DOĞUŞ DATATHON

DOĞUŞ OTOMOTİV – FİNAL SUNUMU

Anıl Öztürk

ANIL ÖZTÜRK

- İstanbul Teknik Üniversitesi - Bilgisayar Mühendisliği, Yüksek Lisans
- Makine Öğrenmesi Mühendisi

PROBLEM

Aktif ruhsat sahipliği bulunan bir müşterinin gelecek 3 ay içerisinde satış dosyası açtırma olasılığı nedir?

- Tahmin yapılan dönemde aktif ruhsatı olmayan kullanıcı gözardı edilmeli
- Satış dosyasının adedi önemli değil
- Müşteri başına bir adet tahmin gerekli
- Tahminler olasılık formatında olmalı

LİTERATÜR

MEVCUT YÖNTEMLER

1. ÖZELLİK ÇIKARIMI

- Unsupervised yöntemler ile müşterileri gruplama amaçlı özellik çıkarımı
- Her bir kullanıcının geçmiş verileri üzerinden istatistiksel özellik çıkarımı (min, max, mean, median, kurtosis, skewness vb.)
- Unsupervised yöntemler ile geçmiş verilerin istatistikleri üzerinden gruplama amaçlı özellik çıkarımı

2. TAHMİN MODELİ TÜRÜ

- Satış miktarını tahmin eden regresyon modelleri
- Satış zamanını tahmin eden regresyon modelleri
- Satış durumunu tahmin eden sınıflandırma modelleri

3. TAHMİN MODELİ MİMARİSİ

- İstatistiksel modelleme
- Gelişmiş karar ağacı algoritmaları
- Ardışıl yapay sinir ağları (RNN, GRU, LSTM)

- Daskalova, Nina & Dragiev, Dragoslav. (2013). Using Classification Methods to Predict Market Demand for Products with Short Sales History.
- https://towardsdatascience.com/predicting-sales-611cb5a252de
- Predicting Future Sales of Retail Products using Machine Learning (https://arxiv.org/pdf/2008.07779.pdf)

ÇÖZÜM AŞAMALARI

VERİ TEMİZLİĞİ

Verideki anomalilerin giderilmesi, formatın standardize edilmesi

ÖZELLİK ÇIKARIMI

Kullanıcıların ve araçlarının geçmişe dayalı verilerine dayanarak yeni sentez verilerin oluşturulması

TAHMİN MODELİ SEÇİMİ

Model mimarisine, hyper-parameter setlerine ve eğitim sürecine karar verilmesi

ÇAPRAZ DOĞRULAMA

Modelin güvenilirliğinin test edilebilmesi adına değişken eğitim-test setleri üzerindeki performanslarının değerlendirilmesi

VERİ TEMİZLİĞİ

MÜŞTERİ BİLGİLERİ

Tek bir müşteri için farklı;

- Cinsiyet
- Şehir
- Meslek
- Medeni Durum

→ **Mod**

RUHSAT BİLGİLERİ

Henüz sonlanmamış ruhsatlar → **2050**

SATIŞ DOSYALARI

Satış dosyasındaki her araç için ayrı bir satır var

- **Satış dosyası özelinde içerdiği araç sayısı**
- **Her satış dosyasına ait tek bir satır**
- **Satış dosyaları sadece tarihler ve içerdikleri araç sayısı ile saklandı**

VERİ TEMİZLİĞİ

ARAÇ DETAYLARI

```
vehicle_df.replace(
    {"Kurşunsuz benzin": "Benzin",
     "Kurşunsuz Benzin": "Benzin",
     "?Benzin": "Benzin",
    "Seçiniz": np.nan,
     "Benzin/Hybrid": "Hibrit",
    "Benzin/Hibrit": "Hibrit",
     "Dizel/Hibrit": "Hibrit",
    "Diesel": "Dizel"},
    inplace=True)
```

```
vehicle_df.replace(
    {
    "Otomaik": "Otomatik",
     "?automatisch": "Otomatik",

    "Otomatik (DSG Şanzuman)": "Otomatik",
     "Otomatik(DSG)": "Otomatik",
    "Otomatik (DSG)": "Otomatik",

    "Düz": "Manuel",
    "manuel": "Manuel",
     "?Handschaltung": "Manuel",
    "Mekanik": "Manuel",

    "Sürekli değişken": "CVT",
    "Sürekli Değişken": "CVT",
    "Otomatik Değişken": "CVT",
    },
    inplace=True)
```

Belirsiz yakıt ve şanzıman tipi → **Unknown**

SERVİS BİLGİLERİ

```
servis_df["IS_MAINTENANCE"].replace(
    {
    1: "Bakım",
    0: "Servis",
    },
    inplace=True)
```

ÖZELLİK ÇIKARIMI

- **Her işlem için ayrı ‘CUSTOMER_ID’**

Bütün işlemler unique ‘BASE_CUSTOMER_ID’ ile eşlendi

- **Verilerde geçmiş-tahmin dönemi ayrımının oluşturulması**

Ocak	Şubat	Mart	Nisan	Mayıs	Haziran	Temmuz	Ağustos	Eylül	
Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin				1. Pencere
Geçmiş	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin			2. Pencere
Geçmiş	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin		3. Pencere
									...

