VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi

Mehmet BOZ[1], Erokan CANBAZOĞLU[2], Zeki ÖZEN[3]*, Sevinç GÜLSEÇEN[3]

[1]*Şişli Mesleki ve Teknik Anadolu Lisesi, İstanbul*
[2]*Akdeniz Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Antalya*
[3]*İstanbul Üniversitesi, Enformatik Bölümü, İstanbul*

Özet

Konaklama hizmeti veren otellerin maksimum kâr elde edebilmesi için doluluk oranlarının yüksek olması gerekmektedir. Bu sebeple oteller rezervasyon sistemleri aracılığıyla sınırlı sayıdaki odalarını doğru zamanda, doğru müşteriye tahsis etmelidir. Ancak rezervasyonlar çeşitli nedenlerle müşteri tarafından iptal edilebilmektedir. Oteller açısından iptal edilen rezervasyonlar doğru politikalar izlenmezse gelir kaybına neden olabilmektedir. Bu sebeple iptallerin önceden tahmin edilmesi büyük önem taşımaktadır.
Bu çalışmada, makine öğrenmesi teknikleriyle beş farklı otele ait toplam 38.826 kayıttan oluşan otel rezervasyon verisi kullanılarak otellerin gelecekteki rezervasyonlarının iptal durumları tahmin edilmeye çalışılmıştır. Çalışmada sınıflandırma algoritmalarından Rastgele Orman Algoritması, Destek Vektör Makineleri, k-En Yakın Komşu Algoritması ve C4.5 Karar Ağacı Algoritması kullanılarak dört farklı model oluşturulmuş ve modellerin performans karşılaştırmaları yapılmıştır. En iyi sonuç %73 doğruluk oranı ile C4.5 Karar Ağacı Algoritmasından elde edilmiştir.

Makale Bilgisi

Başvuru:
30/11/2018
Kabul:
24/12/2018

Anahtar Kelimeler: *Otel rezervasyon iptali, Makine öğrenmesi, Danışmanlı öğrenme*

Predicting Hotel Reservation Cancellation by Using Machine Learning Methods

Abstract

In order to maximize profit for hotels, occupancy rates must be high. For this reason, hotels should allocate a limited number of their rooms to the right customer at the right time using reservation systems software. However, reservations may be cancelled by the customer for various reasons. Cancellations may result for hotels in loss of income if the right policies are not processed. For this reason, it is very important to estimate reservation cancellations.
In this study, the hotel reservation data set consisting of 38,826 records of five different hotels were analyzed by machine learning algorithms to estimate the cancellation of future bookings of hotels. In this context, four different models were formed in this study by using Random Forest Algorithm (RF), Support Vector Machines (SVM), k-Nearest Neighbor (kNN) Algorithm and C4.5 Decision Tree Algorithm and then, performance comparisons were made among these models. The best result was obtained from C4.5 decision tree algorithm with 73% accuracy.

Keywords: *Hotel reservation cancellation, Machine learning, Supervised learning*

* İletişim e-posta: zekiozen@istanbul.edu.tr

1 Giriş

Gelir yönetimi son yıllarda şirketlerin gelirlerini ve kârlılıklarını artırmak için kullandıkları bir yöntemdir. Gelir yönetimi; doğru müşteriye, doğru kapasitenin, doğru zamanda, doğru fiyatla sunulabilmesi için bilgi teknolojileri ve fiyatlama stratejilerinden faydalanılması olarak tanımlanmaktadır [1]. Konaklama işletmelerinde ise gelir yönetimi doğru konuk için, doğru odayı, doğru fiyata, doğru zamanda ve doğru kanalla erişilebilir hale getirilmesi şeklinde tanımlanmaktadır [2].

Konaklama işletmelerinden olan otellerin maksimum kâr elde edebilmeleri için sınırlı sayıdaki odalarını doğru zamanda doğru müşteriye tahsis etmeleri gerekir. Bu amaçla rezervasyon sistemleri kullanılmaktadır. Rezervasyonlar müşteri ile otel arasında yapılan bir sözleşmedir. Rezervasyon ile müşteri belirtilen tarihler arasında otelde konaklamayı, otel yönetimi ise o tarihlerde müşterinin talep ettiği odanın boş olmasını ve müşterinin kullanımını sağlamayı taahhüt eder. Bu yolla otel açısından risklerin minimize edilmesi, müşteri açısından ise talep edilen hizmetin herhangi bir aksaklık olmadan alınması sağlanır. Ancak rezervasyonlar çeşitli nedenlerle müşteri tarafından iptal edilebilmektedir.

Oteller açısından iptal edilen rezervasyonlar doğru politikalar izlenmezse gelir kaybına neden olabilmektedir. Oteller iptallerden doğacak riski minimize etmek için çeşitli iptal politikaları ve çifte rezervasyon (*overbooking*) yöntemleri uygulamaktadır. Fakat her iki uygulama da otel için zararlı olabilmektedir. Bir odanın birden fazla müşteriye rezerve edilmesini ifade eden çifte rezervasyon uygulaması sebebiyle diğer müşteri hizmet alamayacağı için müşteri memnuniyetsizliği olacaktır. Rezervasyon iptallerinin önlenmesine yönelik oluşturulan sıkı iptal politikaları ise rezervasyon sayısının azalmasına neden olabilir [3].

