16.06.2025

Python ile Veri Bilimi

Amazon Satış Verileri ile Sipariş Başarısı Tahmini

Kadir Doğru 132230024

# 1. Giriş

Bu projede, Amazon üzerinde gerçekleşen siparişlerin başarıyla teslim edilip edilmediğini tahmin etmeye yönelik bir sınıflandırma modeli geliştirilmiştir. Hedef, bir siparişin “Shipped” olup olmayacağını veri madenciliği teknikleriyle önceden tahmin etmektir. Bu sayede işletme, iptal riski yüksek siparişleri önceden belirleyip aksiyon alabilecektir.  
Projede Random Forest sınıflandırma algoritması kullanılmış, çeşitli veri ön işleme adımları uygulanmış ve model performansı ölçülmüştür.

# 2. Veri Seti Tanıtımı

# Bu çalışmada kullanılan veri seti, [Kaggle](https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data/data) platformu üzerinden temin edilmiştir. Veri seti, farklı satış kanallarından elde edilen ve Amazon gibi büyük e-ticaret platformlarında gerçekleşen satışlara dair detaylı bilgileri içermektedir. Çeşitli ürün kategorilerini, sipariş durumlarını, teslimat bilgilerini ve müşteri alışkanlıklarını kapsayan bu veri seti, özellikle perakende analitiği ve dijital pazarlama alanlarında kampanya etkinliğini değerlendirmek için değerli bir kaynak sunmaktadır.

# Veri seti e-ticaret, satış yönetimi, lojistik analiz ve müşteri davranışları gibi birçok işlevsel alanı kapsamaktadır. Yapısal olarak 128.975 satır ve 24 sütundan oluşmakta olup, yüksek hacimli ve çok boyutlu veri analizi için uygundur. Aşağıda veri setinden örnek bir satır sunulmuştur:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| order\_id | date | Status | fulfilment | sales\_channel | ship-service-level | category | qty | amount | ship-city | ship-state |
| 405-8078784-5731545 | 405-8078784-5731545 | Cancelled | Merchant | Amazon.in | Standard | Set | 0 | 647.62 | MUMBAI | MAHARASHTRA |

# 4. Veri Ön İşleme

# Veri analizi sürecine geçilmeden önce veri seti detaylı biçimde incelenmiş ve çeşitli ön işleme adımları uygulanmıştır. İlk olarak, veri setinde eksik değerlerin bulunduğu tespit edilmiştir. fulfilled-by ve promotion-ids gibi sütunlarda %50’nin üzerinde eksik veri bulunması nedeniyle bu sütunlar analizden çıkarılmıştır. Sayısal özelliklerden amount ve ship-postal-code, ortanca (median) değeri ile doldurulmuş; böylece aşırı uç değerlerin etkisi azaltılmıştır. Kategorik sütunlar arasında yer alan courier\_status, currency, ship-city, ship-state ve ship-country alanlarında eksik değerler, en sık gözlenen kategori (mode) kullanılarak tamamlanmıştır.

# Veri setinde aykırı değer analizi, sayısal sütunların dağılımına bakılarak gerçekleştirilmiş; ancak bu değerlerin model performansına olumsuz etkisi olmadığı gözlemlendiğinden müdahale edilmemiştir. Kategorik değişkenler, makine öğrenmesi algoritmalarının işleyebilmesi için one-hot encoding yöntemiyle sayısal forma dönüştürülmüştür. Ayrıca, analiz açısından bilgi taşımayan veya çok yüksek benzersiz değere sahip olduğu için öğrenmeye katkısı olmayan order\_id, asin, style gibi sütunlar veri setinden çıkarılmıştır. Bu adımlar sonucunda modelin öğrenme sürecini kolaylaştıracak, tutarlı ve anlamlı bir veri yapısı elde edilmiştir.

# 5. Yöntem ve Uygulama

# Bu projede sınıflandırma problemi olarak ele alınan sipariş durumu tahmini için makine öğrenmesi algoritmalarından Random Forest Classifier tercih edilmiştir. Modelin eğitimi sırasında veri seti %80 eğitim ve %20 test olmak üzere ikiye ayrılmış, bu sayede hem modelin öğrenme başarımı hem de genellenebilirliği değerlendirilmiştir.

# Eğitim sonucunda model oldukça yüksek performans göstermiştir. Elde edilen başarı metrikleri doğrultusunda; doğruluk oranı (Accuracy) %99.88, F1 skoru ise 0.9993 olarak hesaplanmıştır. Modelin tahmin gücü, 21.944 doğru pozitif (TP), 3.822 doğru negatif (TN), sadece 24 yanlış negatif (FN) ve 5 yanlış pozitif (FP) vaka ile oldukça dengeli ve güvenilir sonuçlar üretmiştir.

# Ayrıca modelin hangi değişkenleri daha fazla dikkate aldığı da özellik önem skoru (feature importance) ile analiz edilmiştir. En yüksek bilgi kazancı sağlayan 15 değişken arasından özellikle ship-service-level\_Standard, fulfilment\_Merchant, sales\_channel\_Amazon ve category\_Set gibi özelliklerin tahmin sürecine en fazla katkı sağladığı görülmüştür. Bu tür analizler, hem modelin şeffaflığı hem de işletmenin karar alma süreçlerine yön vermesi açısından büyük önem taşımaktadır.

