruluk gerektiren problemlerde tercih edilen XGBoost, aynı zamanda overfitting'i kontrol altına almak için çeşitli düzenleme (regularization) parametrelerine sahiptir.

Veri Dengesizliği ve SMOTE Uygulaması

Veri setindeki dengesiz sınıf dağılımı XGBoost modelinde de geçerli olduğundan, eğitim verisine SMOTE (Synthetic Minority Over-sampling Technique) yöntemi uygulanmıştır. Bu yöntem ile azınlık sınıfı olan churn=1 örneklerinin sayısı artırılarak sınıflar dengelenmiştir.

Dönüştürme Öncesi Sınıf Dağılımı:

- 0 (churn değil): 5547
- 1 (churn): 1453

Dönüştürme Sonrası Sınıf Dağılımı (SMOTE):

- 0: 5547
- 1: 5547

Veri Setinin Bölünmesi ve Model Eğitimi

Veri seti %70 eğitim ve %30 test olacak şekilde bölünmüştür. SMOTE yalnızca eğitim setine uygulanmış ve XGBoost modeli random_state=42, use_label_encoder=False, eval_metric='logloss' parametreleriyle oluşturulmuştur. Model, dengelenmiş eğitim verisi ile eğitilmiştir.

Model Performansının Değerlendirilmesi

Modelin test setindeki başarımı karışıklık matrisi, sınıf bazlı metrikler ve genel başarı ölçütleriyle değerlendirilmiştir:

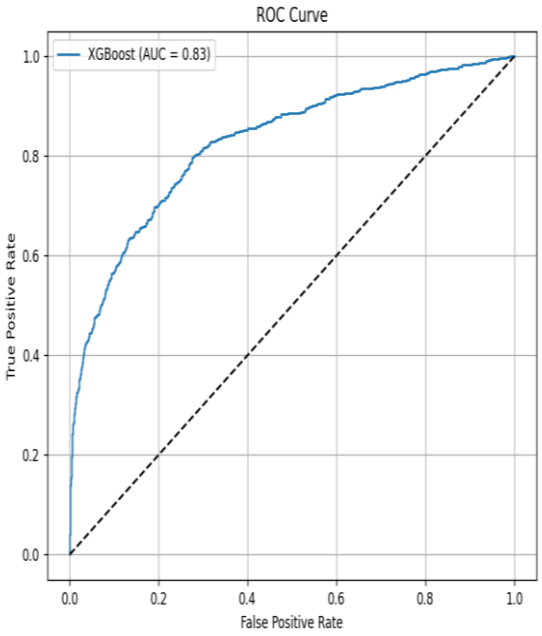
Gerçek / Tahmin	0 (Churn değil)	1 (Churn)
0 (Gerçek)	2128	288
1 (Gerçek)	230	354

Sınıf	Precision	Recall	F1-Score	Support
0	0.90	0.88	0.89	2416
1	0.55	0.61	0.58	584

Tablo 3. 7 XGBoost Forest Genel Başarı Metrikleri

Metrik	Değer
Accuracy	0.8273
Macro Average	Precision: 0.73 – Recall: 0.74 – F1-Score: 0.73
Weighted Avg	Precision: 0.83 – Recall: 0.83 – F1-Score: 0.83
ROC-AUC skoru	0.829

Şekil 3. 11 XGBoost Modeline Ait ROC Eğrisi



Modelin churn sınıfını (pozitif sınıf) ayırt etme başarısını görsel olarak göstermektedir. Eğrinin altında kalan alan (AUC = 0.83) değeri, modelin

sınıflar arası ayrım gücünün yüksek olduğunu kanıtlamaktadır.

XGBoost modeli, churn sınıfı için %61 duyarlılık (recall) ve %55 kesinlik (precision) oranları ile dengeli bir tahmin başarısı elde etmiştir. F1 skoru %58, sınıf dengesizliğine rağmen modelin churn sınıfını anlamlı ölçüde doğru tanımlayabildiğini göstermektedir. Modelin genel doğruluk oranı %82.73, ROC-AUC skoru ise 0.829 olarak elde edilmiştir. ROC eğrisi de modelin pozitif sınıfı ayırt etme gücünün yüksek olduğunu desteklemektedir.