Rezervasyon iptalleri, tüm otellerde toplam rezervasyonların %20'sini [4], havalimanına yakın olan veya yol üstünde bulunan otellerde ise toplam rezervasyonların %60'ını bulmaktadır [5]. Rezervasyon iptallerinden doğan olumsuz etkilerin ortadan kaldırılabilmesi için iptallerin önceden tahmin edilebilmesi büyük önem taşımaktadır.

Bu çalışmada beş otele ait rezervasyon verisi kullanılarak, gelecekteki rezervasyon iptallerini tahmin edecek bir model geliştirilmesi amaçlanmıştır. Oluşturulacak modelin rezervasyonlarla ilgili bir ön uyarı sistemi olarak çalışması amaçlanmaktadır. Böylece otel yöneticileri iptal edilme olasılığı yüksek olan rezervasyonlarla ilgili önleyici tedbirler alabilir veya iptal politikalarını kişilere özel olarak ayarlayabilirler.

Modelin oluşturulması için makine öğrenmesi tekniklerinden faydalanılmıştır. Mitchell (1983) makine öğrenmesini şu şekilde ifade etmiştir [6]:

"*Eğer bir bilgisayar programının G görevlerinde P ile ölçülen performansı deneyim D ile artıyorsa, o bilgisayar programının bazı G görevlerinin sınıflarına ve performans ölçüsü P'ye göre deneyim D'den öğrendiği söylenmektedir.*"

Makine öğrenmesi yöntemleri literatürde çeşitli kategorilerde verilmekle birlikte temelde danışmanlı öğrenme ve danışmansız öğrenme olarak iki ana stratejiden söz edilmektedir. Danışmanlı öğrenme yönteminde çıktıları bilinen örnekler ile model eğitilir. Daha sonra eğitilmiş modele daha önce görmediği örnekler verilerek modelin bu örneklerin sınıflarını doğru tahmin etmesi sağlanır. Danışmansız öğrenmede ise modele herhangi bir hedef verilmeden modelin veriyi kümelemesi amaçlanır [7]. Sınıflandırma ve regresyon uygulamaları danışmanlı öğrenme yöntemini kullanırken, kümeleme uygulamaları danışmansız öğrenme yöntemini kullanır.

2 Literatür taraması

Sayfa düzeni için aşağıdaki kurallara uyulmalıdır. Makale sunumu hazırlanırken bu belgenin şablon olarak kullanılması yazım düzeni koşullarının yerine getirilmesi açısından önerilmektedir.

Oteller için rezervasyon iptalleri gelir yönetimi konusundaki en büyük risklerdendir. Yüksek iptal oranları otel işletmelerinin mevcut kapasitelerini tam olarak kullanamamasına neden olmaktadır. Bu ise istihdam edilen personelin ve diğer kaynakların doğru yönetilememesini ve gelir kaybını beraberinde getirmektedir.

Başarılı bir gelir yönetim sisteminin en önemli göstergelerinden biri, iptal oranlarının doğru tahmin edilmesidir. Rezervasyon iptal modelleri için kabaca iki yaklaşım vardır: Birincisi, genel iptal oranlarının tahmin edilmesi, ikincisi ise her bir rezervasyonun ayrı ayrı sınıflandırılmasıdır. Pölt (1998)'e göre tahmin hatalarında sağlanacak %20'lik bir azalma, işletme gelirlerinde %1'lik bir artış sağlamaktadır [8].

Antonio vd. (2017) tarafından otel rezervasyonlarının iptalinin tahmin edildiği çalışmada dört farklı otel verisi kullanılmıştır [9]. Çalışmada her bir otel için ayrı modeller geliştirilmiş ve ortalama %90'ın üzerinde doğruluk değeri elde edilmiştir.

Bu çalışmada sınıflandırma algoritmalarından Rastgele Orman Algoritması, Destek Vektör Makineleri, k-En Yakın Komşu Algoritması ve C4.5 Karar Ağacı Algoritması kullanılarak modeller geliştirilmiştir.

2.1 Rastgele orman algoritması

Rastgele Orman (*Random Forest - RF*) algoritması ilk olarak 1995 yılında Tin Kam Ho tarafından önerilmiştir [10]. Bilinen haliyle RF algoritmasının tanımı ise Bierman (2001) tarafından "yapılandırılmış bir karar ağacı koleksiyonundan oluşan bir sınıflandırıcıdır" şeklinde yapılmıştır [11]. Burada amaç tek bir karar ağacıyla sonuca ulaşmak yerine birden fazla karar ağacı türeterek daha güçlü sonuçlar elde etmektir.

