

Veri Bilimi ile Otel Rezervasyon İptallerinin Tahmini: SHAP Tabanlı Açıklanabilirlik ve Performans Analizi

Predicting Hotel Reservation Cancellations with Data Science: SHAP-Based Explainability and Performance Analysis

Kamil Abdullah EŞİDİR

Dr., Fırat Kalkınma Ajansı, Elazığ Yatırım Destek Ofisi,
abdullahesidir@yahoo.com
<https://orcid.org/0000-0002-8106-1758>

Makale Başvuru Tarihi: 14.07.2025

Makale Kabul Tarihi: 08.10.2025

Makale Türü: Araştırma Makalesi

ÖZET

Anahtar Kelimeler:

Otel Rezervasyon İptali,
Makine Öğrenimi,
SMOTE,
SHAP Analizi,

Turizm sektöründe gelir yönetimi ve operasyonel planlamanın etkinliği, rezervasyon iptallerinin doğru şekilde tahmin edilmesine bağlıdır. Bu çalışmada, otel rezervasyon iptallerinin makine öğrenmesi temelli modeller aracılığıyla öngörülmesi amaçlanmıştır. Veri kaynağı olarak, 2015–2017 dönemine ait 119.390 gözlemden oluşan açık erişimli “Hotel Booking Demand” veri kümesi kullanılmıştır. Sınıf dengesizliği problemi SMOTE yöntemiyle giderilmiş; Random Forest, XGBoost, LightGBM ve CatBoost modelleri ile tahminlemeler gerçekleştirilmiştir. Modeller; doğruluk, AUC, F1 skoru ve özellikle eğitim süresi gibi performans metrikleriyle karşılaştırılmıştır. En yüksek doğruluk oranı (%81) ve AUC skoru (0,89) Random Forest modeliyle elde edilmiştir. Modelin karar süreçleri SHAP analiziyle açıklanmış; özellikle “lead time”, “adr” ve “hotel” değişkenlerinin belirleyici olduğu tespit edilmiştir. Bulgular, rezervasyon iptallerinin önceden tahmin edilmesinin, turizm sektöründe karar destek sistemlerine entegre edilebilecek anlamlı öngörüler sunduğunu göstermektedir.

ABSTRACT

Keywords:

Hotel Reservation
Cancellation,
Machine Learning,
SMOTE,
SHAP Analysis,

The effectiveness of revenue management and operational planning in the tourism sector largely depends on the ability to accurately predict reservation cancellations. This study aims to predict hotel reservation cancellations using machine learning-based models. An open-access dataset entitled “Hotel Booking Demand”, consisting of 119,390 observations from 2015 to 2017, is employed as the primary data source. To address the class imbalance problem, the SMOTE technique is applied. Prediction tasks are carried out using four models: Random Forest, XGBoost, LightGBM, and CatBoost. These models are compared based on multiple evaluation metrics including accuracy, AUC, F1-score, and training time. The Random Forest model achieves the highest performance with an accuracy of 81% and an AUC score of 0.89. The model’s decision-making processes are interpreted using SHAP analysis, which reveals that variables such as “lead_time”, “adr”, and “hotel” are particularly influential. The findings indicate that early prediction of reservation cancellations can provide valuable insights for the integration of data-driven decision support systems in the tourism industry.

Önerilen Alıntı (Suggested Citation): EŞİDİR, Kamil Abdullah (2025), “Veri Bilimi ile Otel Rezervasyon İptallerinin Tahmini: SHAP Tabanlı Açıklanabilirlik ve Performans Analizi”, *Uluslararası Yönetim Akademisi Dergisi*, S.8(3), ss.767-785, Doi: <https://doi.org/10.33712/mana.1737947>

1. GİRİŞ

Son yıllarda turizm ve konaklama sektöründe veri bilimi uygulamaları, stratejik karar alma süreçlerinde giderek daha belirleyici bir rol üstlenmektedir. Müşteri davranışlarının öngörülmesine yönelik veri temelli yaklaşımlar; otel işletmeleri açısından gelir yönetimi, personel planlaması ve oda tahsisi gibi operasyonel kararların etkinliğini artırma potansiyeli taşımaktadır. Dijitalleşmenin yaygınlaşmasıyla birlikte, otel rezervasyonlarının iptal edilme olasılığı sektördeki belirsizlikleri artırmakta ve öngörüye dayalı karar destek sistemlerine olan ihtiyacı açıkça ortaya koymaktadır (Caggia, 2021).

Rezervasyon iptallerinin geleneksel istatistiksel yöntemlerle sınırlı doğruluk düzeyinde tahmin edilebilmesi, makine öğrenmesi algoritmalarını bu alanda öne çıkarmaktadır. Son yıllarda geliştirilen Destek Vektör Makineleri (SVM), Random Forest, XGBoost, LightGBM ve CatBoost gibi algoritmalar; büyük ve çok değişkenli veri setlerinde yüksek sınıflandırma başarısı sunmakta, aynı zamanda modelin açıklanabilirliğini artırma ve sınıf dengesizliği sorununu aşma gibi avantajlara da sahiptir (Jishan vd., 2024).

Literatür incelendiğinde, otel rezervasyon iptallerine yönelik birçok çalışmanın yalnızca doğruluk oranı veya AUC gibi temel performans ölçütlerine odaklandığı; ancak modellerin karar mekanizmalarını şeffaflaştıran açıklanabilir yapay zekâ tekniklerinin sınırlı düzeyde kullanıldığı görülmektedir. Oysa SHAP gibi yöntemler, modelin karar alma sürecini kullanıcıya izah edilebilir biçimde sunmakta ve bu sayede otel yöneticileri için daha eyleme geçirilebilir öngörüler sağlamaktadır. Ayrıca sınıf dengesizliği problemi - örneğin iptal edilmeyen rezervasyonların sayıca baskın olması - çoğu çalışmada yeterince dikkate alınmamakta ve modellerin güvenilirliğini azaltmaktadır (Caggia, 2021).

Sunulan çalışma, hem açıklanabilir yapay zekâ tekniklerini hem de sınıf dengesizliği sorununa yönelik SMOTE tabanlı çözümleri bir araya getirerek literatürdeki boşluklara doğrudan katkı sunmayı hedeflemektedir. Otel rezervasyon iptallerine ilişkin öngörüler, Random Forest, XGBoost, LightGBM ve CatBoost modelleri kullanılarak elde edilmiş; bu modellerin doğruluk, AUC skoru, F1 skoru ve eğitim süresi gibi çoklu metrikler çerçevesinde performansları karşılaştırılmıştır. Model kararlarının şeffaflaştırılması amacıyla SHAP analizi uygulanmış ve özellikle öne çıkan değişkenlerin karar sürecine katkıları detaylı olarak analiz edilmiştir. Böylece yalnızca yüksek sınıflandırma başarısına sahip modeller geliştirilmemiş, aynı zamanda otel yöneticileri için karar destek sistemlerine entegre edilebilecek yorumlanabilir modeller sunulmuştur. Çalışmanın ikinci bölümünde, otel rezervasyon iptallerine yönelik literatürdeki temel çalışmalar özetlenmiştir. Üçüncü bölümde, veri kümesi, değişkenler, veri ön işleme süreci, uygulanan makine öğrenmesi modelleri ve modelleme yaklaşımı detaylandırılmıştır. Dördüncü bölümde, modellerin sınıflandırma performansları doğruluk, AUC, F1 skoru ve eğitim süresi gibi ölçütler açısından karşılaştırılmış, SHAP analizi ile açıklanabilirlik değerlendirmesi yapılmıştır. Beşinci bölümde, bulgular literatür ışığında tartışılmıştır. Son bölümde ise genel sonuçlara yer verilmiş ve gelecekte yapılacak benzer araştırmalar için öneriler sunulmuştur.

2. LİTERATÜR TARAMASI

Turizm sektöründe dijitalleşmenin artması ve çevrimiçi rezervasyon sistemlerinin yaygınlaşmasıyla birlikte, müşteri davranışlarını önceden tahmin etmeye yönelik veri odaklı analizlerin önemi giderek artmaktadır. Özellikle otel rezervasyon iptallerinin öngörülebilirliğini konu alan çalışmalar, gelir yönetimi, oda tahsisi ve personel planlaması gibi operasyonel süreçlerde karar destek sistemlerine katkı sunmayı hedeflemektedir.

Rezervasyon iptali tahmini konusuna odaklanan birçok akademik çalışma, çeşitli makine öğrenmesi modellerinin bu alanda yüksek başarı sağladığını göstermektedir. Caggia (2021), makine öğrenmesiyle iptal tahmininin gelir yönetiminde kritik rol oynadığını vurgulamaktadır. Boz vd. (2018) tarafından yapılan çalışmada, beş farklı otele ait 38.826 kayıttan oluşan veri kümesi üzerinden dört farklı makine öğrenmesi modeli (*Random Forest*, *SVM*, *k-NN* ve *C4.5 Karar Ağacı*) uygulanmıştır. En iyi performans %73 doğruluk oranıyla C4.5 algoritmasından elde edilmiştir. Bu çalışma özellikle yerli bir veri kümesi ile gerçekleştirilmesi bakımından önem taşımaktadır. Ancak sınıflar arası dengesizlik SMOTE gibi modern yöntemlerle değil, undersampling ile giderilmeye çalışılmıştır ve açıklanabilirlik boyutu ele alınmamıştır.

Antonio vd. (2019), Portekiz'deki iki farklı otelden elde ettikleri rezervasyon verileriyle oluşturdukları veri kümesi aracılığıyla, konaklama sektöründe iptal tahmini gibi karar destek süreçlerine yönelik önemli bir kaynak sunmuşlardır. Söz konusu veri kümesi, her bir rezervasyonu otele girişten bir gün önceki bilgilerle temsil ederek, makine öğrenmesi modellerinde yaygın bir sorun olan bilgi sızıntısını (data leakage) önlemektedir.

Böylece geliştirilen modeller, özellikle iptallerin öngörülmesinde daha güvenilir hale gelmektedir. Bu veri kümesine dayalı olarak geliştirilen ilk uygulamalardan biri Antonio vd. (2017) tarafından gerçekleştirilmiş olup, otel rezervasyon iptallerinin tahmini amacıyla çeşitli sınıflandırma modelleri kullanılmıştır. Naive Bayes, Karar Ağaçları, Random Forest ve AdaBoost gibi modeller karşılaştırılmış ve doğruluk oranlarının %80'in üzerine çıktığı belirtilmiştir. Böylece rezervasyon iptallerinin, müşteri türü, rezervasyon süresi, kanal tercihi gibi değişkenlerle öngörülebilir olduğu ortaya konmuştur. Çalışmada, aynı veri kümesinin sadeleştirilmiş bir sürümü kullanılarak rezervasyonların iptal edilip edilmeyeceği tahmin edilmiştir. SMOTE yöntemi ile veriler dengelenmiş, modellerin başarısı ROC eğrisi, F1 skoru ve SHAP değeri analizleri ile değerlendirilmiştir.

Ercan ve Irmak (2022) tarafından gerçekleştirilen çalışmada, otel rezervasyon iptallerinin sınıflandırılmasına yönelik olarak karar ağaçları, lojistik regresyon ve destek vektör makineleri gibi farklı makine öğrenmesi algoritmaları uygulanmış; modellerin doğruluk, hata oranı ve ROC-AUC değerleri karşılaştırılmıştır. Çalışmada, özellikle destek vektör makinelerinin yüksek sınıflandırma başarısı bulunduğu belirtilmiştir.

Şahinbaş ve Doğuç (2022) tarafından gerçekleştirilen çalışmada, 80.000 civarı rezervasyon kaydı için Karar Ağaçları ve Rastgele Orman makine öğrenmesi modelleri ile tahminlemeler yapılmıştır. En yüksek doğruluk oranı %86,7 ile Rastgele Orman modeli tarafından elde edilmiştir. Depozito tipi ve önceki iptal durumu değişkenlerinin tahminleme sürecinde belirleyici olduğu tespit edilmiştir. İptal davranışının özniteliklere dayalı olarak güçlü biçimde modellenebileceği yönündeki bulgular desteklenmektedir.

