Flight Delays

Context

Business intelligence and data-driven decision making

It was not possible to build a predictive model using multiple linear regression techniques to offer insights into the impact of weather on departure delays.

Further work was performed to investigate whether decision trees and random forests would reveal information which could be used by the business. It found that flight departure delays were more likely based on the time of day a flight departed and the time of year, which was also evident from the exploratory analysis.

Domain knowledge and the business context

I was hired by Newark International Airport to investigate the effect of weather on flight departure delays.

The business believes that poor weather conditions are causing too many delays and want to invest in improving facilities, so that aircraft can take off in more types of weather. However, they do not fully understand how serious weather related delays are, and are not sure what type of weather they should be most concerned about. As part of investigating the effect of weather, other factors were explored to understand how important weather is in comparison to them.

For context, the Federal Aviation Administration (FAA), which is the transportation authority for civil aviation in the United States, consider a flight to be delayed when it is 15 minutes later than its scheduled arrival or departure time.

In 2019, the FAA estimated the annual cost of flight delays in the United States to be $33 billion.

Data

Internal and external data sources

The data received was entirely sourced from the business, no external data sources were used in the project.

Types of data

planes.csv:

- tailnum - character
- year - integer
- type - character
- manufacturer - character
- model - character
- engines - integer
- seats - integer
- speed - integer
- engine - character

airports.csv:

- faa - character
- name - character
- lat - numeric
- lon - numeric
- alt - numeric
- tz - numeric
- dst - character
- tzone - character

airlines.csv:

- carrier - character
- name - character

flights.csv:

- year - integer
- month - integer
- day - integer
- dep_time - integer
- sched_dep_time - integer
- dep_delay - numeric
- arr_time - integer
- sched_arr_time - integer
- arr_delay - numeric
- carrier - character
- flight - integer
- tailnum - character
- origin - character
- dest - character
- air_time - numeric
- distance - numeric
- hour - numeric
- minute - numeric
- time_hour - POSIXct / POSIXt

weather:

- origin - character
- year - integer
- month - integer
- day - integer
- hour - integer
- temp - numeric
- dewp - numeric
- humid - numeric
- wind_dir - numeric
- wind_speed - numeric
- wind_gust - numeric
- pressure - numeric
- visib - numeric
- time_hour - POSIXct / POSIXt

Data formats

All data used in the project was in the form of .csv files.

- aircraft.csv
- airlines.csv
- flights.csv
- planes.csv
- weather.csv

Data quality and bias

The data provided related only to flights flying internally within the United States. It would not therefore be possible to draw any conclusions from the project findings regarding international flights from the United States.

Ethics

Ethical issues in data sourcing and extraction

Technically the data used for this project does not fall within the scope of GDPR as it relates to flights taken internally within the United States, who do not have a federal privacy laws like GDPR.

Additionally, there is no information relating to passengers, only the movement of flights.

Ethical implications of business requirements

There are no ethical implications of the business requirements, which is to improve the facilities at Newark International Airport.

Analysis

Stages in the data analysis process

- Understand the scope of the project and the problem
- Data cleaning and transformation
- Exploratory data analysis
- Model building
- Reporting

Tools for data analysis

This project was carried out entirely using R.

Descriptive, diagnostic, predictive and prescriptive analysis

- Linear regression: Linear regression is predictive analysis method which is a way of predicting future events between a dependent variable, in this case departure delays, and independent or predictor variables.
- Decision trees & random forests: Decisions trees and random forests are both predictive modelling methods.