# 6. Sonuç ve Yorumlar

# Modelin, veri setindeki dengesiz sınıf dağılımına rağmen (siparişlerin yaklaşık %85’i gönderilmiş, %15’i iptal edilmiştir) son derece başarılı bir performans sergilediği gözlemlenmiştir. Random Forest algoritması, hem azınlık sınıfı olan iptal edilen siparişleri hem de çoğunluk sınıfı olan gönderilen siparişleri %99.8’in üzerinde doğrulukla tahmin edebilmiştir. Bu durum, modelin yalnızca baskın sınıfa yönelmeden tüm sınıfları etkin bir şekilde öğrenebildiğini ve genellenebilirliğinin yüksek olduğunu göstermektedir.

# Bu analiz sonucunda geliştirilen tahmin modeli, işletmelere stratejik düzeyde önemli katkılar sunabilecek niteliktedir. Özellikle iptal riski taşıyan siparişlerin önceden belirlenebilmesi, hem müşteri memnuniyetini artırmak hem de lojistik süreçlerde zaman ve maliyet tasarrufu sağlamak açısından değer taşımaktadır. Ayrıca modelin sunduğu çıktılar, operasyonel planlamaların daha veriye dayalı yürütülmesine ve proaktif karar mekanizmalarının geliştirilmesine olanak tanımaktadır.

# 7. Kaynakça

# - Scikit-learn Belgeleri: https://scikit-learn.org

# - Pandas Belgeleri: https://pandas.pydata.org

# - NumPy Belgeleri: https://numpy.org/doc

# - Seaborn Belgeleri: https://seaborn.pydata.org

# - Matplotlib Belgeleri: https://matplotlib.org/stable/contents.html

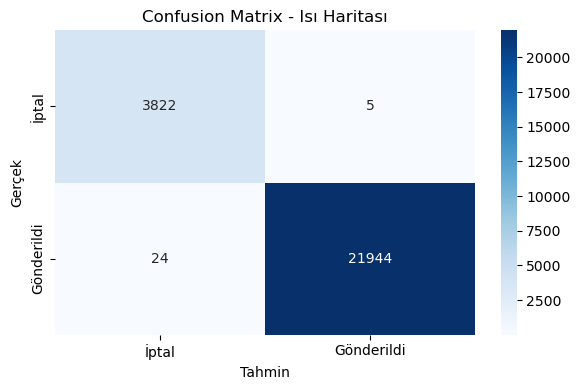
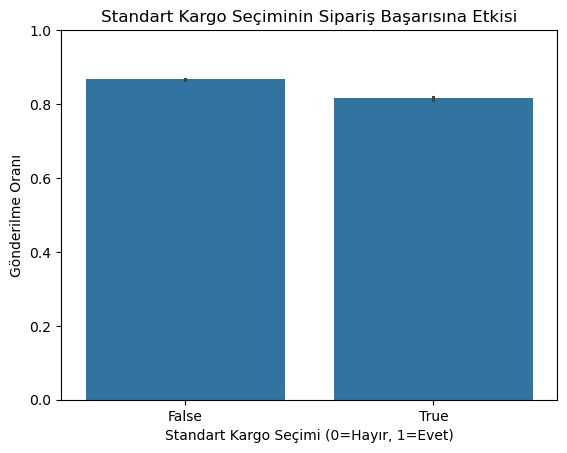
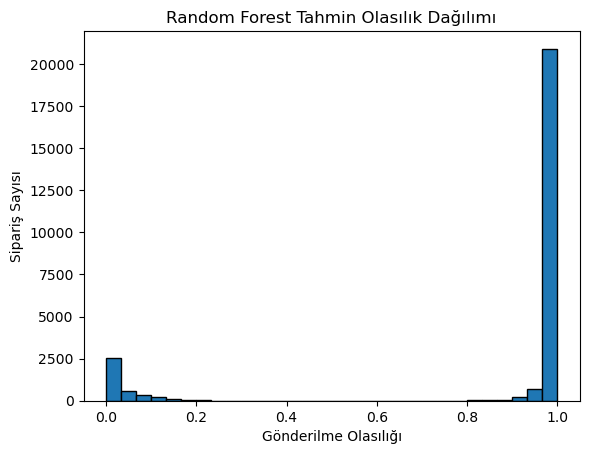
# - Kullanılan Veri Seti: https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data

# - Géron, A. (2019). \*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow\* (2nd ed.). O'Reilly Media.

# - Jupyter Project: https://jupyter.org/

# 8.Ekler

# Şekil: Random Forest Algoritmasına Göre En Önemli 5 Özellik

Yukarıdaki görselde, Random Forest modelinin siparişin başarılı (Shipped) olup olmayacağını tahmin ederken en fazla bilgi kazancı sağladığı 5 özellik yer almaktadır. Modelin karar sürecinde en yüksek ağırlığa sahip değişkenin qty (sipariş adedi) olduğu görülmektedir. Bu, sipariş miktarının doğrudan sipariş durumu üzerinde belirleyici bir etkisi olduğunu göstermektedir. Ardından gelen courier\_status\_Shipped ve courier\_status\_Unshipped gibi lojistikle ilgili değişkenler de modelin tahmin gücünü artıran önemli faktörler arasındadır. amount (sipariş tutarı) ve ship-postal-code gibi diğer sayısal değişkenler ise daha düşük ağırlıklara sahip olmakla birlikte tahmin sürecine katkı sağlamaktadır. Bu analiz, işletmenin hangi değişkenleri operasyonel olarak daha fazla göz önünde bulundurması gerektiğine dair içgörü sunmaktadır.  
  