Eşidir (2025a), TÜİK'in 2022 yılına ait Yetişkin Eğitimi Araştırması mikro verilerini kullanarak, bireylerin örgün eğitime katılım durumlarını çeşitli makine öğrenimi modelleri ile analiz etmiştir. Çalışmada, Gradient Boosting, XGBoost, CatBoost, LightGBM ve Random Forest modelleri karşılaştırılmıştır. CatBoost modeli, en uygun model olarak ön plana çıkmıştır. Elde edilen sonuçlar, makine öğrenimi modellerinin eğitim verileri üzerindeki etkinliğini göstermektedir.

Karaca ve Akın (2025), Türkiye'de turizm alanındaki makine öğrenmesi yayınlarının bibliyometrik analizini yapmıştır. Yazarlar, 1995-2024 döneminde yayınlanan çalışmalar temelinde yıllara göre makale sayısı, araştırmacı profili ve metodolojik eğilimler üzerine değerlendirme sunmuş; rezervasyon iptali gibi davranışsal tahmin konularının artan akademik ilgiyle ele alındığını belirtmiştir. Bununla birlikte bibliyometrik analiz, veri dengesizliği, hiperparametre optimizasyonu ve model açıklanabilirliği gibi teknik unsurlara yeterli vurgu yapılmadığını ortaya koymuştur.

3. ARAŞTIRMANIN YÖNTEMİ

Bu araştırmada, otel rezervasyon iptallerinin tahmin edilmesine yönelik olarak denetimli makine öğrenmesi yaklaşımı benimsenmiştir. Sınıflandırma problemi olarak kurgulanan çalışmada bağımlı değişken, rezervasyonun iptal edilip edilmediğini gösteren *is_canceled* değişkenidir. Bağımsız değişkenler ise otel türü, müşteri profili, rezervasyon tarihi, oda tipi, kanal bilgisi ve fiyatlandırma gibi faktörleri içeren 17 açıklayıcı değişkenden oluşmaktadır. Veri dengesizliği problemi nedeniyle, model eğitiminden önce SMOTE (*Synthetic Minority Oversampling Technique*) yöntemi kullanılarak sınıflar arası dağılım dengelenmiştir. Tahmin sürecinde, Random Forest, XGBoost, LightGBM ve CatBoost modelleri karşılaştırılmıştır. Model çıktıları ROC eğrisi, karışıklık matrisi, F1 skoru gibi klasik başarı metrikleriyle değerlendirilmiş; ayrıca modelin karar mekanizması SHAP yöntemi ile yorumlanmıştır. Bu yapı sayesinde hem sınıflandırma doğruluğu sağlanmış hem de modelin kararlarına açıklık kazandırılmıştır.

3.1. Yazılım ve Donanım

Veri analizleri ve makine öğrenmesi model tahminlerinin yüksek doğruluk ve verimlilik ile gerçekleştirilebilmesi için güçlü bir donanım altyapısı ve güncel yazılımlar kullanılmıştır. İşletim sistemi olarak Windows 11 Pro (64-bit) tercih edilmiştir. Sistemde Intel64 Family 6 Model mimarisine sahip bir işlemci yer almaktadır. Bilgisayar, büyük ölçekli veri kümelerinin işlenmesi ve hesaplama açısından yoğun modellerin eğitilmesi için yeterli olan 16 GB RAM barındırmaktadır.

Yazılım tarafında ise; veri ön işleme, modelleme ve değerlendirme aşamalarında Python 3.12 sürümü kullanılmıştır. NumPy, Pandas, Scikit-learn, TensorFlow ve Keras gibi kütüphanelerden istifade edilerek analizler gerçekleştirilmiştir. Böylece; veri manipülasyonu, model eğitimi ve performans değerlendirme süreçleri etkin şekilde yürütülmüştür. Analizlerin daha etkileşimli ve görselleştirmeye açık bir ortamda

sürdürülmesi amacıyla Jupyter Notebook 7.2 platformu kullanılmıştır. Bu platform, kodlama ve açıklamaların bütüncül biçimde sunulmasına imkân tanımaktadır.

3.2. Veri Kümesi

Kaggle platformunda kamuya açık şekilde kullanıma sunulmuş olan “*Hotel Booking Demand*” isimli veri kümesi, Jesse Mostipak tarafından 2018 yılında yayınlanmıştır. Veri kümesinde, Portekiz’deki iki otele ait olan 2015–2017 yılları arasındaki 119.390 rezervasyon kayıtları bulunmaktadır. Veri kümesi, otel yönetimi karar süreçlerinde kullanılmak üzere temizlenmiş demografik ve operasyonel bilgileri içermektedir. Gereksiz ve eksik değerlerin fazlaca bulunduğu sütunlar, analizlere katkı düzeylerinin sınırlı olması sebebi ile veri kümesinden çıkartılmıştır. Veri kümesi, Tatil (Resort) Otel’ne ait 40.060 ve Şehir (City) Otel’ne ait 79.330 gözlem olmak üzere toplamda 119.390 rezervasyon kaydı içermektedir. Her bir gözlem, bir rezervasyon kaydını temsil etmektedir. 44.224 adet rezervasyon iptal edilmiş ve 75.166 adet rezervasyon ise gerçekleşmiştir. Temizlenmiş veri kümesinde herhangi bir eksik veri bulunmamaktadır. Analizlerde kullanılan veri kümesinde 119.390 adet satır ve 18 adet sütun bulunmaktadır. Veri kümesi, ileri düzey analizler için oldukça uygundur. Veri kümesi; otel türü, rezervasyon iptali, geliş ve kalış tarihleri, misafir sayısı, yemek planı, pazar segmenti, dağıtım kanalı, önceki rezervasyon geçmişi ve ortalama günlük gelir (ADR) gibi temel demografik ve operasyonel değişkenleri içermektedir (Mostipak, 2018).

Analizlerde kullanılan ve temizlenmiş veri kümesinde bulunan değişkenler, Tablo 1’de gösterilmiştir. Otel rezervasyonlarına ait 18 değişkenin anlamları kısaca özet olarak açıklanmıştır. Değişkenler arasında, rezervasyonun yapıldığı otel türü, iptal durumu, müşterinin otele geliş süresi, konaklama gün sayısı, müşteri tipi, kanal bilgisi ve ortalama günlük ücret benzeri tahminleme sürecinde belirleyici olan nitelikler bulunmaktadır.

Tablo 1. Çalışmada Kullanılan Değişkenler ve Açıklamaları

Değişken Adı	Açıklama
hotel	Rezervasyonun yapıldığı otelin türü (Şehir oteli veya tatil oteli)
is_canceled	Rezervasyonun iptal edilip edilmediği (1 = iptal, 0 = gerçekleştirilmiş)
lead_time	Rezervasyon tarihi ile giriş tarihi arasındaki gün sayısı
arrival_date_year	Müşterinin otele giriş yaptığı yıl
arrival_date_month	Otele varışın gerçekleştiği ay
stays_in_weekend_nights	Hafta sonu (Cumartesi-Pazar) kalınan gece sayısı
stays_in_week_nights	Hafta içi (Pazartesi-Cuma) kalınan gece sayısı
adults	Rezervasyonda yer alan yetişkin sayısı
meal	Seçilen yemek (ör. BB = kahvaltı dahil)
market_segment	Rezervasyonun yapıldığı pazar segmenti (ör. Online, grup, doğrudan)
distribution_channel	Rezervasyonun iletildiği kanal (ör. seyahat acentesi, doğrudan)
is_repeated_guest	Müşterinin tekrar eden bir misafir olup olmadığı
previous_cancellations	Müşterinin daha önceki iptal sayısı
previous_bookings_not_canceled	Müşterinin daha önce iptal edilmeyen rezervasyon sayısı
reserved_room_type	Rezervasyon sırasında talep edilen oda tipi
days_in_waiting_list	Bekleme listesinde geçirilen gün sayısı
customer_type	Müşteri tipi (ör. Transient, Grup, Kontratlı müşteri)
adr	Ortalama günlük oda ücreti (Average Daily Rate)

Kaynak: Yazar tarafından oluşturulmuştur.

Her bir satır, bireysel bir otel rezervasyon kaydını temsil etmektedir. Hem sayısal hem de kategorik veri türleri bulunmaktadır. Veri kümesinde 10 değişken sayısal (lead_time, adr, previous_cancellations gibi) ve 7 değişken kategorik (hotel, meal, market_segment gibi) yapıdadır. Kapsamlı veri kümesinin, otel rezervasyon iptallerini tahmin etmeye yönelik makine öğrenmesi uygulamaları açısından güzel bir örnek olduğu söylenebilir.

3.3. Çalışmanın Konusu, Amacı ve Önemi

Çalışmanın konusu, otel rezervasyonlarının iptal edilip edilmeyeceğini makine öğrenmesi tabanlı sınıflandırma modelleri aracılığı ile öngörmektir. Konaklama sektöründe rezervasyon iptalleri, özellikle dönemsel talep

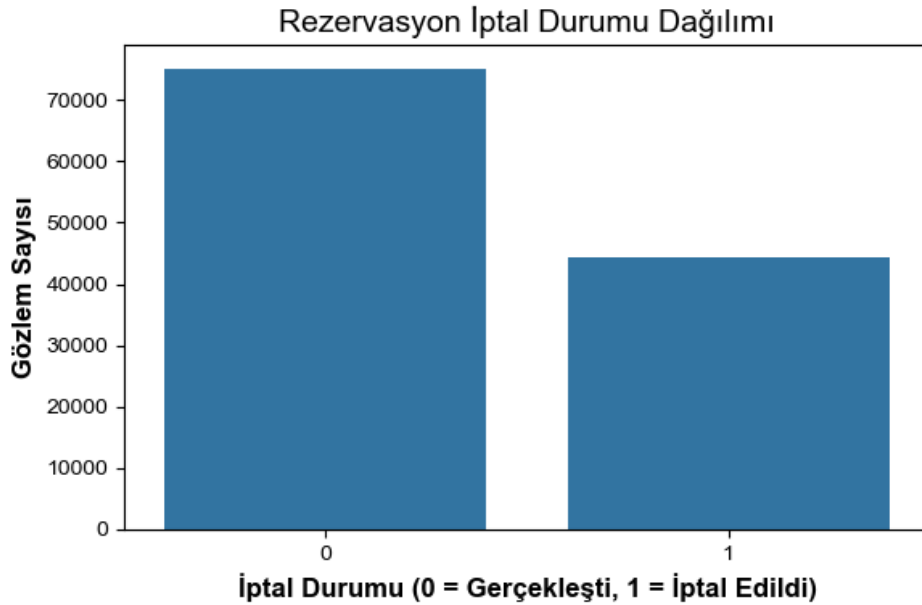
dalgalanmalarının yaşandığı dönemlerde otel gelirlerini doğrudan etkileyen kritik bir sorun alanı oluşturmaktadır. Rezervasyon iptallerinin öngörülebilirliği, sadece oda doluluk oranlarının değil, aynı zamanda operasyonel planlamaların, insan kaynağı tahsisinin ve fiyatlandırma stratejilerinin etkin bir biçimde yönetilmesine katkı sağlamaktadır. Bu çerçevede çalışmanın temel amacı, kamuya açık bir otel rezervasyon veri kümesi üzerinde CatBoost algoritmasını kullanarak iptal tahmini gerçekleştirmek, veri dengesizliğini SMOTE yöntemiyle gidermek ve model performansını ROC eğrileri, F1 skoru ve SHAP analizi gibi ileri düzey tekniklerle değerlendirmektir. Ayrıca modelin başarımı görselleştirilerek ve açıklanabilir yapay zekâ araçları ile yorumlanarak otel yöneticileri ve sektörel karar vericiler için pratik faydalar üretmek hedeflenmektedir.