RF algoritması hem sınıflandırma hem de regresyon problemlerinde kullanılmaktadır [12]. Sınıflandırma problemlerinde ağaçlardan elde edilen tahminlerin sayıca üstünlüğüne göre karar verilirken, regresyon problemlerinde elde edilen sonuçların ortalaması alınarak karar verilmektedir.

2.2 k-En yakın komşu algoritması

k-En Yakın Komşu (*k-Nearest Neighbor - kNN*) sınıflandırması oldukça basit bir mantığa dayanmaktadır: Veri setindeki nesneler en yakın komşusunun sınıf sayısının çokluğuna göre sınıflandırılır. Sınıflandırma için birden fazla komşu hesaba katıldığında kNN algoritması elde edilmiş olur. Buradaki k değeri sınıflandırma için bakılacak komşu sayısını ifade eder [13]. Veri setindeki bir örneğin sınıfı, etrafındaki k komşusunda en çok bulunan sınıf değeri olarak belirlenir. Eşitlik olması durumunda rastgele karar verilir. Bu durumun olmaması için k değeri genellikle tek sayılardan seçilir [14].

2.3 Destek vektör makineleri algoritması

Destek Vektör Makineleri (*Support Vector Machines - SVM*), bir sınıflandırma modeli olarak Vapnik ve Lerner tarafından 1963 yılında geliştirilmiştir [15]. SVM ile sınıflandırma işlemi doğrusal sınıflandırma ve doğrusal olmayan sınıflandırma olmak üzere iki farklı yöntemle yapılabilir. Doğrusal sınıflandırmada amaç, verideki sınıfları en iyi şekilde ayıran bir karar çizgisi veya hiper düzlem elde etmektir. Verideki sınıfları birbirinden ayıran birçok karar çizgisi veya hiper düzlem olabilir. Bu düzlem veya çizgilerden en iyi olanı veri kümeleri ile kendisi arasındaki mesafe en büyük olanıdır. Böylelikle sınıflandırma, gürültülü veya hatalı veriye karşı dirençli hale gelir [16].

2.4 C4.5 Karar ağacı algoritması

Karar ağaçları pek çok alanda kullanılan bir sınıflandırma algoritmasıdır. C4.5 karar ağacı algoritması 1993 yılında Quinlan tarafından geliştirilmiştir [17]. Bir karar ağacı kök, düğüm, dal ve yaprak bileşenlerinden oluşur. Ağaç yapısında en altta kalan kısım yaprak en üstte olan kısım ise kök olarak adlandırılır. Veri setinde bulunan her bir nitelik ise düğüm noktalarını temsil etmektedir. Düğümler birbirlerine dallar ile bağlanır [18].

Quinlan (1993) tarafından daha önce geliştirilen ID3 algoritmasında bölme kriteri olarak bilgi kazancı kullanılırken, C4.5 algoritmasında kazanç oranı (*gain ratio*) kullanılmaktadır. C4.5 algoritması varsayılan bölme kriteri olarak, bilgi tabanlı bir ölçü olan kazanç oranını kullanmaktadır. En yüksek kazanç oranını sağlayan nitelik ayrımın yapılacağı nitelik olarak seçilmektedir. C4.5 algoritması hem kategorik hem de nümerik değerler alan nitelikler ile çalışabilmektedir [19].

3 Materyal ve metod

Bu çalışmada Veri Madenciliği İçin Çapraz Endüstri Standart Süreç Modeli (*CRISP-DM: Cross-Industry Standard Process for Data Mining*) adımları takip edilmiştir. Çalışmanın problemi Giriş bölümünde ifade edildiği için yöntem bölümü veriyi anlama adımı ile başlamıştır.

3.1 Veriyi anlama

Tablo 1'de veri setindeki niteliklerin isimleri, tipleri ve açıklamaları verilmiştir. (K: Kategorik, M: Metin, N: Nümerik, T: Tarih)

Tablo 1. Veri setinde bulunan niteliklerin tipleri ve açıklamaları

Nitelik	Tip	Açıklama
otel_kodu	K	Otellerin kod numarası
giriş_tarihi	T	Rezervasyon başlangıç tarihi
cikis_tarihi	T	Rezervasyonun bitiş tarihi
yetiskin	N	Kişi sayısı
cocuk1	N	2-6 yaş arası çocuk sayısı
cocuk2	N	6-12 yaş arası çocuk sayısı
bebek	N	0-2 yaş bebek sayısı

Nitelik	Tip	Açıklama
oda_sayisi	N	Rezerve edilen oda sayısı
kayit_tarihi	T	Rezervasyonun oluşturulma tarihi
ozel_istek	M	Müşterilerin ekstra talepleri
toplam_fiyat	N	Toplam rezervasyon tutarı
milliyet	K	Milliyet
rezervasyon_durumu	K	Rezervasyon durum kodu
gece_sayisi	N	Konaklanacak gece sayısı
ücretli_kisi_sayisi	N	Ücrete tabi kişi sayısı
hesaplanan_gece	N	Ücrete tabi kişi sayısı ile konaklanacak gece sayısının çarpımından elde edilen değer
konaklama_tipi	K	Konaklama tipi
fiyat_tipi	K	Fiyat tipi
iptal_tarihi	T	Rezervasyon iptal tarihi
toplam_rez_say	N	Müşterinin toplam rezervasyon sayısı
toplam_iptal_say	N	Rezervasyonu yapan müşterinin toplam iptal sayısı

