Otel Rezervasyonu İptal Tahmini

Ezgi CİNKILIÇ
Yapay Zekâ Mühendisliği
TOBB Ekonomi ve Teknoloji Üniversitesi
e.cinkilic@etu.edu.tr

***Öz*– Bu makalede, çifte rezervasyon stratejilerini daha etkin bir şekilde yönetmek amacıyla rezervasyonların iptal edilip edilmeyeceğini tahmin eden farklı makine öğrenmesi modelleri eğitilmiş ve performansları kıyaslanmıştır.**

Anahtar Kelimeler – Makine öğrenmesi, Çifte rezervasyon, Karar Ağaçları, Rastgele Orman, XGBoost, Lojistik Regresyon

I. GİRİŞ

Otel rezervasyonlarının iptal edilme durumu, otel işletmeciliğinde önemli bir sorundur. Çifte rezervasyon stratejileri, otellerin odalarını en verimli şekilde kullanabilmeleri için yaygın olarak kullanılmaktadır. Ancak, bu stratejilerin etkili bir şekilde yönetilebilmesi için rezervasyonların iptal edilme olasılığının doğru bir şekilde tahmin edilmesi gerekmektedir. Bu çalışmada, farklı makine öğrenmesi modelleri kullanılarak otel rezervasyonlarının iptal edilip edilmeyeceği tahmin edilmeye çalışılmıştır. Bu modellerin doğruluğu ve etkinliği, otel yönetiminde daha bilinçli kararlar alınmasına ve müşteri memnuniyetinin artırılmasına katkı sağlayacaktır.

II. LİTERATÜR TARAMASI

Rezervasyon iptallerinin tahmin edilmesi, sadece otel yönetimi alanında değil, rezervasyon ile çalışan tüm sektörler (ulaşım, eğlence vb.) gibi sektörler için de önemli bir araştırma konusu olmuştur. Literatürde, bu sorunun çözümüne yönelik çeşitli istatistiksel yöntemler, matematiksel modeller ve makine öğrenmesi modelleri önerilmiştir. Bu çalışmalar, rezervasyon iptallerini tahmin ederek işletmelerin gelir yönetimini optimize etmelerine ve müşteri memnuniyetini artırmalarına yardımcı olmayı amaçlamaktadır.

Özellikle, çifte rezervasyon stratejilerinin optimizasyonu üzerine yapılan çalışmalarda, rezervasyon kazancı ve telafi maliyetini dengeleyerek maksimum geliri hedefleyen istatistiksel modeller geliştirilmiştir [1]. Bu modeller, geçmiş verileri analiz ederek ve iptal olasılıklarını hesaba katarak otellerin en uygun çifte rezervasyon seviyesini belirlemelerine yardımcı olmaktadır.

Son yıllarda, makine öğrenmesi uygulamaları bu alanda daha etkin sonuçlar elde edebilmek için kullanılmaya başlanmıştır. Makine öğrenmesi modelleri, büyük veri kümelerini analiz ederek ve trendleri öngörerek, iptal olasılıklarını daha doğru bir şekilde tahmin edebilmekte ve böylece çifte rezervasyon stratejilerini daha hassas bir şekilde optimize edebilmektedir [2].

Önceki çalışmalarda, lojistik regresyon, rastgele orman, destek vektör makineleri (SVM), XGBoost, C 5.0, k-en yakın komşu ve yapay sinir ağları gibi farklı makine öğrenmesi yöntemlerinin bu problemde kullanıldığı görülmüştür.

Ayrıca, iptal tahmin modellerinin performansını artırmak için özellik mühendisliği, veri dengeleme ve veri ön işleme tekniklerinin önemine de dikkat çekilmiştir. Veri dengelemek için Rastgele Alt Örnekleme (RUS), Rastgele Üst Örnekleme (ROS), Sentetik Azınlık Üst Örnekleme Tekniği (SMOTE) metotları kullanılmıştır [2]. Öznitelik seçimi için filtre bazlı metotların yanı sıra üç sarmal metodu olarak adlandırılan Rastgele Orman, XGBoost ve Lightgbm kullanılmıştır [2].

Mevsimsellik etkisi nedeniyle test ve eğitim verilerinin nasıl ayrılması gerektiği de makalelerde işlenen bir konudur. Zaman serisi yöntemlerinin kullanılıp kullanılmaması veri setinin yanı sıra problem ve gelecekteki rezervasyon tahmini ile de ilgilidir. Gelecekte rezervasyon sayısının değişmesi beklendiğinde ve gelecek tahminleri kısa zamanlı periyotlarda yapıldığında kolaylık esaslı bölme (convenience) tekniği kullanılmıştır [3]. Verilerde az da olsa mevsimsellik etkisi görülmesine rağmen her mevsim talep gören bir konumda olan bir otel için zaman serisi modelleri kullanmak yerine zaman bilgisi öznitelik olarak verilmiş ve mevsimsellik etkisi yok sayılmıştır [4].

III. YÖNTEM

A. Veri Hazırlama

Kullanılacak veri seti 31 adet öznitelik ve 119390 adet girdi içermektedir. Verilerin yapısı Tablo 1'de ayrıntılı görülmektedir.