Bu araştırmanın literatüre sağladığı katkı üç düzeyde değerlendirilebilir. İlk olarak, mevcut çalışmalar çoğunlukla temel sınıflandırma algoritmaları kullanırken, bu çalışmada hem güçlü hem de yorumlanabilir bir model olan CatBoost ile detaylı bir tahminleme süreci gerçekleştirilmiştir. İkinci olarak, modelin karar mekanizması SHAP analizi ile yorumlanmış, böylece değişkenlerin iptal kararına etkisi açık biçimde ortaya konmuştur. Son olarak çalışmanın çıktıları Türkçe olarak bilimsel ve uygulamalı bağlamda ele alınarak, konaklama sektörüne yönelik yerli literatürün genişletilmesine katkı sağlanmıştır.

3.4. Keşifsel Veri Analizleri

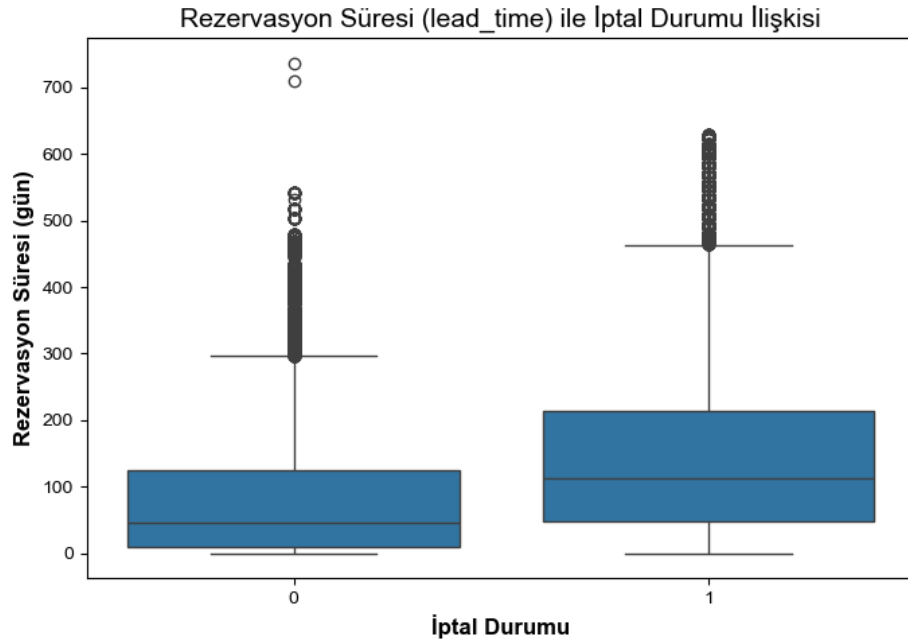
John W. Tukey tarafından 1977 yılında “*Keşif Verileri Analizi*” kitabı yazılmıştır. Böylelikle keşifsel veri analizleri ortaya çıkmıştır. keşifsel veri analizi (*exploratory data analysis; EDA*), genellikle istatistiksel grafikler ve diğer veri görselleştirme yöntemlerini kullanarak temel özelliklerini özetlemek için veri kümelerini analiz etme yaklaşımıdır. eşfedici veri analizi, sağlam istatistikler, parametrik olmayan istatistikler ve istatistiksel programlama dillerinin geliştirilmesine yarar sağlamıştır. Ayrıca istatistikçilerin bilimsel ve mühendislik problemleri üzerindeki çalışmalarını kolaylaştırmıştır (Chatfield, 1995; Fernholz ve Morgenthaler, 2000).

Şekil 1. Otel Rezervasyonlarının İptal Durumuna Göre Dağılımı



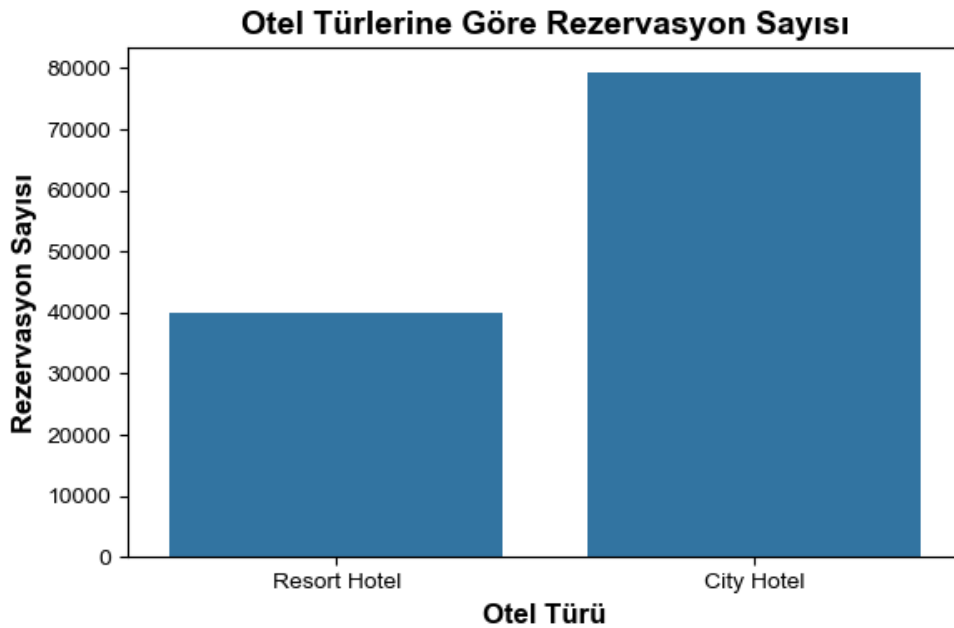
Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 1’de otel rezervasyonlarının iptal durumlarına göre dağılımı gösterilmektedir. Görüldüğü üzere, toplam rezervasyonların yaklaşık %63’ü gerçekleşmiş, %37’si ise iptal edilmiştir. Dolayısıyla, veri kümesinde sınıf dengesizliği bulunmakta ve sınıflandırma modellerinde mevcut dengesizliğin dikkate alınması gerekmektedir. İptal edilen rezervasyonların oranı, sektör açısından önemli bir gelir kaybı riskini yansıtmaktadır. Bu nedenle, iptal olasılığı yüksek olan rezervasyonların erkenden tespit edilmesi ve önleyici stratejilerin geliştirilmesi, otel yönetimleri açısından kritik öneme sahiptir.

Şekil 2. Rezervasyon Süresi ile İptal Durumu Arasındaki İlişki

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 2’de, rezervasyonun yapıldığı tarih ile konaklama tarihi arasındaki süreyi temsil eden lead_time değişkeni ile rezervasyonun iptal edilip edilmediğini gösteren is_canceled değişkeni arasındaki ilişki boxplot yöntemiyle görselleştirilmiştir. Görsel incelendiğinde, iptal edilen rezervasyonların ortalama ve medyan lead_time değerlerinin, iptal edilmeyen rezervasyonlara kıyasla belirgin düzeyde daha yüksek olduğu görülmektedir. Ayrıca iptal edilen grupta varyans daha geniştir ve daha fazla uç değer (outlier) gözlemlenmiştir. Bu bulgular, erken yapılan rezervasyonların iptal edilme olasılığının daha yüksek olduğunu ve lead_time değişkeninin rezervasyon iptallerini öngörmede anlamlı bir belirleyici olduğunu göstermektedir. Özellikle konaklama tarihine yakın yapılan rezervasyonlarda iptal oranlarının düşük olması, otel yönetimlerinin rezervasyon politikalarını bu eğilime göre yeniden şekillendirmelerini gerekli kılmaktadır.

Şekil 3. Otel Türlerine Göre Toplam Rezervasyon Sayılarının Dağılımı

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 3’te, otel türlerine göre toplam rezervasyon sayılarına ilişkin dağılım yer almaktadır. Verilere göre, “City Hotel” olarak sınıflandırılan şehir otelleri toplam 79.330 rezervasyonla, “Resort Hotel” olarak tanımlanan tatil

otellerine kıyasla (40.060 rezervasyon) daha fazla talep görmüştür. Bu durum, şehir otellerinin yıl boyunca daha istikrarlı bir doluluk oranına sahip olması, iş seyahatleri ve bireysel konaklamalar gibi süreklilik arz eden talep kaynaklarına hizmet etmesiyle açıklanabilir. Buna karşılık resort otellerin büyük ölçüde mevsimsel turizm hareketlerine bağlı olarak tercih edilmesi, toplam rezervasyon sayısının görece düşük kalmasına neden olmuştur. Elde edilen bu bulgu, otel türü ile müşteri talebi arasındaki ilişkinin arz-talep dengesini şekillendiren farklı pazar dinamikleriyle belirlendiğini göstermektedir.

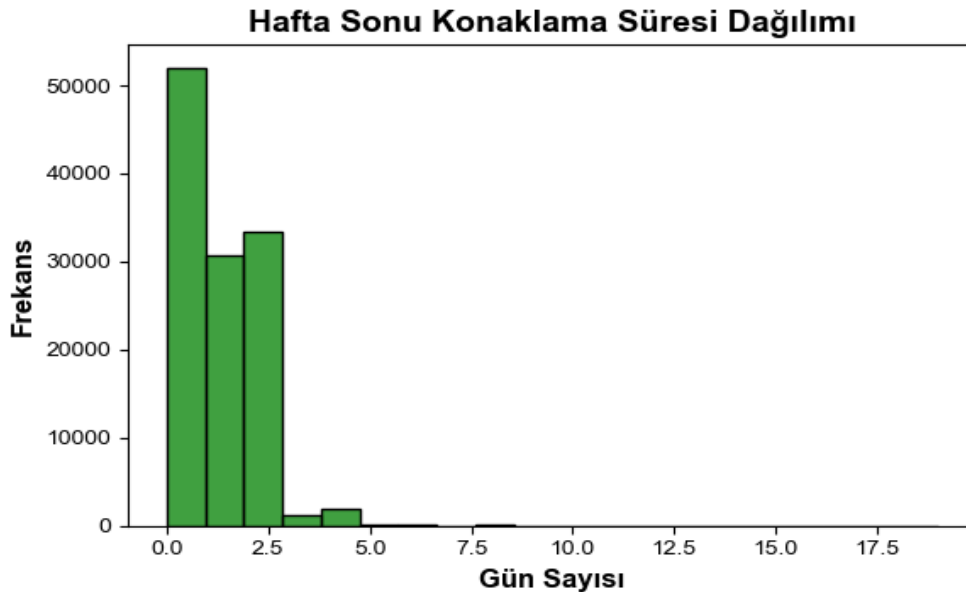
Şekil 4. Hafta İçi Konaklama Süresi Dağılımı



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 4'te, hafta içi gerçekleştirilen konaklamaların süresine ilişkin dağılım gösterilmiştir. Histogram analizine göre, konaklamaların büyük çoğunluğu 1 ila 3 gece arasında yoğunlaşmakta olup sırasıyla 30.310, 33.684 ve 22.258 rezervasyonla en yüksek frekansa sahiptir. Bu durum, kısa süreli iş seyahatleri veya bireysel amaçlı konaklamaların yaygınlığını yansıtmaktadır. Özellikle 1 gecelik konaklamaların yüksekliği, şehir otellerinin hafta içi kısa süreli konaklamalar için sıklıkla tercih edildiğini göstermektedir. 10 gece ve üzeri konaklamalar ise yalnızca 324 rezervasyon ile oldukça sınırlı düzeydedir ve bu da uzun süreli konaklamaların istisnai olduğunu ortaya koymaktadır. Elde edilen bulgular, hafta içi konaklama davranışlarının büyük ölçüde kısa vadeli olduğunu göstermekte ve otel işletmelerinin fiyatlandırma, promosyon ve hizmet paketlerini bu eğilim doğrultusunda yeniden yapılandırmaları gerektiğine işaret etmektedir.