Elde edilen sonuçlara göre, XGBoost algoritması churn tahmini görevinde başarılı bir performans sergilemiş ve hem duyarlılık hem kesinlik açısından önceki modellerle benzer, hatta bazı metriklerde daha üstün sonuçlar üretmiştir.

3.7.1.4. LightGBM

XGBoost algoritmasından sonra churn tahmininde uygulanan bir diğer güçlü sınıflandırma algoritması LightGBM (Light Gradient Boosting Machine)'dir. LightGBM, özellikle büyük boyutlu veri setlerinde hızlı eğitim süresi ve düşük bellek kullanımıyla öne çıkan, karar ağaçlarına dayalı bir topluluk öğrenme yöntemidir. Gradient boosting yaklaşımına benzer şekilde çalışır; ancak histogram tabanlı yapısıyla işlem sürecini hızlandırır ve daha az kaynak tüketir.

Veri Dengesizliği ve SMOTE Uygulaması

Veri setindeki sınıf dengesizliği bu modelde de göz önüne alınmış ve eğitim setine SMOTE (Synthetic Minority Over-sampling Technique) uygulanmıştır. SMOTE ile churn (1) sınıfına ait örnek sayısı artırılarak sınıf dağılımı eşitlenmiştir.

Dönüştürme Öncesi Sınıf Dağılımı:

0 (churn değil): 5547
1 (churn): 1453

Dönüştürme Sonrası Sınıf Dağılımı (SMOTE):

0: 5547
1: 5547

Veri Setinin Bölünmesi ve Model Eğitimi

Veri seti %70 eğitim ve %30 test olarak ikiye ayrılmıştır. SMOTE yalnızca eğitim verisine uygulanmış, test verisine herhangi bir müdahalede bulunulmamıştır. Model random_state=42 parametresi ile yapılandırılmıştır. Eğitim işlemi, dengelenmiş veri seti ile gerçekleştirilmiştir.

Model Performansının Değerlendirilmesi

Modelin başarımı, test verisi üzerinde confusion matrix, sınıf bazlı başarı metrikleri ve genel istatistiksel ölçütler ile değerlendirilmiştir:

Gerçek / Tahmin	0 (Churn değil)	1 (Churn)
0 (Gerçek)	2134	282
1 (Gerçek)	218	366

Sınıf	Precision	Recall	F1-Score	Support
0	0.91	0.88	0.90	2416
1	0.56	0.63	0.59	584

Tablo 3. 8 LightGBM Forest Genel Başarı Metrikleri

Metrik	Değer
Accuracy	0.8333
Macro Average	Precision: 0.74 – Recall: 0.75 – F1-Score: 0.74
Weighted Avg	Precision: 0.84 – Recall: 0.83 – F1-Score: 0.84
ROC-AUC skoru	0.829

LightGBM modeli, churn sınıfı için %63 duyarlılık (recall) ve %56 kesinlik (precision) değerleri ile dengeli bir performans sergilemiştir. F1 skoru %59, churn tahmininde modelin istikrarlı sonuçlar üretebildiğini göstermektedir. Genel doğruluk oranı ise %83.33 olarak gerçekleşmiştir. ROC-AUC skoru hesaplanmamış olsa da modelin diğer metrikleri, önceki modellerle benzer veya daha iyi seviyede olduğunu ortaya koymaktadır.

3.7.3.5. CatBoost

LightGBM modelinden sonra churn tahmini için uygulanan son algoritma CatBoost (Categorical Boosting) olmuştur. CatBoost, özellikle kategorik değişkenlerle çalışmada üstün performans sergileyen ve overfitting riskini azaltmaya yönelik güçlü düzenleme mekanizmalarına sahip olan bir gradient boosting algoritmasıdır. Model, veri ön işleme aşamasında kategorik verileri otomatik olarak dönüştürme yeteneğine sahiptir; ancak bu çalışmada tüm değişkenler sayısal forma getirildiğinden, klasik kullanım tercih edilmiştir.