TABLO I
VERİ YAPILARI

Veri Adı	Veri Yapısı
Hotel	İki sınıflı sayısal olmayan
Meal	Sıralı çok sınıflı sayısal olmayan
Country	Sırasız çok sınıflı sayısal olmayan
Market Segment	Sırasız çok sınıflı sayısal olmayan
Distribution Channel	Sırasız çok sınıflı sayısal olmayan
Reserved Room Type	Sırasız çok sınıflı sayısal olmayan
Assigned Room Type	Sırasız çok sınıflı sayısal olmayan
Deposit Type	Sırasız çok sınıflı sayısal olmayan
Customer Type	Sırasız çok sınıflı sayısal olmayan
Reservation Status	Sırasız çok sınıflı sayısal olmayan
Arrival Date Month	Sıralı çok sınıflı sayısal olmayan
Is canceled	İki sınıflı sayısal olmayan
Lead Time	Sayısal ordinal
Arrival Date Year	Sayısal ordinal

Arrival Date Week Number	Sayısal ordinal
Arrival Date Day of Month	Sayısal ordinal
Stays in Weekend Nights	Sayısal ordinal
Stays in Week Nights	Sayısal ordinal
Adults	Sayısal ordinal
Children	Sayısal ordinal
Babies	Sayısal ordinal
Is Repeated Guest	İki sınıflı sayısal olmayan
Previous Cancellations	Sayısal ordinal
Previous Bookings Not Canceled	Sayısal ordinal
Booking Changes	Sayısal ordinal
Agent	Sayısal nominal
Company	Sayısal nominal
Days in Waiting List	Sayısal ordinal
Adr (Avarage Daily Rate)	Sayısal ordinal
Required Car Parking Spaces	Sayısal ordinal
Total of Speciel Requests	Sayısal ordinal
Reservation Status Date	Düz metin

Eksik veriler analiz edilmiştir. Eksik verilerin analiz sonucu Tablo 2'de gösterilmektedir.

TABLO 2
EKSİK VERİ ADLARI, SAYILARI VE ORANLARI

Veri Adı	Eksik Veri Sayısı	Eksik Veri Yüzde Oranı
Children	4	%0.003350
Country	488	%0.408744
Agent	16340	%13.686238
Company	112593	%94.306893

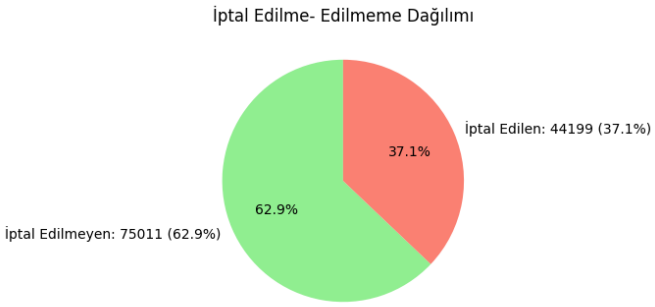
Eksik verileri gidermek için "Children" özniteliğindeki eksik değerler medyan değeri ile, "Country" özniteliğindeki değerler mod değeri ile doldurulmuştur. "Agent" özniteliğindeki boş değerler bir ajansın olmadığını belirttiği için "0" ile değiştirilmiştir. "Company" özniteliği %90'dan fazla boş olduğu için bu öznitelik silinmiştir.

Rezervasyon yapan kişi sayısının sıfır olduğu satırlar, "Adult", "Children" ve "Baby" değerlerinin sıfıra eşit olduğu durum, silinmiştir.

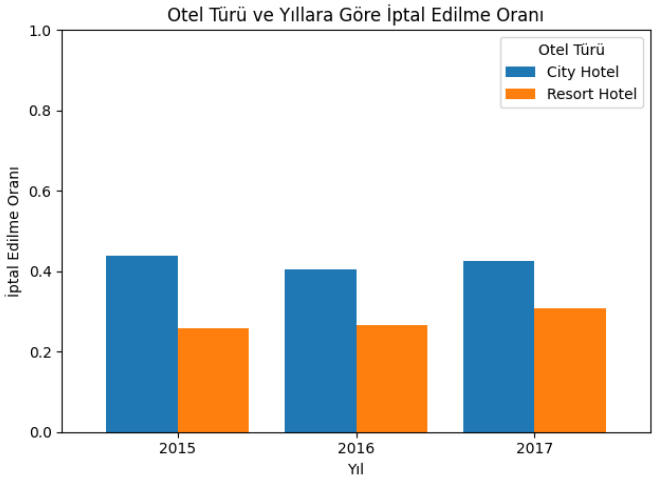
Nesne türünde tutulan öznitelikler uygun şekilde etiket kodlama veya tek bir aktif vektör teknikleri kullanılarak sayısal değerlere dönüştürülmüştür. Sayısal değerler normalize edilmiştir.