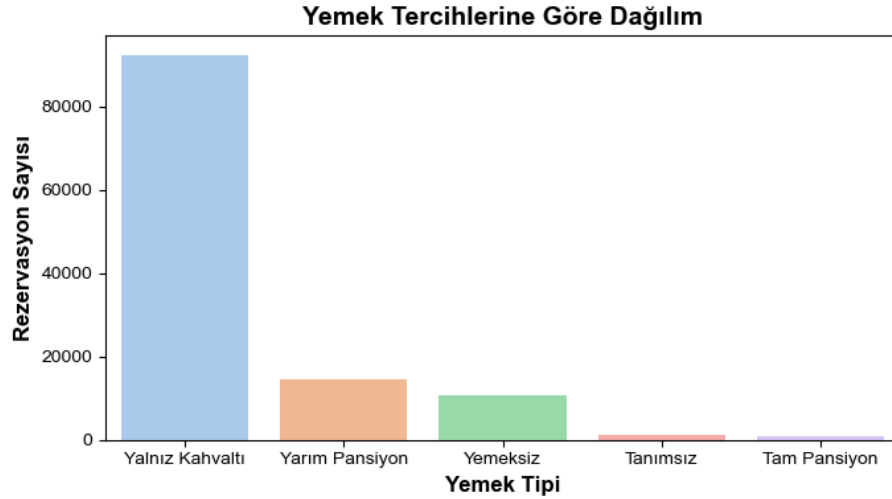
Şekil 5. Hafta Sonu Konaklama Süresi Dağılımı



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 5’te, hafta sonu gerçekleşen konaklamaların gece sayısına göre dağılımı sunulmuştur. Verilere göre, rezervasyonların önemli bir bölümü 1 (n=30.626) veya 2 (n=33.308) gecelik konaklamalardan oluşmaktadır. Hafta sonu tatillerinin büyük ölçüde kısa süreli planlandığı ve bireylerin kısa zaman dilimlerinde seyahat etmeyi tercih ettiği anlaşılmaktadır. Öte yandan, 3 gece ve üzeri konaklamalar oldukça sınırlı düzeyde gerçekleşmiş (n=3.478) ve toplam rezervasyonların yalnızca yaklaşık %2,9’unu oluşturmuştur. Bu bulgu, hafta sonu konaklama davranışlarının büyük ölçüde kısa süreli olduğunu ortaya koymakta; otel işletmeleri açısından promosyon, fiyatlandırma ve hizmet stratejilerinin bu tüketim eğilimi doğrultusunda yapılandırılması gerektiğine işaret etmektedir.

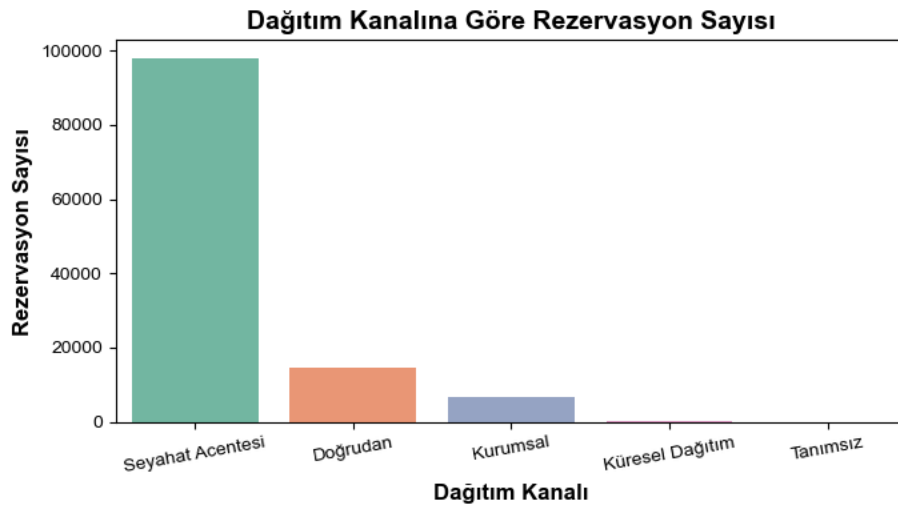
Şekil 6. Müşteri Rezervasyonlarına Göre Yemek Tercihlerinin Dağılımı



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 6’da, otel müşterilerinin tercih ettikleri yemek türlerinin rezervasyon sayısına göre dağılımı sunulmuştur. Verilere göre, “*Yalnız Kahvaltı*” seçeneği açık ara en fazla tercih edilen yemek türü olup 92.310 rezervasyon ile öne çıkmaktadır. Bu bulgu, müşterilerin genellikle daha esnek ve ekonomik konaklama seçeneklerine yöneldiğini göstermektedir. “*Yarım Pansiyon*” ve “*Yemeksiz*” tercihlerinin sırasıyla 14.463 ve 10.650 adet rezervasyon ile orta düzeyde tercih edildiği görülmektedir. Öte yandan “*Tam Pansiyon*” ve “*Tanımsız*” kategorileri oldukça sınırlı sayıda gözlem içermektedir. “*Tanımsız*” kategorisinde bulunan 1.169 adet rezervasyon, veri giriş sürecindeki eksiklikten veya sistemsel sorunlardan meydana gelmiş olabilir. Bu bulgular, otel işletmeleri için müşteri eğilimlerini analiz ederek yemek seçeneklerini yeniden yapılandırma konusunda stratejik bir fırsat sunmaktadır.

Şekil 7. Dağıtım Kanallarına Göre Toplam Rezervasyon Sayılarının Dağılımı



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 7’de, otel rezervasyonlarının dağıtım kanallarına göre sayısal dağılımı sunulmuştur. Rezervasyonların büyük çoğunluğunun “*Seyahat Acentesi*” aracılığı ile gerçekleştirildiği görülmektedir (n=97.870). Otel sektöründe geleneksel tur operatörleri ve seyahat acentelerinin hâlen belirleyici bir rol oynadığı anlaşılmaktadır.

İkinci sırada yer alan “*Doğrudan*” seçeneği (n=14.645), bireysel müşterilerin otel ile doğrudan iletişime geçerek veya çevrimiçi rezervasyon sistemleri ile işlemler yaptıklarını göstermektedir. “*Kurumsal*” (n=6.677) ve “*GDS*” (n=193) gibi kanallar ise daha spesifik müşteri segmentlerine hitap etmekte olup, toplam rezervasyonlar içindeki payları sınırlı kalmıştır. “*Tanımsız*” olarak kodlanan az sayıdaki kayıt (n=5) ise veri girişinde yaşanan eksikliklerden kaynaklanabilir. Bu dağılım, otel yönetimlerinin rezervasyon kaynaklarını analiz ederek kanal bazlı stratejik kararlar geliştirmesinin önemine işaret etmektedir.

4. MAKİNE ÖĞRENMESİ MODELLERİ

Makine öğrenmesi, bilgisayar sistemlerinin veri analizi yoluyla öğrenmesini ve bu süreçte elde edilen bilgiler doğrultusunda tahmin yapabilme yetisi kazanmasını amaçlayan bir bilim dalıdır. Çözüm süreçlerinde istatistiksel ve matematiksel modeller kullanılarak veri kümesinin yapısına ve problem türüne uygun algoritmalar seçilmektedir (Eşidir, 2025b).

4.1. Random Forest (Rastgele Orman)

Random Forest algoritması, topluluk öğrenme (ensemble learning) yaklaşımına dayalı olarak geliştirilen, sınıflandırma ve regresyon problemlerinde yaygın şekilde kullanılan güçlü bir makine öğrenmesi yöntemidir. Bu model, farklı alt veri setleri üzerinde inşa edilen çok sayıda karar ağacının bir araya getirilmesiyle oluşturulur. Böylelikle her bir ağacın bireysel önyargıları ve varyansı dengelenerek daha sağlam, kararlı ve genellenebilir bir model elde edilmektedir (Oguine vd., 2021).

Algoritmanın temelinde bootstrap örnekleme yöntemi yer almakta olup, bu yöntem sayesinde eğitim verisinden çok sayıda alt örneklem (subsample) oluşturulmaktadır. Her bir alt örneklem, orijinal veri kümesinden rastgele seçilen örneklerden oluşmakta ve bu alt kümeler üzerinde bağımsız karar ağaçları inşa edilmektedir. Her bir ağacın eğitimi sırasında, düğüm bölünmelerinde tüm özellikler yerine rastgele seçilen bir alt küme dikkate alınır. Bu yaklaşım, modelin varyansını azaltmakta, karar ağaçları arasındaki korelasyonu düşürmekte ve böylece aşırı öğrenme (overfitting) riskini önemli ölçüde azaltmaktadır. Sınıflandırma problemlerinde nihai tahmin, genellikle her ağacın verdiği sınıf tahminine dayalı olarak çoğunluk oylaması (majority voting) yöntemiyle belirlenmektedir. Bu yapı sayesinde Random Forest, karmaşık ve yüksek boyutlu veri kümelerinde dahi yüksek doğruluk oranları ile etkili sonuçlar üretebilmektedir. Random Forest algoritması, parametre ayarlamalarına duyarsızlığı, yüksek boyutlu verilerle etkili çalışabilmesi ve aşırı öğrenmeye karşı dirençli yapısıyla öne çıkmaktadır. Ayrıca değişken önemini belirleme yeteneği de önemli bir avantajdır. Ancak modelin yorumlanabilirliği düşüktür ve çok sayıda ağacın birlikte çalışması nedeniyle özellikle büyük veri setlerinde hesaplama maliyeti artabilmektedir.

4.2. XGBoost (Extreme Gradient Boosting)

XGBoost, gradyan artırma algoritmalarının geliştirilmiş bir versiyonu olarak, yüksek doğruluk oranları ve hızlı hesaplama kapasitesi ile öne çıkan güçlü bir makine öğrenmesi yöntemidir. Sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılmakta olup, özellikle büyük ölçekli veri analizlerinde ve rekabetçi veri bilimi projelerinde sıklıkla tercih edilmektedir. Spam e-posta tespiti, reklam eşleştirme, dolandırıcılık analizi ve anomali tespiti gibi birçok farklı alanda başarılı sonuçlar sunması, XGBoost’un modern makine öğrenmesi uygulamalarındaki önemini artırmaktadır (Chen vd., 2016).

XGBoost, topluluk öğrenmesi (ensemble learning) yaklaşımı ile birden fazla zayıf öğreniciyi ardışık biçimde birleştirerek tahmin doğruluğunu artırmayı amaçlamaktadır. “*Extreme Gradient Boosting*” ifadesinin kısaltması olan XGBoost, karar ağaçları temelinde çalışan bir boosting algoritmasıdır. Modelin temel hedefi, her adımda hata fonksiyonunu minimize eden yeni ağaçlar üreterek mevcut öğrencilerin hatalarını düzeltmektir. Yüksek esnekliği ve hesaplama verimliliği sayesinde, özellikle büyük veri setlerinde etkili sonuçlar veren ölçeklenebilir bir çözüm sunmaktadır (Wang vd., 2020).

4.3. LightGBM (Light Gradient Boosting Machine)

LightGBM modeli, Microsoft firması tarafından geliştirilmiştir. Büyük veri setleri üzerinde hızlı ve verimli çalışabilen bir modeldir. Geleneksel gradyan artırma algoritmalarına kıyasla daha az bellek tüketmekte ve

eğitim süresi daha kısadır. Bu performans avantajı, histogram tabanlı bölünme yöntemi ile derinlik yerine yaprak odaklı (leaf-wise) büyüme stratejisinin birlikte kullanılmasından kaynaklanmaktadır. Bu strateji, modelin genel doğruluğunu artırırken hesaplama maliyetini azaltmaktadır. LightGBM, özellikle yüksek boyutlu ve seyrek metin verilerinden oluşan veri setlerinde etkili sonuçlar veren güçlü bir sınıflandırma algoritması olarak öne çıkmaktadır (Ke vd., 2017).