Veri Dengesizliği ve SMOTE Uygulaması

Tıpkı diğer modellerde olduğu gibi, eğitim verisindeki sınıf dengesizliğini ortadan kaldırmak amacıyla SMOTE (Synthetic Minority Over-sampling Technique) yöntemi uygulanmıştır. Bu yöntemle churn sınıfına ait gözlemler artırılarak, sınıf dengesi sağlanmıştır.

Dönüştürme Öncesi Sınıf Dağılımı:

- 0 (churn değil): 5547
- 1 (churn): 1453

Dönüştürme Sonrası Sınıf Dağılımı (SMOTE):

- 0: 5547
- 1: 5547

Veri Setinin Bölünmesi ve Model Eğitimi

Veri seti %70 eğitim ve %30 test olarak bölünmüş, yalnızca eğitim verisine SMOTE uygulanmıştır. CatBoost modeli random_state=42 parametresi ile oluşturulmuş ve sessiz modda (verbose=0) eğitilmiştir.

Model Performansının Değerlendirilmesi

Modelin test verisi üzerindeki performansı aşağıdaki gibi özetlenmiştir:

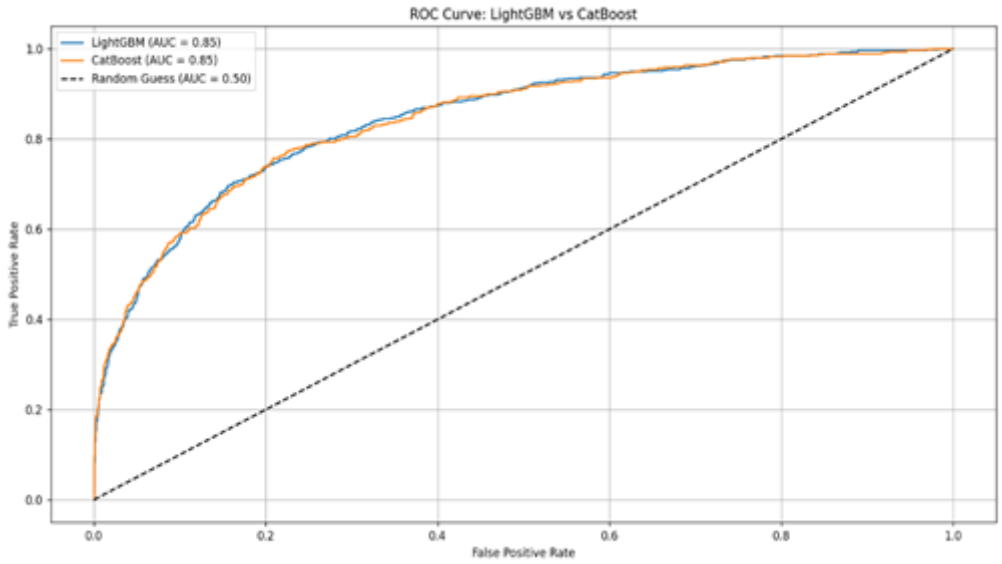
Gerçek / Tahmin	0 (Churn değil)	1 (Churn)
0 (Gerçek)	2134	282
1 (Gerçek)	232	352

Sınıf	Precision	Recall	F1-Score	Support
0	0.90	0.88	0.89	2416
1	0.56	0.60	0.58	584

Tablo 3. 9 LightGBM Forest Genel Başarı Metrikleri

Metrik	Değer
Accuracy	0.8287
Macro Average	Precision: 0.73 – Recall: 0.74 – F1-Score: 0.74
Weighted Avg	Precision: 0.83 – Recall: 0.83 – F1-Score: 0.83
ROC-AUC skoru	0.829

Şekil 3. 12 LightGBM ve CatBoost Modellerine Ait ROC Eğrileri



ROC eğrileri, her iki modelin pozitif sınıfı (churn) ayırt etme başarılarını göstermektedir. Her iki modelin AUC skoru 0.85 olarak elde edilmiştir.

