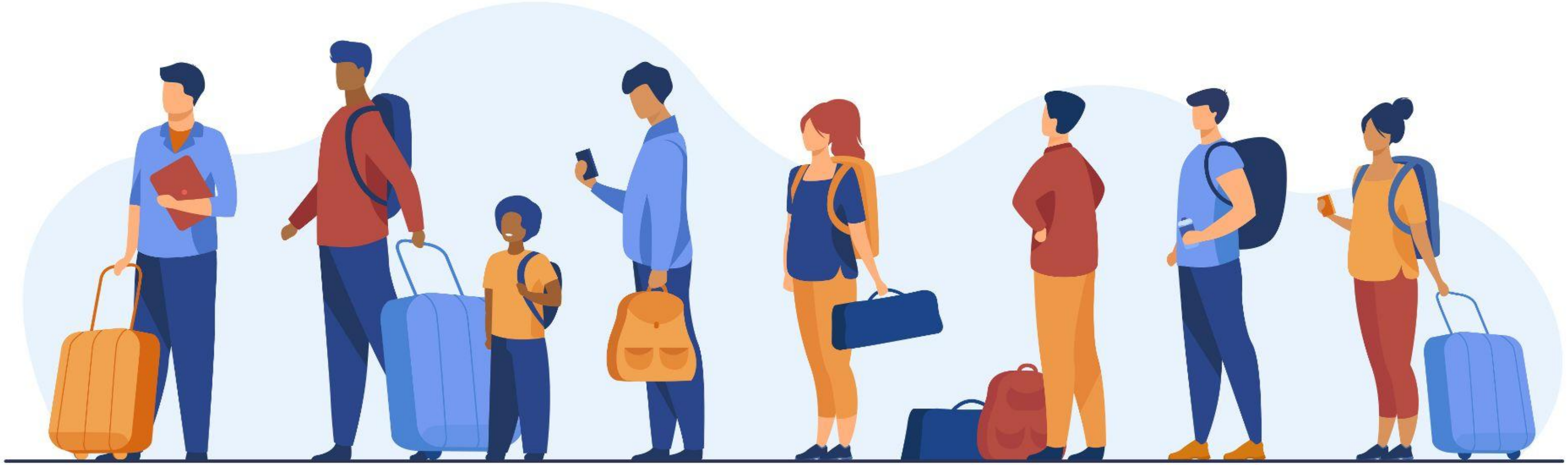


Prediction of Flight Delays

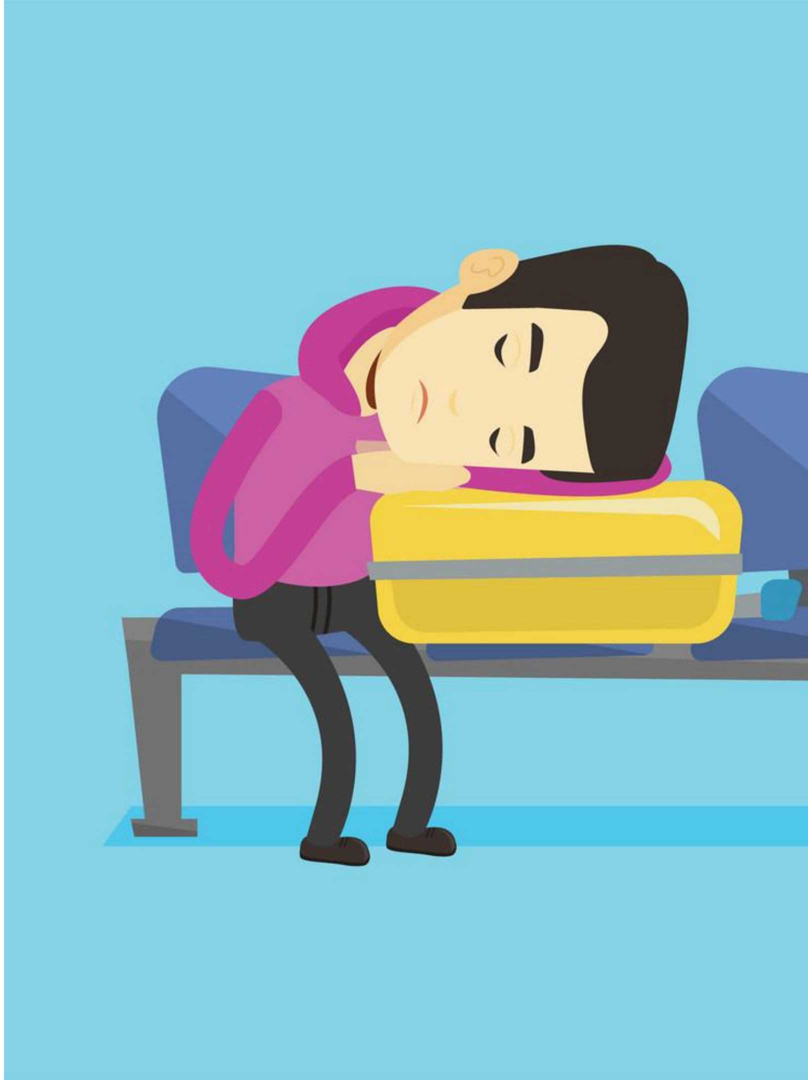
Felix, Ludmila, Chang-Ming

Figures Designed by Freepik

Introduction

No one likes flight delays:

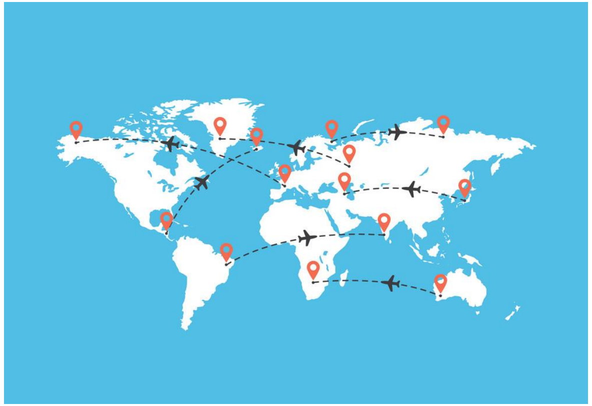
- Distressful waiting for passengers, missed transfer opportunities
- Extra costs for airlines
- Lower efficiency for airports (increased organizational effort)