4.4. CatBoost (Categorical Boosting)

CatBoost modeli, Yandex firması tarafından geliştirilmiş bir gradyan artırma algoritmasıdır. Bu model, özellikle kategorik değişkenlerle etkili bir biçimde çalışacak şekilde optimize edilmiştir. CatBoost'un en dikkat çekici yönlerinden biri, diğer geleneksel boosting algoritmalarına kıyasla aşırı öğrenme (overfitting) eğiliminin daha düşük olmasıdır.

Modelin başarısının ardında yatan temel etkenler arasında sıralı öğrenme (ordered boosting) ve simetrik ağaç yapıları yer almaktadır. Bu yapılar sayesinde CatBoost, veri sırasına duyarlı olmayan ve daha kararlı sonuçlar üretebilen bir yapı ortaya koymaktadır. Buna karşılık, özellikle yüksek oranda kategorik değişken içeren veri setlerinde CatBoost modeli, sınıflandırma ve regresyon görevlerinde önemli avantajlar sunmaktadır (Prokhorenkova vd., 2018).

4.5. Modelleme Süreci

Otel rezervasyon iptallerinin tahminine yönelik olarak makine öğrenmesi tabanlı modeller kullanılarak sınıflandırma işlemleri gerçekleştirilmiştir. Modelleme süreci; veri ön işleme, eğitim-test ayrımı, hiperparametre optimizasyonu, çapraz doğrulama, model değerlendirme ve açıklanabilirlik analizleri olmak üzere çok katmanlı bir yapıda tasarlanmıştır.

İlk olarak, veri kümesi üzerinde eksik gözlemler barındıran kayıtlar temizlenmiş ve değişkenler türlerine göre sınıflandırılmıştır. Sayısal ve kategorik değişkenlerin ayrıştırılmasının ardından, sayısal değişkenler üzerinde standartlaştırma işlemi gerçekleştirilmiştir. Kategorik değişkenler tekil kodlama (One-Hot Encoding) yöntemi ile dönüştürülmüştür. Bu işlemler, ColumnTransformer yapısı içinde yapılandırılan bir boru hattı (pipeline) aracılığıyla uygulanmıştır. Modelleme sürecine başlamadan önce, hedef değişken olan “*is_canceled*” değişkeni dikkate alınarak veri kümesi %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Bu ayrım, sınıf dağılımını koruyacak biçimde stratified olarak gerçekleştirilmiştir.

Dört farklı makine öğrenmesi modeli (Random Forest, XGBoost, LightGBM ve CatBoost) ile analizler gerçekleştirilmiştir. Her bir model için, RandomizedSearchCV ile hiperparametre optimizasyonu yapılmıştır. Bu süreçte, işlem süresini verimli kullanmak adına yinleme sayısı 15 ile sınırlandırılmıştır. Hiperparametre ayarlamaları sırasında 5 katlı StratifiedKFold çapraz doğrulama kullanılarak her katmanda eğitim işlemleri tekrarlanarak modelin genel performansı test edilmiştir.

Modellerin başarısı, test verisi üzerinde çeşitli değerlendirme metrikleriyle analiz edilmiştir. Doğruluk oranı, ROC AUC skoru, karışıklık matrisi vb. kapsamlı bir performans analizi gerçekleştirilmiştir. ROC eğrileri her model için ayrı ayrı görselleştirilmiş, ayrıca F1 skorları bulunmuştur. Model kararlarının açıklanabilirliğini sağlamak için SHAP yöntemi kullanılmıştır. Modelleme süreçlerinin tamamı Python ortamında geliştirilmiştir.

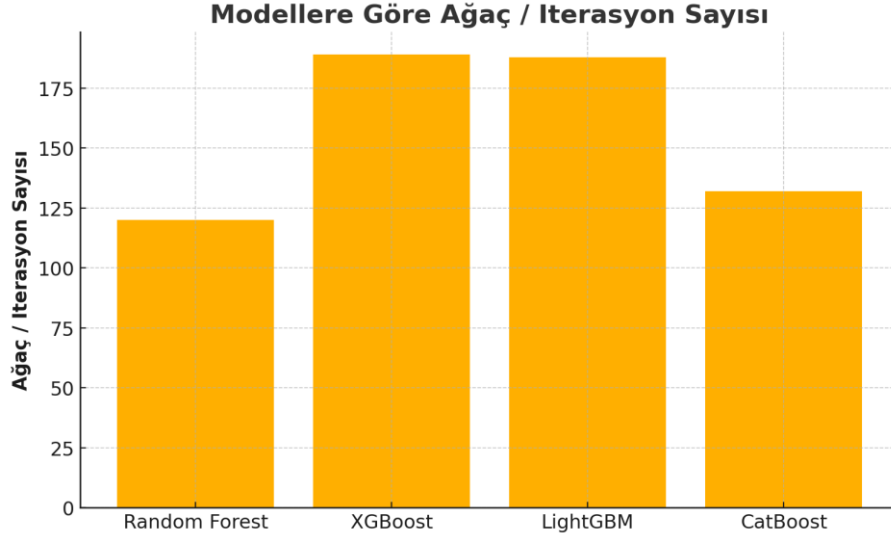
4.6. Model Geliştirme ve Hiperparametre Ayarlamaları

Çalışmada kullanılan dört farklı makine öğrenmesi modeli için, en yüksek sınıflandırma performansına ulaşmak amacıyla hiperparametre ayarlamaları gerçekleştirilmiştir. Bu kapsamda, her modelin kendi yapısal özelliklerine uygun biçimde belirli parametreler farklı değerlerde denenmiş ve en iyi sonuçları veren ayar kombinasyonları tercih edilmiştir.

Random Forest modelinde, karar ağaçlarının sayısı artırılmış ve her bir ağacın bölünme koşulları ile ilgili alt sınırlar optimize edilmiştir. Ayrıca modelin karar verirken kullanacağı değişken sayısı da sınırlandırılarak aşırı öğrenmenin önüne geçilmiştir. XGBoost ve LightGBM modellerinde, özellikle öğrenme oranı, ağaç derinliği ve örneklem oranları gibi temel parametreler üzerinde durulmuştur. Bu ayarlamalar sayesinde, modellerin hem hızlı öğrenmesi hem de genel geçer sonuçlar üretmesi sağlanmıştır.

CatBoost modelinde ise yine öğrenme hızı ve ağacın derinliği gibi temel ayarlar dengelenmiş; bunun yanı sıra, modelin farklı veri alt kümeleri üzerinde tutarlı çalışmasını sağlayan ek düzenlemeler yapılmıştır. Yapılan bu hiperparametre ayarlamaları sonucunda, her bir model en verimli performansı gösterecek biçimde yapılandırılmıştır. Bu da sınıflandırma doğruluğunu ve modelin farklı veri setlerinde benzer başarıyı tekrar etme yetisini (genellenebilirliğini) artırmıştır. Ayarlamaların amacı yalnızca yüksek başarı elde etmek değil, aynı zamanda her modelin adil ve dengeli biçimde karşılaştırılabilmesini sağlamak olmuştur.

Şekil 8. Modellere Göre Ağaç/İterasyon Sayısı



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 8’de görüldüğü üzere, XGBoost ve LightGBM modelleri yaklaşık olarak 190 iterasyon ile en fazla karar ağacı oluşturan modellerdir. Bu modeller daha fazla yineleme yaparak daha ayrıntılı örüntüler öğrenmeye çalışmıştır. CatBoost modeli ise 132 iterasyonla orta düzeyde bir yapı sergilerken, Random Forest ise 120 ağaç ile en az sayıda modelleme ögesi kullanmıştır. Ancak Random Forest, daha düşük iterasyon sayısına rağmen yüksek başarı oranları elde ederek dengeli bir yapı ortaya koymuştur. Yapılan karşılaştırma, modellerin yalnızca doğruluklarına göre değil, aynı zamanda öğrenme sürecinde ne kadar kaynak tükettiğine göre de değerlendirilmesi gerektiğini ortaya koymaktadır.

5. ANALİZ SONUÇLARI VE ELDE EDİLEN BULGULAR

Dört makine öğrenmesi modelinin (Random Forest, XGBoost, LightGBM, CatBoost) sınıflandırma performansları çoklu metrikler ile karşılaştırılmıştır. En yüksek doğruluk (%81), AUC skoru (0,89) ve F1 skoru (0,71) Random Forest modeli tarafından elde edilmiştir. Eğitim süresi açısından en verimli model ise LightGBM olmuştur. Sonuçlar, veri dengesizliğine rağmen güçlü ve güvenilir bir sınıflandırma başarısına ulaşıldığını ortaya koymaktadır.

5.1. Modellerin Performanslarının Karşılaştırılması

Makine öğrenmesi modellerinin sınıflandırma başarıları, çok boyutlu performans metrikleri ile karşılaştırılmıştır. Modellerin değerlendirilmesinde doğruluk oranı, eğri altı alan (AUC), kesinlik, duyarlılık (hassasiyet), F1 skoru, çapraz doğrulama skoru ve eğitim süresi gibi ölçütler dikkate alınmıştır. Elde edilen bulgular, model seçiminin yalnızca tek bir performans metriğine dayanarak yapılmasının yanıltıcı olabileceğini göstermektedir. Çoklu ölçütlerin birlikte değerlendirilmesi gerektiği ortaya konulmaktadır.

Tablo 2. Makine Öğrenmesi Modellerinin Performans Karşılaştırması

Modeller	Doğruluk Oranı	AUC Skoru	Kesinlik	Hassasiyet	F1 Skoru	Çapraz Doğrulama Skoru
Random Forest	0,81	0,89	0,84	0,62	0,71	0,89
XGBoost	0,79	0,87	0,82	0,57	0,67	0,87
LightGBM	0,80	0,88	0,82	0,59	0,69	0,88
CatBoost	0,79	0,86	0,82	0,56	0,66	0,86

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 2’de dört farklı modelin çoklu başarı ölçütleri temelinde karşılaştırmalı sonuçları sunulmuştur. En yüksek doğruluk oranı ve eğri altı alan skoru Random Forest modeline ait olup; bu model aynı zamanda F1 skoru ve çapraz doğrulama başarımı açısından da üstünlük göstermektedir. Buna karşın eğitim süresi bakımından en verimli model LightGBM olarak öne çıkmaktadır.

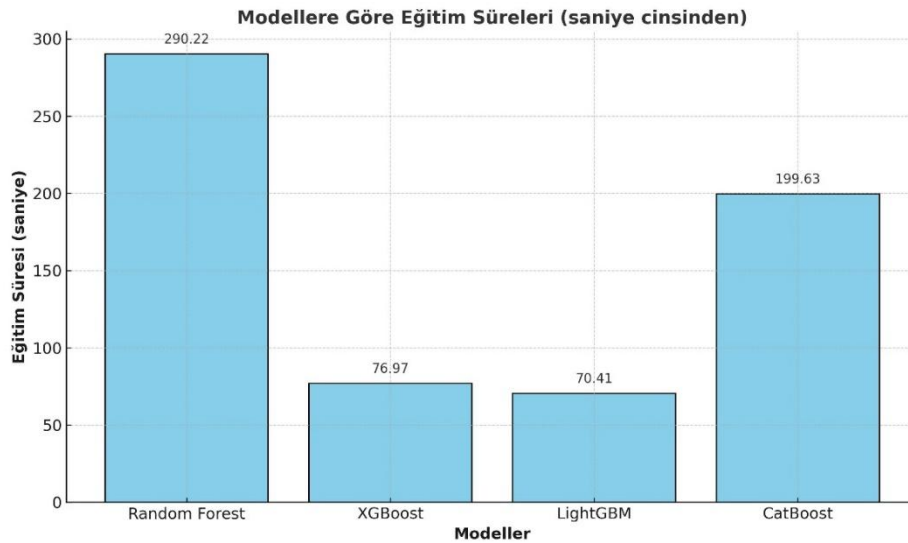
Analiz sonuçlarına göre, Random Forest modeli, %81 doğruluk oranı ve 0,89 AUC skoru ile diğer modellere kıyasla daha yüksek sınıflandırma başarısı sergilemiştir. Bu bulgu, modelin hem doğru sınıflandırma yapma yeteneğinin hem de pozitif ve negatif sınıflar arasındaki ayrımı yüksek doğrulukla gerçekleştirebildiğinin göstergesidir. Aynı model, 0,71 düzeyindeki F1 skoru ile denge ölçütü açısından da üstünlük göstermektedir. F1 skorunun yüksekliği, modelin özellikle dengesiz veri setlerinde hem kesinlik hem de duyarlılığı dengeli biçimde sağladığını ortaya koymaktadır. Ayrıca, 0,89’luk çapraz doğrulama skoru, Random Forest’ın farklı veri alt kümelerinde de tutarlı ve güvenilir sonuçlar verdiğini göstermekte, modelin genellenebilirliğini desteklemektedir.

