TECHCAREER.NET

DATA SCIENCE BOOTCAMP

MEZUNİYET PROJESİ

OTEL REZERVASYONLARININ İPTALİ TAHİMİNİ İÇİN SINIFLANDIRMA ÇALIŞMASI

HAZIRLAYAN: BUĞRA AKOĞLU

EĞİTMEN: ÇAĞLA ŞAHİN

07 EKİM 2023

İçindekiler

Veri Araştırması:

Mezuniyet projesinin kapsamını karşılayacak bir veri setinin araştırılması sırasında kullanılacak olan model tipinin sınıflandırma modeli olmasına daha önceden karar verdim. Sınıflandırma modellerinin düzgün çalışabilmesi açısından dikkat ettiğim bazı unsurlar şu şekildedir:

1. En az 10 özellik içermesi
2. Nominal ve numerik değişkenleri barındırması ve numerik değişkenlerin nominal değişkenlere göre bir miktar daha fazla olması.
3. Hedef değişkenin özellikler ile mantıklı bir şekilde ilişkili olması.
4. Dikey boyut bakımından verinin barındırdığı her bir özellik başına 1000 örnek barındırması.
5. Yapılacak olan çalışmanın eğitim kapsamında öğrendiklerimi kullanabilme kabiliyetimi göstermesi gerektirdiğinden ötürü verini aşırı temiz ve işlenmiş olmaması.
6. DateTimeIndex kullanılabilmesi açısından en azından sadece görselleştirme yapılabilecek kadar uygun olacak şekilde zaman verisi barındırması.

Yukarıdaki bahsettiğim koşullar ile Kaggle ve UCI üzerinden yaptığım araştırmada Ahsan Raza tarafından Kaggle platformunda paylaşılan Hotel Reservations Dataset isimli veri setinin gerekliliklerin çok büyük bir kısmını karşılamasından dolayı proje için uygun veri seti olarak kullanılmasına karar verdim. Bu veri setini kullanmaya karar vermeden önce incelediğim bazı veri setleri aşağıdaki gibidir.

1. Wine Quality
2. Credit Card Customers
3. Company Bankruptcy Prediction

Wine Qulatiy isimli veri seti, Hotel Reservations veri setini kullanmaya karar vermeden önce en çok incelediğim veri seti oldu. Farklı şarapların kimyasal değerlerinin ölçümlerini barındıran veri seti hedef değişken olarak şarap örneklerinin uzamanlar tarafından incelendikten sonra belirlenen kalite düzeyi bilgisini içermektedir. Ancak Nominal değerlerin eksikliği ve neredeyse tamamen numerik veri barındırmasıyla birlikte verinin görece çok temiz olmasından dolayı tercih etmeyip başka bir veri seti araştırmak zorunda kalmama neden olmuştur.

Verinin İncelenmesi ve Yüklenmesi:

Kaggle'da yayımlanmış olan Hotel Reservations verisi içerdiği özelliklerin ne anlama geldiğini ortalama bir yeterlilikte açıklamaktadır ancak turizm sektörüne olan yabancılığı birkaç özelliğin tam olarak ne ifade ettiğini anlamamda bir zorluk yarattı. Özellik isimleri üzerinde yaptığım kısa bir araştırma bu özelliklerin ne anlama geldiklerini ve nasıl bir ölçüm birimine sahip oldukları konusunda bilgi sahibi olmamı sağladı. Örnek vermem gerekirse, lead_time isimli özelliğin rezervasyonun bağlandığı tarih ile rezervasyon tarihi arasındaki sürenin gün bakımından gösterimi olduğunu öğrendim ki bu özelliğin ilerleyen görseller ve korelasyon tablosu üzerinden incelendiğinde hedef değişken üzerinden görmezden gelinmeyecek bir etkisi olabileceğini yorumladım. Verini bir ilişkisel veri tabanına ait tablo gibi durması ise veri hakkında geçmiş deneyimlerimden kaynaklanarak iki varsayım oluşturmamı sağladı.

1. Bu veri gerçek bir veri olup maskelenerek yayımlandı.
2. Bu veri yapay bir veri olup SQL üzerinden rasgele olarak belirli parametreler ile oluşturuldu.

Betimsel istatistiklerin incelenmesi sırasında denk geldiğim bazı durumlar bu iki varsayımımın çok olası olduğunu gösterir niteliktedir.

