Analysis on U.S. Flight Delays

G1 Group 6

Contents

1 Introduction

As of 2015, the U.S. Domestic Aviation Industry is worth $135 billion It is expected to grow to $150 billion within the next 4 years (IBISWorld, n.d.). With 696.2 million domestic passengers in 2015, the industry witnessed a 5% rise in the number of total passengers from 2014. System-wide demand measured also grew by 5.5% in 2015 (Bureau of Transportation Statistics, 2017). Amongst the competitive industry, American Airlines and Southwest Airlines, emerged with the greatest market shares.

Despite the growth, the airline industry faces high cost pressures due to the high fixed costs alongside the many variable ones. As these expenses are important, airlines look to avoid other such expenditures, specifically those incurred from flight delays. As these expenses are important, airlines look to avoid other such expenditures, specifically those incurred from flight delays. In 2007, flight delays across the U.S. airspace system, defined by the Federal Aviation Administration (FAA) as a flight that departs **more than 15 minutes past its scheduled time**, had cost the U.S. economy $32.9 billion (CAPA, 2010). Though passengers are estimated to bear more than half of the cost, airlines are estimated to have shouldered $8.3 billion in costs from flight delays alone in 2007. Therefore, our analysis will focus on creating a model and solution for the market leader in the airline industry to minimize costs related to flight delays.

2 Business Use Case and Motivation

As aforementioned, the airline industry incurs high fixed and variable costs every day. Focusing on the leaders of the domestic industry, we would be analyzing data of Southwest Airlines to understand reasons for departure delays and thereby look to find ways to better pre-empt and manage them.

By minimizing delays, airlines could reduce costs for both the business and passengers. The average cost of aircraft block time for U.S. passenger airlines has been estimated to be at $74.24 per minute (Airlines for America, 2020). Meanwhile, the average value of a passenger's time is estimated to be $47 per hour, thereby the economic cost of delay per minute to be $75.

3 Dataset

Our dataset is obtained from "2015 Flight Delays and Cancellations" published by the U.S. Department of Transportation on Kaggle. This dataset contains information of domestic flights of major air carriers in the U.S., including details of on-time, delayed, cancelled and diverted flights.

4 Understanding the Data

4.1 Data Loading and Preprocessing

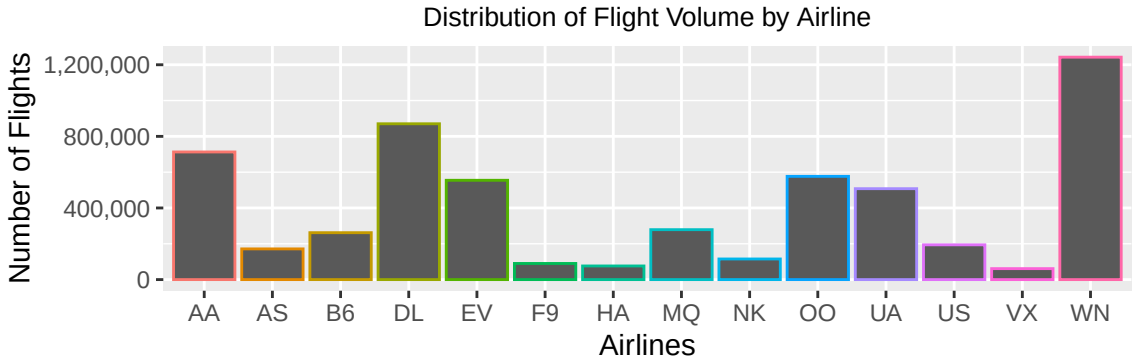
Before exploring the data, some general data preprocessing and cleaning is done, including mapping airport ID to IATA formats and imputing missing values. Moreover, as the focus of this project is on flight delays, flights that are cancelled or diverted are removed. Certain features are also converted to factor/character types. We create `DEP_DELAY` and `ARR_DELAY` variables following FAA's definition of flight delays. Irrelevant features such as the year are dropped.

4.2 Exploratory Data Analysis (EDA)

Our EDA process will firstly examine the general airline industry performance as well as compare Southwest's to its competitors where applicable. Subsequently, we explore Southwest's performance in terms of delays and the factors that might contribute to delays.