LightGBM modeli %80 doğruluk oranı, 0,88 AUC skoru ve 0,69 F1 skoru ile sınıflandırma performansı açısından da tatmin edici sonuçlar sunmuştur. XGBoost ve CatBoost modelleri ise genel olarak benzer doğruluk ve AUC skorları üretmiştir. Her iki model de %79 doğruluk ve 0,86–0,87 arasında değişen AUC değerleriyle istikrarlı performans göstermiştir. Ancak bu iki modelin duyarlılık düzeyleri diğer modellere göre daha düşüktür (XGBoost: 0,57; CatBoost: 0,56). Pozitif sınıfların tespitinde bu modeller daha düşük başarı sağlamıştır.

5.2. Hesaplama Süresi Açısından Model Verimliliği

Şekil 9’da modellerin eğitim süreleri verilmiştir. LightGBM ve XGBoost modelleri, kısa eğitim süreleriyle dikkat çekerken; Random Forest modeli en uzun eğitim süresine sahip olmuştur. Modellerin sınıflandırma performansları kadar eğitim süreleri de uygulama açısından önem arz etmektedir.

Şekil 9. Modellerin Eğitim Sürelerinin Karşılaştırılması



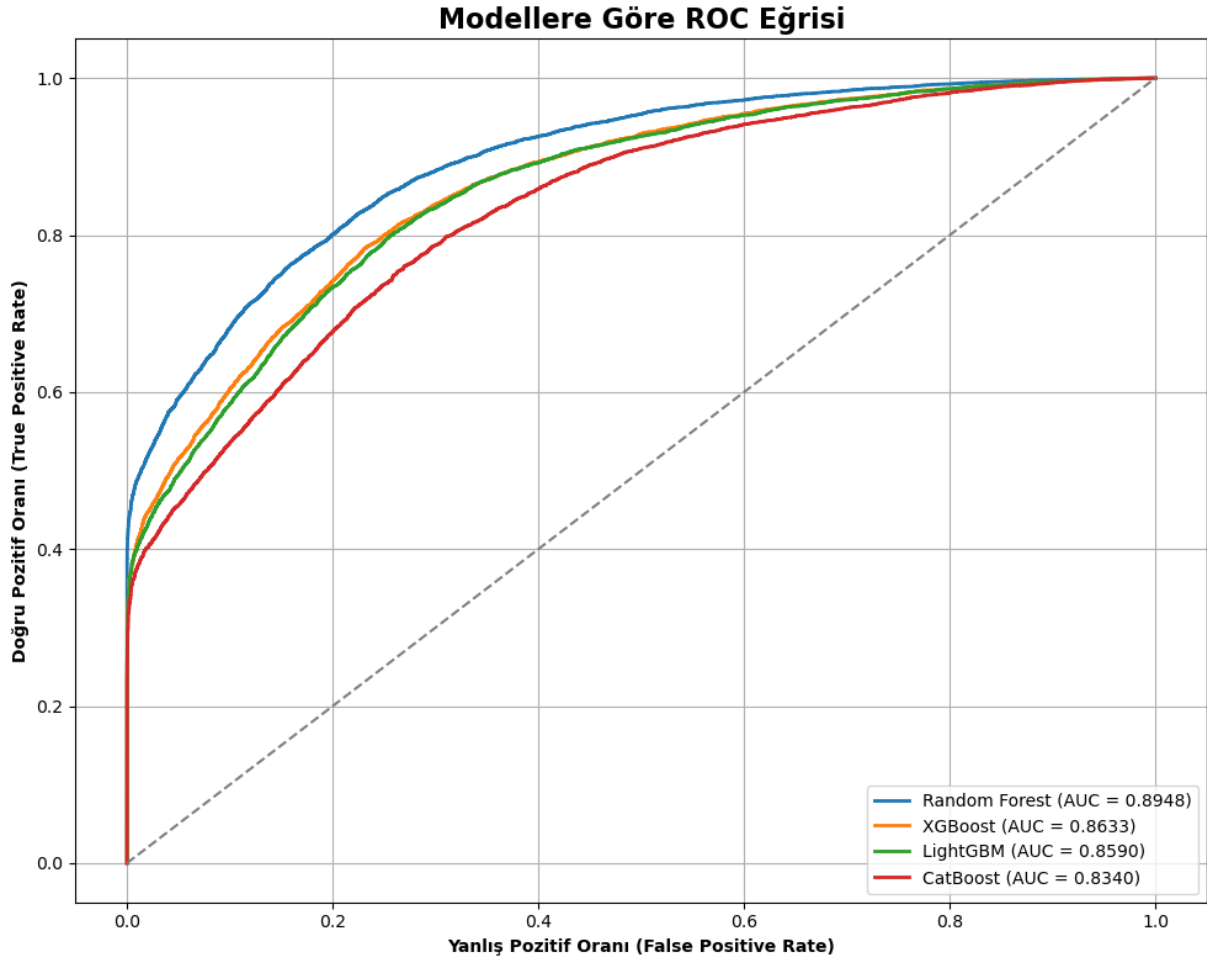
Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 9’da görüldüğü üzere, Random Forest modeli 290,22 saniye ile en uzun eğitim süresine sahipken; LightGBM ve XGBoost sırasıyla 70,41 ve 76,97 saniyelik sürelerle çok daha hızlı şekilde eğitilmiştir. CatBoost ise 199,63 saniyelik süresiyle orta düzeyde bir eğitim süresi sergilemiştir. Yüksek başarı oranına rağmen Random Forest modeli zaman ve kaynak açısından daha maliyetli bir seçenek olurken, LightGBM ve XGBoost modelleri özellikle zaman kısıtlı uygulamalar için daha uygun alternatifler sunmaktadır.

5.3. Grafikselsel Performans Analizleri

Modellerin performansları, ROC eğrileri ve F1 skoru karşılaştırma grafikleri ile görselleştirilmiştir. ROC eğrileri, modellerin sınıflandırma gücünü farklı eşik değerlerinde değerlendirme imkanı sunarken, AUC değeri model genel başarısını özetlemektedir. F1 skoru grafikselsel karşılaştırması ise modellerin dengesiz veri yapısı üzerindeki performansını açıkça ortaya koymaktadır. Bu görseller, okuyucunun model karşılaştırmasını kolayca kavramasını sağlamak ve makalenin görsel anlatımını güçlendirmektedir.

Şekil 10. Modellerin ROC Eğrileri

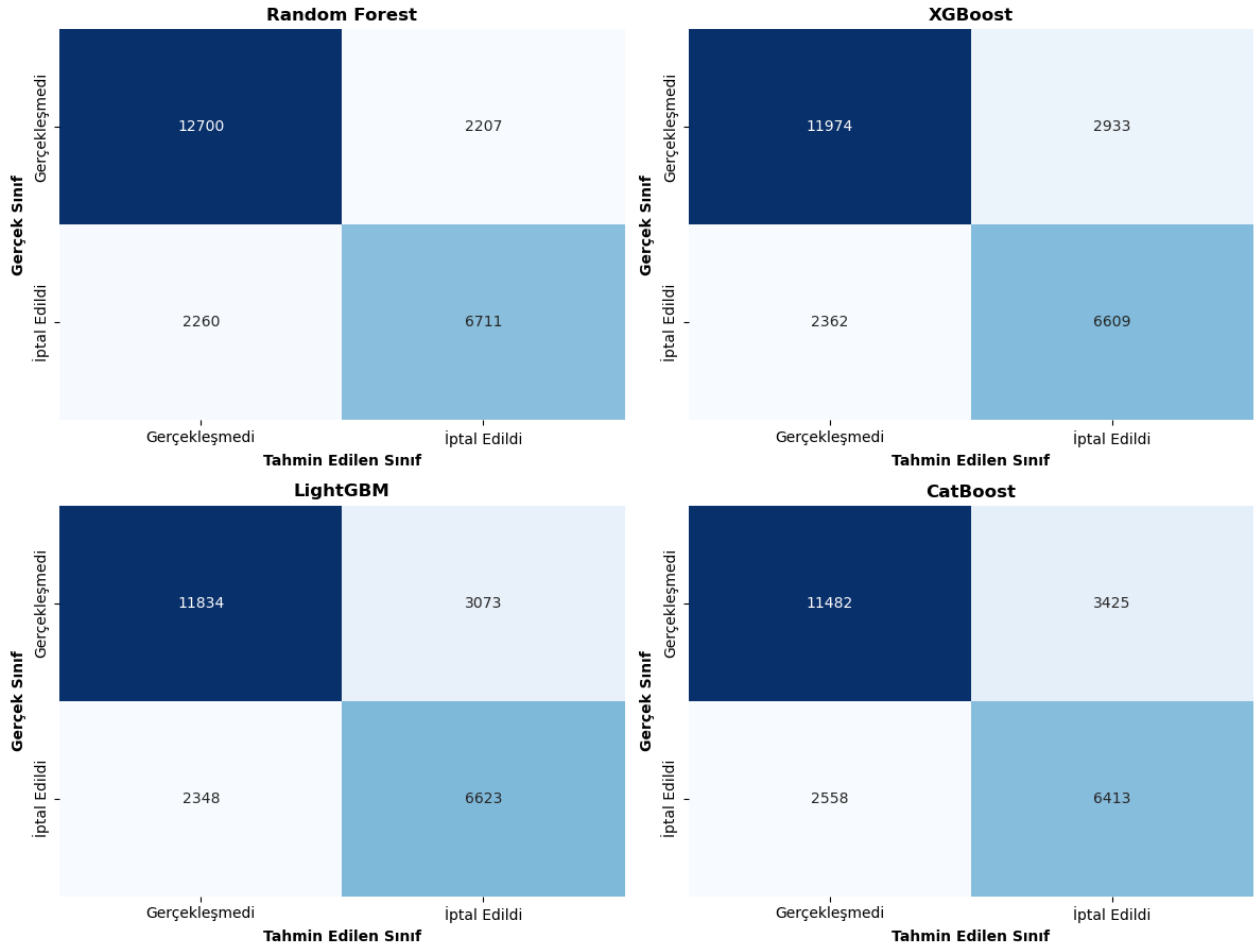


Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 10'da modellerin ROC eğrileri görülmektedir. Modellerin sınıflandırma başarıları, ROC eğrileri ve AUC skorları üzerinden karşılaştırılmıştır. ROC eğrisi, modelin farklı eşik değerlerindeki duyarlılık (True Positive Rate) ile özgüllük (False Positive Rate) arasındaki dengeyi görsel olarak sunmakta ve genel ayırtırma gücünü AUC (Area Under the Curve) skoru ile ifade etmektedir. Elde edilen AUC değerleri, Random Forest (0.8948), XGBoost (0.8633), LightGBM (0.8590) ve CatBoost (0.8340) şeklindedir.

Sonuçlara göre, Random Forest modeli en yüksek AUC skoruna ulaşarak iptal tahmini görevinde en başarılı model olarak öne çıkmıştır. XGBoost ve LightGBM benzer performans gösterirken, CatBoost modeli nispeten daha düşük bir AUC değeri sunmuştur. Bulgular, tüm modellerin ayırt edici güce sahip olduğunu göstermektedir. Random Forest modeli, otel rezervasyonlarında iptal olasılığını tahmin etme konusunda en dengeli performansa sahip model olarak ön plana çıkmaktadır.