Datasets

Records of international flights:

- Totally ~100k flights
- 3 years of data (Jan. 2016 - Dec. 2018) incl. time of dep. & arr., airports of dep. & arr., etc.
- Geographical data for airports
- Target: delay time



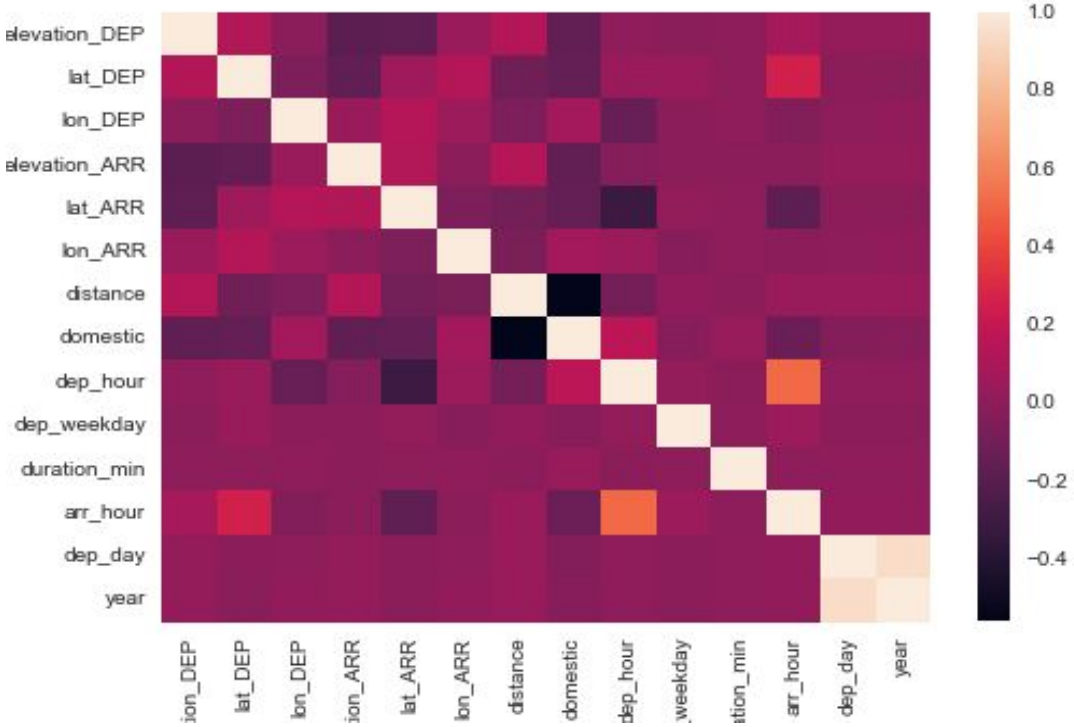
Flight data: https://zindi.africa/competitions/ai-tunisia-hack-5-predictive-analytics-challenge-2