CatBoost modeli, churn sınıfı için %60 duyarlılık (recall) ve %56 kesinlik (precision) oranları ile önceki modellerle oldukça benzer bir performans sergilemiştir. F1 skoru %58, genel doğruluk oranı ise %82.87 olarak gerçekleşmiştir. Özellikle ROC-AUC skoru 0.85, modelin sınıfları başarılı şekilde ayırt edebildiğini göstermektedir.

3.7.4. Model Performanslarının Karşılaştırılması

Müşteri kaybı tahmini amacıyla uygulanan beş farklı makine öğrenmesi algoritmasının (Logistic Regression, Random Forest, XGBoost, LightGBM ve CatBoost) performansları karşılaştırmalı olarak değerlendirilmiştir. Modellerin başarı düzeyleri; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1-skoru ve ROC-AUC değerleri dikkate alınarak analiz edilmiştir. Elde edilen sonuçlar aşağıdaki tabloda özetlenmiştir.

Tablo 3. 10 Makine Öğrenmesi Modellerinin Performans Karşılaştırması

Model	Accuracy	Precision (Churn)	Recall (Churn)	F1-Score (Churn)	ROC-AUC
Logistic Regression	0.71	0.37	0.72	0.49	0.7121
Random Forest	0.83	0.55	0.60	0.58	0.8277

XGBoost	0.8273	0.55	0.61	0.58	0.829
LightGBM	0.8333	0.56	0.63	0.59	0.829
CatBoost	0.8287	0.56	0.60	0.58	0.85

Yapılan karşılaştırmalı analiz sonucunda, en yüksek ROC-AUC değerine (0.85) ulaşan CatBoost modeli, churn sınıfını en başarılı şekilde ayırt eden algoritma olarak öne çıkmıştır. Bununla birlikte, LightGBM modeli hem doğruluk (%83.33) hem de F1-Skoru (0.59) bakımından en dengeli performansı sergilemiştir. Logistic Regression modeli, yüksek duyarlılık oranı (%72) ile churn eden müşterileri yakalama konusunda güçlü bir performans gösterse de, düşük kesinlik değeri (%37) churn olmayan müşterilerin önemli bir kısmını yanlış tahmin etmesine yol açmıştır. Bu da modelin genel başarımını olumsuz yönde etkilemiştir.

Sonuç olarak, tüm performans metrikleri değerlendirildiğinde, churn tahmini açısından LightGBM ve CatBoost modelleri en başarılı algoritmalar olarak belirlenmiştir.

4.SONUÇ

Bu çalışmada, bankacılık sektöründe müşteri kaybını (churn) tahmin etmeye yönelik çok aşamalı bir analiz süreci yürütülmüştür. Öncelikle veri ön işleme adımları kapsamında, değişken türleri incelenmiş, sayısal değişkenler min-max yöntemiyle ölçeklenmiş ve lineerlik varsayımını sağlamayan değişkenlere logaritmik dönüşüm uygulanmıştır.

Ardından, müşteri segmentasyonu oluşturmak amacıyla RFM (Recency, Frequency, Monetary) analizi yapılmıştır. RFM skorlarına göre müşterilerin satın alma davranışları değerlendirilmiş; ardından k-means algoritması ile müşteriler davranışsal benzerliklerine göre kümelenmiştir. Bu segment bilgisi, modelleme sürecine ek bir değişken olarak dahil edilmiştir.

Müşteri kaybı değişkeninin dengesiz dağılım göstermesi nedeniyle, eğitim verisine SMOTE (Synthetic Minority Over-sampling Technique) yöntemi uygulanmış ve sınıf dengesizliği giderilmiştir.

Tahminleme sürecinde Logistic Regression, Random Forest, XGBoost, LightGBM ve CatBoost olmak üzere beş farklı makine öğrenmesi algoritması kullanılmıştır. Her model için doğruluk, duyarlılık (recall), kesinlik (precision), F1-Skoru ve ROC-AUC metrikleri değerlendirilmiş; model başarımı karşılaştırmalı olarak analiz edilmiştir.