4.2.1 Distribution of Flight Volume by Airline

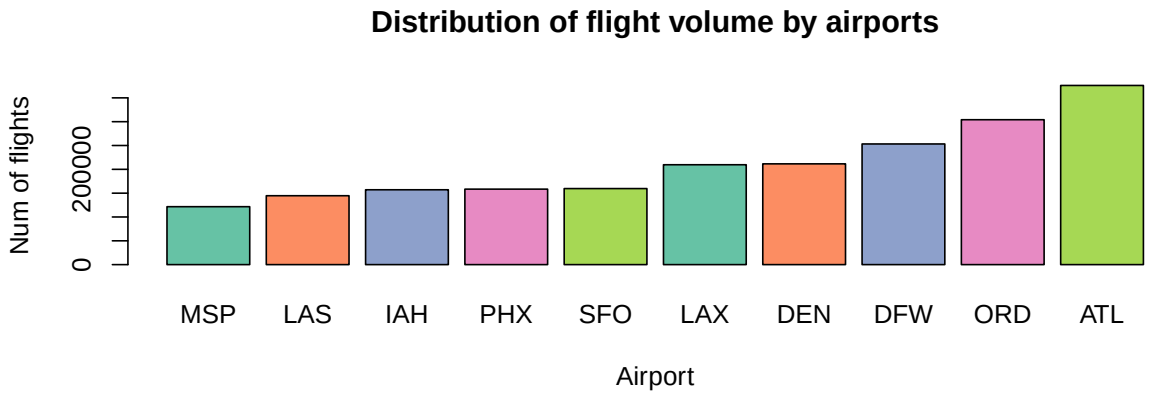
As Southwest Airlines (WN) has the highest market share in terms of revenue in the U.S. domestic airline industry, we can examine how it compares with its other competitors in terms of flight volume in 2015. The distribution of the Airlines by volume is as follows:



From the barplot, we can see that Southwest Airlines indeed has the most number of flights in 2015 and the difference with the next most popular airline, Delta Airlines (DL), is quite significant. **With the highest flight volume in the domestic industry, Southwest has more at stake to ensure satisfactory performance in order to maintain its lead**.

4.2.2 Top 10 busiest airports

We also would like to see the distribution by airports to identify the top 10 busiest airports.



The busiest airport by volume is ATL, Hartsfield–Jackson Atlanta International Airport.

4.2.3 Departure and Arrival Delays

There are two main categories of delays in the airline industry that affect passengers:departure delay and arrival delay. Below shows the proportion of flights for each type of delay. We have similarly defined a delayed arrival as having arrived more than 15 minutes past its scheduled time. The number of delayed flights in departure and arrival is similar, around 17.8%

```
## Class distribution for departure delay:  17.7 %
## Class distribution for arrival delay: 17.91 %
```

We hypothesize that there is a strong correlation between departure delay and arrival delay, such that a flight that departs late is likely to arrive late by a similar extent.

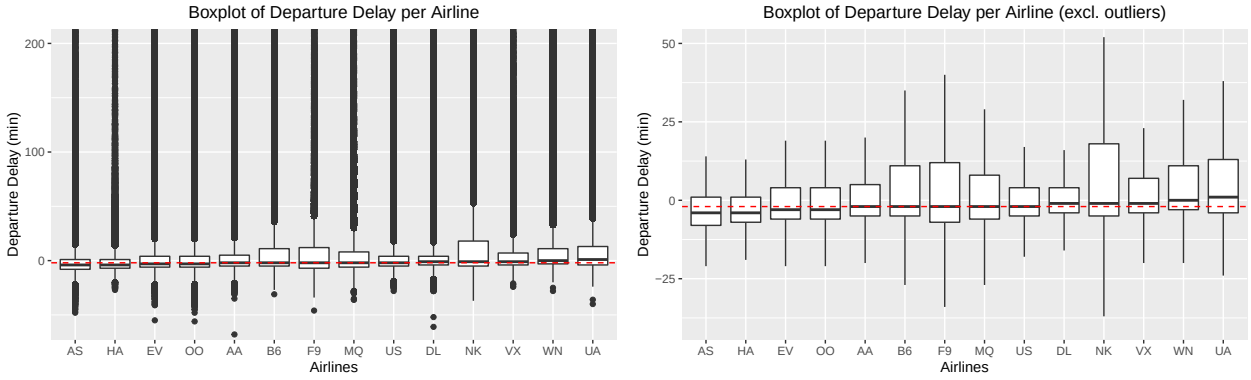
```
cor(data$ARRIVAL_DELAY, data$DEPARTURE_DELAY)
```

```
## [1] 0.9446715
```

The correlation between departure delay and arrival delay is very high at 0.9447, which proves our hypothesis true. This is unsurprising given that **late departure is most likely to lead to late arrival**. As such, our analysis should focus on finding factors that contribute to departure delays.

