Airline On-Time Performance Analysis

Memory Alpha

Dhvani Patel
dhvani.patel@mail.rit.edu
Rochester Intitute of Technology

Rohit Keswani
rohit.keswani@mail.rit.edu
Rochester Intitute of Technology

ABSTRACT

With so many airlines having overlapping routes, it has become increasingly important for the customer to know about the performance statistics of the specific flight they intent to select. In this paper, we have performed visualization and data mining on data for the performance statistics of the airlines. These visualizations and mining techniques enable us to understand better about the tendency of the flight being delayed or being on-time.

1. INTRODUCTION

The airline industry is a dominant sector in transportation. Various statistics from around the world conclude that a lot of people are switching to use airlines as their preferred form of travel. There are 5000 to 5500 airlines in the world. In the US, there are 18 airlines out of which 10 hold a major share in US Airspace. US Airline industry is worth $1.7 trillion and transports around 2.4 million passengers every day.

The data in consideration here has been obtained from the Bureau of Transportation Statistics, US [2] and Data.world [1]. They are two datasets that have data of statistics about major airlines in the US. These statistics include the origin, destination, delay in take-off, delay in landing, distance traveled, the reason for delay/cancellation/diversion. Each dataset has about 600K instances with over 30 attributes. Following are the few of significant attributes:

1. Year: The year of the concerned flight.
2. Month: The month of the concerned flight.
3. Day: The date of the concerned flight
4. DayOfWeek: The day of the week of the concerned flight.
5. CarrierID: The airline of the flight.
6. FlightNumber: The flight number of the airplane.
7. OriginAirport: Source of the flight.
8. DestinationAirport: Destination of the flight.
9. ScheduledDeparture: The timestamp for flight scheduled to leave the gate.
10. ActualDepartureTime: The timestamp for flight when it departed.
11. DepartureDelay: Delay(in time) in leaving the gate.
12. TaxiOutTime: Time taken between leaving the gate to takeoff.
13. TaxiInTime: Time taken between landing and reaching the gate.
14. ScheduledArrival: The timestamp for flight scheduled to arrive.
15. ActualArrivalTime: The timestamp for flight when it arrived.
16. ArrivalDelay: Delay(in time) in arriving at gate.
17. AirTime: Time the fight was in air.
18. Distance: Total distance covered by the flight from Source to Destination.
19. isDelayed?: If there is an arrival or departure delay, the value is set, else it is unset.

Among the following sections, Section 2 presents a motivation for the selection of airline performance for the term project, and Section 3 presents with tasks that are performed to clean the data and load it in database server. Section 4 presents the visualizations performed on the dataset and two data mining approaches performed on the dataset. Section 5 and Section 6 present the lessons learned and the scope of the future work of this project.

2. MOTIVATION

Since we are a team of international students we have traveled a lot overseas and have experienced quite a lot of delays while flying back home from the US and vice-versa. There have been instances when we have also missed our connecting flights. Issues like these not only lead to a wastage of time by the delay but also frustrates travelers. A lot of portals while booking the tickets mention things like, "this flight fills fast", "Only 1 ticket left to book", etc. However, there rarely is a portal which could inform travelers whether a flight is bound to delay by 2 hours or this flight usually is canceled or they might miss the connecting flight. This effort is in the direction so it could be extended and travelers can book their tickets with much more knowledge about the flight statistics.

3. DATA MANAGEMENT

3.1 Data Cleaning and Preparation

Data cleaning was performed using Microsoft Excel. Few attributes like the tail number were deleted as it did not seem to be useful for this project. A new attribute isDelayed? is added which has the value of 1, if the flight has either arrival delay or departure delay, else 0. This attribute is added to perform data mining tasks. Details about other attributes are mentioned as follows:

1. Records with missing values for ActualDepartureTime and ActualArrivalTime were deleted (4000 records) as this value could not be replaced with mean or anything else. Without this information, it made no sense to have these records.
2. Missing values for DepartutreDelay and ArrivalDelay attribute was replaced with 0 for records whose scheduled and actual departure/arrival time was the same.
3. Missing values for TaxiOutTime (100 records) were replaced by the mean of other values of this attribute.
4. Missing values for TaxiInTime (300 records) were replaced by the mean of other values of this attribute.