Verinin Yüklenmesi ve İşlenmesi:

Verinin csv formatında olmasından dolayı yüklenmesi konusunda bir sorun ile karşılaşmadım. Verinin temel yapısını incelediğimde numerik değerlerin int64 veya float64 şeklinde tanımlandığını ve nominal değişkenleri object tipinde tanımlandığını gördüm ancak bu değişkenleri daha dikkatli incelediğimde iki nominal değişkenin binary olarak girişi yapıldığı için int64 olarak tanımlandığını farkettim ve uygun olacak şekilde string değerler tanımlayarak bu değişkenler object tipini dönüştürdüm. Bu dönüşümleri yapmadan önce veri seti yüklendiği anda raw_data isminde bir DataFrame olarak kaydettim ve betimsel işlemler, zaman serisi işlemleri, korelasyon işlemler ve model eğitimi öncesinde uygun isimler belirleyerek bu DataFrame'den kopyalarak sakladım. Bunu yapmamın sebebi bu farklı başlıklar altında yapacağım değişikliklerin birbirlerine kötü etkisinin olabilmesinden kaynaklanıyor ki zaman serileri üzerine DateTimeIndex işlemi yapılması diğer bütün işlemleri etkileyebilecek bir etmendir.

Veri setinin tekrarlı temerrüt kayıt barındırıp barındırmadığı durumunun kontrolü bir veri tabanı sistemine ait bir tabloya benzetmeme neden olan Booking_ID adlı sütunun bir unique key olmasından dolayı çok rahat bir şekilde kontrol edildi ve temerrüt kayıt barındırmadığını gördüm. Ayrıca her bir sütun için Na ve NaN geçersiz değerlerinin varlığını kontrol ettiğimde böyle bir değer ile karşılaşmadım. Verinin bir yanlışlık barındırmadığı ve kullanıma hazır olduğuna kanaat getirdikten sonra betimsel istatistiklerin incelenmesi çalışmasına başladım.

Verinin İşlenmesi ve Betimsel İstatistiklerin İncelenmesi:

Bir rezervasyonun kaç hafta içi gece ve kaç hafta sonu geceye sahip olduğu verisinin bir lineer birleşimi olarak toplam kaç gece rezerve edildiğine dair bir özellik oluşturmanın uygun olacağını ve böyle bir görselleştirmenin yapılmasının iyi olacağını düşündüğüm için bu iki özelliği basitçe toplayarak no_of_total_nights isminde yeni bir özellik elde ettim ve DataFrame içerisindeki konumunu bu iki değişkenin hemen bitişiğine yerleştirerek daha uygun bir görsel görünüm elde ettim.

Verinin barındırdığı tarihsel veri yıl, ay ve gün bilgilerini ayrı sütunlarda göstermektedir. Bu bilgilerin birleşimi ile tam bir tarih özelliği elde etmek istedim ve combine_dates isimli bir fonksiyon yazarak bu verileri birleştirdim. Ancak bu işlemi yaparken bir hata ile karşılaştım ki bu hata day is out of range for the month şeklinde bir uyarı vermiştir. Uyarıyı dikkate alarak ilk olarak 2017 ve 2018 yıllarının şubat aylarının günlerinin unique değerlerini sorguladım ve şüphelendiğim gibi 29 değerine rastladım ki bu yıllar artık yıl değildir. Bu durumun 37 kayıt barındırması çok büyük bir veri kaybı yaşamadan bu kayıtların silinebileceğini göstermektedir. Ben bu durumun yanlış bir simülasyon veya düşünülmeden yapılmış bir maskeleme olduğuna inanıyorum. Bu 37 kayıtın silinmesindense 28 Şubat olarak düzenlenmesinin bilgi bakımından bir zarar vermeyeceğini kanaat getirerek düzenleme işlemini gerçekleştirdim.

Verinin Görselleştirilmesi:

Yapılacak olan görselleştirmelere ilk olarak hedef değişkenin dağılımının bar grafiği ile gösterilmesiyle başladım. Binary bir yapıya sahip olan değişkende rezervasyon iptalinin olmamasının daha baskın bir durumda olduğunu gördüm ki bu durum model oluştururken accuracy metriğinin yeterli olmayacağını ve yanlı öğrenime sebep olabileceğine bir işarettir. Kullanmayı düşündüğüm logistik regresyon, svm ve random forest modellerinin bagging classifier kullanarak eğitilmesinin bu durumun üstesinden geleceğine karar verdim.

Daha sonraki görselleri her bir değişkenin hedef değişken bakımından incelenmesiyle devam ettim. Böylece hedef değişkenin belirli sonuçlarının ağır bastığı ve belirleyici olabilecek değişkenleri görmeyi amaçladım. Ancak genel olarak benzer dağılım eğilimlerini rastladım. Sadece bir rezervasyonun tekrarlı (eski) bir müşteriye ait olmasının rezervasyon iptalinin olmaması durumuna etki ettiğini gördüm.