Araştırma kapsamında kullanılan otel verisi beş farklı otelin 2016 yılı Nisan ayından 2018 yılı Nisan ayına kadar olan iki yıllık rezervasyon kayıtlarından oluşmaktadır. Otellerden ikisi İstanbul, ikisi Antalya, biri ise İzmir ilinde bulunmaktadır. Araştırmada verisi kullanılan oteller kullanım amacı bakımından farklılık göstermektedir. Antalya ve İzmir'deki oteller tatil oteli olarak hizmet verirken, İstanbul'daki oteller hem tatil hem de iş seyahatleri için tercih edilmektedir. Çalışmada bu durumdan kaynaklanabilecek farklılıklar göz ardı edilmiştir.

Rezervasyon verisi çalışmacılara anonim halde verilmiştir. Yani veride herhangi bir müşterinin kimliğini tespit etmeyi sağlayacak bir veri alanı bulunmamaktadır. Verinin analizine geçmeden önce veriyi anlamak adına veri sağlayıcılarla ve alan uzmanlarıyla görüşmeler yapılmıştır.

Ön işleme aşamasından öncelikle temin edilen veri setindeki nitelik adları Türkçeleştirilmiştir. Tablo 2'de veri setindeki otellerin her birinin toplam kayıt sayısı, otele verilen numara ve otelin bölgesi yer almaktadır.

Tablo 2. Farklı Hold-out oranları için elde edilen ortalama performans değerleri.

Otel Kodu	Toplam Kayıt Sayısı	Bölge
1	8131	İstanbul
2	8110	İzmir
3	5980	Antalya
4	6853	Antalya
5	9570	İstanbul
Toplam	38826	-

3.2 Veri önişleme

Literatürde veri önişleme süreci adımları Han ve Kamber (2006) tarafından veri özetleme, veri temizleme, veri bütünleştirme ve dönüştürme, veri indirgeme, veri ayrıklaştırma ve kavram hiyerarşisi oluşturma olarak sıralanmıştır [19]. Bu çalışmada veri setini analize hazır hale getirmek üzere gerçekleştirilen işlemler madde halinde açıklanmıştır.

3.2.1 Hedef niteliğin oluşturulması

Hedef niteliği olan rezervasyon iptal bilgisini oluşturmak üzere *iptal_durumu* adında yeni bir nitelik eklenmiştir. *iptal_durumu* niteliği *iptal_tarihi* ve *rezervasyon_durumu* alanlarının birleştirilmesi ile elde edilmiştir.

3.2.2 Verinin temizlenmesi ve tekdüze edilmesi

Bu aşamada sırasıyla aşağıdaki işlemler gerçekleştirilmiştir:

•Veri setinde "deneme", "test" gibi sözcükler içeren satırlar ve tüm alanlarının NULL değere sahip olduğu kayıtlar veri setinden çıkarılmıştır.

•Veri setinde konaklama tarihi, rezervasyon kayıt tarihinden önce olan kayıtlar olduğu görülmüştür. Bu kayıtlar gerçek bir rezervasyon kaydı olmadığı düşüncesiyle veri setinden çıkarılmıştır.

•*toplam_fiyat* ve *gece_sayisi* alanlarında 0 veya NULL değere sahip kayıtlar veri setinden çıkarılmıştır.

•*milliyet* alanında çok fazla kategori bulunduğundan bu alandaki veri "yerli" ve "yabancı" olmak üzere iki kategoriye indirgenmiştir.

•*fiyat_tipi* alanında aynı anlama gelen ifadeler Türkçeleştirilerek birleştirilmiştir.

•*rezervasyon_durumu* hedef nitelikle ilgili bilgi barındırdığından veri setinden çıkarılmıştır.

•Konaklanacak gün sayısı veri setindeki *gece_sayisi* alanında bulunduğundan *cikis_tarihi* niteliği veri setinden çıkarılmıştır.

3.2.3 Veri ayrıklaştırma

Veri setinde bulunan *giriş_tarihi* ve *kayit_tarihi* alanları gün, ay ve yıl olarak ayrılmıştır. Her iki nitelikteki gün alanı, haftanın günlerini içeren kategorik veri olarak düzenlenmiştir.