B. Veri Analizi

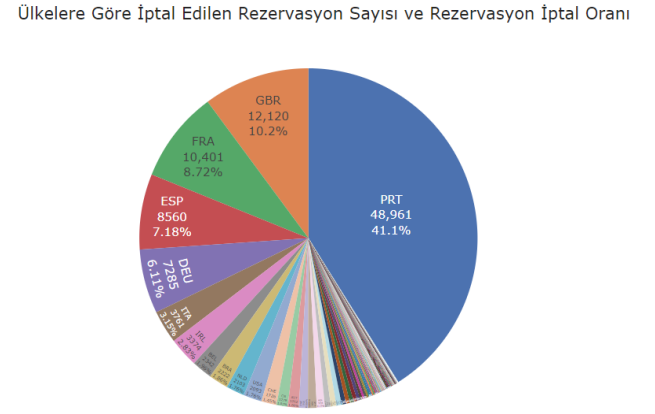
Veri analizi kapsamında çeşitli değişkenlerin iptal oranları, otel türlerine göre verilerin dağılımı, zamana göre verilerin dağılımı incelenmiştir.



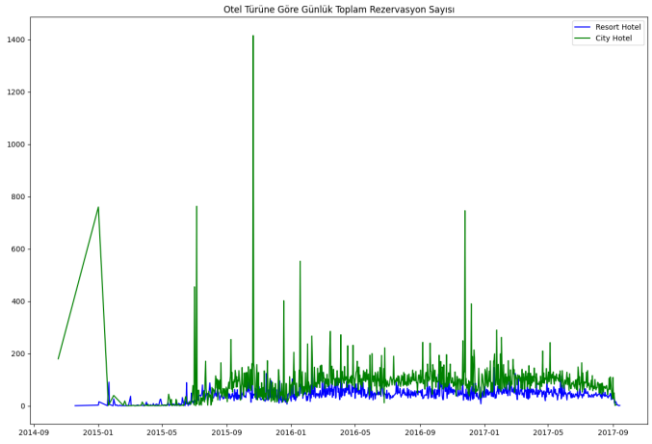
Şekil. 1 Rezervasyonların İptal Durumu Dağılımı



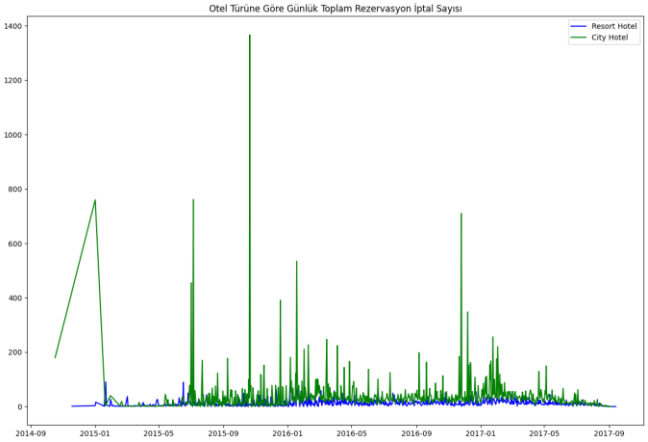
Şekil. 2 Otel Türlerine ve Yıllara Göre Rezervasyon İptal Dağılımı



Şekil. 3 Ülkelere Göre İptal Edilen Rezervasyon Sayısı ve İptal Oranı



Şekil. 4 Otel Türlerine Göre Günlük Toplam Rezervasyon Sayısı



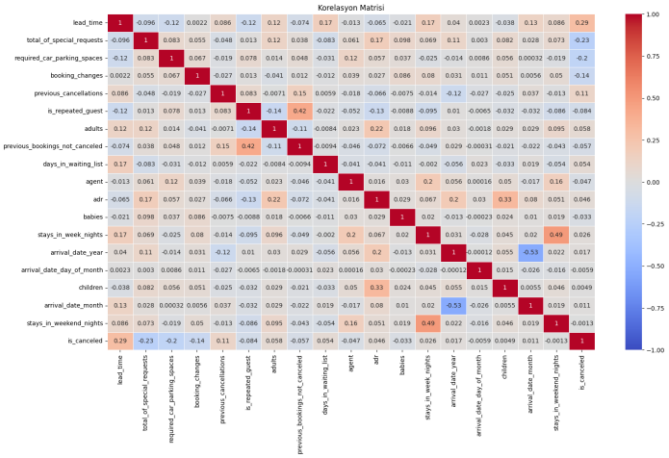
Şekil. 5 Otel Türlerine Göre Günlük Toplam Rezervasyon İptal Sayısı

Yapılan veri analizi kapsamında oluşturulan birkaç grafik yukarıda yer almaktadır. Şekil 1 incelendiğinde rezervasyonların %62.9'unun iptal edilmediği, %37.1'inin iptal edildiği görülmektedir. Dağılımlar eşit olmadığı için veri sınıflarının dengesiz olduğu söylenebilir. Şekil 2 incelendiğinde iptal oranlarının yıllara göre değişmediği çıkarımı yapılabilir. Şekil 3'te rezervasyon yaptıran kişilerin ülkelerinin rezervasyon iptal durumundaki etkisini göstermektedir. Ancak rezervasyonların çoğu otellerin bulunduğu ülke olan Portekiz vatandaşları tarafından yapıldığı için verilerde dengesizlik vardır. Şekil 4 ve 5 mevsimsellik etkisini ve trendini analiz etmek için oluşturulmuştur. Ancak verilerin sapmasının çok olması nedeniyle etkili bir analiz yapılamaz.