Airport data: https://pypi.org/project/airportsdata/

Exploratory Data Analysis

Hypotheses: delay depends on:

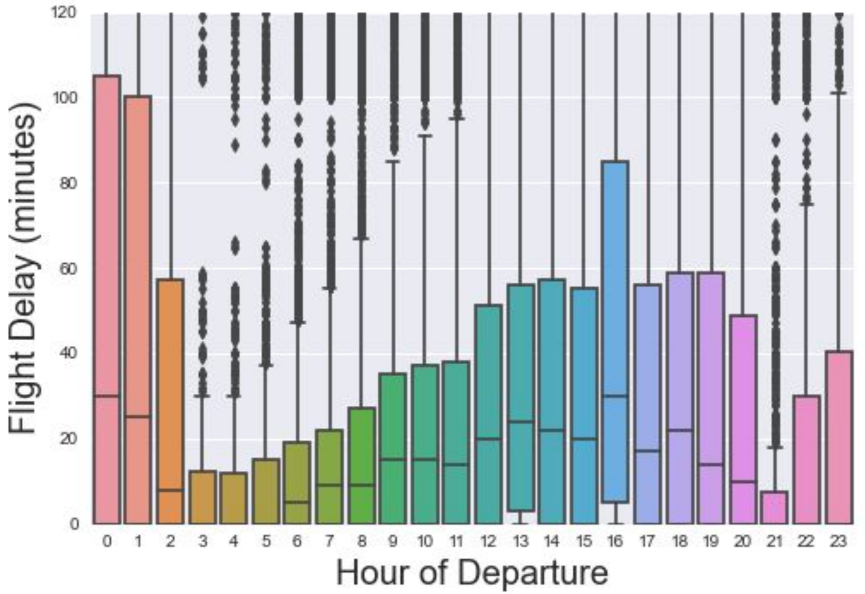
- Time of departure in a day
- Airport of departure & arrival
- Domestic or international travel



Delay vs. Time of Departure

Rationales:

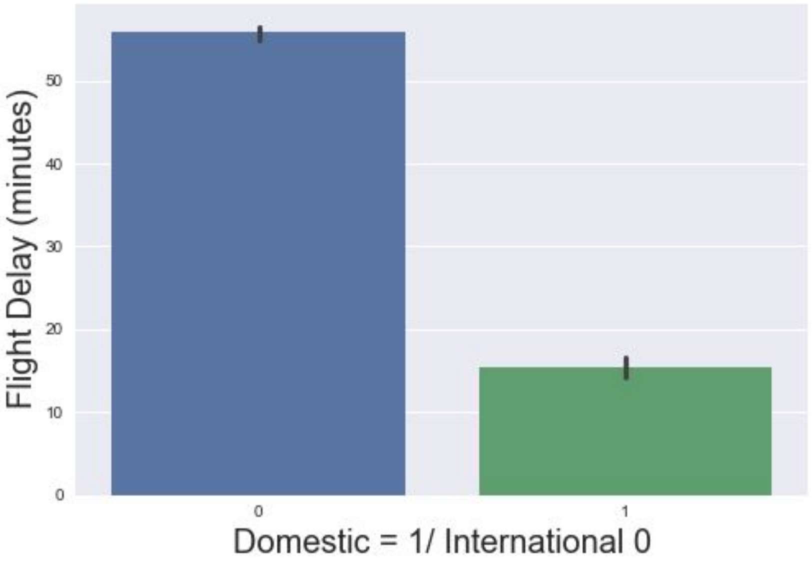
- Night flights not allowed
- Traffic accumulates from morning



Domestic or international flight

Rationales:

