FET445 Veri Madenciliği

UÇUŞ GECİKME TAHMİNİ

GRUP: TURBULENCE

Youtube link:

https://www.youtube.com/watch?v=obqgXc4Q7BA

Tarih:21.12.2025

UÇUŞ GECİKME TAHMİNİ

Problemin Açıklaması:

Havacılık sektöründe rötarlar, hem operasyonel maliyetleri artırıyor hem de yolcu memnuniyetini düşürüyor. Bizim amacımız; uçuş gerçekleşmeden önce, mevcut verileri sınıflandırma yöntemiyle uçağın rötar yapıp yapmayacağını önceden tahmin etmek.

Veri Seti Açıklaması:

Linki: https://www.kaggle.com/datasets/usdot/flight-delays

Veri Seti: ABD Ulaştırma Bakanlığı'na ait (US DOT) gerçek "Airline On-Time Performance" veri setini kullandık.

Yapı: Üç tablo flights.csv, airlines.csv, airports.csv birleştirilerek kullanılmıştır.

Boyut: 5.8 milyon satır ve 40 sütun.

```
Flights tablosunun boyutu : (5819079, 31)
Airports tablosunun boyutu : (322, 7)
Airlines tablosunun boyutu : (14, 2)
```

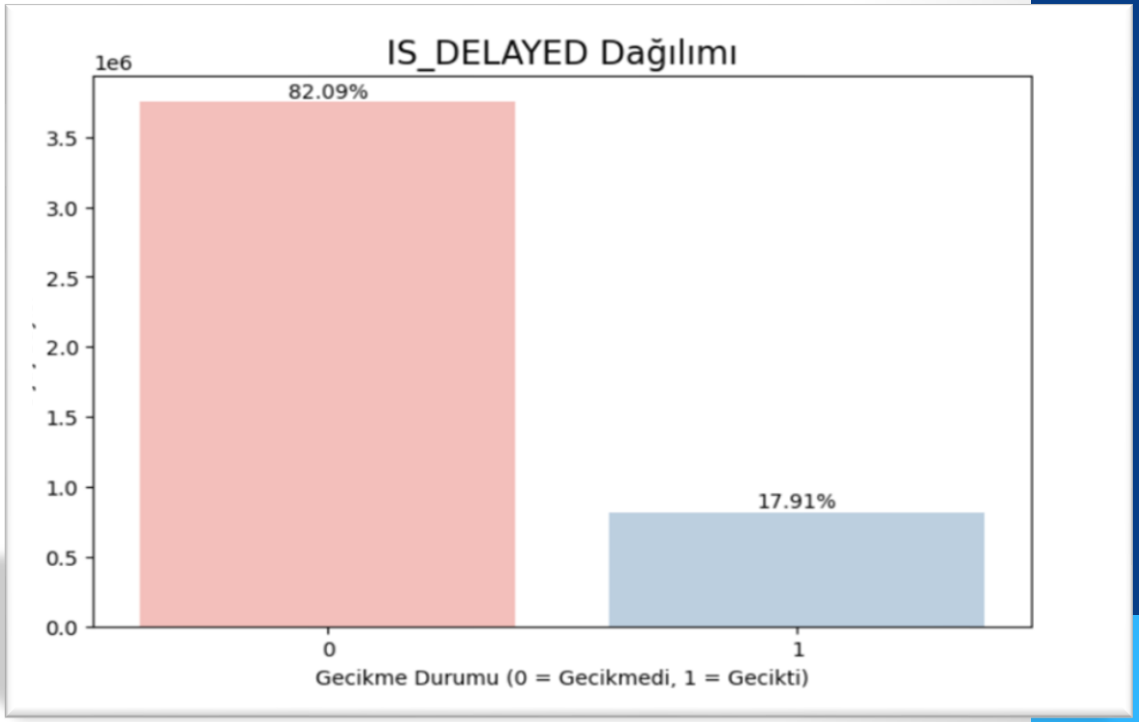
Size, özellik tipi

```
]:  1  df_flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5819079 entries, 0 to 5819078
Data columns (total 31 columns):
 #   Column               Dtype
---  ------               -----
 0   YEAR                 int64
 1   MONTH                int64
 2   DAY                  int64
 3   DAY_OF_WEEK          int64
 4   AIRLINE              object
 5   FLIGHT_NUMBER        int64
 6   TAIL_NUMBER          object
 7   ORIGIN_AIRPORT       object
 8   DESTINATION_AIRPORT  object
 9   SCHEDULED_DEPARTURE  int64
 10  DEPARTURE_TIME       float64
 11  DEPARTURE_DELAY      float64
 12  TAXI_OUT             float64
 13  WHEELS_OFF           float64
 14  SCHEDULED_TIME       float64
 15  ELAPSED_TIME         float64
 16  AIR_TIME             float64
 17  DISTANCE             int64
 18  WHEELS_ON            float64
 19  TAXI_IN              float64
 20  SCHEDULED_ARRIVAL    int64
 21  ARRIVAL_TIME         float64
 22  ARRIVAL_DELAY        float64
 23  DIVERTED             int64
 24  CANCELLED            int64
 25  CANCELLATION_REASON  object
 26  AIR_SYSTEM_DELAY     float64
 27  SECURITY_DELAY       float64
 28  AIRLINE_DELAY        float64
 29  LATE_AIRCRAFT_DELAY  float64
 30  WEATHER_DELAY        float64
dtypes: float64(16), int64(10), object(5)
memory usage: 1.3+ GB
```