C. *Öznitelik Seçimi*

Öznitelik seçiminde korelasyon matrisi (heat-map), çift değişkenli dağılım grafiği (pair plot) ve daha önceki çalışmalardan faydalanılmıştır. Sayısal değerler için korelasyon matrisi oluşturulmuştur. Ancak matris sonucunda doğrudan iptal bilgisine etki eden bir öznitelik bulunamamıştır.

Her bir özelliğin diğer tüm özelliklerle olan ilişkisini incelemek için çift değişkenli dağılım grafiği kullanılmıştır. Çift değişkenli dağılım grafiği için özniteliklerden rastgele 7 tane seçilmiştir. Tüm özniteliklerin kullanılmamasının sebebi maliyetinin yüksek olmasıdır.



Şekil. 6 Sayısal Veriler İçin Korelasyon Isı Matrisi

Yapılan analizler sonucunda "Arrival Date Month" özniteliği ile "Arrival Date Week Number" öznitelikleri aynı bilgiyi verdiği için "Arrival Date Week Number" özniteliği çıkarılmıştır.

"Country" özniteliği verilerin alındığı otelin konumu ile bağlantılı olması nedeniyle sonuçlarda yanlılığa sebep olmaması nedeniyle çıkarılmıştır [3].

"Reservation Status" özniteliği "is_canceled" özniteliği ile aynı bilgiyi içerdiği için çıkarılmıştır.

D. *Modellerin Eğitilmesi*

Karar Ağacı, Rastgele Orman, XGBoost ve Lojistik Regresyon modelleri seçilmiştir.

Veri setindeki dengesizliğin giderilmesi için Rastgele Alt Örnekleme (RUS), Rastgele Üst Örnekleme (ROS) ve Sentetik Azınlık Üst Örnekleme Tekniği (SMOTE) metotları kullanılmış ve performansları çeşitli metrikler ile karşılaştırılmıştır. Dengelenmiş veri seti sadece Lojistik Regresyon modelinde kullanılmıştır. Bunun nedeni diğer modellerin dengesiz veriler ile baş edebilmesi ve özellikle karar ağaçlarında performansı azaltıcı etki edebilmesidir. Ayrıca verilerin çıkartılması bilgi kaybına neden olarak modelin genelleme yeteneğini düşürebilmektedir.

Kullanılan veri seti büyük olduğu için K-katlama doğrulama ile doğrulama yapmaya gerek yoktur. Çünkü tek eğitim-test verisi tüm veri dağılımını temsil edebilir. Bunu kanıtlamak için örnek olarak Rastgele Orman ve XGBoost modellerinde K-katlama doğrulama kullanılmış ve ayrık (hold-out) doğrulama ile sonuçlar karşılaştırılmıştır. Beklendiği gibi benzer sonuçlar elde edilmiştir.

Model performanslarını test etmek için doğruluk, kesinlik, geri çağırma, F1-skoru ve alıcı işletim eğrisi altında kalan alan (EAA) değerleri kullanılmıştır.

1) Karar Ağacı

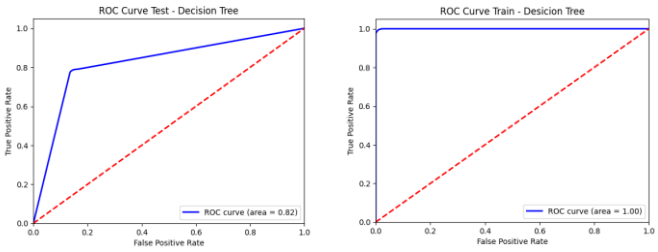
Karar ağaçları, veriyi dallarına ayırarak sınıflandırma veya regresyon yapabilen bir denetimli öğrenme algoritmasıdır. Her bir dal, bir özelliğin bir değerine göre veri kümesini böler, bu da modelin veri içinde karar vermesini sağlar. Karar ağaçları, genellikle hızlı ve anlaşılabilir modeller oluştururken, fazla büyüyen ağaçlar aşırı öğrenmeye neden olabilir. Önemli parametreleri maksimum derinlik, düğümü bölmek için gerekli minimum örnek sayısı ve yapraklarda bulunması gereken minimum örnek sayısıdır.

Karar ağaçları dengesiz veriler ile doğal olarak başa çıkabildiği için eğitilirken dengelenmiş veriler kullanılmamıştır.

Test ve eğitim verileri için performans metrikleri Tablo 3'te yer almaktadır.