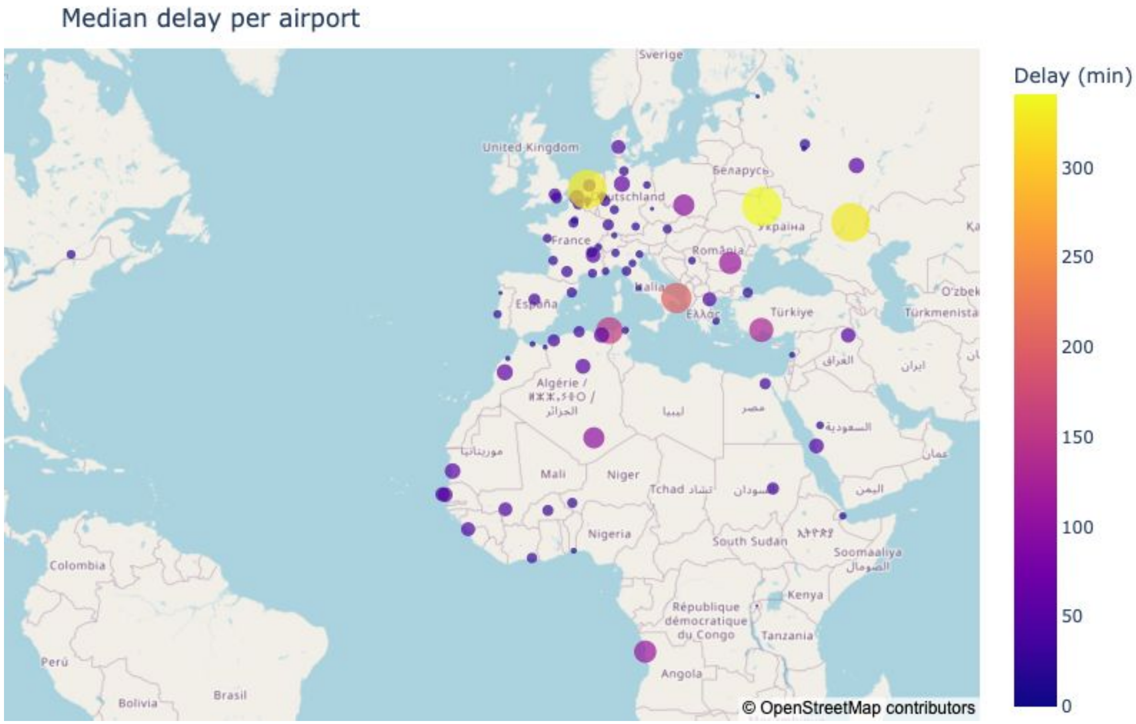
- More delays on international flights



Delay vs. Airport

Airport greatly influences flight delay

- Highest delays for Aiports in Russia, Ukraine and Netherlands (Rotterdam)



Modeling Delay

Models used:

- KNN
- XGBoost
- Linear Regression

Baseline Model & Metric

Baseline model:

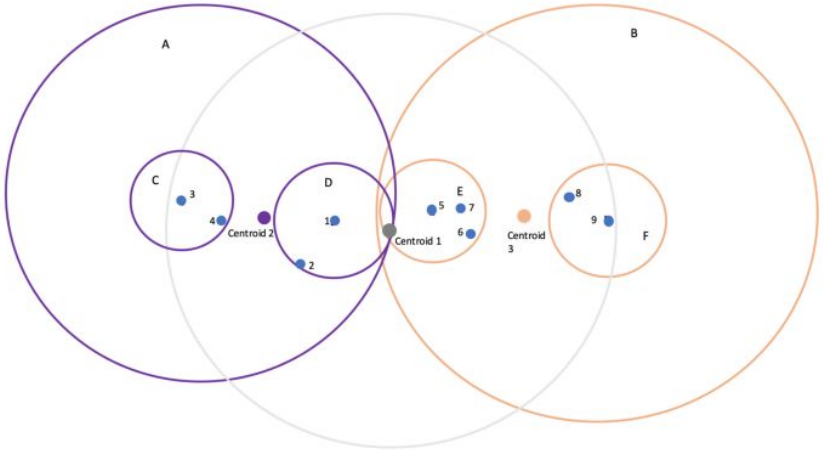
- Predicted delay = Median of delay

Evaluation metric:

- Root-mean-squared error: 106

KNN Machine Learning Model

- Following parameters were found using GridSearchCV:
  - ball tree algorithm
  - euclidean metric
  - 12 neighbors
  - uniform weights
- Very slow, calculation time >1h with a train size of 60%, RMSE 113, no optimization

