US Domestic Flight Analysis

Can we predict flight delays?

Jaskaran Garha
214314439
garhajas
garhajas@my.yorku.ca

David Geller
214404255
dgeller
dgeller@my.yorku.ca

Gilbert Andze Tsoungui
211262656
christ18
chris18@my.yorku.ca

William Zhen
214305171
will636
will636@my.yorku.ca

Abstract

The impact of the airline industry on a country's economy is huge to say the least. In this report we analyze on-time performance of US domestic flights. The goal here is to highlight some of the key factors such as delays and cancellations that can have a vast impact on the industry and economy. By looking at past historical flight data, companies can improve their services thereby reducing losses and increasing revenues. In addition, customers have the added benefit of making informed decisions about best airlines and possible delays/cancellations when planning to travel. Various big data technologies were leveraged to achieve desired results. Some of these technologies include Apache Spark, Parquet, HDFS, and basic Machine Learning libraries for predicting delays.

Introduction and Motivation

The International Air Transportation Association (IATA) forecasts the number of passengers to double to 8.2 billion by the year 2037 [1]. An increasing number of people are choosing to travel the world and therefore the airline industry will need to evolve accordingly. The focus of the report is to analyze flight data from the past couple years and highlight inconsistencies such as delays, cancellations, and airport air-traffic which could help airline companies improve their services and thereby improve the customer experience. In addition, consumers can use the results gathered by our analysis to make well informed decisions about purchasing airline tickets from reputable airline companies. Furthermore, we are going a step further to see if we can predict if the average number of delays from past years will improve based on the data from past years. In order to achieve useful results, we used the following questions as a guide for our analysis:

- What is the most common cause of flight delays and cancellations?
- Which airline companies have the most/least delays and cancellations?
- What times of the year has a high probability of delays/cancellations?
- What are the busiest and most/least delayed flight routes?

By combining results from the above questions, one can gain insight into which airline companies provide better service on a given day and flight route.

An important thing to note here is that our data is limited to US domestic flights. We examined US domestic flight data from recent years (2016-2019) provided by the Bureau of Transportation Statistics (BTS) [2]. In the future our analysis could be extended to international flights provided the data is of the same structure.

Data and Data Analysis

The data is available as a CSV file on the BTS website under the "On-Time: Marketing Carrier On-Time Performance" section [2]. BTS provides a clean and well-structured dataset with minimal data cleaning required. In this project no data cleaning was done thereby eliminating any complications that arise due to unclean data. The size of data is 5.5GB (referred to as "medium" size in the industry) for the four years of data analyzed, 2016 to 2019. The sink rate is low because popular transportation statistics agencies release data on a monthly basis. To get around this we downloaded and combined the data of each month for each year we were looking at, 2016 to 2019. BTS is one of the statistical federal agencies in the United States therefore the data is high quality. In addition, BTS allows you to select the fields you want in the dataset thereby reducing the size of data. Lastly, the dataset is semi-complete in its current state because though some of the fields are mandatory, others can be inferred or calculated from the mandatory fields. Some of these mandatory fields include destination and origin of flight, airport names, elapsed, arrival and departure times, and lastly the causes of delay or cancellation of the flight along with other miscellaneous information (non-mandatory fields).

Most of the data analysis employed was text analysis along with a little bit of predictive analysis. Simple Statistical methods such as count, average, and linear regression were used to generate results.

Architecture

Data was downloaded from BTS in csv file format. Since we knew early on that we would be using a lot of queries, we decided to make use of columnar storage in order to increase the performance of our system. To accomplish this, we used apache parquet files for data processing. Simple script transformed the csv data file to a parquet file, providing a significant reduction in size of file from 1.2GB to 143MB. After transforming the data, we used data frames and SQL queries to process and analyze our data. This step was easy since it was very similar to working with pandas. Lastly, we plotted the results on graphs for better visualization and understanding. Figure 1 shows the pipeline of our architecture.

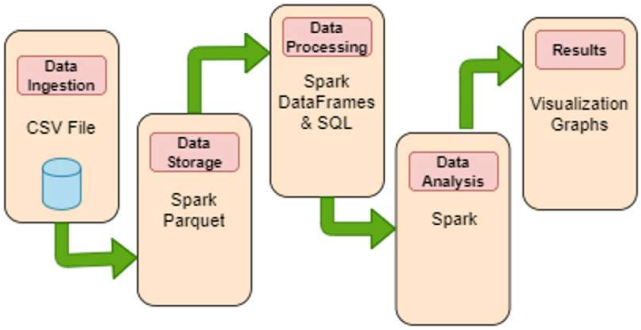

Figure 1: Depiction of architecture and data flow in the system

There were some limitations when using this approach. For instance, for our predictive analysis we tried to use the pyspark.ml package to train our data and perform basic machine learning. However, we faced several issues related to the data format, probably the parquet data files. We were not able to debug the issues at this time however, to compensate for this we used Microsoft's Azure Machine Learning studio tool to predict delays, discussed in detail later in report.