Class Distribution (sınıflandırma ise)

Sınıf Dağılımı: Veri seti yüksek oranda dengesiz (imbalanced). Uçuşların çoğu zamanında kalkıyor, bu da modelin gecikmeleri öğrenmesini zorlaştırıyor.



Temel Kullanılan Teknikler

Zaman Segmentasyonu
Sürekli sayısal veriler, 5 farklı kategorik zaman dilimine (`TIME_OF_DAY`) dönüştürülmüştür. Bu işlem, modelin günün bölümlerine göre gecikme örüntülerini daha iyi öğrenmesini sağlar

Target Encoding:
Yüksek boyutlu değişkenler için ortalama gecikme oranı hesaplanarak sayısal bir değere dönüştürülmüştür. Bu sayede veri setinde yüzlerce yeni sütun oluşması engellenerek kategorik bilgi korunmuştur.

One-Hot Encoding:
Düşük sınıf sayısına sahip olan değişkenler, modelin bu kategoriler arasındaki farkı net bir şekilde öğrenebilmesi için binary vektörlere dönüştürülmüştür

Veri Entegrasyonu: Kullanılan tablolar birleştirilerek havayolu ve havalimanı özellikleri analize dahil edilmiştir.

```
silinecek_sutunlar = [
      #sızıntı yapanlar
    'DEPARTURE_TIME',
    'DEPARTURE_DELAY',
    'TAXI_OUT',
    'WHEELS_OFF',
    'ELAPSED_TIME',
    'AIR_TIME',
    'WHEELS_ON',
    'TAXI_IN',
    'ARRIVAL_TIME',
    'ARRIVAL_DELAY',
    #gereksizler hepsi aynı değer
    'YEAR',
    'COUNTRY_ORIGIN',
    'COUNTRY_DEST',
    #gürültüye sebebiyet verenler
    'TAIL_NUMBER',
    'FLIGHT_NUMBER',
    #gereksiz değişkenler
    'SCHEDULED_DEPARTURE',
    'SCHEDULED_HOUR',
    'SCHEDULED_ARRIVAL'
```

```
    #eklenen zenginleştirmeler
    'AIRLINE_NAME',
    'AIRPORT_ORIGIN',
    'CITY_ORIGIN',
    'STATE_ORIGIN',
    'LATITUDE_ORIGIN',
    'LONGITUDE_ORIGIN',
    'AIRPORT_DEST',
    'CITY_DEST',
    'STATE_DEST',
    'LATITUDE_DEST',
    'LONGITUDE_DEST',

]
```

Train -test split oranı nedir?

