Project Plan

Anne Braae

12/09/2021

Flight delays

Context

Business intelligence and data-driven decision making This is a data-driven exploratory analysis to provide insight on the impact of weather on flight delay. Consideration will also be given to other factors such as time of flight, airline carrier, frequency of flights and departure airport.

The report will allow Newark airport to decide if it is worth investing in improving facilities to allow aircraft to take off in the types of weather which are most affecting departure delays. By considering other factors, the report will also show if there might be other areas which would benefit from investment which could also reduce flight delays.

Domain knowledge and the business context Newark Airport (or Newark Liberty International Airport) is one of the three major airports in the New York metropolitan area. Along with the other two major New York airports, John F. Kennedy International Airport and LaGardia Airport, Newark Airport is owned by the Port Authority of New York and New Jersey. The Port Authority owns 5 airports in total.

Recently, $2.7 billion was invested to build a new Terminal One to replace Terminal A at Newark Airport. Other parts of the redevelopment program are focused on increasing airfield paving, redesigning roadways and improving parking.

Reducing flight delays is important to ensure a high quality of service for the airport. Flight delays are a key performance indicator for the airport. The Bureau of Transportation Statistics (BTS) counts flights as delayed if they depart 15 minutes or later than their scheduled time. For the airport authority, knowing what the primary cause for delays is can provide insight as to what measures need investing in. Improving infrastructure or improving facilities to deal with poor weather conditions.

An American study on 2007 flight data estimated that delayed flights cost the US economy a total of nearly 33 billion dollars a year. Flights may be delayed due to reasons within the air carrier's control such as maintenance or issues with staffing or crew or for reasons outside the air carrier's control such as extreme weather delay, national aviation system delay. Over a third of the main cost of delay is fuel costs. With the aviation industry focusing on improving sustainability reducing unnecessary fuel use is paramount.

Optimizing processes and efficiently using resources are key to reducing flight delays. Airlines and airports are using data analytics to improve on-time performance.

sources: Bureau of Transportation Statistics (BTC) Port Authority

Data

Internal and external data sources

- Internal data sources (provided by Newark Airport):
- Internal data from several CSV files. Details below:

Filename	Description
`flights.csv`	Information on all flights per day for Newark, John F. Kennedy and LaGuardia airports. Includes departure and arrival times, delay times, destination airports
`airports.csv`	Information on airports in the flights data including names, geocoordinates, altitude, timezone, daylight savings.
`airlines.csv`	Information on airline carrier code and airline carrier full name.
`planes.csv`	Information on planes in the flights data including year of make, type, model and cruising speed.
`weather.csv`	Information on weather for the three New York airports including temperature, wind measurements, precipitation, visibility and time of the recording per hour.

- External data sources:
- Additional weather data was sourced from The National Centers for Environmental Information. This was also in csv format. This data included measurements per day for temperature, precipitation, wind and snow.

Types of data The data consists of categorical (ordinal), categorical (nominal), numeric (continuous), numeric (discrete) and datetime variables. Examples of each are as follows:

categorical (ordinal) - `year`, `month`, `weekday` categorical (nominal) - `dest_airport`, `engine` (type of engine) numeric (continuous) - `wind_speed`, `temperature` numeric (discrete) - `seats`, `engines` (number of engines)

Data formats The data was all in downloaded flat files (csv) or flat files received from the organisation.

Data quality and bias The data consists of every flight per day for the year of 2017 departing from three New York airports, it also consists of weather data obtained for the three airports for the same time period.

In general the data quality for the flights data was good, there was only missing information for some destination airports. The data provided on weather had a large amount of missing values for temperature measurements. The data that was externally sourced only provided information per day and not per hour. This could have biased the data as the flight informationw as per hour.

Destination information was only provided for US domestic airports, this may bias the analysis. It may be relevant to the organisation to include all destination airports for the purposes of this analysis.

Ethics

Ethical issues in data sourcing and extraction There are no ethical concerns with the data sourcing. All information was openly and freely available. There are no identifying variables in the data provided by the airport, however I would ensure that this data was kept private if the airport requested this.

Ethical implications of business requirements The data should be communicated back to airport passengers and airline carriers. Predicting flight delays will allow the airport to provide better services for their passengers. There was no passenger information provided in the data used for this analysis. however, should future analyses expand the analysis to include passenger data then this information should be protected.

Analysis

Stages in the data analysis process

1. Planning
2. Data collection for missing weather data
3. Setting up a GitHub project repository
4. Cleaning the data
5. Data wrangling
6. Exploratory data analysis
7. Creating basic visualisations and graph selection
8. Model building and refinement
9. Documenting the analysis
10. Presenting business insights

Tools for data analysis Data analysis tools used: Python, Jupyter Notebooks and Visual Studio Code.

Project planning tools used: Trello, Excalidraw (star diagram for joining datasets, timeline)

Descriptive, diagnostic, predictive and prescriptive analysis **Descriptive Analytics** tells you what happened in the past.

Diagnostic Analytics helps you understand why something happened in the past.

Predictive Analytics predicts what is most likely to happen in the future.

Prescriptive Analytics recommends actions you can take to affect those outcomes.

The analysis performed for Newark Airport investigating departure delays focused on Descriptive Analytics, Diagnostic Analytics and Prescriptive analytics.

The initial exploratory data analysis was very descriptive, falling under Descriptive Analytics. This analysis focused on examining different variables and using data visualisations to see how flight delays changed with each different variable.

The modelling was both Diagnostic and Prescriptive. This analysis focused on finding explanatory variables for departure delays which had happened in 2017 which is Diagnostic Analysis. Using the results from the model, I was able to communicate back to the business the required actions to take. This is Prescriptive Analytics. The recommended actions to take were additional investigation into other factors involved in departure delay.

PDA Outcomes

Working with Data (J4Y6 35)

1. Plan an analysis to provide business intelligence

- 1.1 Business intelligence and data-driven decision making
- 1.2 Domain knowledge and the business context
- 1.4 Internal and external data sources
- 1.5 Data quality
- 1.6 Stages in the data analysis process
- 1.7 Descriptive, diagnostic, predictive and prescriptive analysis
- 1.9 Ethical implications of business requirements
- 1.10 Tools for data analysis

2. Extract data from a variety of sources

- 2.1 Tools for querying data sources
- 2.2 Types of data (categorical and numerical data and their sub-types)
- 2.3 Data formats
- 2.6 Data quality including data bias
- 2.7 Ethical issues in data sourcing and extraction

4. Analyse data to provide business intelligence

- 4.7 Role of domain knowledge in interpreting analyses