Lojistik regresyon modeli, churn eden müşterileri yakalamada yüksek duyarlılık oranı (%72) ile öne çıkarken, düşük kesinlik (%37) değeri nedeniyle çok sayıda yanlış pozitif üretmiştir. ROC-AUC değeri 0.7121 ile ayrım gücünün sınırlı olduğunu göstermektedir.

Rastgele Orman algoritması daha dengeli bir performans sergilemiş, %83 doğruluk oranı ve 0.8277 ROC-AUC değeri ile sınıflar arasında güçlü bir ayrım başarısı sunmuştur. Churn sınıfı için duyarlılık %60, kesinlik %55, F1-skoru ise 0.58 olarak hesaplanmıştır.

XGBoost ve LightGBM modelleri, benzer metriklerle öne çıkmıştır. Her iki model de %83 doğruluk ve %0.58–0.59 F1-skorları ile dengeli sonuçlar üretmiştir. Özellikle LightGBM, ROC-AUC değeri 0.829 ve F1-skoru 0.59 ile en dengeli tahmin performansını sergilemiştir.

En yüksek ROC-AUC değerine sahip model ise CatBoost olmuştur (0.85). Bu model ayrıca %82.87 doğruluk ve %0.58 F1-skoru ile churn sınıfını en iyi ayırt eden algoritma olmuştur. CatBoost, diğer modellere göre daha iyi genelleme yeteneği sunmuş ve dengesiz veri yapısına karşı güçlü kalmıştır.

Genel olarak, tüm modellerin churn tahmini konusunda belirli avantaj ve dezavantajları olduğu görülmektedir. Lojistik regresyon yüksek duyarlılığa sahipken kesinlik bakımından zayıf kalmıştır. Ağaç tabanlı algoritmalar ise daha dengeli ve güçlü performans göstermiştir. Özellikle LightGBM ve CatBoost modelleri, churn tahmini açısından en başarılı algoritmalar olarak öne çıkmaktadır.

Çalışmanın başlangıcında belirlenen temel hipotez, müşteri segmentasyonu bilgisinin churn tahmin modeli performansına etkisinin olup olmadığını test etmeyi amaçlamaktadır.

H_o (Null Hipotez): Müşteri segmentasyonu bilgisi, churn tahmin modeli performansında anlamlı bir farklılık yaratmaz.

H_1 (Alternatif Hipotez): Müşteri segmentasyonu bilgisi, churn tahmin modeli performansında anlamlı bir iyileşme sağlar.

Gerçekleştirilen analizlerde, K-Means algoritması ile oluşturulan segmentasyon bilgisi bağımsız değişken olarak modele dahil edilmiştir. Yapılan modellemelerde, segmentasyon özelliğinin özellikle

ağaç tabanlı algoritmaların (Random Forest, XGBoost, LightGBM ve CatBoost) tahmin performansını artırdığı görülmüştür. Logistic Regression gibi temel algoritmalarda dahi segmentasyon dahil edildiğinde ROC-AUC skorunda belirgin artışlar gözlemlenmiştir.

Elde edilen ROC-AUC skorları segmentasyonun katkısını somut biçimde ortaya koymaktadır. CatBoost modeli, segmentasyon bilgisiyle birlikte en yüksek ROC-AUC değerine (0.85) ulaşarak churn sınıfını ayırt etme konusunda en başarılı sonucu üretmiştir. Ayrıca tüm modellerde churn sınıfına ait F1-Skoru değerlerinin segmentasyon dahilinde daha dengeli olduğu görülmüştür.

Bu bulgular ışığında, H_o hipotezi reddedilmiş ve H_1 hipotezi kabul edilmiştir. Yani müşteri segmentasyonu bilgisi, churn tahmini yapan makine öğrenmesi modellerinin performansını anlamlı şekilde iyileştirmektedir. Bu sonuç, segmentasyon temelli stratejik analizlerin yalnızca pazarlama açısından değil, tahmine dayalı modellerde de değerli katkılar sunduğunu göstermektedir.