  
  
  
Aşağıda yer alan confusion matrix, geliştirilen Random Forest modelinin test veri seti üzerindeki tahmin performansını detaylı olarak göstermektedir. Görseldeki değerler, modelin 21.944 gönderilen siparişi doğru bir şekilde sınıflandırdığını (True Positive), 3.822 iptal edilen siparişi ise doğru bir şekilde iptal olarak tahmin ettiğini (True Negative) ortaya koymaktadır. Sadece 24 gönderilen sipariş yanlışlıkla iptal (False Negative), 5 iptal edilen sipariş ise yanlışlıkla gönderildi (False Positive) olarak tahmin edilmiştir. Bu sonuçlar, modelin hem azınlık hem de çoğunluk sınıflarını yüksek başarıyla ayırt edebildiğini ve dengesiz veri yapısına rağmen güçlü genelleme kabiliyeti sunduğunu göstermektedir. Confusion matrix aynı zamanda hataların yönünü analiz ederek modelin hangi durumlarda iyileştirilebileceğine dair değerli içgörüler sunar.  
  
Şekil: Confusion Matrix - Model Sınıflandırma Başarımı  
  
  
  
  
  
  
  
  
  
  
  
Bu görsel, müşterilerin kargo hizmeti tercihlerinin siparişin başarıyla tamamlanma olasılığı üzerindeki etkisini analiz etmektedir. Görüldüğü üzere, standart kargo hizmeti seçmeyen müşterilerde gönderim oranı daha yüksekken, standart kargo tercihi yapan siparişlerde bu oran bir miktar daha düşmektedir. Bu durum, lojistik süreçlerde yaşanan gecikmelerin ya da operasyonel farklılıkların siparişin tamamlanma durumunu etkileyebileceğine işaret etmektedir.İşletme açısından bu bulgu, standart kargo hizmeti ile gönderilen siparişlerin daha yakından takip edilmesi gerektiğini ve alternatif teslimat seçeneklerinin değerlendirilebileceğini göstermektedir. Ayrıca, gönderim oranlarındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için daha ileri analizler yapılması önerilir.  
  
  
Şekil: Standart Kargo Seçiminin Sipariş Başarısına Etkisi  
  
  
  
  
Şekil: Random Forest Tahmin Olasılık Dağılımı  
  
  
Bu grafik, Random Forest modelinin test verisi üzerinde her bir sipariş için "gönderilme" sınıfına (status = 1) ait olasılık tahminlerini göstermektedir. Gözlemler büyük oranda 0 veya 1 olasılık değerlerine yakın kümelenmiştir. Bu durum, modelin karar verirken yüksek güvenle (deterministik şekilde) sınıflandırma yaptığını göstermektedir.

Modelin, siparişleri yüksek bir ayrım gücüyle sınıflandırabildiği, kararsız aralıkta (örneğin 0.4–0.6) çok az sayıda tahmin yaptığı görülmektedir. Bu, modelin hem ayrıştırıcı özelliğinin güçlü olduğunu hem de sınıflar arasındaki sınırları net şekilde belirleyebildiğini göstermektedir. Dolayısıyla, bu dağılım hem yüksek doğruluk oranını desteklemekte hem de modelin tahminlerine güven duyulabileceğini göstermektedir.

Notlar:  
Ek olarak, modelin overfitting (aşırı öğrenme) riski göz önüne alınarak ikinci bir model olarak Logistic Regression algoritması da uygulanmıştır. Logistic Regression, daha yalın bir yapıya sahip olmasına rağmen benzer başarı oranları sunmuş ve sınıflar arasındaki ayrımı yüksek doğrulukla gerçekleştirmiştir. Modelin doğruluk oranı (Accuracy) %99.87, F1 skoru ise 0.9993 olarak hesaplanmıştır. Confusion matrix çıktısına göre 3.820 doğru negatif, 21.943 doğru pozitif; sadece 7 yanlış pozitif ve 25 yanlış negatif tahmin yapılmıştır. Bu sonuçlar Logistic Regression modelinin genelleme kabiliyetinin yüksek olduğunu ve Random Forest ile benzer doğruluk seviyelerine ulaşabildiğini göstermektedir. Bu karşılaştırmalı yaklaşım, farklı model yapılarını değerlendirerek daha güvenilir bir tahmin sistemi oluşturulmasına katkı sağlamıştır.