Zaman serisi grafiklerine geçmeden önce her bir değişkenin nasıl dağıldığını görmek için bar ve pie grafikleri kullandım. Böylece değişkenler içerisinde inbalance durumunu ve olası modeli etkileyebilecek durumları gözden geçirdim. Görsellerde en çok dikkatimi çeken bulgu misafir edilen çocuk sayısının çok yüksek olmasıydı. Alan hakkında yaptığım araştırma ve turizm alanında çalışan arkadaşlarıma danışmam sonucunda otellerde okul gezisi veya herhangi bir etkinlik yaraşma için yapılan rezervasyonlarda bu tür durumlar ile karşılaşıldığını gösterdi. Ayrıca otelin hafta sonu müşteri barındırmaktan ziyade ağırlık ile hafta içi müşteri barındırdığı bilgisine ulaşarak bu otelin tatil tipi bir oteldense işlevsel bir otel olabileceği kanısına vardım.

Rezervasyonların toplam kaç gece olduğu özelliğini incelediğimde ortalamanın yaklaşık olarak 3 olduğunu gördüm ve dağılım 0 ile 24 aralığında sağa çarpık bir yapıdadır.

Benzer dağılım karakteri diğer özelliklerde de çok farklılık olmadan gözlemlenmiştir.

Zaman bakımından veri incelenmeden önce üç ayrı sütundan elde edilen tarih özelliği ile DateTimeIndex işlemi yapılmıştır. Verinin çok fazla gün barındırması sebebiyle yapılan gün bazlı grafikler okunmayacak biçimde oluşmuştur bu yüzden resample işlemi yapılarak ay bazında incelenmesini uygun gördüm. 2017 Temmuz – 2018 Aralık tarihleri arasında misafir sayıları, rezervasyon sayıları ve ağırlanan eski müşteri sayıları incelendiğinde:

1. 2017 ve 2018 yılları arasında müşteri sayısı bakımından bariz bir fark gözükmektedir.
2. Otel sıcak aylarda daha fazla rezervasyon almıştır.
3. Rezervasyonlarda çocuk ve yetişkin sayıları pozitif lineer ilişki göstermiştir.
4. Eski müşteriler otelin daha boş olduğu ara sezon olarak adlandırılabilecek dönemlerde daha fazla görülmüştür. (Özellikle 2018 Aralık ayında pik gözlemlenmiştir ki total rezervasyon sayısı o ayda düşüş trendindedir.)
5. Mevsimsellik sektör gereği bariz bir şekilde varlığını ortaya koymaktadır.
6. 2017 ve 2018 yıllarına ait rezervasyon sayıları bakımından bariz bir artış trendi görülmektedir. Çalışmanın içerisinde bulunmasa bile mevsimsellikten arındırılarak bu trend daha iyi bir şekilde görülebilir.
7. Ay ve yıl bakımından az gruplanması sebebiyle zaman serisi analizi yapılmak istenirse gün bazından alınması örnek boyutunun yeterli olması açısından daha iyi olacaktır.
8. ACF ve PACF grafiklerinin incelenmesiyle uygun model parametreleri çıkarılabilir.

Korelasyon:

Korelasyon katsayılarının elde edilmesi işlemi için farklı metriklerin kullanılması gerekmektedir.

1. Numerik – Numerik ilişki: Pearson
2. Nominal – Nominal ilişki: Cramers’s V
3. Nominal – Numerik ilişki: Corelation Raiton

Bu değerlerin ayrı ayrı elde edilmesi ve sonra tekbir matris üzerinde birleştirerek tek bir ısı grafiği elde etmeyi amaçladım ancak daha sonra bunu yapan bir kütüphane olduğunu öğrendim. Dython isimli kütüphanenin hem nominal sütunları ayırt etmesi için bir fonksiyona sahip olması hem de associations adındaki fonksiyonunun bu ayrımı yaparken her bir değişkenin tanımlanan tipine göre uygun ilişki katsayısını kullanarak ısı grafiğini oluşturması mükemmel bir kolaylık sağladı. Sahip olduğu çok sayıdaki parametre renk ayarlarından hesaplama yaparken kaç fiziksel çekirdek kullanmasını istediğime kadar karar verebilmeme olanak tanıdı. Bu ısı grafiğini oluştururken yaşadığın tek sıkıntı ise annot=True olarak ayarlamama rağmen ilk satır haricinde hücre değerlerinde sayısal değerleri gösterememem oldu ki elle hesaplamaları yapıp seaborn kütüphanesinin heatmap fonksiyonunu kullanarak oluşturmaya çalıştığımda aynı durum ile karşılaştım. Bunun benim Python ortamımım veya kullandığım vscode editörü ile alakalı olabileceğini düşünmekteyim.