TABLO 3
KARAR AĞACI TEST VE EĞİTİM PERFORMANS DEĞERLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
Karar Ağacı (Test)	0.8317	0.8321	0.8317	0.8319	0.8223
Karar Ağacı (Eğitim)	0.9923	0.9923	0.9923	0.9923	0.9998



Şekil. 7-8 Karar Ağacı İçin Test (Sol) ve Eğitim (Sağ) ROC Eğrileri

Tablo 3 incelendiğinde modelin eğitim setinde başarısının çok yüksekken test setinde yani daha önce görmediği verilerde aynı başarıyı yakalayamadığı görülmektedir. Bunun nedeni modelin eğitim verisine aşırı uyum sağlamasıdır. Aşırı uyumu engellemek için maksimum derinlik azaltılabilir veya daha iyi özellik seçimi yapılarak öznitelik sayısı azaltılabilir.

Bu çalışma kapsamında aşırı uyumu azaltmak için torbalama ve arttırma yöntemleri kullanılmıştır.

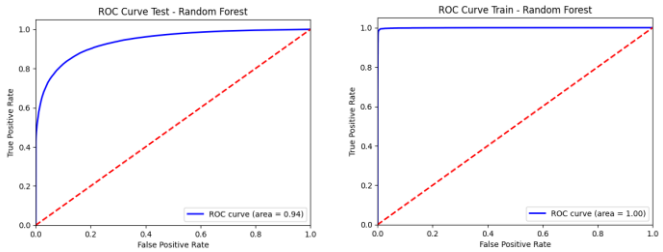
2) Rastgele Orman

Rastgele Ormanlar, birçok karar ağacının topluca çalıştırıldığı bir öğrenme yöntemidir. Model, farklı veri alt kümeleri ve özelliklerle eğitilen ağaçların oylarını birleştirerek nihai tahminini yapar. Bu yöntem, karar ağaçlarının aşırı öğrenme riskini azaltır ve genellikle daha iyi genel performans sağlar. Torbalama tekniği kullanılarak, her bir ağaç farklı bir veri alt kümesi üzerinde eğitilir ve sonuçlar birleştirilir, böylece modelin dengesiz veri setlerinde performansı artırılır. Torbalama, özellikle dengesiz veri setlerinde, azınlık sınıfın daha iyi temsil edilmesini sağlayarak genel performansı iyileştirir ve aşırı öğrenme riskini önemli ölçüde azaltır. Bu nedenle kullanılan veri dengelenmemiştir. En önemli parametreleri ormandaki ağaç sayısı, her ağaç işin kullanılacak maksimum özellik sayısı ve kullanılan özelliklerin yeniden örneklenip örneklenmeyeceğidir.

Test ve eğitim verileri için k-katlama doğrulama kullanılan performans metrikleri Tablo 4'te yer almaktadır.

TABLO 4
RASTGELE ORMAN TEST VE EĞİTİM PERFORMANS DEĞERLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
Rastgele Orman (Test)	0.8741	0.8740	0.8741	0.8723	0.9383
Rastgele Orman (Eğitim)	0.9923	0.9923	0.9923	0.9923	0.9992



Şekil. 9-10 Rastgele Orman İçin Test (Sol) ve Eğitim (Sağ) ROC Eğrileri

Tablo 4 incelendiğinde eğitim setindeki metriklerin test setindekilerden çok daha iyi olduğu görülmektedir. Bu da modelin eğitim verisi üzerinde çok iyi performans gösterdiğini ancak genel test verisi üzerinde performansının düştüğünü göstermektedir. Modelin aşırı öğrenme yaşadığını ve eğitim verisinin özelliklerini fazla ezberlediğini, dolayısıyla test verisinde genelleme yeteneğinin azaldığı söylenebilir. Bunların yanı sıra genel olarak modelin performansının iyi olduğu, özellikle yüksek EAA değeriyle sınıfları iyi ayırdığı söylenebilir.

Aşırı öğrenme sorununu çözmek için ağaç sayısı, ağaçların derinliği azaltılabilir, yapraklardaki özellik sayısı arttırılabilir. Veya düzenleme (regularization) parametresine sahip, XGBoost gibi, bir model kullanılabilir. Bu çalışmada XGBoost kullanılmıştır.

3) XGBoost

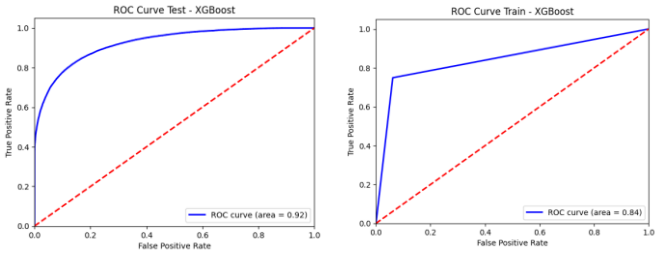
XGBoost güçlü bir sınıflandırma ve regresyon algoritmasıdır ve gradyan artırma yöntemini kullanarak yüksek performanslı ve hızlı bir model oluşturur. XGBoost, karar ağaçlarını iteratif olarak eğiterek, her bir ağacın önceki ağaçların hatalarını düzeltmesini sağlar. Bu yöntem hem doğruluk hem de işlem hızı açısından güçlü performans sağlar ve büyük veri setleri ve karmaşık görevler için uygundur. XGBoost, modelin aşırı öğrenme riskini azaltmak için çeşitli düzenleme teknikleri sunar ve bu sayede daha iyi genel

performans sağlar. En önemli parametreleri ağaç sayısı, öğrenme hızı ve ağaçların maksimum derinliğidir.