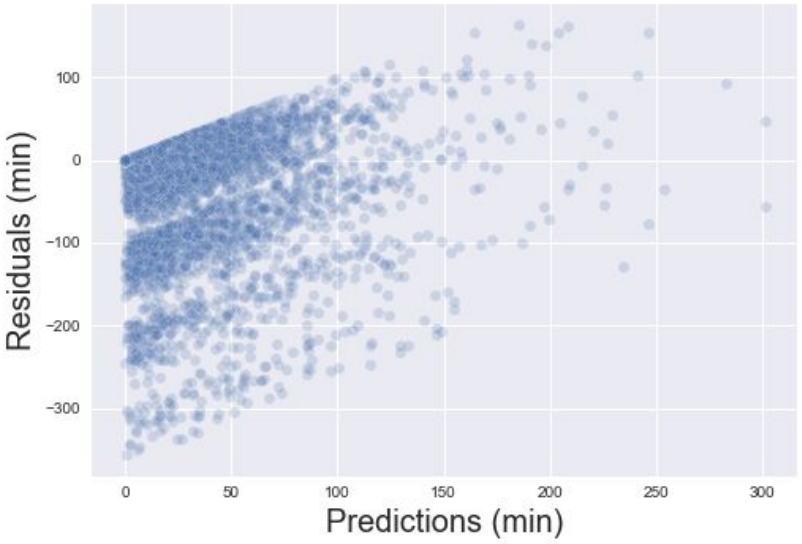


https://towardsdatascience.com/tree-algorithms-explained-ball-tree-algorithm-vs-kd-tree-vs-brute-force-9746debcd940

XGBoost

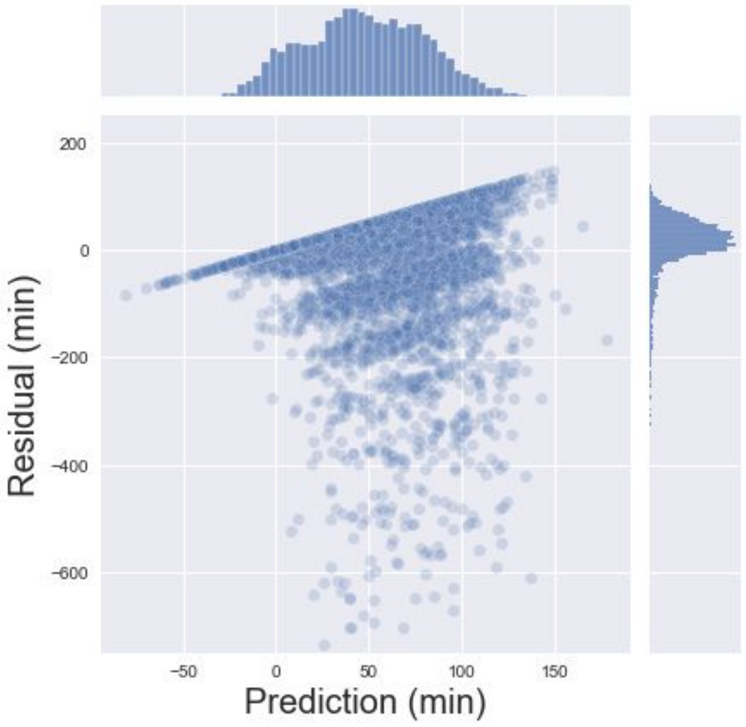
Selected to tackle overfitting of random forest model

- Regularization
  - max_depth
  - subsample
- Log transformation of target
- RMSE 93



Linear Regression

- Elastic net with polynomial features
- Grid search applied for different regularizations
- RMSE: 98 minutes



Predictions

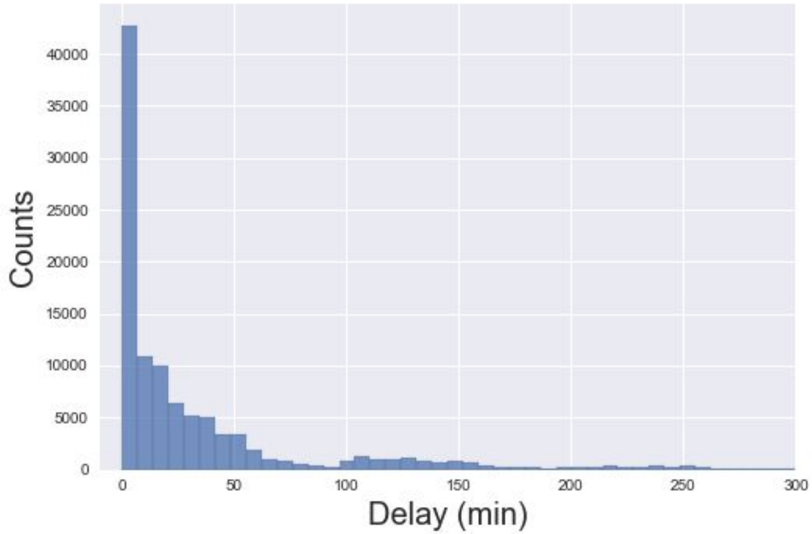
Best result:

- XGBoost
- RMSE = 93 minutes
- Top 130 in the leader board

Future Work

Challenges:

- Target distribution: what about negative delays?!
- Classification (no delay / delay), followed by regression
- Feature importance
- Additional features / data



Summary

We have:

- carried out EDA
- made reasonable predictions using different models
- found the directions for improvement in the future

Backup Slides

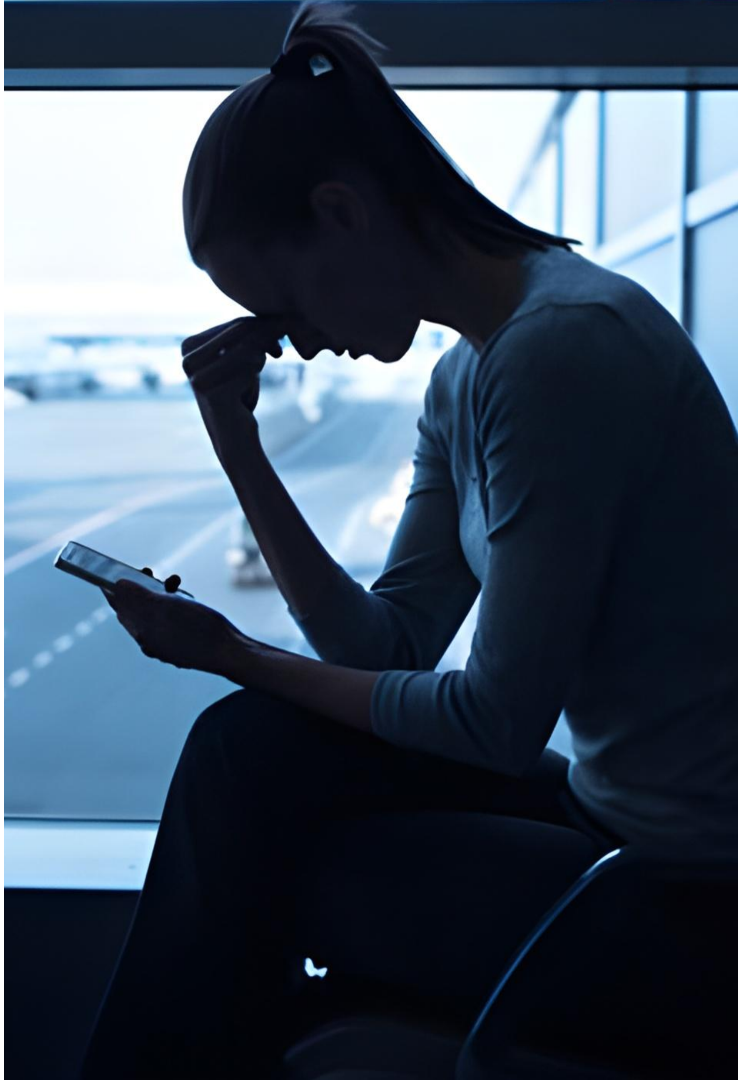
Prediction of Flight Delays

Felix, Ludmila, Chang-Ming

Introduction

No one likes flight delays:

- Distressful waiting for passengers
- Extra costs for airlines
- Lower efficiency for airports



Introduction

No one likes flight delays:

- Distressful waiting for passengers
- Extra costs for airlines
- Lower efficiency for airports

tmp

tmp:

- tmp.

B01
A19
B13
A26
A37
A40
A28
A34
A22
B09
A27
12:40
12:45
12:45
12:50
13:00
13:00
13:10
13:15
13:20
13:30

Modeling Delay Time

Overview of models used:

- Linear regression
- KNN
- Random forest