Text

Description automatically generated

12/30/2021

IDM CLASS PROJECT

Airline Dataset

Members

SYED MUHAMMAD BILAL FAHIM – 19677

ROSHAN TARIQ -- 19781

SYED MUJTABA ALI ZAIDI -- 18619

# **Problem statement:**

The airline dataset contains details data of flights in the US and their Arrival and departure delay information.

We use the following data to find the expected delays in the arrival time of flights through methods of Data processing and analytics.

This information is helpful to airline companies in batter planning of expected delays and arrival of flights. This outputted data with accuracy metrics benefits the airline companies in terms of cost, time, labor, and unexpected outcomes.

**Wonder question:** Have you ever been stuck in an airport because your flight was delayed or canceled and wondered if you could have predicted it if you'd had more data?

**Problem Type:** Since we are interested in the arrival delay in minutes, we will use regression to predict the arrival delay which may be caused due by some underlying factors like if there is a delay in departure of flight then it may have arrival delay too, etc.

**Techniques and strategy:** We will be using correlation filter-based feature selection for our model. We will be using some graphical plots to visualize and understand the data. We will look for the missing values and will eliminate them.

# **About the data:**

The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, but due to limitations of resources, we sampled the data from the years 2007 and 2008 only.

## **Dataset link:**

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7>

## Dataset attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Airline Data**  **Airline.csv** | | | | |
| **Attribute**: | **Input format** |  | **Attribute:** | **Input format** |
| **year:** | year (2000) | crselapsetime | Estimated elapsed time in minutes |
| **month:** | January – December (1-12) | actualelapsetime | Actual elapse time in minutes |
| **dayofweek:** | 1 – 7 (Monday-Tuesday | origin | Origin airport code |
| **crsdeptime** | Scheduled departure time  (Local: hhmm) | dest | Destination airport code |
| **crsarrtime** | Scheduled arrival time  (Local: hhmm) | distance | Distance between the airport in miles |
| **arrtime** | Actual arrival time  (Local, hhmm) | taxiin | Wheels down and arrival at the destination airport gate |
| **uniquecarrier** | Carrier code | cancelled | Flight canceled?  Boolean |
| **fligthnum** | Flight number | cancellationcode | Reason for cancellation:  A = carrier  B = weather  C = NAS  D = security |
| **talinum** | Plane tail number | diverted | Boolean 1 = yes  0 = no |
| **airtime** | Flight time – in minutes | carrierdelay | Delay in minutes caused by the carrier |
| **arrdelay** | The difference in minutes between scheduled and arrival time.  Early arrival shows in negative numbers | wetherdelay | Delay in minutes caused by weather |
| **Depdelay** | The difference in minutes between scheduled and departure times.  Early arrival shows in negative numbers | nasdelay | Delay in minutes caused by National air system NAS |
|  |  | Securitydelay | Delay in minutes caused by security |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Plane data**  **Plane\_data.csv** | | | | |
| Attributes | Input format |  | Attribute | Input format |
| talinum | Aircraft tail number | status | Status |
| type | Usage type | aircraft\_type | Aircraft type |
| manufacturer | Manufacture company | engine\_type | Engine type |
| issue\_date | Data start operating | year | Manufacture year |
| model | Aircraft model |  |  |

## **Missing Data:**

Table

Description automatically generated

We can see that the data has some number of missing values in the last 5 columns. The cancellation Code column, however, has a large amount of missing data therefore, we will discard it while we drop na's from the other columns.

# **Data Exploration and Analysis:**

To explore this data further, we considered some of the questions which we can answer by analyzing the data. These have been answered below:

1. When is the best time of day, day of the week, and time of year to fly to minimize delays?

For this, we gathered the month, year, day and computed their mean. The one with the minimum mean would be the best month, year, day because the delay seems to be less in this period.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

The first graph shows less delay on Friday, the second shows less delay in September while the 3rd graph shows less delay on the 24th day of the month. These are the days, months, and days of the month where delay seems to be less. Also, for the time I fetched the min delay according to the departure time and found that early hours of the day seem to be better for flying.

1. Do older planes suffer more delays?

For this, we joined plane data with airline data using the flight’s tail number. Then plotted the date of manufacture against the delays. Following is the graph.

Chart, line chart

Description automatically generated

The graph shows a highly gradual decline in the maximum arrival delay time for each day which translates to older planes that have been issued earlier suffering more delays than newer planes.

1. How does the number of people flying between different locations change over time?

For this we visualize the change in distance, the number of people traveling and the number of flights over the years. Following are the graphs.

Chart, line chart

Description automatically generatedChart, line chart

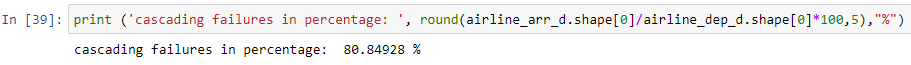
Description automatically generatedChart, line chart

Description automatically generated

From the three graphs above we can see that the average distance traveled has been increasing but the total number of people flying has been decreasing with a peak in August 2007, after which we can see that in the third graph, the count of flights has been decreasing rapidly.

1. Can you detect cascading failures as delays in one airport create delays in others?

For this we subsetted the data with departure time greater than zero and then subsetted this data where arrival time is greater than zero and then calculated the arrival vs departure percentage.



There is an 80.849% chance that a delay on the departure airport will create a delay on the arrival airport as well, therefore cascading failures do seem to exist.

# **Data Pre-Processing and Modelling:**

The data pre-processing includes the downsampling of data, we picked only 5% of the data for our model because the data is so big. The next step was the removal of some unnecessary columns like TaxiIn and TaxiOut. The next step was the feature selection, we used a filter-based approach to pick out the best features for our data. We used correlation as a tool to judge the predictor’s importance towards the target variable which is Arrival delay.

Following is the correlation matrix plot:

Chart

Description automatically generated

To get a clear view of this, we used a correlation matrix that gives the variables which are highly correlated.

Text

Description automatically generatedThe following variables turn out to be the best variables for our model, therefore we only included these in the model.

## **Model:**

Since we need to predict the arrival delays in minutes, we used a linear regression model, later we some metrics to evaluate these.

We split our sampled data into 70/30 train and test data.

Chart

Description automatically generated

The plot shows the actual vs predicted arrival delay values. This graph gives a good representation of the predicted arrival delays. Other than this, we can see that the coefficient of the model here.

### **Model Evaluation:**

We used root mean square, mean square errors and the r squared value to evaluate the model.

Graphical user interface, text, email

Description automatically generated

We can see that the r square value is about 1 which indicates that the model outperformed in predicting the arrival delays.

### **Model Findings:**

From the model, we can perfectly predict the arrival delays using the stated attributes like departure delays, etc. It also shows a very good relationship between the actual data and the test data. The following histogram shows the range of predicted arrival delay values.

Chart, histogram

Description automatically generated

The values less than 0 shows that the flight reached before the arrival time which is the best-case scenario. The values greater than 0 show the arrival delay. The model also shows that the departure delay is one of the big reasons for arrival delay while other factors like weather delay security delays are also some of the causes of high arrival delays.

How long did it take for a single evaluation?

Because we were utilizing the Jupyter notebook, it took 5-6 minutes to completely execute all blocks/cells of code in the notebook.

# **SOFTWARE AND LIBRARIES USED:**

The data was processed in python programming language on the platform Jupyter notebook.  
pythons imports were utilized.

Graphical user interface, text, application

Description automatically generated

THAT WILL BE ALL

THANK YOU