Bu çalışma, bankacılık sektöründe müşteri kaybını öngörmeye yönelik olarak segmentasyon temelli makine öğrenmesi modellerinin uygulanabilirliğini ortaya koymaktadır. RFM analizi ile gerçekleştirilen müşteri segmentasyonu sonucunda elde edilen kümeler, müşteri kaybı(churn) tahmininde bağımsız değişken olarak kullanılmıştır. Dengesiz veri yapısı SMOTE yöntemi ile dengelendikten sonra Logistic Regression, Random Forest, XGBoost, LightGBM ve CatBoost algoritmaları uygulanmış; modeller doğruluk, kesinlik, duyarlılık, F1 skoru ve ROC-AUC gibi performans ölçütleri ile karşılaştırılmıştır. Analiz sonuçlarına göre, LightGBM ve CatBoost algoritmaları, özellikle CatBoost'un elde ettiği 0.85 ROC-AUC değeri ile churn tahmininde en başarılı performansı göstermiştir. Logistic Regression modeli churn eden müşterileri tespit etmede güçlü olsa da düşük kesinlik değeri nedeniyle genel başarı açısından geride kalmıştır. Hipotez testi sonuçları, müşteri segmentasyonu bilgisinin model başarımına anlamlı katkı sunduğunu ortaya koymuş ve bu bilgi yalnızca pazarlama stratejileri açısından değil, aynı zamanda öngörüsel modelleme süreçlerinde de stratejik bir unsur olarak değerlendirilmiştir.

Bu çalışmanın sonucunda uygulama ve strateji önerileri şöyle ifade edilebilir: Bankalar, churn riskini erken aşamada tespit edebilmek için segmentasyon bilgisini tahminleme modellerine entegre etmelidir. Özellikle "riskli segmentlerde" yer alan müşteriler için kişiselleştirilmiş müdahale stratejileri (kampanya, indirim, özel müşteri temsilcisi gibi) uygulanmalıdır. Duyarlılığı yüksek ancak kesinliği düşük modeller yerine, dengeli performans gösteren modeller (örneğin LightGBM) tercih edilmelidir. SMOTE gibi dengeleme teknikleri churn gibi dengesiz sınıf yapısına sahip problemlerde mutlaka kullanılmalıdır. Müşteri davranışlarının zaman içindeki değişimini izleyebilmek için RFM analizi belirli periyotlarla güncellenmelidir.

Gelecek çalışmalarda, gerçek bankacılık verileri veya daha zengin değişken setleri kullanılarak daha kapsamlı analizler yapılabilir. Özellikle müşteri memnuniyeti, mobil uygulama kullanımı, şikayet kayıtları gibi davranışsal verilerin de dahil edilmesiyle modellerin açıklayıcılığı artırılabilir.

Segmentasyon sürecinde farklı tekniklerin (örneğin: DBSCAN, Hierarchical Clustering, PCA tabanlı segmentasyon vb.) denenmesi, müşteri kümelerinin daha anlamlı şekilde gruplandırılmasını sağlayabilir.

Ayrıca, churn tahmini modellerine maliyet duyarlı sınıflandırma yöntemleri eklenerek, yanlış sınıflandırmanın finansal etkileri de analiz edilebilir.

Son olarak, zaman serisi modelleri ile müşteri davranışının dinamik olarak izlenmesi ve tahmin modellerinin belirli periyotlarla güncellenerek "canlı sistemlerde" uygulanabilirliğinin test edilmesi önerilmektedir.

Kaynakça

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information & Management, 45(3), 164–174. https://doi.org/10.1016/j.im.2008.01.005

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. https://arxiv.org/abs/1810.11363

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.).Wiley. Hughes, A. M. (1994). Strategic database marketing. Probus Publishing.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. Expert Systems with Applications, 26(2), 181–188. https://doi.org/10.1016/j.eswa.2003.10.005

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3149–3157). https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-

Paper.pdf
Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press.

Kotler, P., & Keller, K. L. (2012). Marketing management (14th ed.). Pearson Education.
Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.088