

Profit Tahmini - SampleSuperstore

VB_DS Veri Bilimi Projesi

Problem ve Hedef

- Hedef degisen: Profit (kar)
- Gorev: regresyon tahmini
- Neden: kar tahmini is planlama ve kampanya kararlarini destekler

Veri Seti

- Dosya: data/raw/SampleSuperstore.csv
- Temel kolonlar: Sales, Profit, Discount, Quantity
- Kategorikler: Category, Sub-Category, Segment, Region, State, City, Ship Mode
- Tarih varsa: Order Date, Ship Date

Temizleme ve Ozellik Muhendisligi

- Sayisal bos degerler median ile dolduruldu
- Kategorikler mod ile dolduruldu; string + trim uygulandi
- Tarih varsa: order_month, order_dayofweek, shipping_delay
- Tarih yoksa: sales_per_item, discounted_sales, is_high_discount
- Profit hedefi icin profit_margin drop edildi (leakage onlemi)

Modelleme Akisi

- Pipeline: OneHotEncoder + StandardScaler
- Modeller: LinearRegression (baseline), RandomForestRegressor
- Profit negatif olabilir: hedefe log1p donusumu (gerekirse shift)
- Train/test split: random_state=42

Model Sonuçları

- Full LR -> MAE 94.83, RMSE 232.56, R² -0.115
- Full RF -> MAE 42.15, RMSE 156.90, R² 0.492
- No-Geo LR -> MAE 74.62, RMSE 211.84, R² 0.074
- No-Geo RF -> MAE 25.98, RMSE 116.89, R² 0.718

Feature Importance (Top 6)

- num__Sales: 0.212
- num__sales_per_item: 0.178
- num__discounted_sales: 0.172
- num__Discount: 0.045
- cat__City_Lancaster: 0.042
- num__Postal Code: 0.038
- En etkili faktorler: Sales, sales_per_item, discounted_sales
- İndirim etkisi belirgin: Discount,

Sonuc ve Kısa Özeti

- Linear Regression zayıf kaldığı için R² negatif çıktı
- Random Forest ile MAE düşü ve R² 0.49 seviyesine geldi
- No-Geo RF ile R² 0.718 seviyesine çıktı
- Karlilik ilişkileri doğrusal değil; indirim ve kategori etkisi kritik

Sinirlamalar ve Ilteri Isler

- City/Postal Code overfit riski tasiyabilir
- drop_geo testi geo kolonların gurultu olabileceğini gösterdi
- Hiperparametre araması performansı artırabilir
- Kategorik sadeleştirme daha genellenebilir sonuc verebilir

Soru - Cevap

- R² neden negatif? -> Lineer model yetersiz kaldı
- Neden RF? -> Dogrusal olmayan ilişkileri yakalıyor
- profit_margin neden yok? -> Leakage önlemek için çıkarıldı