Evaluation and Results

Graphs were used to effectively understand and visualize the results generated by our analysis. Figure 2 shows the distribution of common flight delays for the year 2018. Similar results were generated for the other three years. We discovered that contrary to popular belief, weather delays only make up a small amount (3%) of the total delays.

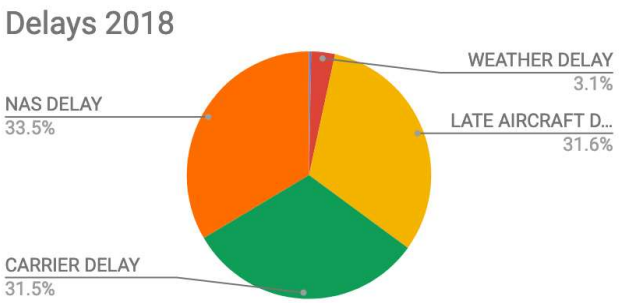

Figure 2: 2018 Flight Delays

Figure 3 shows the reasons for flight cancellations for the year 2016. Again, similar results were generated for the other years. Here we see expected results as most cancellations are due to weather.

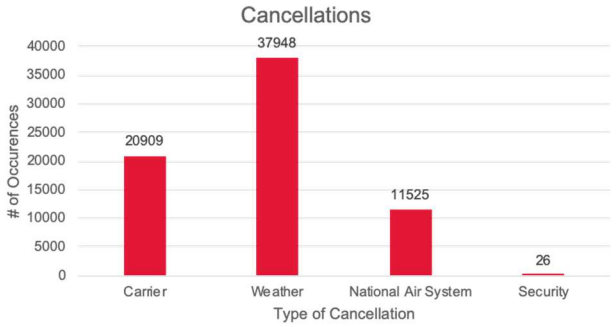

Figure 3: 2016 Cancellations Reasons

Figure 4 shows delays in minutes per month over the four years. There were no surprises here as we expected the average delay to increase during summer vacation and Christmas time.

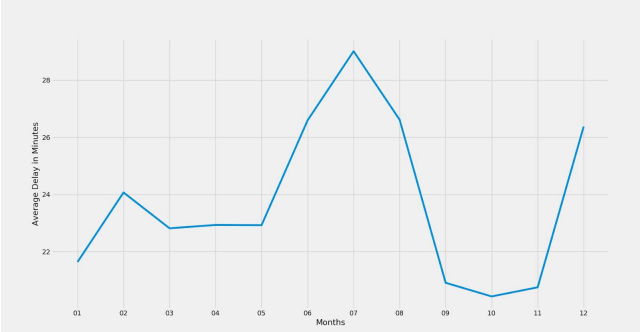

Figure 4: Average Delay per month

In Figure 5, we looked at individual airlines and graphed the total number of delays vs. average minutes of delay. Although not shown here, a similar graph was plotted for individual airline cancellations. These two analyses are crucial for calculating the 5-star rating system discussed later in the report.

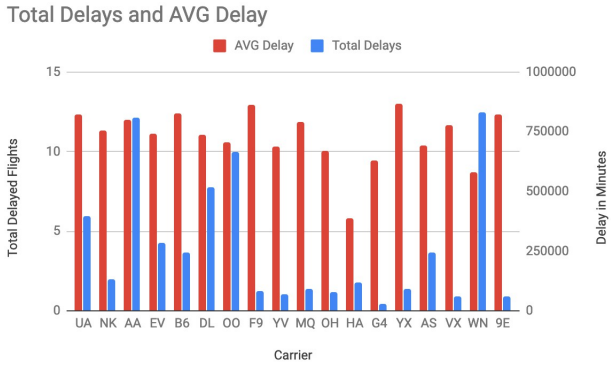

Figure 5: Airline Delays

Next, we need to consider the busiest routes/airports that may contribute to delays. Figure 6 shows the top 10 busiest routes and the final delay occurring on the end of that route.

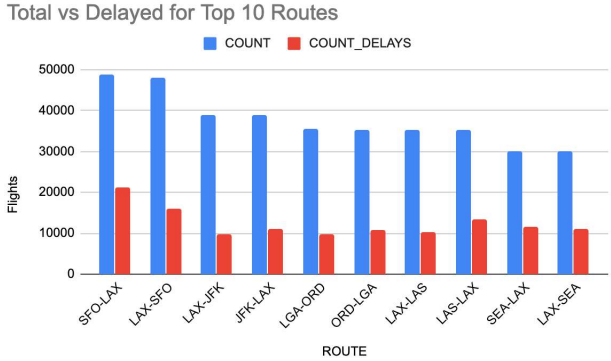

Figure 6: Busiest routes with delays

Lastly, figure 7 compares the top 10 least and most delayed routes. The least delayed routes shown here on the left are filtered to include popular routes (> 1000 flights) so as to have a more representative model.