Şekil 11. Modellerle Göre Karışıklık Matrisleri
Modellere Göre Karışıklık Matrisleri

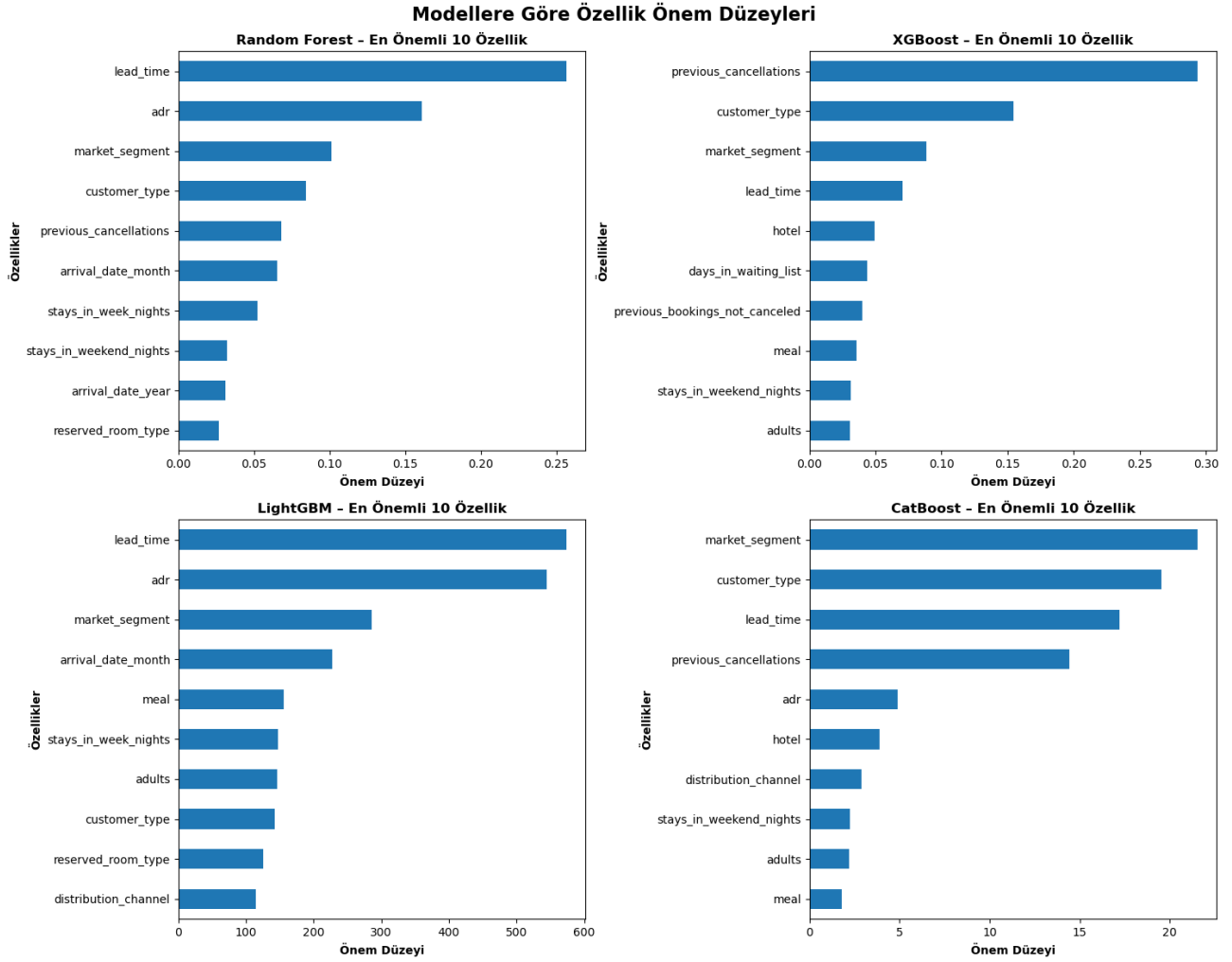


Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 11’de, her bir modele ait karışıklık matrisleri yer almaktadır. Bu matrisler üzerinden modellerin tahmin başarısı nicel olarak karşılaştırılmıştır. Random Forest modeli, doğru sınıflandırma sayısı bakımından (TP=6.711, TN=12.700) en yüksek performansı göstermiştir. Bu model, özellikle "*iptal edildi*" sınıfını ayırt etmede diğer modellere kıyasla daha isabetli tahminler üretmiştir.

XGBoost ve LightGBM modelleri, genel başarı açısından birbirine oldukça yakın değerler sunarken, "*iptal edildi*" sınıfında görece daha fazla yanlış sınıflama yapmıştır. CatBoost modeli ise özellikle her iki sınıfta da diğer modellere göre daha düşük doğru sınıflandırma başarısı göstermiştir. Karışıklık matrislerinden elde edilen bu bulgular, Random Forest modelinin sınıflandırma görevinde daha dengeli ve güvenilir bir performans sergilediğini göstermektedir.

Şekil 12. Modellere Göre İptal Tahmininde En Etkili İlk 10 Öznitelik



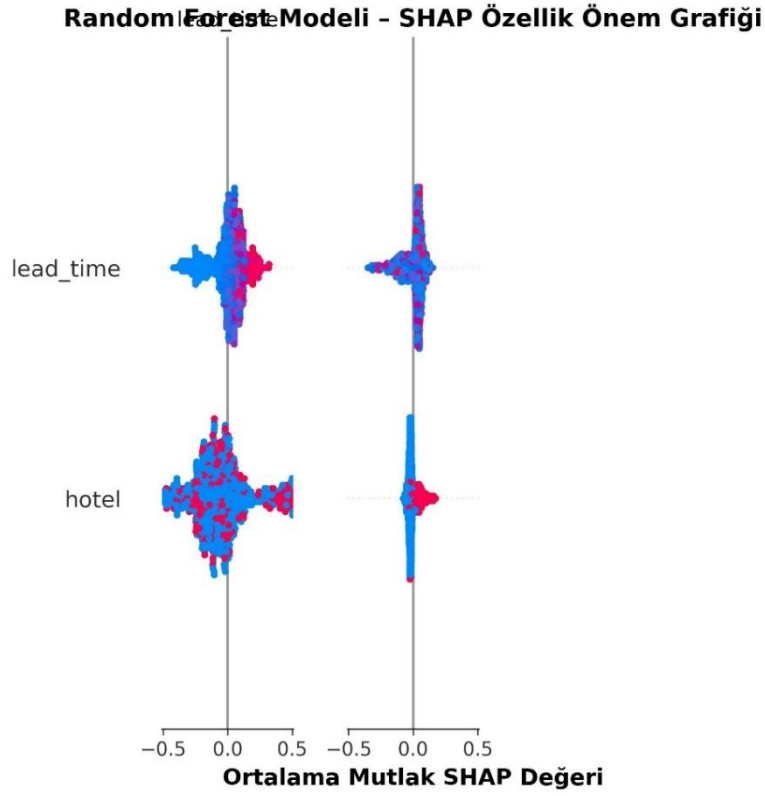
Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 12, dört farklı sınıflandırma modeline göre otel rezervasyon verilerindeki en önemli özniteliklerin sıralamasını göstermektedir. Random Forest ve LightGBM modellerinde “lead_time” (rezervasyon ile giriş tarihi arasındaki süre) en belirleyici öznitelik olarak öne çıkarken, XGBoost modelinde “previous_cancellations” ve CatBoost modelinde ise “market_segment” özniteliği ilk sırada yer almıştır. Tüm modeller arasında ortak olarak öne çıkan bazı öznitelikler ise: “lead_time”, “customer_type”, “market_segment” ve “previous_cancellations” değişkenleridir. Bu özniteliklerin modeller tarafından tutarlı biçimde yüksek önem düzeyinde değerlendirilmesi, iptal tahmininde güçlü belirleyiciler olduklarını göstermektedir. Ayrıca modellerin farklı algoritmik yapıları nedeniyle önem derecelerinde farklılıklar görülmektedir. Bu durum, özellik seçiminde model temelli varyasyonların dikkate alınması gerektiğini ve açıklanabilirliğin model özelinde analiz edilmesinin önemini ortaya koymaktadır.

5.4. En Başarılı Model ve Açıklanabilirlik Analizi

Gerçekleştirilen karşılaştırmalı modelleme sonuçlarına göre, Random Forest algoritması, otel rezervasyon iptallerini tahmin etmede en yüksek başarıyı göstermiştir. Doğruluk oranı %81,28 ve AUC skoru 0.895 olan bu model, aynı zamanda düşük yanlış pozitif (FP) ve yanlış negatif (FN) değerleriyle istikrarlı bir performans sergilemiştir. Bu nedenle, açıklanabilirlik temelli analizler yalnızca bu model için sunulmuştur. Modelin karar mekanizmalarının şeffaf biçimde anlaşılabilmesi amacıyla SHAP analizi gerçekleştirilmiş ve özniteliklerin sınıflandırma üzerindeki katkı düzeyleri incelenmiştir. Şekil 12’de görüldüğü üzere, “lead_time”, “hotel” ve “adr” gibi değişkenler model kararlarını en çok etkileyen faktörler olarak öne çıkmaktadır. SHAP değerlerinin pozitif ya da negatif etkisi, her bir değişkenin sınıflandırma üzerindeki yönünü de açık biçimde ortaya koymaktadır.

Şekil 13. Random Forest Modeli SHAP Özellik Önem Grafiği



Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 13'te, Random Forest modelinde öne çıkan “*lead_time*” ve “*hotel*” değişkenlerinin tahminlere etkisi SHAP analizi ile gösterilmiştir. Uzun rezervasyon süresi (*lead_time*) iptal olasılığını artırırken, otel tipi (*hotel*) değişkeni de sınıflandırma üzerinde belirleyici bir rol oynamaktadır. Yüksek SHAP değerleri modelin kararlarına güçlü katkılar sağlamış, bu da modelin karar verme sürecinin daha şeffaf biçimde anlaşılmasına olanak tanımıştır.

6. TARTIŞMA

Analizlerden elde edilen sonuçlar ve bulgular, literatürdeki çalışmalar ile genelde örtüşmektedir. Örneğin Antonio vd. (2017) ile Şahinbaş ve Doğuç (2022) tarafından yapılan çalışmalarda, rezervasyon iptallerinin öngörülebilir bir yapıya sahip olduğu ve özellikle “*lead_time*” ile “*customer_type*” gibi değişkenlerin belirleyici rol oynadığı vurgulanmıştır.

Elde edilen bulgular, otel rezervasyon iptallerinin makine öğrenmesi modelleri ile yüksek doğrulukta tahmin edilebileceğini bir kez daha ortaya koymaktadır. Özellikle Random Forest modeli, %81,28 doğruluk oranı ve 0,895 AUC skoru ile diğer tüm modellerden daha yüksek bir performans sergilemiştir. Modelin karar mekanizmasının SHAP analizi ile açıklanması, tahmin sürecine etki eden değişkenlerin (*lead_time*, *adr*, *hotel* vb.) görece öneminin açık biçimde ortaya konmasına imkân tanımıştır. Bu durum, geliştirilen modelin yalnızca güçlü sonuçlar üretmekle kalmadığını, aynı zamanda karar süreçlerinin şeffaf ve yorumlanabilir olduğunu da göstermektedir.