A log file was maintained where all the modifications to the dataset were logged. Also before merging the two datasets, attributes with similar details were renamed to represent one attribute instead of two attributes.

3.2 Data Assembly

After performing data cleaning, both the datasets were loaded into a SQL database using MySQL and JDBC connection for java. The code files are attached with the submission.

After loading the dataset in MySQL Server, several SQL queries were executed to better understand the data. A few of the SQL queries and their results are present in Figure 1 and others are attached along with the submission.

```
mysql> SELECT COUNT(*) AS TotalRecords FROM airlinestatistics;
+--------------+
| TotalRecords |
+--------------+
|       137645 |
+--------------+
1 row in set (0.01 sec)
```

Figure 1: Total Number of records in MySQL database

4. DATA MINING AND VISUALIZATIONS

In this Section, several exploratory analysis and two data mining techniques are performed on the dataset.

4.1 Data Visualization

For exploring the database, an exploratory analysis was performed with Tableau [5] and the results for exploratory analysis are given below.

In Figure 2, every bubble and color represents one carrier as mentioned in the legend. The average arrival delay is represented by the size of the bubble. As observed from the graph, Skywest Airlines (OO) has the most arrival delays and Hawaiian Airlines(HA) has minimum delays.

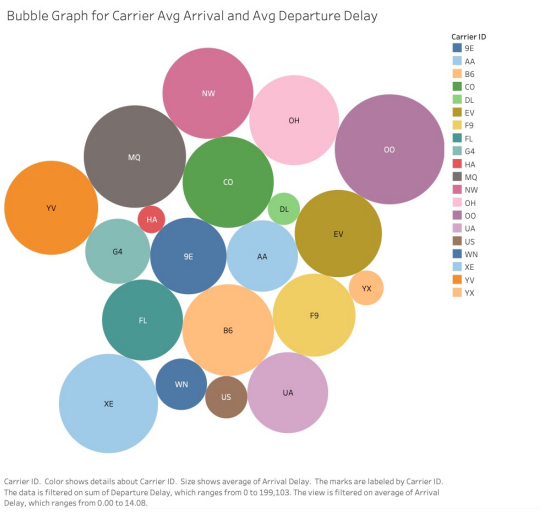

Figure 2: Bubble Graph: Arrival and Departure Delay w.r.t Carrier

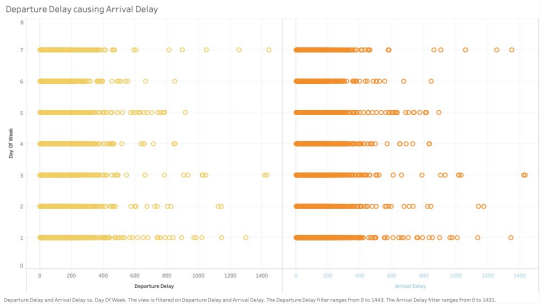

Figure 3: Horizontal Bar Graph: DayOfWeek v/s Departure and Arrival Delay

It is clearly visible from Figure 3 that on day 6, that is Saturdays, there is the least amount of delays in flights. Also, the general trend observed for each day lesser the departure delay in the number of flights, lesser is the arrival delay. It is verified in Figure 4.

Pairwise comparison of attributes in Figure 4 verifies the trend by Figure 3. It is seen that DepartureDelay is positively correlated to ArrivalDelay. Also, AirTime is positively Correlated to Distance, which was expected. We expected to see a relation between TaxiInTime v/s ArrivalDelay and TaxiOutDelay v/s DepartureDelay, however, there does not seem to be a strong relation between them.

4.2 Data Mining

In order to perform data mining, the dataset was randomly split between a training set and a testing set. The training set comprised 70% of the data, while testing set had to remain 30% of the data.