3.2.4 Eksik verinin tamamlanması

Bu aşamaya kadar gerçekleştirilen işlemler sonucunda ortaya çıkan veri setinin özet görüntüsü Şekil 1'de sunulmuştur.

```
otel_kodu        gun             ay          yil          oda_sayisi       yetiskin
1:7261    Cuma      :5290   8       : 5215   2016:   12   Min.    : 1.00   Min.    :0.000
2:8110    Cumartesi:5998   9       : 4013   2017:32363   1st Qu.: 1.00   1st Qu.:2.000
3:5980    Çarşamba :5171   7       : 3778   2018: 3789   Median : 1.00   Median :2.000
4:6853    Pazar    :4478   12      : 3505                Mean   : 1.03   Mean   :1.989
5:7960    Pazartesi:5261   3       : 2902                3rd Qu.: 1.00   3rd Qu.:2.000
          Perşembe :5537   10      : 2884                Max.   :63.00   Max.   :8.000
          Salı     :4429   (Other):13867
    cocuk1           cocuk2             bebek         ozel_istek  toplam_fiyat         kayit_gun
Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   0:32693    Min.   :    1.0   Cuma      :5237
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1: 3471    1st Qu.:  300.0   Cumartesi:4548
Median :0.0000   Median :0.00000   Median :0.0000              Median :  809.3   Çarşamba :5609
Mean   :0.1441   Mean   :0.03998   Mean   :0.0618              Mean   : 1794.8   Pazar    :3332
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000              3rd Qu.: 2280.3   Pazartesi:5615
Max.   :6.0000   Max.   :3.00000   Max.   :2.0000              Max.   :63246.4   Perşembe :5833
                                                                                 Salı     :5990
   kayit_ay     kayit_yil       milliyet      gece_sayisi     ucretli_kisi_sayisi hesaplanan_gece
3       : 4123   2017:34096   Yabancı:25261   Min.   :  1.00   Min.   :1.000       Min.   :  1.000
11      : 4105   2018: 2068   Yerli  :10903   1st Qu.:  1.00   1st Qu.:2.000       1st Qu.:  2.000
2       : 3666                               Median :  3.00   Median :2.000       Median :  6.000
7       : 3368                               Mean   :  3.38   Mean   :2.004       Mean   :  7.331
12      : 3316                               3rd Qu.:  5.00   3rd Qu.:2.000       3rd Qu.: 10.000
8       : 3289                               Max.   :180.00   Max.   :8.000       Max.   :360.000
(Other):14297
konaklama_tipi          fiyat_tipi    toplam_rez_say   toplam_iptal_say   iptal_durumu
AI : 5965      Esnek Fiyat     :   13   Min.   :  1.000   Min.   :  0.0000   0:25151
BB :14790      FiyatTipi 1     : 6317   1st Qu.:  1.000   1st Qu.:  0.0000   1:11013
HB :   15      İade Edilebilir:11766   Median :  1.000   Median :  0.0000
OB : 2798      İade Edilemez   : 5236   Mean   :  2.492   Mean   :  0.9119
RO : 5743      Kontrat         : 6852   3rd Qu.:  1.000   3rd Qu.:  0.0000
UAI: 6853      Standard        : 5980   Max.   :196.000   Max.   :124.0000
```

Şekil 1. Veri seti özet görüntüsü

Şekil 1 incelendiğinde veri setinde *konaklama_tipi* ve *fiyat_tipi* alanlarında NULL değerler bulunduğu ve kayıp değerler olduğu görülmektedir. Kayıp değer içeren nitelik eğer nümerikse kayıp değerlerin o niteliğe ait ortalama değeriyle, eğer nitelik kategorikse nitelikte en çok tekrar eden değer ile doldurulması uygulanabilecek en basit yöntemlerdendir [14]. *konaklama_tipi* ve *fiyat_tipi* alanlarındaki kayıp veri otel bazında en çok tekrar eden değerle doldurulmuştur.

3.2.5 Aykırı değerlerin ve tekrarlı verinin temizlenmesi

Nümerik değer içeren *toplam_fiyat* alanında ilk yüzde 25'lik ve son yüzde 25'lik dilimde olan kayıtlar veri setinden çıkarılarak aykırı değerler temizlenmiştir. Veri setinde tekrar eden veri R dilindeki fonksiyonlar kullanılarak temizlenmiştir.

3.2.6 Veri setinin dengeli hale getirilmesi

Veri setinde hedef nitelikteki sınıf sayıları arasında bir orantısızlık olması durumunda veri seti "dengesiz veri seti" olarak nitelendirilmektedir [20]. Dengesiz veri setleri ile yapılan analizlerde yalnızca doğruluk değerine bakılması sağlıklı değerlendirme yapılmasını güçleştirecektir [20]. Dengesiz veri setlerinde performans ölçütü olarak doğruluk değerinin yanında duyarlılık, belirleyicilik ve F-ölçütü diğer performans değerlerine de bakılması gerekmektedir. Ayrıca analizlerden önce çeşitli tekniklerle veri setini dengeli hale getirmek de mümkündür.

Hedef niteliği olan *iptal_durumu* alanına bakıldığında iptal edilen rezervasyon sayısının 6631, iptal edilmeyen rezervasyon sayısının 18420 olduğu görülmüştür. Bu yönüyle veri setinin hedef nitelik bakımından dengesiz bir yapıda olduğu söylenebilir.

