Predicting flight delays

Team Members
Rainer Vana
Kaarel Kõomägi
Jaan Otter

Task 1. Setting up

GitHub: https://github.com/importb/predicting-flight-delays

Our GitHubs are:

importb : Rainer Vana
KaarelKoo : Kaarel Kõomägi
JAAN555 : Jaan Otter

Our dataset :
https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data.
We will be using data from 2013.

Task 2. Business understanding

Identifying our business goals

Project background

Our aim is to find correlation between certain attributes of flights that we are aware of prior to the flight landing, to predict how a flight lands. We work on a Kaggle dataset that previously has been used to study relationships between its features. Some machine learning methods have also been applied to this dataset (such as Logistic Regression, KNN and Random Forest).

Business goals

- Predict whether a flight will be delayed and for how long.
- Analyze the impact of departure / arrival airports on delays.
- Investigate common causes of delays and their average durations.

Business success criteria

Being able to correctly predict for most (>75 %) of the flights whether one of the binary attributes (whether the flight was cancelled or diverted) is true or false.

Assessing our situation

Inventory of resources

We have 3 rookie data scientists and data on over 6 million different flights with 28 attributes for each flight. We each work on our own hardware or by using cloud based hosting services such as Google Colaboratory.

Requirements, assumptions, and constraints

Our project lacks important limitations or constraints that are unique to our work.
We have the same hard time limits as every other project: we must complete the research and make a poster of our results and conclusions by the 9th of December 2024.

Risks and contingencies

Realistically the biggest risks include:

- Bad time management on our part.
- Random issues such as electric outage or networking / internet issues.
- The whole dataset is too large to work on within a reasonable time frame. A solution for that is that we start off by working on a smaller subset. Once a reasonable model is developed, we can attempt using it on the larger dataset.

Terminology

Our project, in general, lacks extremely specific terminology that would require definition.

Costs and benefits

Electricity costs, depending on the current cost in Estonia.
The time we all spend on it (approximately 30 hours per person).
We benefit by becoming more experienced in the field and getting a more intimate understanding of our project matter: flight delays.

Defining our data-mining goals

Data-mining goals

Developing a model that allows us to predict how a flight will go depending on previously known facts.
The culmination of our work will be a short and easily understandable visual representation of these facts: a poster that also explains some of the most important correlations.

Data-mining success criteria

We will see our project as a success if our developed model is able to predict for most flights, whether they will be delayed or not (>75 % accuracy) and to what extent (within 5 minutes of the actual delay.)

Task 3. Data understanding

Gathering Data

Outline Data Requirements

The primary objective of this project is to understand the factors influencing flight delays. So the data we need is departure times, arrival times, flight duration, airline and route information, delay types and durations.

Verifying Data Availability

Looking at the dataset 2013.csv, it can be noted that it has all the aforementioned fields needed and even some extra fields.

Define Selection Criteria

For the analysis, we will be focusing on flights that weren't cancelled or diverted, because these cases could introduce anomalies.

Describing Data

The dataset we have contains over 6.3 million rows, each representing a unique flight during the year of 2013. Each flight (row) has 28 attributes. There are four categorial features, 2 numerical features, 2 binary features, one timing feature for the date and the rest are all timing features: a grand total of 19 features denoting either certain times or delays.
The features are:

- **OP_CARRIER_FL_NUM** : Unique number for the flight. (Numerical feature, identification feature)
- **FL_DATE** : The date of the flight. (Timing feature)
- **DISTANCE** : Distance between the two airports (in miles). (Numerical feature)
- **OP_CARRIER** : Airline who's operating the flight. (Categorical feature)
- **ORIGIN** : Airport code where the flight begins. (Categorical feature)
- **DEST** : Airport code where the flight ends. (Categorical feature)
- **CRS_DEP_TIME** : Planned time for the flight's departure. (Timing feature)
- **DEP_TIME** : Actual time when the flight departed. (Timing feature)
- **DEP_DELAY** : Actual time when the flight departed difference from planned departure time. (Timing feature)
- **CRS_ARR_TIME** : Planned time for the flight's arrival. (Timing feature)
- **ARR_TIME** : Actual time of arrival. (Timing feature)
- **ARR_DELAY** : Difference between planned flight arrival time and actual time of arrival. (Timing feature)
- **CRS_ELAPSED_TIME** : How much time the flight was planned to take. (Timing feature)
- **ACTUAL_ELAPSED_TIME** : How long the flight actually lasted (AIR_TIME + TAXI_OFF + TAXI_ON | time spent in the air, time spent on taking off and landing). (Timing feature)
- **ARR_DELAY** : Difference between planned flight arrival time and actual time of arrival. (Timing feature)
- **AIR_TIME** : The time duration between wheels_off and wheels_on time. (Timing feature)