Test ve eğitim verileri için k-katlama doğrulama kullanılan performans metrikleri Tablo 5'te yer almaktadır.

TABLO 5
XGBOOST TEST VE EĞİTİM PERFORMANS DEĞERLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
XGBoost (Test)	0.8536	0.8543	0.8536	0.8505	0.9179
XGBoost (Eğitim)	0.8686	0.8693	0.8686	0.8661	0.8443



Şekil. 11-12 XGBoost İçin Test (Sol) ve Eğitim (Sağ) ROC Eğrileri

Tablo 5 incelendiğinde eğitim ve test verileri üzerinde dengeli bir performans sergilediği görülmektedir. Test metrikleri biraz düşük olmasına rağmen EAA değerinin yüksek olması nedeniyle güçlü bir sınıflandırıcı olduğu söylenebilir. Eğitim ve test verilerindeki performans farkı, modelin eğitim aşırı uyuma eğilimli olabileceğini belirtmektedir.

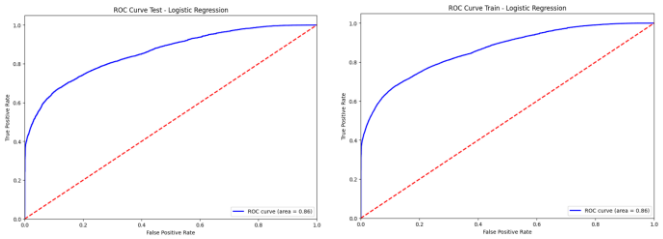
4)Lojistik Regresyon

Lojistik regresyon, ikili sınıflandırma problemlerinde yaygın olarak kullanılan lineer bir modeldir. Model, bir örneğin belirli bir sınıfa ait olma olasılığını tahmin etmektedir. En önemli parametreleri düzenleme parametresi ve algoritmanın konverjansı için gerekli maksimum iterasyon sayısıdır.

Test ve eğitim verileri için performans metrikleri Tablo 6'da yer almaktadır.

TABLO 6
LOJİSTİK REGRESYON TEST VE EĞİTİM PERFORMANS DEĞERLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
Lojistik Regresyon (Test)	0.7743	0.7753	0.7743	0.7641	0.8056
Lojistik Regresyon (Eğitim)	0.7755	0.7754	0.7755	0.7656	0.8055



Şekil. 13-14 Lojistik Regresyon İçin Test (Sol) ve Eğitim (Sağ) ROC Eğrileri

Tablo 6 incelendiğinde lojistik regresyonun hem test hem de eğitim setlerinde beklenen performansı göstermediği görülmüştür. Bunun nedeni iterasyon limitinin aşılması ve dolayısıyla konverjansın erken sonlanmasıdır. Bu sorunu çözmek için hiperparemetre ayarlaması yapılmıştır. Izgara arama tekniği kullanılarak en iyi parametreler bulunmuştur. En iyi parametreler ile oluşturulan modelin performans değerleri Tablo 7'de yer almaktadır.

TABLO 7
LOJİSTİK REGRESYON (EN İYİ PARAMETRELER) PERFORMANS DEĞERLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
Lojistik Regresyon (Test)	0.8114	0.8155	0.8114	0.8035	0.8598
Lojistik Regresyon (Eğitim)	0.8091	0.8125	0.8091	0.8011	0.8586

Lojistik regresyonda eğer veri setinde bir sınıf diğerinden çok daha fazla örneğe sahipse yani veriler dengesizse, model bu sınıfa karşı daha yüksek bir hata payı gösterebilir. Bu nedenle dengelenmiş veri setlerinin kullanılması tavsiye edilir. Veri setindeki dengesizliğin giderilmesi için en iyi parametreler ile oluşturulan modelde Rastgele Alt Örnekleme (RUS), Rastgele Üst Örnekleme (ROS) ve Sentetik Azınlık Üst Örnekleme Tekniği (SMOTE) metotları kullanılmıştır.

Rastgele Alt Örnekleme (RUS), veri setindeki çoğunluk sınıfından rastgele örnekler çıkararak veri dengesizliğini azaltmayı amaçlayan bir tekniktir. Rastgele Üst Örnekleme (ROS), azınlık sınıfındaki örnekleri çoğaltarak veri dengesizliğini giderir, bu sayede sınıflar arasındaki dağılımı daha dengeli hale getirir. Sentetik Azınlık Üst Örnekleme Tekniği (SMOTE) ise, azınlık sınıfındaki örnekler arasında yeni sentetik örnekler oluşturarak veri dengesizliğini ortadan kaldırmaya çalışır, böylece modelin azınlık sınıfını daha iyi öğrenmesini sağlar.

Performans karşılaştırmaları Tablo 8'de yer almaktadır.