Veri setini dengeli hale getirmek için veriyi yeniden örnekleme yöntemlerinden biri hedef niteliğin sayıca çok olan sınıf etiketi sayısının, sayıca az olan sınıf etiketine yaklaştırılması yani örneklem imdirgeme (*undersampling*) tekniğidir [20]. Bu çalışmada veri setini undersampling tekniğiyle dengeli hale getirmek için ROSE [21] paketi kullanılmıştır. Undersampling işlemi sonrasında

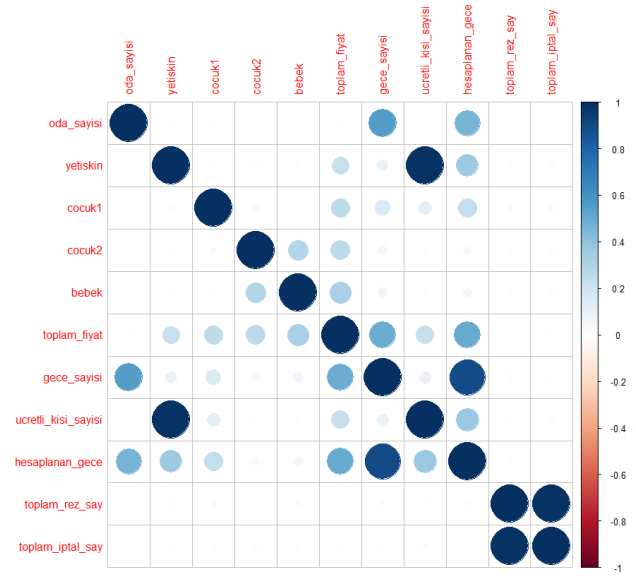
iptal edilen ve edilmeyen rezervasyon sayıları 6631 olmuştur.

3.2.7 Nitelik seçimi

Analize dâhil edilecek nitelikleri belirlemek üzere öncelikle nümerik alanlar arasındaki korelasyon değerlerine bakılmıştır. Şekil 2'de nitelikler arasındaki korelasyon ilişkisi görülmektedir. Dairelerin çapı ve rengi korelasyonun büyüklüğünü ve yönünü göstermektedir.

Korelasyon katsayısı ∓0,80 ve üzerinde olan nitelikler analize dâhil edilmemiştir. Şekil 2 incelendiğinde *ucretli_kisi_sayisi* ile *yetiskin*, *hesaplanan_gece* ile *gece_sayisi* ve *toplam_iptal_say* ile *toplam_rez_say* niteliklerinin yüksek korelasyona sahip olduğu görülmektedir. Bu nedenle *ucretli_kisi_sayisi*, *hesaplanan_gece* ve *toplam_rez_say* nitelikleri veri setinden çıkarılmıştır.

Niteliklerin tahmin sonuçları üzerinde ne kadar etkili olduğunu tespit etmek üzere RF algoritmasından faydalanılmıştır. RF algoritmasından elde edilen çıktılardan biri de nitelik önemidir. RF algoritması, veri setindeki satırların yerleri değiştirildiğinde tahmin hatasının ne kadar olduğunu inceleyerek niteliklerin önemini hesaplamaktadır [14].



Şekil 2. Nümerik niteliklere ait korelasyon grafiği

Nihai modeller oluşturulmadan önce RF algoritması tüm veri seti için çalıştırılarak niteliklerin tahminler üzerindeki etkileri belirlenmiştir. Tahminler üzerinde etkisinin düşük olduğu görüldüğünden *milliyet, yil, kayit_yil ve oda_sayisi* nitelikleri analize dâhil edilmemiştir. Tablo 3'te nitelik önem değerleri görülmektedir.

Tablo 3. Niteliklerin tahmin sonuçları üzerindeki etki gücü

Nitelik	Önem	Oran (%)
toplam_fiyat	1021,77	19,80
fiyat_tipi	605,36	11,73
kayit_ay	544,06	10,54
kayit_gun	534,52	10,36
gun	518,11	10,04
ay	486,07	9,42
toplam_iptal_say	465,32	9,02
gece_sayisi	387,64	7,51
otel_kodu	231,00	4,48
konaklama_tipi	170,67	3,31
milliyet	85,88	1,66
yil	47,41	0,92
kayit_yil	43,47	0,84
oda_sayisi	18,40	0,36

3.3 Model kurma

Makine öğrenmesi alanında pratik yapanlar, karmaşık ve çok sayıda farklı nitelikten oluşan ham veriyi birtakım yöntemler kullanarak modellemektedir. Geliştirilen model, verinin içindeki olası gizli düzen ya da örüntüyü yakalamak için kullanılmaktadır [19].

Makine öğrenmesi çalışmalarında genellikle birden fazla model oluşturulmakta ve modellerin performansları karşılaştırılmaktadır. Farklı algoritmalar kullanılarak modeller oluşturulabileceği gibi, bir algoritmanın parametreleri değiştirilerek de yeni modeller oluşturulabilmektedir.

