# Predicting Flight Delays

**Rainer Vana, Kaarel Kõomägi, Jaan Otter**
Institute of Computer Science, University of Tartu

UNIVERSITY OF TARTU
1632

## INTRODUCTION

The aviation industry relies heavily on understanding factors that affect flight outcomes, such as delays or punctuality. We analyzed a Kaggle dataset containing extensive information about flights, focusing on attributes available before a flight lands.

Our goals:

- Identify patterns and correlations between attributes and landing outcomes.
- Create a good model that somewhat accurately predicts flight delays.

## THE DATA

The dataset, sourced from Kaggle, contains flight data from 2009 to 2018 with over 6.3 million rows and 28 attributes per flight. We used the 2013 dataset.

It includes categorical features like airline operator, origin, and destination airports, and numerical features such as flight number and distance. Binary features indicate cancellations and diversions, while 19 timing attributes cover planned and actual times and delays.

## DATA PREPROCESSING

We performed extensive preprocessing on the dataset, which consists of 6,369,482 rows and 28 attributes. This process was time-consuming due to the dataset's size and complexity.

```
Dimensions of the dataset are: (6369482, 28)
```

Procedures we completed during data preprocessing:

- Standardized numerical values to Int64.
- Removed the unnecessary column 'Unnamed: 27'.
- Transformed time columns into HH:MM format.
- Converted 'CANCELLED' and 'DIVERTED' flags to boolean type.
- Split 'FL_DATE' into 'YEAR', 'MONTH', 'DAY', and 'DAY_OF_WEEK'.
- Resolved missing values, particularly for cancelled and diverted flights.
- Filled missing 'CRS_ELAPSED_TIME' values.
- Calculated 'ARR_DELAY', 'ACTUAL_ELAPSED_TIME', and 'AIR_TIME'.
- Added 'DEP_Rush_Hour' feature to identify rush hours.
- Generated summary statistics for numerical features.
- Saved cleaned DataFrames for machine learning algorithms.

## EXPLORATORY DATA ANALYSIS

Correlation Analysis

- DEP_DELAY and ARR_DELAY have high correlation.
- WEATHER_DELAY has low correlation to other factors.

Group-Level Analysis

- **Delay types**:

  There are a total of five different delay types. Carrier delays are among the biggest contributors, while security delays are minimal.

- **Airports**:

  Airports with more departures tend to have lower average delays, while those with fewer departures experience higher average delays.
  The average delay is quite consistent for airports that have frequent departures. (Less than 10 minutes)

## MODEL BUILDING

The dataset is really big, which means we can't train the models on the full dataset, so we sampled a subset of the whole dataset (frac=0.0015). We created the first model using Random Forest Regression (RFR) which yielded decent results. We also created three other models using Linear Regression, Gradient Boosting (GBM) and LightGBM to get an idea of which algorithm is the best.

Models have access to about half of the attributes during training, mainly consisting of the flight carrier, origin, destination and duration of the flight, planned departure, arriving and elapsed time, the date and the actual departure time.

## MODEL EVALUATION

$R^2\ Score$ using sample data.

| **Model** | $R^2\ Score$ |
|---|---|
| Random Forest Regression | 0.8163 |
| Linear Regression | -0.2033 |
| GBM | 0.6926 |
| LightGBM | 0.7844 |

Prediction success rate if flights are delayed (frac=0.0015)