FET445 Veri Madenciliği

Otel Rezervasyon İptal Tahminlemesi ve Müşteri Davranış Analizi

Grup:Team100

Hüseyin Günbeldek – 22040301024 – huseyingunbeldek@stu.topkapi.edu.tr

Berat Cidacı – 24040301050 - beratcidaci@stu.topkapi.edu.tr

Ahmad Alhouri – 22040301159 - ahmadalhouri@stu.topkapi.edu.tr

Ousama Aldaya Fawaz – 22040301232 - ousamaaldayafawaz@stu.topkapi.edu.tr

Github Link: https://github.com/BeratCdc/Team100_VeriMadenciligi.git

Youtube Video link:

Problem Tanımı

Otel endüstrisinde rezervasyon iptalleri, gelir yönetimi ve operasyonel planlama için büyük bir sorundur. Bu proje, bir müşterinin rezervasyonunu iptal edip etmeyeceğini (booking_status) rezervasyon detaylarına (fiyat, kalış süresi, özel istekler vb.) dayanarak tahmin etmeyi amaçlamaktadır.

Amaç: Müşteri demografisi ve rezervasyon detaylarını kullanarak bir rezervasyonun "Canceled" (İptal) mı yoksa "Not_Canceled" (İptal Edilmedi) mi olacağını önceden tahmin etmek.

Hedef: İptal olasılığı yüksek olan rezervasyonlar için otel yönetiminin depozito veya teyit gibi proaktif önlemler almasını sağlamak.

Veri Seti

Veri Seti: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset

Boyut: 36,275 satır, 19 sütun.

Özellik Tipleri:

Sayısal: Ortalama oda fiyatı (avg_price_per_room), rezervasyon süresi (lead_time), yetişkin sayısı.

Kategorikal: Yemek planı (type_of_meal_plan), pazar segmenti (market_segment_type).

Sınıf Dağılımı: Veri setinde iptal edilmeyenler (%67) çoğunlukta olsa da, iptal edilenlerin oranı (%33) modelin öğrenmesi için yeterli seviyededir.

Ön İşleme ve Metrikler

- **Temizlik:** Booking_ID gibi ayırt edici olmayan sütunlar kaldırıldı.

- **Encoding:** Kategorikal değişkenler için **One-Hot Encoding** (LabelEncoder'a ek olarak) uygulandı.

- **Ölçeklendirme:** Modellerin (özellikle KNN ve SVM tabanlıların) kararlı çalışması için **StandardScaler** ve **RobustScaler** kullanıldı.

- **Metrikler:** Sınıf dağılımı nedeniyle sadece **Accuracy** (Doğruluk) değil, asıl başarımızı gösteren **F1-Score** ve **ROC-AUC** değerlerine odaklanıldı.

Geliştirilen En İyi Modeller

Best Model 1: XGBoost Classifier
Kullanılan Yaklaşımlar: Gradyan artırma algoritması kullanılarak hataların ardışık olarak düzeltilmesi sağlandı.
Aşırı öğrenmeyi (overfitting) engellemek için **L1 ve L2 regülarizasyonu** uygulandı.
Sınıf dengesizliğini yönetmek amacıyla scale_pos_weight parametresi göz önünde bulunduruldu.
Hyper-parameter Tune Tekniği: **GridSearchCV** kullanılarak en iyi parametre kombinasyonları belirlendi.
Optimize Edilen Parametreler: learning_rate: 0.1, max_depth: 10, n_estimators: 300, subsample: 0.8.
Feature Set: lead_time, avg_price_per_room, no_of_special_requests, arrival_year, market_segment_type (One-Hot Encoded).

Best Model 2: Bagging Classifier
Kullanılan Yaklaşımlar: Varyansı düşürmek amacıyla Bootstrap yöntemiyle veriden rastgele örneklemler alınarak 100 farklı Karar Ağacı eğitildi.

Her bir ağacın tahmini "çoğunluk oylaması yöntemiyle birleştirilerek nihai karar verildi. Modelin kararlılığını artırmak için ağaçların birbirinden bağımsız öğrenmesi sağlandı.
Hyper-parameter Tune Tekniği: Manual Tuning & Cross-Validation
Parametreler: n_estimators: 100, base_estimator: DecisionTreeClassifier, bootstrap: True.
Feature Set: Veri setindeki tüm sayısal ve kategorikal değişkenler ile kullanılmıştır.

Modellerin Karşılaştırılması

Model Adı	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	**0.9005**	**0.9117**	**0.9421**	**0.9267**	**0.9573**
XGBoost + SelectKBest	0.8992	0.9118	0.9400	0.9257	0.9573
Bagging Classifier	0.9000	0.9000	0.9000	0.9000	0.9547
Gradient Boosting	0.8864	0.8953	0.9396	0.9169	0.9464
AdaBoost Classifier	0.8900	0.8900	0.8900	0.8900	0.9462
Stacking Classifier	0.8799	0.8917	0.9335	0.9121	0.9417
Stacking + RFE	0.8787	0.8916	0.9314	0.9111	0.9412
Extra Trees	0.8620	0.8597	0.9480	0.9017	0.9278

Sonuç ve Değerlendirme

- **En Güçlü Tahminleyici:** Geliştirilen 8 farklı ensemble model arasında **XGBoost**, karmaşık müşteri davranışlarını ve doğrusal olmayan ilişkileri en iyi yakalayan model olarak seçilmiştir (**0.9573 ROC-AUC**).
- **Kritik Değişkenler:** Rezervasyonun iptal edilme olasılığını en çok etkileyen faktörlerin **"Lead Time"** (bekleme süresi) ve **"Average Price per Room"** (oda fiyatı) olduğu saptanmıştır. Bu, otel yönetiminin özellikle aylar öncesinden yapılan yüksek fiyatlı rezervasyonlar için depozito şartı getirmesi gerektiğini kanıtlamaktadır.
- **Operasyonel Verimlilik:** Geliştirilen bu modeller, otellerin "Overbooking" (fazladan rezervasyon) stratejilerini veri odaklı yönetmelerine olanak tanıyarak gelir kaybını minimize etme potansiyeline sahiptir.