Bu çalışmada rezervasyon iptallerini tahmin etmek üzere veri seti RF, kNN, SVM ve C4.5 algoritmaları ile analiz edilmiş ve dört farklı model oluşturulmuştur. Oluşturulan modellerin performansları, belirlenen performans değerlendirme kriterleri açısından karşılaştırılmış ve en iyi model tespit edilmiştir.

Analizler R dilinde gerçekleştirilmiştir. R, istatistiksel hesaplamalar ve grafikler oluşturmak için kullanılan ücretsiz bir programlama dilidir [22]. Kodlama ortamı olarak RStudio kullanılmıştır. Veri ön işleme adımları için R dilinin yanında Microsoft Excel yazılımından da faydalanılmıştır.

Bu çalışmadaki analizlerde RF algoritması için randomForest [23] paketi, kNN algoritması için dprep [24] paketi, SVM algoritması için e1071 [25] paketi, C4.5 algoritması için RWeka [26] paketi kullanılmıştır.

3.4 Performans değerlendirme yönteminin belirlenmesi

Bu çalışmada oluşturulan modellerin performansını ölçmek üzere dışarıda tutma yöntemi (*Holdout*) kullanılmıştır. Holdout yönteminde veri seti eğitim ve test olmak üzere rastgele iki gruba ayrılır. Eğitim ve test veri setlerinde, hedef niteliğin tüm sınıflarının orantılı bir şekilde dağılması için tabakalı örnekleme yönteminden faydalanılır. Oluşturulan model eğitim veri setiyle eğitilir ve test veri setiyle de modelin performansı ölçülür [14].

Veri seti caret [27] paketi kullanılarak rastgele %75 eğitim ve %25 test olarak ikiye ayrılmıştır. Oluşturulan modeller eğitim veri setiyle eğitilmiş, ardından modellerin performansları test veri setiyle ölçülmüştür.

3.5 Performans değerlendirme ölçütlerinin belirlenmesi

Bu çalışmada oluşturulan modellerin performansını karşılaştırmak üzere doğruluk, duyarlılık, belirleyicilik ve F-ölçütü değerleri kullanılmıştır.

4 Bulgular ve tartışma

Rezervasyon iptali üzerinde rezervasyonun toplam tutarının en yüksek öneme sahip olduğu görülmektedir (Tablo 3).

Oluşturulan modellerden elde edilen performans sonuçları Tablo 4'te sunulmuştur. Performans sonuçları incelendiğinde rezervasyon iptallerinin ortalama %70 doğruluk oranı ile tahmin edildiği görülmektedir. C4.5 algoritması ile oluşturulan modelin en iyi doğruluğu (%73) verdiği görülmektedir.

Tablo 4. Karşılaştırmalı performans tablosu

Algoritma	Doğruluk	Duyarlılık	Belirleyicilik	F-ölçütü
RF	0,72	0,67	0,78	0,71
SVM	0,67	0,58	0,76	0,64
kNN	0,67	0,62	0,71	0,65
C4.5	0,73	0,67	0,79	0,71

Model performansları değerlendirilirken rezervasyonun iptal edilmesi durumu pozitif sınıf olarak seçilmiştir. Tüm modellerde duyarlılık değerinin belirleyicilik değerinden düşük çıkması, modellerin pozitif sınıfı doğru tahmin etme konusunda, negatif sınıfa oranla daha başarısız olduğunu göstermektedir. Yani modeller iptal edilmeyen rezervasyonları daha yüksek oranda doğru tahmin ederken, iptal edilen rezervasyonları daha düşük oranda doğru tahmin etmektedir.

5 Sonuçlar

Rezervasyon sistemleri müşterilere hizmet alımını garanti altına alma imkânı, otellere de kaynaklarını doğru yönetme olanağı sunar. Oteller aldıkları rezervasyon doğrultusunda personel istihdam eder, ihtiyaç duydukları alt yapıyı hazırlar, malzeme ve gıda stoklarını oluşturur. Oteller açısından bu konudaki en büyük risk, yapılan rezervasyonların müşteri tarafından iptal edilmesidir. Müşteri kaynaklı iptallerin nedenleri bilinemediğinden önüne geçmek pratik olarak mümkün değildir. Bu nedenle oteller her halükârda yapılan rezervasyonların bir bölümünün iptal edilmesi durumuyla karşı karşıya kalmaktadır. Bu noktada iptallerden doğacak riski minimize etmenin en kolay yolu, iptal edilecek rezervasyonların önceden tahmin edilmesidir.

Bu çalışmada makine öğrenmesi tekniklerinden faydalanılarak otel rezervasyon iptalleri tahmin edilmeye çalışılmıştır. Kurulan modeller otel rezervasyon iptallerini %73 doğrulukla tahmin edebilmektedir. Çalışmada kullanılan veri seti tek bir kaynaktan sağlanmamıştır. Bu ise veri setinin anlaşılmasını ve analizini zorlaştırmıştır. Ayrıca bu durumun performansın nispeten düşük çıkmasında etkili olduğu düşünülmektedir.