Bu sonuçlar, literatürde benzer veri kümesi ile gerçekleştirilen çalışmalarla karşılaştırıldığında belirgin farklılıklar sunmaktadır. Örneğin Ercan ve Irmak (2022) tarafından yürütülen çalışmada, aynı veri kümesi kullanılarak destek vektör makineleri, lojistik regresyon ve karar ağaçları gibi modeller uygulanmış ve en yüksek başarı SVM ile %78,69 doğruluk ve 0,835 AUC skoru olarak elde edilmiştir. Ancak ilgili çalışmada veri dengesizliği sorunu giderilmemiş ve model karar mekanizmasının açıklanmasına yönelik herhangi bir yöntem (örneğin SHAP) kullanılmamıştır. Bunun yanı sıra değişken önem düzeylerine ilişkin yorumlamalara da yer verilmemiştir.

Çalışmada SMOTE yöntemiyle veri dengesizliği sorunu giderilmiş ve sınıflandırma başarısı artırılmıştır. Dört farklı makine öğrenmesi modeli ile karşılaştırma yapılmış ve en başarılı model belirlenmiştir. SHAP analizi kullanılarak modelin karar mekanizması görselleştirilmiştir. Dolayısı ile çalışma, hem tahmin performansı hem de açıklanabilirlik yönüyle literatürdeki mevcut çalışmalardan ayrılmaktadır.

7. SONUÇ

Turizm sektöründe gelir yönetimi ve operasyonel planlamanın başarısı, rezervasyon iptallerinin yüksek doğrulukla tahmin edilmesine doğrudan bağlıdır. Bu çalışmada, otel rezervasyon iptallerini öngörmek amacıyla veri madenciliği temelli, gelişmiş bir makine öğrenmesi sınıflandırma modeli geliştirilmiştir. Elde edilen sonuçlar, Antonio vd. (2017) ile Şahinbaş ve Doğuç (2022) gibi önceki çalışmalarla karşılaştırıldığında, önerilen yaklaşımın daha ileri düzey modelleme teknikleri sunduğunu ortaya koymaktadır. SHAP analizi ile modelin karar süreçleri şeffaf biçimde açıklanmıştır. Veri dengesizliği problemi SMOTE yöntemiyle giderilmiş, böylece modellerin her iki sınıf için de dengeli ve adil tahminler üretebilmesi sağlanmıştır. Sonuç olarak çalışma, yalnızca yüksek performanslı bir tahmin modeli sunmakla kalmamış, aynı zamanda sektörel karar mekanizmalarının kanıta dayalı ve yorumlanabilir biçimde desteklenmesine katkı sağlamıştır.

Çalışmanın literatüre önemli katkılarından biri, otelcilik sektöründe iptal kararlarının tahminine yönelik ileri düzey makine öğrenmesi modellerinin Türkçe akademik çalışmada uygulanmasıdır. Uluslararası bir veri kümesinin kullanılması, yerli veri kaynaklarının sınırlı olduğu bu alanda farklı pazarlara yönelik kıyaslamalı araştırmalar için güçlü bir metodolojik temel sunmaktadır. Bununla birlikte, çalışmanın bazı sınırlılıkları bulunmaktadır. Kullanılan veri kümesi, yalnızca Portekiz'deki iki otele ait rezervasyon verilerini içermekte olup farklı ülkelerdeki müşteri davranışlarını temsil etme kapasitesi sınırlıdır. Ayrıca, iptal kararlarının gerekçelerini içeren metin tabanlı verilerin analizlere dâhil edilmemesi, modelin yalnızca nicel davranışsal örüntüler üzerinden tahmin üretmesine yol açmıştır.

Keşifsel veri analizleri, veri kümesinde iptal sınıfının daha az temsil edildiğini ve modelleme öncesinde dengelenmesinin gerekli olduğunu ortaya koymuştur. En başarılı model Random Forest olmuştur. Karşılaştırmalı değerlendirmede bu model, doğruluk oranında %81'e, AUC değerinde 0,89'a ve F1 skorunda 0,71'e ulaşarak genel performans bakımından diğer modellere kıyasla üstünlük göstermiştir. Modelin karar süreçleri, SHAP analizi aracılığıyla ayrıntılı biçimde açıklanmış ve özellikle "lead_time", "adr" ve "hotel" değişkenlerinin iptal tahminlerinde belirleyici rol oynadığı saptanmıştır. Analiz bulguları, erken uyarı sistemleri, esnek iptal politikaları ve bekleme listesi yönetimi gibi operasyonel karar mekanizmalarına doğrudan girdi sağlayabilecek potansiyele işaret etmektedir. Elde edilen sonuçlar, otelcilik sektöründe rekabet avantajı oluşturacak, esnek ve öngörüye dayalı yönetim modellerinin geliştirilmesi için güçlü bir temel sunmaktadır.

Gelecekte yürütülecek çalışmalarda, rezervasyon iptallerine ilişkin müşteri yorumları ve açıklama metinlerinin analizlere dâhil edilmesi, modelin karar mekanizmalarının daha derinlemesine anlaşılmasına katkı sağlayacaktır. Bu bağlamda doğal dil işleme (NLP) ve duygu analizi gibi yöntemlerin entegrasyonu, tahmin doğruluğunu ve yorumlanabilirliği artırma potansiyeline sahiptir. Ayrıca otel türü, sezon, müşteri profili ve coğrafi konum gibi değişkenlere göre ayrıştırılmış alt modellerin geliştirilmesi, sektörel farklılıkların daha ayrıntılı biçimde ortaya konulmasına ve daha hedefe yönelik stratejilerin belirlenmesine imkân tanıyacaktır.

YAZAR BEYANI / AUTHORS' DECLARATION:

Bu makale Araştırma ve Yayın Etiğine uygundur. Beyan edilecek herhangi bir çıkar çatışması yoktur. Araştırmanın ortaya konulmasında herhangi bir mali destek alınmamıştır. Makale yazım ve intihal/benzerlik açısından kontrol edilmiştir. Makale, "en az iki dış hakem" ve "çift taraflı körleme" yöntemi ile değerlendirilmiştir. Yazar, dergiye imzalı "Telif Devir Formu" belgesi göndermişlerdir. Mevcut çalışma için mevzuat gereği etik izni alınmaya ihtiyaç yoktur. Bu konuda yazar tarafından dergiye "Etik İznine Gerek Olmadığına Dair Beyan Formu" gönderilmiştir. Yazar, çalışmanın tüm bölümlerine ve aşamalarına tek başına katkıda bulunmuştur. / This paper complies with Research and Publication Ethics, has no conflict of interest to declare, and has received no financial support. The article has been checked for spelling and plagiarism/similarity. The article was evaluated by "at least two external referees" and "double blinding" method. The author sent a signed "Copyright Transfer Form" to the journal. There is no need to obtain ethical permission for the current study as per the legislation. The "Declaration Form Regarding No Ethics Permission Required" was sent to the journal by the author on this subject. The author contributed to all sections and stages of the study alone.

KAYNAKLAR

- ANTONIO, Nuno, ALMEIDA, Ana de ve NUNES, Luis (2017), “*Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model*”, **16th IEEE International Conference on Machine Learning and Applications (ICMLA)**, 18-21 December 2017 – Cansun, IEEE Publisher, New York, ss.1049-1054.
- ANTONIO, Nuno, ALMEIDA, Ana de ve NUNES, Luis (2019), “*Hotel Booking Demand Datasets*”, **Data in Brief**, S.22, ss.41-49.
- ASLAN, Kaan (2025), **Yapay Zekâ, Makine Öğrenmesi ve Veri Bilimi Kursu, Sınıfta Yapılan Örnekler ve Özet Notlar**, C ve Sistem Programcıları Derneği Yayını, İstanbul.
- BOZ, Mehmet, CANBAZOĞLU, Erokan, ÖZEN, Zeki ve GÜLSEÇEN, Sevinç (2018), “*Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi*”, **Veri Bilimi**, S.1(1), ss.7-14.
- CAGGIA, Caitlyn (2021), “*Forecasting Hotel Reservation Cancellations with Machine Learning*”, **Lodging Magazine** (E-Article), 10 March 2021, <https://www.lodgingmagazine.com/forecasting-hotel-reservation-cancellations-with-machine-learning/> (Erişim Tarihi: 14.07.2025).
- CHATFIELD, Christopher (1995), **Problem Solving: A Statistician’s Guide**, Chapman & Hall Publisher, Londra, 2nd Edition.
- CHEN, Tianqi ve GUESTRIN, Carlos (2016), “*XGBoost: A scalable Tree Boosting System*”, **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 13-17 August 2016 – San Francisco, Association for Computing Machinery Publisher, New York, ss.785-794.
- ERCAN, Uğur ve IRMAK, Sezgin (2022), “*Turizm Endüstrisinde Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmini*”, **Journal of Gastronomy, Hospitality and Travel**, S.5(1), ss.45-54.
- EŞİDİR, Kamil Abdullah (2025a), “*Makine Öğrenimi Modelleri İle Yetişkin Eğitimi Analizi: Modellerin Karşılaştırmalı Performansı*”, **Elektronik Sosyal Bilimler Dergisi**, S.24(2), ss.946-964.
- EŞİDİR, Kamil Abdullah (2025b), “*Video Oyunları Kullanıcı Yorumlarının Makine Öğrenmesi İle Tahmini: Veri Bilimi Perspektifinden Bir İnceleme*”, **MTÜ Sosyal ve Beşeri Bilimler Dergisi**, S.5(1), ss.97-108.
- FERNHOLZ, Luisa T. ve MORGENTHALER, Stephan (2000), “*A Conversation with John W. Tukey and Elizabeth Tukey*”, **Statistical Science**, S.15(1), ss.79-94.
- JISHAN, Md Asifuzzaman., SINGH, Vikas, GHOSH, Ayan Kumar, ALAM, Md Shahabub, MAHMUD, Khan Raqib ve PAUL, Bijan (2024), “*Hotel Booking Cancellation Prediction Using Applied Bayesian Models*”, **arXiv Preprint** (E-Article), 21 October 2024, <https://arxiv.org/abs/2410.16406> (Erişim Tarihi: 14.07.2025).
- KARACA, Şükran ve AKIN, Mehmet Halit (2025), “*Turizmde Makine Öğrenmesi ile İlgili Ulusal Yazındaki Çalışmaların Bibliyometrik Analizi*”, **TURAR Turizm ve Araştırma Dergisi**, S.14(1), ss.169-195.
- KE, Guolin, MENG, Qi, FINLEY, Thomas, WANG, Taifeng, CHEN, Wei, MA, Weidong, YE, Qiwei ve LIU, Tie-Yan (2017), “*Light GBM: A Highly Efficient Gradient Boosting Decision Tree*”, **Advances in Neural Information Processing Systems**, S.30, ss.3146-3154.
- MOSTIPAK, Jesse (2018), “*Hotel Booking Demand*”, **Kaggle** (E-Article), <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand> (Erişim Tarihi: 14.07.2025).
- OGUINE, Ozioma Collins ve OGUINE, Kanyifechukwu Jane, OKORIE, Chukwundindu I. ve OGUINE, Munachimso Blessing (2021), “*Comparative Analysis and Forecasting on the Death Rate of COVID-19 Patients in Nigeria Using Random Forest and Multinomial Bayesian Epidemiological Models*”, **Journal of Clinical Case Studies, Reviews & Reports**, S.3(4), ss.1-7.

PROKHORENKOVA, Liudmila, GUSEV, Gleb, VOROBEOV, Aleksandr, DOROGUSH, Anna Veronika ve GULIN, Andrey (2018), “CatBoost: Unbiased Boosting with Categorical Features”, **Advances in Neural Information Processing Systems**, S.31, ss.6639-6649.

ŞAHİNBAŞ, Kevser ve DOĞUÇ, Özge (2022), “Otel Rezervasyon İptal Tahmin Modelinin Veri Madenciliği Algoritmaları ile Uygulanması”, **Selçuk Turizm ve Bilişim Araştırmaları Dergisi**, S.1(2), ss.15-24.

WANG, Liyang., WANG, Xiaoya, CHEN, Angxuan, JIN, Xian ve CHE, Huilian (2020), “Prediction of Type 2 Diabetes Risk and its Effect Evaluation Based on the Xgboost Mode”, **Healthcare**, S.8(3), ss.(247).