Dönem penceresinden çıkarılmış özellikler

1 veya **0**

ÖZELLİK ÇIKARIMI

- **3 AY İÇİNDE SATIŞ DOSYASI AÇTIRMA DURUMU:** {0,1}
- **Aktif ruhsatlı araç sayısı**
- **Pasif ruhsatlı araç sayısı**
- **Araca sahip olunduğundan beri geçen zaman**
- **Aracın trafiğe çıktığından beri geçen zaman**
- **Aracın markası**
- **Aracın modeli**
- **Aracın motor yakıt tipi**
- **Aracın şanzıman tipi**
- **Bu zamana kadar Doğuş'tan aldığı toplam araç sayısı**
- **Mevcut aracı Doğuş'tan alıp almadığı**

→ **Ruhsat bazlı istatistikler**

→ **Araç ve sahiplik durumu detayları**

→ **Müşteri ve aracının Doğuş ile ilişkisi**

ÖZELLİK ÇIKARIMI

- **Son açtırılan satış dosyasından beri geçen zaman**
- **Son açtırılan satış dosyasının ayı**
- **Bu zamana kadar açtırdığı toplam satış dosyası sayısı**
- **Şimdiye kadar açtırdığı satış dosyalarının içerdiği ortalama araç sayısı**
- **Ortalama satış dosyası periyodu**

→ **Satış dosyası istatistikleri**

- **Araç için şimdiye kadar harcadığı toplam servis bedeli**
- **Araç için şimdiye kadar servise gitme sayısı**
- **Aracın son servis gördüğünden beri geçen süre**
- **Araç için şimdiye kadar harcadığı toplam bakım bedeli**
- **Araç için şimdiye kadar bakıma gitme sayısı**
- **Aracın son bakıma girdiğinden beri geçen süre**

→ **Aracın bakım ve servis istatistikleri**

ÖZELLİK ÇIKARIMI

Birden Fazla Aracı Olan Müşteriler

- Araç özelliklerinin mod'u ve medyanı
- Son alınan aracın özellikleri

ÖZELLİK ÇIKARIMI

- **Cinsiyet**
- **Şehir**
- **Meslek**
- **Medeni durum**
- **Yaş** *(Her dönem için mevcut yıla göre dinamik hesaplanır)*

→ **Müşteri bilgileri**

- **USD/TRY aylık fark oranı**
- **USD/TRY aylık fark**
- **TÜFE aylık fark oranı**
- **TÜFE aylık fark**

→ **Harici parite ve yüzde bilgileri**

ÖZELLİK ÇIKARIMI

BACKLOG

- **Satış dosyası tarihleri üzerinden zaman-serisi oto-korelasyon özellikleri**
- **CORONA:** {0,1}
- **Merkez Bankası faiz bilgileri**