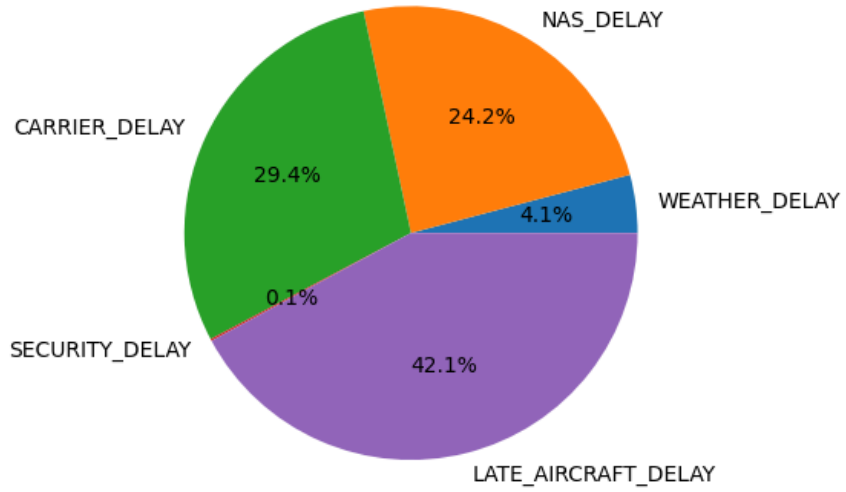
- **TAXI_IN** : The time duration elapsed between wheels-on and gate arrival at the destination airport. (Timing feature)
- **TAXI_OUT** : The time duration elapsed between departure from the origin airport gate and wheels off. (Timing feature)
- **WHEELS_OFF** : The time point that the aircraft's wheels leave the ground. (Timing feature)
- **WHEELS_ON** : The time point that the aircraft's wheels touch on the ground. (Timing feature)
- **WEATHER_DELAY** : Delay caused by weather. (Timing feature)
- **NAS_DELAY** : Delay caused by the National Air System. (Timing feature)
- **CARRIER_DELAY** : Delay caused by the airline. (Timing feature)
- **SECURITY_DELAY** : Delay caused by security. (Timing feature)
- **LATE_AIRCRAFT_DELAY** : Delay caused by aircraft. (Timing feature)
- **DIVERTED** : Flag for whether the flight was diverted or not. (Binary feature)
- **CANCELLED** : Flag for whether the flight was cancelled or not. (Binary feature)
- **CANCELLATION_CODE** : Reason why the flight was cancelled. (Categorical feature)

Exploring Data

Numerical feature's statistics

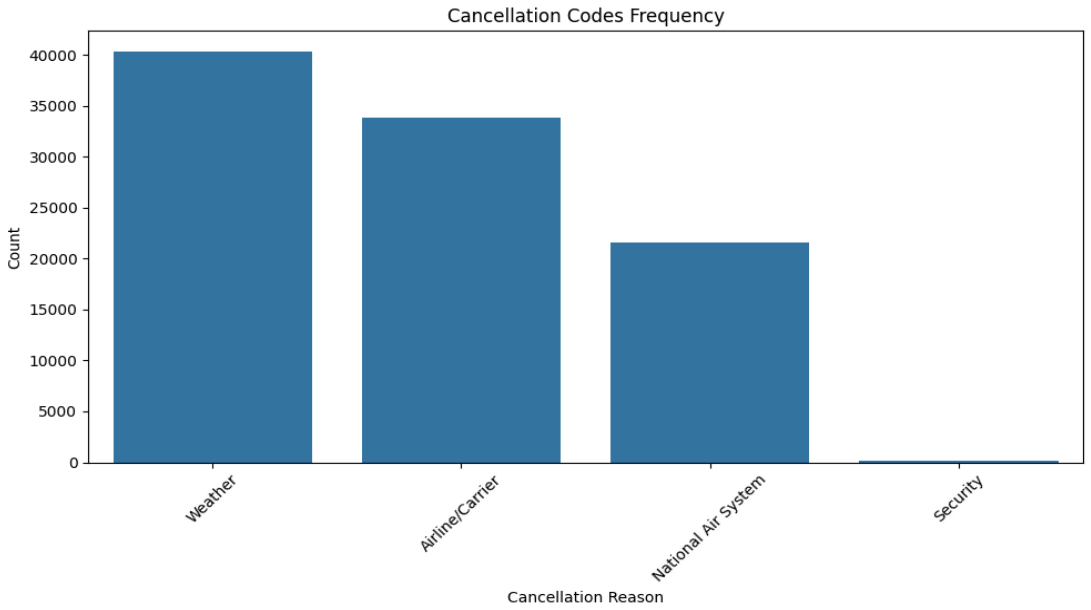
The average delay of all the flights (delay from planned lift off to actual lift off) is 6 minutes. The average planned length of the flights is 132 minutes and 42 seconds, while the average actual flight time is about 129 minutes and 20 seconds. The average flight distance is 762 miles.