Veri madenciliği sürecinde modelin başarısını tarafsız bir şekilde ölçebilmek için veri seti **Training** ve **Test** olmak üzere iki ana bölüme ayrılmıştır.

5.8 milyon satırlık büyük bir veri setiyle çalışıldığı için, modelin daha fazla veriyle öğrenmesini sağlamak ve yüksek varyans riskini azaltmak amacıyla **%80 Eğitim - %20 Test** stratejisi benimsenmiştir.

```
Flights tablosunun boyutu : (5819079, 31)
Airports tablosunun boyutu : (322, 7)
Airlines tablosunun boyutu : (14, 2)
Flights Train boyutu: (4655263, 31)
Flights Test boyutu: (1163816, 31)
```

```
Train Veri Boyutu: (4571137, 10)
Test Veri Boyutu: (1142871, 10)
```

Performance metrikler nelerdir?

Accuracy

Yapılan tahminlerden kaç tanesinin doğru olduğunu gösteren en temel metriktir

ROC AUC

Geciken ve gecikmeyen sınıfların birbirinden ayırt etme yeteneğini ölçer

F1 Score

Precision ve Recall değerlerinin harmonik ortalamasıdır

PR AUC

Modelin gecikmeleri ne kadar kaliteli tahmin ettiğini doğrudan gösterir

En İyi Model: RandomForestClassifier

Geliştiren: Zekeriya Deniz Uğurlu.

Yaklaşım: Bagging yöntemi kullanılarak çok sayıda karar ağacı ile aşırı öğrenme (overfitting) riski minimize edilmiştir.

Hiperparametre Optimizasyonu: `RandomizedSearchCV` yöntemi ve `ROC-AUC` optimizasyon metriği kullanılmıştır.

Sınıf Dengesizliği Yönetimi: Azınlık sınıfı (gecikmeler) olan "1" sınıfına 3 kat daha fazla ağırlık verilerek (`class_weight: {0:1, 1:3}`) modelin gecikmeleri yakalama hassasiyeti artırılmıştır.

Özellik Seti: PCA ve RFE gibi yöntemlerle gürültüden arındırılmış veri setleri üzerinde test edilmiştir.

Modellerin Karşılaştırıldığı Tablo

Grup üyelerinin en iyi ana modelleri:

Üye	Model	accuracy	F1 Score	precision	recall	ROC AUC	PR AUC	MAE
Emine Güneş	Gradient Boosting Classifier +CA	82.27	0.05	0.61	0.03	67.85	32.40	27.46
Zekeriya Deniz Uğurlu	RandomForestClassifier	81.70	0.41	0.48	0.35	74.78	43.56	18.30
Muhammed Mert Oruç	QDA+ANOVA F	79.12	0.14	0.26	0.09	61.49	24.35	20.89
Oğuzhan Özdemir	Extra Tree Classifier+ Selection	81.24	0.25	0.0015	0.75	68.08	31.85	18.01
Muhammet Enes İnal	Stacking + SelectFromModel	68.88	0.29	0.54	0.68	33.53	38.59	31.05

Sonuç ve Değerlendirme

Başarı Kriterleri: RandomForestClassifier; %80 doğruluk, 0.20 F1 ve 0.70 ROC AUC barajlarını aynı anda geçen tek model olmuştur.

Sınıf Dengesizliği Etkisi: Sadece doğruluğa (Accuracy) odaklanmanın yanıltıcı olduğu; örneğin Emine Güneş'in modelinin yüksek doğruluğa rağmen çok düşük F1 skoruna sahip olmasıyla kanıtlanmıştır.

Veri Sızıntısı: Sızıntı sütunların temizlenmesi, modelin gerçek dünya verileriyle çalışabilmesi için kritik bir adım olmuştur.

Son Söz: Modelimiz, yüksek boyutlu ve dengesiz bir veri setinde uçuş gecikmelerini ayırt edebilecek kararlı bir performans sergilemiştir.

Dinlediğiniz İçin Teşekkür Ederiz

Emine Güneş- 22040101036
Mert Oruç-22040101035
Enes İnal-22040101023
Deniz Uğurlu-22040101034
Oğuzhan Özdemir-22040101017