TAHMİN MODELİ

Problem Tipi

Satış dosyası zamanı tahmini

~3.7 ay ortalama sapma

Satış dosyası durumu tahmini

~83% AUC

TAHMİN MODELİ

Model Mimarisi

One-Hot Encoding

- SVM
- Logistic Regression
- XGBoost

Unique Categorical Encoding

- CatBoost 
- LightGBM
- Keras - TensorFlow - NN

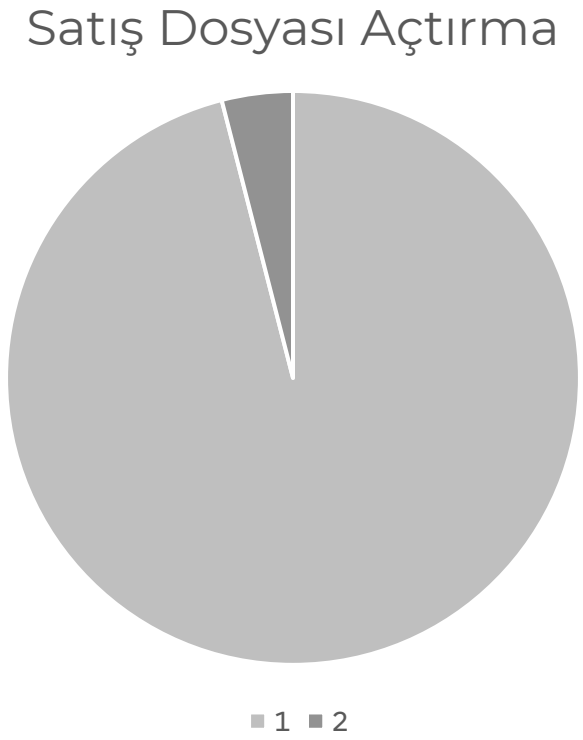
Stacked-Ensemble

- Logistic Regression Head
- Random Forest Head

TAHMİN MODELİ

Sınıf Dengesizliği

- Class Weighting
- Square-Root Class Weighting
- Random Undersampling
- Random Oversampling
- Focal Loss



TAHMİN MODELİ

Hyper-Parameter Tuning

- **Optuna**
- **AUC maksimizasyonu**
- Depth
- Random Strength
- Bagging Temperature
- Learning Rate
- Border Count

ÇAPRAZ DOĞRULAMA

	Ocak	Şubat	Mart	Nisan	Mayıs	Haziran	Temmuz	Ağustos	Eylül
Eğitim	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin			
	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin		
Test	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin	

- Group KFold (müşteri bazında gruplama)
- Time Series Splitting with gaps

TIME SERIES SPLITTING WITH GAPS

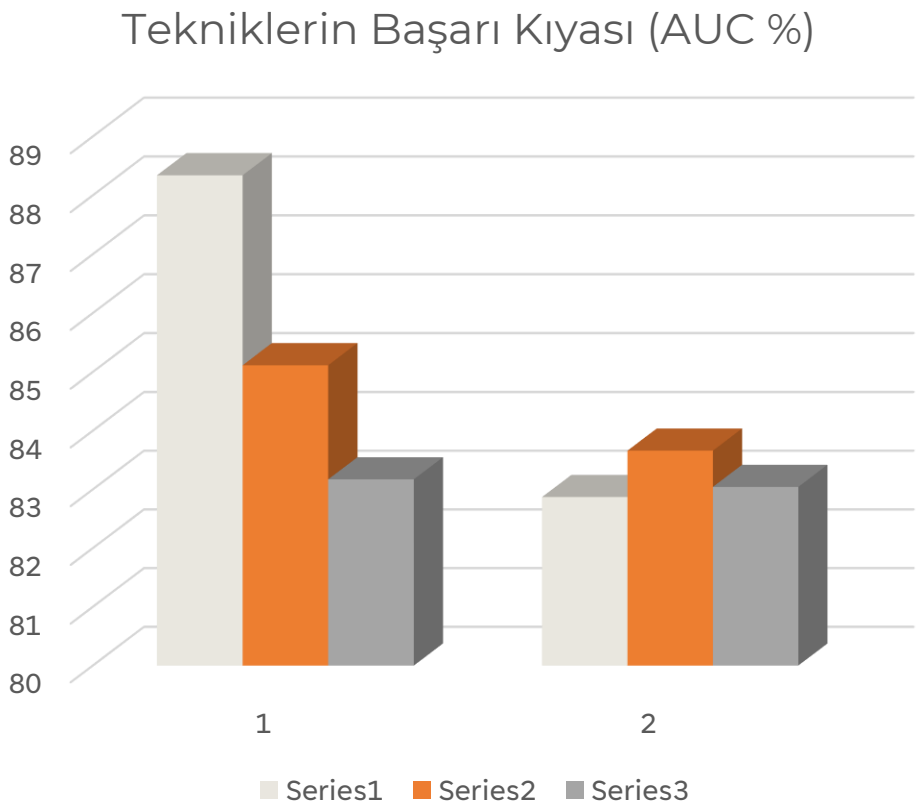
FOLD 1

Ocak	Şubat	Mart	Nisan	Mayıs	Haziran	Temmuz	Ağustos	Eylül	
Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin				1. Pencere
Geçmiş	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin	2. Pencere

FOLD 2

Şubat	Mart	Nisan	Mayıs	Haziran	Temmuz	Ağustos	Eylül	Ekim	
Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin				1. Pencere
Geçmiş	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Geçmiş	Tahmin	Tahmin	Tahmin	2. Pencere

ÇAPRAZ DOĞRULAMA



TAHMİN AŞAMASI

- Çapraz doğrulama ile model eğitimi
- Her fold modeli ile tahmin
- Modellerin tahminlerin ortalamasının alınması

TEŞEKKÜRLER

Anıl Öztürk

anilozturk96@gmail.com