TABLO 8
LOJİSTİK REGRESYON (DENGELENMİŞ VERİLER) PERFORMANS DEĞERLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
Lojistik Regresyon (Test) – Dengelenmemiş	0.8114	0.8155	0.8114	0.8035	0.8598
Lojistik Regresyon (Eğitim) – Dengelenmemiş	0.8091	0.8125	0.8091	0.8011	0.8586
Lojistik Regresyon (Test) – RUS	0.7861	0.7885	0.7861	0.7870	0.8628
Lojistik Regresyon (Eğitim) – RUS	0.7748	0.7767	0.7748	0.7744	0.8621
Lojistik Regresyon (Test) – ROS	0.7861	0.7886	0.7861	0.7871	0.7761
Lojistik Regresyon (Eğitim) – ROS	0.7739	0.7757	0.7739	0.7735	0.7739
Lojistik Regresyon (Test) – SMOT	0.7965	0.7942	0.7965	0.7946	0.7741
Lojistik Regresyon (Eğitim) – SMOT	0.8121	0.8152	0.8121	0.8116	0.8121

Lojistik regresyon modelinin performansı, kullanılan veri dengeleme tekniklerine bağlı olarak önemli ölçüde değişiklik göstermiştir. En yüksek EAA değerine ve dolayısıyla en iyi sınıf ayrım yeteneğine sahip olan RUS modelidir. Ancak diğer metriklerde performans kaybı vardır. Bunun nedeni bilgi kaybı olabilir. SMOT kullanılan durumda eğitim setinde performans artışı yaşanmasına rağmen test setinde beklenen başarı sağlanmamıştır. RUS yöntemi yüksek EAA değeri sunmasına rağmen, modelin genelleme yeteneğini artırmak için SMOTE'un daha dengeli ve sürdürülebilir performans sağladığı düşünülmektedir. Dengelenmemiş veri seti ile eğitilen modelin sonuçları hem dengeli hem de sınıf ayrım yeteneği yüksek olduğu için tercih edilmiştir.

E. Öznitelik Seçimi

En başarılı iki model olan Rastgele Orman ve XGBoost modellerinin öznitelik seçimleri incelenmiştir. Öznitelik seçimlerine modeldeki ağırlıklarına, f-skorlarına bakılmıştır. Bu öznitelikler seçilerek boyut azaltılmış, bu sayede modellerin performansının artması, modellerin karmaşıklığının azalarak daha hızlı eğitilmesi ve aşırı öğrenmenin azaltılması hedeflenmiştir. Eğitilen yeni modellerin başarımında ciddi bir değişim olmamıştır. Öznitelik azaltma işlemi Rastgele Orman modelinin aşırı öğrenmesini azaltmamış ve genelleme yeteneğini arttırmamıştır.

TABLO 9
EN ÖNEMLİ 10 ÖZNİTELİK TABLOSU

XGBoost	Rastgele Orman
Lead Time	Lead Time
Adr (Avarage Daily Rate)	Adr (Avarage Daily Rate)
Arrival Date Day of Month	Deposit Type – No Deposit
Agent	Arrival Date Day of Month
Arrival Date Month	Deposit Type – Non-Refund
Stays in Week Nights	Total of Speciel Requests
Total of Speciel Requests	Arrival Date Month
Arrival Date Year	Stays in Week Nights
Booking Changes	Agent
Stays in Weekend Nights	Previous Cancellations

F. Model Başarımını Arttırmak

Model başarımını arttırmak için çeşitli teknikler vardır. Bu tekniklerden birkaçı torbalama (bagging), arttırma (boosting), yığma (stacking) ve hiperparametre ayarlamadır.

Torbalama, aynı modelin farklı versiyonlarını eğitmek için veri setlerinden farklı ön yükleme örnekleri kullanarak model varyansını azaltmayı amaçlayan bir tekniktir. Kullandığımız modellerden rastgele orman, torbalama tekniğinin bir uygulamasıdır. Bu model, birçok karar ağacını ön yükleme örnekleri üzerinde eğitir ve bunların sonuçlarını birleştirerek daha iyi bir genel performans sağlar.

Arttırma, zayıf modellerin (genellikle karar ağaçları) art arda eğitildiği ve her birinin önceki modelin hatalarını düzeltmeye çalıştığı bir yöntemdir. Arttırmanın amacı, yüksek doğruluk elde etmek için bu zayıf modelleri birleştirerek güçlü bir model oluşturmaktır. Kullandığımız modellerden XGBoost, arttırmanın gelişmiş bir versiyonudur ve zayıf modelleri art arda eğiterek güçlü bir model oluşturur.

Hiperparametre ayarlama, bir modelin performansını optimize etmek için hiperparametrelerin ayarlanması işlemidir. Bu işlem, genellikle ızgara arama veya rastgele arama gibi yöntemlerle gerçekleştirilir. Hiperparametre ayarlama, modelin genel performansını artırmak için oldukça önemlidir. Lojistik Regresyon modelinde maksimum iterasyon sayısını ve c katsayısını belirlemek için kullanılmıştır. Lojistik regresyonda modelin başarısının arttığı gözlemlenmiştir. Ayrıca Rastgele Orman ve XGBoost modelleri için de hiperparametre ayarlama işlemi yapılmıştır. Bu işlem en yüksek doğruluk değerini hedeflediği için ve Rastgele Orman modeli eğitim verilerini ezberlediği için model başarımlarında iyileşme görülmemiştir.