Data mining approaches such as Logistic Regression [4] and K-NN Classifier [3] were performed on the training set and its accuracy was calculated on the test set. The target attribute for both the mining tasks was "isDelayed?". With the help of data from other attributes, the model predicts if the flight is going to be delayed. The classification report which consists of details such as accuracy, precision, recall

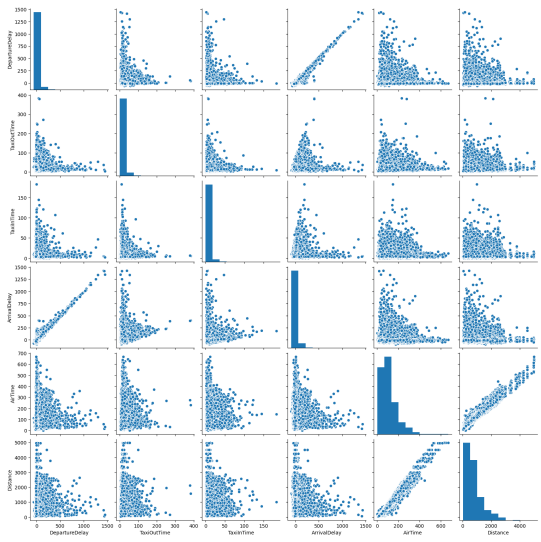

Figure 4: Pairwise Comparison of Attributes

and F1-score of each of the models is shown in Figure 5 and Figure 6.

```
----------------Logistic Regression-----------------
Confusion Matrix
 [[16700   992]
 [ 1419 15746]]
Accuracy of Logistic Regression on training set: 0.93
Accuracy of Logistic Regression on test set: 0.93
              precision    recall  f1-score   support

           0       0.92      0.94      0.93     17692
           1       0.94      0.92      0.93     17165

    accuracy                           0.93     34857
   macro avg       0.93      0.93      0.93     34857
weighted avg       0.93      0.93      0.93     34857
```

Figure 5: Logistic Regression

5. CONCLUSION

With the help of a step-by-step approach of cleaning, preparing and visualizing the data, it is evident interesting knowledge can be discovered from the data. The visualization enables us to understand hidden patterns within the data, such as DepartureDelay being the most important cause of delay. It is a bottleneck that needs to be catered and most of the flights will be on-time then.

Logistic Regression [4] over the analysis of Airline Performance data provides an accuracy of 93% over the test set. K-NN [3] provides an accuracy of 69% over the test set. These techniques seem promising for predicting new data inputs over the model.

6. FUTURE WORK

It would be great to analyze the difference of accuracies between the two models. Also, creating a web application or a browser plugin that is trained on the data, and user

```
--------------------KNN Classifier--------------------
Confusion Matrix
 [[13479  4222]
 [ 6564 10592]]
Accuracy of K-NN classifier on training set: 0.80
Accuracy of K-NN classifier on test set: 0.69
              precision    recall  f1-score   support

           0       0.67      0.76      0.71     17701
           1       0.71      0.62      0.66     17156

    accuracy                           0.69     34857
   macro avg       0.69      0.69      0.69     34857
weighted avg       0.69      0.69      0.69     34857
```

Figure 6: K-NN Classifier

can invoke it while shortlisting his flight to understand if the delay should be expected for the flight.

7. REFERENCES

[1] Airlines delay. `https://data.world/data-society/airlines-delay`. Accessed: 2020-03-10.

[2] Bureau of transportation statistics. `https://www.bts.gov`. Accessed: 2020-02-15.

[3] G. Guo, H. Wang, D. Bell, and Y. Bi. Knn model-based approach in classification. 08 2004.

[4] J. Peng, K. Lee, and G. Ingersoll. An introduction to logistic regression analysis and reporting. *Journal of Educational Research - J EDUC RES*, 96:3–14, 09 2002.

[5] R. M. G. Wesley, M. Eldridge, and P. Terlecki. An analytic data engine for visualization in tableau. *SIGMOD*, June 2011.