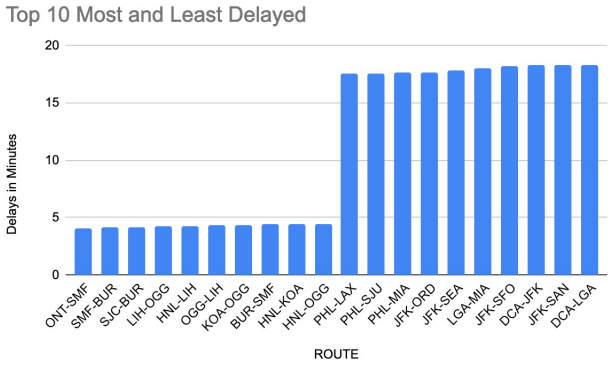

Figure 7: Least vs. Most delayed routes

Predictive Analysis

Using Microsoft's Azure Machine Learning studio, we were able to build a predictive model for predicting the delay of a flight for the year 2019. The target value (the value to predict) chosen was "Arrival Delay" in minutes. Arrival delay is considered as the final delay by consumers which is why we chose that as our target value. A linear regression algorithm was used because we are trying to predict a number: delay in minutes. 75% of the data was used to train the model and remaining 25% was used to test the model. The results are shown below in figure 8. We got a mean absolute error (MAE) of 9.3 which means the average difference between actual arrival delay vs. predicted arrival delay is about 9 minutes. Although in the real world this is not a very good predictive model, we can simply improve it by providing more training data and using different machine learning features. Features here refer to the columns that are said to be measurable properties for predicting target value.

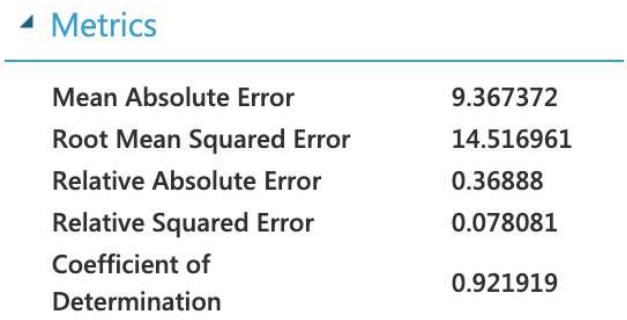

Figure 8: Results of Azure ML when predicting arrival delay 2019

Conclusion

Some of the interesting findings include:

- Weather delays only account for 3% of total delays whereas cancellations account for most of the cancellations
- Most of the airlines have negative delay over the four years, meaning they are mostly on-time or early
- From the delayed airlines, Hawaiian airlines had the worst delay over the four years and delta airlines had the best delay time
- We predicted the arrival delay with a MAE of 9 minutes, which we intend to decrease in the future with more training data and analysis.

By combining the above results, one can create a model for rating the airlines based on their various service attributes such as total number of flights, percentage of delays, cancellations and on-time arrival plus average time of delay. Furthermore, based on the route and date of travel, one can create an application that tells customers which airline is to serve them best based on historical data. Although the price is not included in this model, in the future our analysis could be extended to gather real-time airline prices. In summary a finished product would look something like google flights, where different layers of data would coordinate to deliver the best experience to consumers. Such layers would include streaming data from weather stations and

airports in real time to predict delays and cancellations ahead of time.

References

[1] “IATA Forecast Predicts 8.2 billion Air Travelers in 2037,” *IATA*, Oct. 2018. [Online]. Available: https://www.iata.org/en/pressroom/pr/2018-10-24-02
[2] Bureau of Transportation Statistics, Sep-2019. [Online]. Available: https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline On-Time Performance Data&DB_Short_Name=On-Time.

Additional References

- “Apache Spark Tutorial,” *Tutorials Point*. [Online]. Available: https://www.tutorialspoint.com/apache_spark/.
- “Quick Start,” *Quick Start - Spark 2.4.4 Documentation*. [Online]. Available: https://spark.apache.org/docs/latest/quick-start.html.
- “pyspark.ml package” *pyspark.ml package - PySpark 2.4.4 documentation*. [Online]. Available: https://spark.apache.org/docs/latest/api/python/pyspark.ml.html.
- “Jupyter Notebook for Beginners Tutorial,” *Dataquest*, 11-Sep-2019. [Online]. Available: https://www.dataquest.io/blog/jupyter-notebook-tutorial/.
- “Quickstart: Create a data science experiment - ML Studio (classic) - Azure,” *Microsoft Azure*. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio/create-experiment#prepare-the-data.