IV. SONUÇ

Bu çalışmada, otel rezervasyon verilerini kullanarak çifte rezervasyon tahminleri yapmak amaçlamış ve çeşitli makine öğrenme modelleri ile elde edilen performansları karşılaştırmıştır. Kullanılan modeller arasında Karar Ağacı, Rastgele Orman, Lojistik Regresyon ve XGBoost bulunmaktadır. Performans değerlendirmelerinde doğruluk, kesinlik, geri çağırma, F1-skoru ve EAA metrikleri kullanılmıştır.

TABLO 10
MODELLERİN PERFORMANS ÖLÇÜMLERİ

Model	Doğruluk	Kesinlik	Geri Çağırma	F1-Skoru	EAA
Karar Ağacı (Test)	0.8317	0.8321	0.8317	0.8319	0.8223
Karar Ağacı (Eğitim)	0.9923	0.9923	0.9923	0.9923	0.9998
Rastgele Orman (Test)	0.8741	0.8740	0.8741	0.8723	0.9383
Rastgele Orman (Eğitim)	0.9923	0.9923	0.9923	0.9923	0.9992
XGBoost (Test)	0.8536	0.8543	0.8536	0.8505	0.9179
XGBoost (Eğitim)	0.8686	0.8693	0.8686	0.8661	0.8443
Lojistik Regresyon (Test)	0.8114	0.8155	0.8114	0.8035	0.8598
Lojistik Regresyon (Eğitim)	0.8091	0.8125	0.8091	0.8011	0.8586

XGBoost ve Rastgele Orman, tüm modeller arasında en yüksek performansları göstererek en iyi doğruluk, kesinlik, geri çağırma, F1 skoru ve EAA ile öne çıkmıştır. Rastgele Orman modelinin aşırı öğrenmesi nedeniyle kullanılacak yöntem olarak XGBoost seçilmiştir.

Özellik seçimi sürecinde, Rastgele Orman ve XGBoost modelleri kullanılarak özniteliklerin önem dereceleri belirlenmiş ve bu önemli özniteliklerle en iyi performansa sahip model yeniden eğitilmiştir. Bu süreç, model performansını iyileştirmek için etkili bir yöntem olarak değerlendirilmiştir.

Hiperparametre ayarlama işlemleri, modellerin performansını optimize etmek amacıyla gerçekleştirilmiştir. Izgara arama ve rastgele arama gibi yöntemler kullanılarak, Lojistik Regresyon, Rastgele Orman ve XGBoost için en uygun parametreler belirlenmiştir.

Çalışmada öznitelik seçimi yapılırken mevcut bilgileri kaybetmemek için az sayıda öznitelik çıkartılmış ve modellerin öznitelikler için belirlediği ağırlıklar kullanılarak en önemli olanlar belirlenmiştir. Başka bir yaklaşım olarak veri setinin boyutu küçültülebilir, bu sayede hem çalışma maliyeti hem de aşırı öğrenme ihtimali azaltılabilirdi. Ayrıca literatürde bu problemler için yapay sinir ağları ve derin öğrenme modelleri kullanılmaktadır. İncelenen makalelerde model performansları yüksek olduğu için bu modeller de kullanılabilirdi.

Gelecek çalışmalar, gerçek otel verilerini kullanarak mevsimsel değişimlerin ve rezervasyon trendlerinin analizi üzerine odaklanabilir. Bu bağlamda, çifte rezervasyon tahminlerinin mevsimsel olarak sınıflandırılması ve bu verilerin tahmin doğruluğuna etkisi araştırılabilir. Bu çalışmada rezervasyon sayılarının dengesiz olması nedeniyle bu analizler verimli bir şekilde gerçekleştirilememiştir. Ayrıca, çifte rezervasyonların kazanç ve telafi maliyetleri üzerinde yapılacak bir kar analizi, otel işletmelerinin rezervasyon yönetimi stratejilerini daha etkin bir şekilde planlamalarına yardımcı olabilir. Sabit çifte rezervasyon oranı ile performans karşılaştırması yapılarak, çeşitli stratejilerin etkinliği değerlendirilebilir ve bu doğrultuda öneriler geliştirilebilir. Bu tür analizler, otel işletmelerinin rezervasyon süreçlerini optimize etmelerine ve müşteri memnuniyetini artırmalarına yönelik önemli bilgiler sağlayabilir.

KAYNAKÇA

[1] J. A. Fitzsimmons and M. J. Fitzsimmons, Service Management: Operations, Strategy, Information Technology, 7th ed. New York, NY, USA: McGraw-Hill, 2011, pp. 270-273.

[2] Zhai, Q., Tian, Y., Luo, J., & Zhou, J. (2023). Hotel overbooking based on no-show probability forecasts. Computers & Industrial Engineering, 180, 109226.

[3] N. Antonio, A. de Almeida and L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 1049-1054, doi: 10.1109/ICMLA.2017.00-11.

[4] Sánchez-Medina, A. J., & C-Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. International Journal of Hospitality Management, 89, 102546.