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

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 K-Means 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 1.7 million different flights with 15 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 flight_delays.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 1.7 million rows of flight information with each row having 16 columns. There are four numerical features (one of them also being identification feature), six categorical features, four timing features and 2 binary features.

The features are:

- **FlightID** : A unique identifier for each record. (Numerical feature, identification feature)
- **Airline** : Airline who's operating the flight. (Categorical feature)
- **FlightNumber** : Unique number for the flight. (Numerical feature)
- **Origin** : Airport code where the flight begins. (Categorical feature)
- **Destination** : Airport code where the flight ends. (Categorical feature)
- **ScheduledDeparture** : Planned date and time for the flight's departure. (Timing feature)
- **ActualDeparture** : Actual date and time when the flight departed. (Timing feature)
- **ScheduledArrival** : Planned date and time for the flight's arrival. (Timing feature)
- **ActualArrival** : Actual date and time when the flight arrived. (Timing feature)
- **DelayMinutes** : Difference in minutes between the scheduled and actual departure / arrival times. (Numerical feature)
- **DelayReason** : Cause of the delay. (Categorical feature)
- **Cancelled** : True / False based on if the flight was cancelled. (Binary feature)
- **Diverted** : True / False based on if the flight was diverted. (Binary feature)
- **AircraftType** : Model of the aircraft. (Categorical feature)
- **TailNumber** : Registration number of the aircraft. (Categorical feature)
- **Distance** : Distance (in miles) between the origin and destination airports. (Numerical feature)

Exploring Data

Numerical feature’s statistics

Doing some analysis on the first 50,000 rows in the dataset it can be determined that if a flight gets delayed the average delay is about 15 minutes. The average distance of a flight is approximately 1,500 miles.

Categorical feature’s statistics

The most common reason a flight is delayed is because of “Weather”. The majority of the flights are using Boeing 737.

Correlation analysis

When doing correlation analysis between AircraftType and DelayMinutes it can be noted that there really isn’t any correlation between them, meaning aircraft type may not be a major factor in delays.

Verifying data quality

Checking for invalid values

Analyzing the dataset for missing values, the only values that are missing are from “DelayReason”, which is expected. When looking for more suspicious values, it can be noted that the feature “DelayMinutes” sometimes has negative values, which means the flight departed earlier than expected.

Data Consistency

Timestamps are all formatted in the samy format, delay’s reasons are categorized in certain groups. Origin and destination are always 3 letter abbreviations. So overall the data is really consistent and doesn’t require modifying the values in any way.

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 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).