Çalışmada verisi kullanılan oteller farklı şehirlerde olduğundan tercih edilme sebepleri de buna bağlı olarak değişmektedir. Bu nedenle otel rezervasyon iptalleri tahmin edilirken tüm veri seti üzerinden değil otel bazlı model kurulması daha sağlıklı sonuçlar verebilir.

Sonuçların daha da iyileştirilmesi için mevcut veri seti üzerinde yapay sinir ağları ve diğer makine öğrenmesi algoritmaları kullanılabilir ve çalışmada kullanılmayan farklı dengeleme işlemleri veri setine uygulanabilir.

Kaynaklar

[1] Kimes SE, Wirtz J. "Has revenue management become acceptable? Findings from an international study on the perceived fairness of rate fences". *J. Serv. Res.*, 6(2), 125–135, 2003.

[2] Mehrotra R, Ruttley J. "*Revenue management*". (2nd ed.). Washington, DC, American Hotel and Lodging Association, 2006.

[3] Smith SJ, Parsa HG, Bujisic M, van der Rest JP. "Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry". *J. Travel Tour. Mark.*, 32(7), 886–906, 2015.

[4] Morales DR, Wang J. "Forecasting cancellation rates for services booking revenue management using

data mining". *Eur. J. Oper. Res.*, 202(2), 554–562, 2010.
[5] Liu PH. "Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods". *Revenue Manag. Pricing Case Stud. Appl.*, 91–101, 2004.
[6] Carbonell CG, Michalski RS, Mitchell TM. "An overview of machine learning". In *Machine Learning*, San Francisco, CA, Morgan Kaufmann, 3-23, 1983.
[7] Sullivan W. "*Machine learning Beginners Guide Algorithms Supervised & Unsupervised learning, Decision Tree & Random Forest Introduction*". USA, CreateSpace Independent Publishing Platform, 2017.
[8] Pölt S, "Forecasting is difficult–especially if it refers to the future". In *AGIFORS- Reservations and Yield Management Study Group Meeting Proceedings*, 61–91, 1998.
[9] Antonio N, de Almeida A, Nunes L. "Predicting hotel booking cancellations to decrease uncertainty and increase revenue", *Tour. Manag. Stud.*, 13(2), 2017.
[10] Ho TK. "Random decision forests". In *Proceedings of the third international conference on Document analysis and recognition*, 278–282, 1995.
[11] Breiman L. "Random forests". *Mach. Learn.*, 45(1), 5–32, 2001.
[12] Louppe G. "Understanding Random Forests: From Theory to Practice". Doktora Tezi, University of Liege, Belgium, 2014.
[13] Cunningham P, Delany SJ, "k-Nearest neighbour classifiers", *Mult. Classif. Syst.*, 34(8), 1–17, 2007.
[14] Balaban ME ve Kartal E, "*Veri madenciliği ve makine öğrenmesi temel algoritmaları ve R dili ile uygulamaları*". 2. Baskı. İstanbul, Çağlayan Kitabevi, 2018.
[15] Vapnik V, Lerner A. "Pattern recognition using generalized portrait method". *Autom. Remote Control*, 24, 774-780, 1963.
[16] Nayak J, Naik B, Behera H. "A comprehensive survey on support vector machine in data mining tasks: applications & challenges". *Int. J. Database Theory Appl.*, 8(1), 169–186, 2015.
[17] Quinlan JR. "*C4. 5: programs for machine learning*", CA, Morgan Kaufmann Publishers, 1993.
[18] Çölkesen İ. "Uzaktan algılamada ileri sınıflandırma tekniklerinin karşılaştırılması ve analizi", Yüksek Lisans Tezi, Gebze Yüksek Teknoloji Enstitüsü, Kocaeli, 2009.
[19] Kartal E. "Sınıflandırmaya dayalı makine öğrenmesi teknikleri ve kardiyolojik risk değerlendirmesine ilişkin bir uygulama". Doktora Tezi, İstanbul Üniversitesi, İstanbul, 2015.
[20] Kartal E, Özen Z. "Dengesiz Veri Setlerinde Sınıflandırma". İçinde *Mühendislikte Yapay Zekâ Uygulamaları*, Sakarya, 109-131, 2017.
[21] Lunardon N, Menardi G, Torelli N, "ROSE: a Package for Binary Imbalanced Learning". *The R Journal*, 6(1), 82–92, 2014.
[22] R Core Team, "R: A Language and Environment for Statistical Computing". https://www.R-project.org/ (01-Eki-2018).
[23] Liaw A, Wiener M. "Classification and Regression by randomForest". *R News*, 2(3), 18-22, 2002.
[24] Acuna E, The CASTLE Research Group "dprep: Data Pre-Processing and Visualization Functions for Classification". 2015.
[25] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. "e1071: Misc Functions of the Department of Statistics". Probability Theory Group (Formerly: E1071), TU Wien. 2017.
[26] Hornik K, Buchta C, Zeileis A. "Open-Source Machine Learning: R Meets Weka", *Comput. Stat.*, 24(2), 225–232, 2009.
[27] Kuhn M., caret: Classification and Regression Training. 2018.