

# Profit Tahmini - SampleSuperstore

VB\_DS Veri Bilimi Projesi

# Problem ve Hedef

- Hedef degisken: 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 Sonuclari

- 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 faktörler: Sales, sales\_per\_item, discounted\_sales
- Indirim etkisi belirgin: Discount,

# Sonuc ve Kisa Ozet

- Linear Regression zayif kaldigi icin  $R^2$  negatif cikti
- Random Forest ile MAE dustu ve  $R^2$  0.49 seviyesine geldi
- No-Geo RF ile  $R^2$  0.718 seviyesine cikti
- Karlilik iliskileri dogrusal degil; indirim ve kategori etkisi kritik

# Sinirlamalar ve Ileri Isler

- City/Postal Code overfit riski tasiyabilir
- drop\_geo testi geo kolonlarin gurultu olabilecegini gosterdi
- Hiperparametre aramasi performansi artirabilir
- Kategorik sadelestirme daha genellenebilir sonuc verebilir

# Soru - Cevap

- $R^2$  neden negatif? -> Lineer model yetersiz kaldı
- Neden RF? -> Doğrusal olmayan ilişkileri yakalıyor
- profit\_margin neden yok? -> Leakage önlemek için çıkarıldı