Modelin Eğitilmesi, Tahmin ve Karşılaştırma:

Sınıflandırma probleminin çözümü için daha önce kullandığım ve altyapısına hakim olduğum 6 modeli kullanmayı uygun gördüm. Bu modeller:

1. Lojistik Regresyon
2. Destek Vektör Makineleri
3. Rastgele Orman
4. AdaBoost
5. XGBoost
6. CatBoost

Şeklindedir. Bu modeller içerisinden ilk üçü verinin inbalance durumuna karşın zayıf kalabileceğinden ve diğer 3 modelin Boosting metodu kullanarak ensemble modeller oluşturacağından dolayı performans bakımından yetersiz kalabileceğini düşündüğüm için bagging metodu ile kullanarak ensemble modeller oluşturmayı amaçladım. Zayıf öğrenicilerin boostrap kullanarak görece daha küçük örneklemler üzerinden edindikleri farklı bilgi birikimlerini birleştirerek daha kuvvetli bir tahmin modeli oluşturmalarını amaçladım. Grandyant temmelli modellerin içerisinde catboost modelinin kategorik verileri daha iyi işlemesi ve grandyant temelli çalışma prensibinde zayıf öğrenicileri ağırlıklandırma işlemini daha iyi yapacağını düşünerek sonradan ekledim.

Model içerisinde anlamlı bilgi barındırmadığını düşündüğüm özellikleri düşürerek yeni bir dataframe oluşturduktan sonra CatBoost modelinde çalışacak eğitim ve test verilerini ayrı diğer modellerde çalışacak verileri ayrı oluşturdum. Bunun sebebi Catboost modelinin diğer modellerden farklı olarak nominal verilerin encode edilmesine ihtiyaç duymamasıdır ki bir sonraki adımda diğer modellerin kullanacağı veriler label encoding kullanarak encode edilmiştir. One hor encoding kullanılmamasının sebebi üstesinden gelemediğim index sebepli olduğunu düşündüğüm bir NaN değerlerin türemesi problemi oluşmasından kaynaklanmaktadır.

Her bir model eğitilirken ortalama 250 tahmin edici türetileceğinden dolayı kullanacakları zamanın farklılık yaratabileceğini düşündüğüm içim time kütüphanesini kullanarak runtime verilerini performans metrikleri ile kaydettim.

Model parametrelerini araştırmak için GridSearchCV kullandım böylece C, depth, n_estimator, learning_rate gibi parametrelerden en uygunları araştırarak oluşabilecek en iyi modelleri oluşturmayı amaçladım. Model performanslarının hassasiyetini daha iyi görebilmek için ROC Curve görselleştirmelerini ekledim. Modellerin eğitilmesi, tahmin değerlerinin oluşturulması ve test edilmesi ile birlikte performans metriklerini ve sürelerini ayrı ayrı kaydedip bir tablo oluşturdum. Bu tabloyu yorumlayarak en iyi performans sahibi modelin tüm metriklerde öne çıkan CatBoost algoritması olduğunu, ikinci en iyi model olarak görece az farklılıklar ama çok daha hızlı çalışma süresiyle XGboost algoritmasını olduğu kanaatine vardım. Modeller içerisinde logistik regresyon ve destek vektör makineleri en kötü sonuçları veren modeller olmakla birlikte destek vektör makineleri uzun runtime süresi ile pratik bir kullanıma uzaklığıyla dikkat çekmiştir. SVM modelinin bu tür bir sıkıntı oluşturmasının hiperparametre optimizasyonu ve verinin boyutsal büyüklüğü ile ilgili olduğunu düşünmekteyim. Modelin performansının uygun kernelin kullanılmasında aşırı bağımlı olması kullanımında aşırı dikkatli olmayı gerektirmektedir. Ancak hiperparametre optimizasyonu için harcadığım sürenin runtime süresi bakımından saatler alması sebebiyle yeterli miktarda üstüne düşemedim. Sonuç olarak verinin sahip olduğu özellikler kümesi bir otel rezervasyonunun iptal olup olmayacağını yüksek bir kesinlik ile tahmin edebilme yetisine sahiptir. Bu tahmin işlemini kullanılan modeller arasında en iyi catboost ve xgboost sağlamıştır. Yapılabilecek ANOVA ve farklı analizler ile özelliklerin hedef değişkene etkisi ve birbirlerine etkisi araştırılabilir.