The most common cause of flight delay is delay caused by the aircraft itself (it arriving too late to the starting airport, taking off too late because of gas refill, staff etc.).



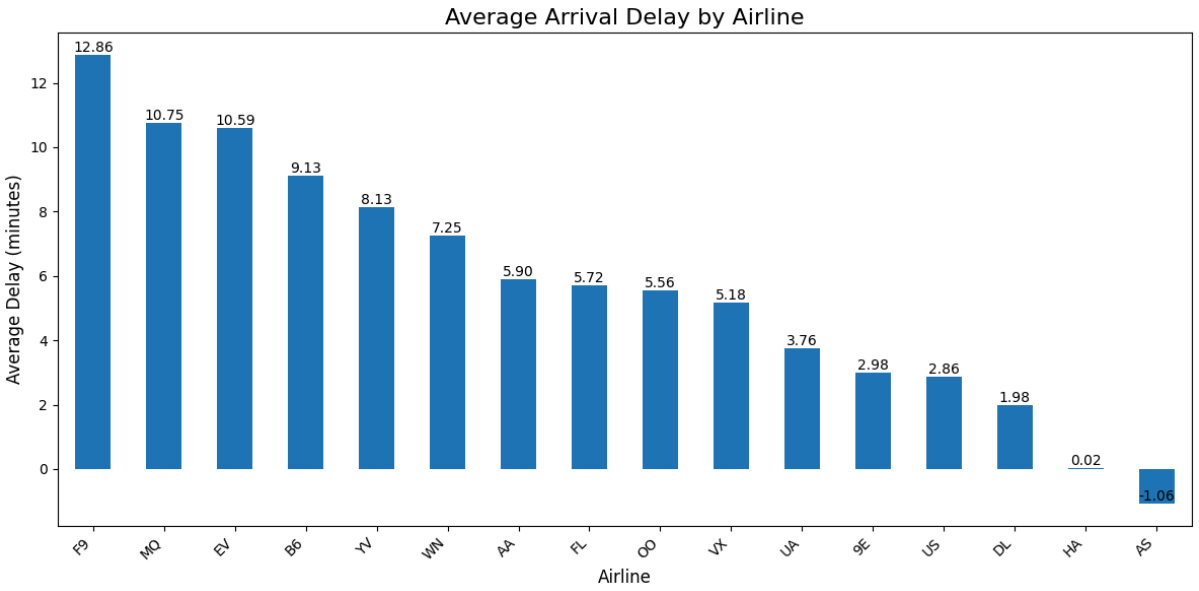
Categorical feature's statistics

The most common reason for flight cancellation is the weather, the actual statistics are here:



Correlation analysis

There are plenty of correlations in between all of the timing features: of course all of the delays will have an effect on the arrival and departure time.
More interestingly, there is a correlation between the average arrival delay and the airline:



Verifying data quality

Checking for invalid values

As far as we can tell, there are seemingly no “invalid” values, as most values are timing related, some such as the arrival time will of course be missing if the flight was cancelled. Some delays can actually have negative values, denoting that something happened earlier than planned (take off or arrival).

Data Consistency

Timestamps are formatted in the "Military Time" or "24-Hour Time Format", meaning that for example 1445 = 14:45, 10 = 00:10. Delays are formatted like regular numbers, denoting minutes, meaning that e.g ARR_DELAY = 83 means that the arrival was delayed by 83 minutes. Origin and destination are always 3 letter abbreviations, airline (OP_CARRIER) is a two letter abbreviation, the binary features CANCELLED and DIVERTED are truly binary so either a 0 or a 1 and the CANCELLATION_CODE is either A, B, C or D denoting one of the four delay reasons.

Task 4. Planning your project

Our project consists of several tasks:

- **Understanding the problem.**
  Our initial task is to understand what we are doing and define goals for the project. Each team member is expected to spend 4 hours on this task.

- **Data exploration and preparation.**
  In this task, we plan to inspect the data, clean it and apply feature engineering. Each team member is expected to spend 4 hours on this task.

- **Performing exploratory data analysis.**
  In this task, we will be summarising the main characteristics of relevant features and also using data visualization techniques. Each team member is expected to spend 7 hours on this task.

- **Building and testing models.**
  In this task, we plan to try out different machine learning models for predicting the flight delay. We will evaluate each model's performance and make conclusions. Each team member is expected to spend 9 hours on this task.

- **Reporting.**
  In this task, we plan to summarize our work and prepare a poster which will introduce our project. Each team member is expected to spend 6 hours on this task.

Methods and tools

For programming, we are going to use Python. We will be using Google Colaboratory and Visual Studio Code to work together for the project. We will be using various machine learning methods (such as Gradient Boosting, Random Forest, etc.) for a classification problem (whether a flight will be delayed or not) and some other methods (such as XGBoost, Decision Trees, etc.) for a regression problem (how long the delay will be). As much of this has already been performed on this dataset, we will be also experimenting with different parameters within the models to try to find the best model.