4.2.4 Departure Delay Performance of Airlines

We now use `DEPARTURE_DELAY` as a measure to assess how well the airlines perform in terms of delays. We use a boxplot to represent this information, with the overall industry median departure delay as a reference point as to compare each airline to the industry level.



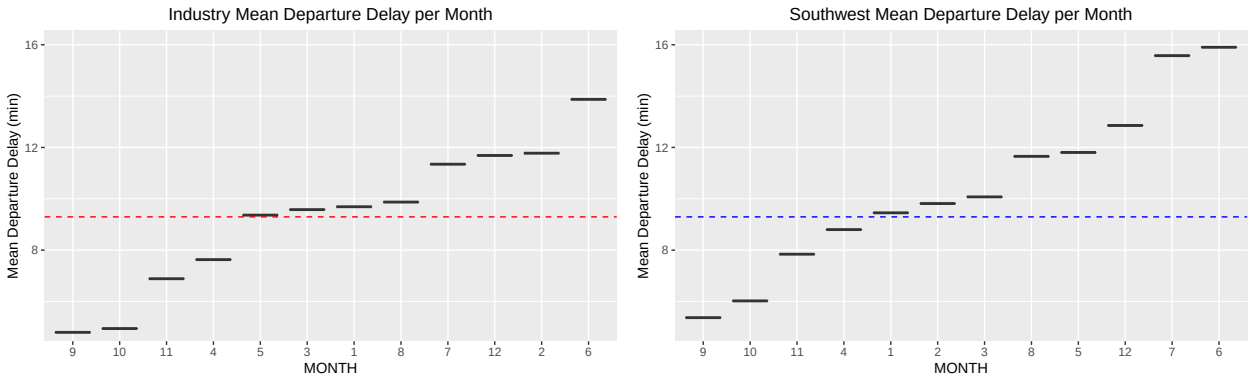
We discovered that there were too many outliers for all airlines to make the boxplot useful in investigating the departure delay of the majority of the flights. As such, we exclude the outliers in our boxplot. After hiding the outliers, we have a better view of the distribution of departure delays for each airline. Compared to the overall median departure delay (red dashed line), we observe that 9/14 airlines have a median departure delay below or almost equal to the overall median departure delay, showing that most airlines performances are similar to the industry average.

However, for Southwest airlines, its median delay is above the industry median. As an industry leader, we believe Southwest Airlines should aim to improve its departure on-time performance to maintain its lead as as well as to protect its reputation in the eyes of passengers and ensure that their passengers remain satisfied with their service.

We chart our course in trying to find additional factors that might contribute to departure delays.

4.2.5 Impact of Seasonality on Departure Delay

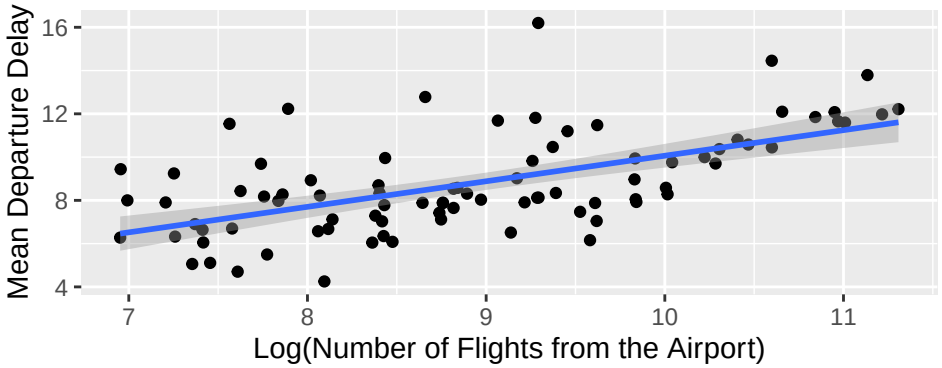
We now want to explore if the month of the flight would affect departure delay likelihood. This is done by comparing the average departure delay to the monthly average departure delay. We found that **the peak holidays season (i.e. December, January, February, June, July) usually affects the departure delay**. These are the winter break as well as the summer break periods. However, the mean values are not significantly large such that it goes beyond the FAA's definition of 15 minutes. (Note: dashed lines represent the industry average across the year)



Additionally, by comparing Southwest's performance with the industry averages, we can see that in the months of 4, 9, 10, and 11 where average delay is below the yearly industrial average, Southwest on average experiences slightly higher departure delays. In the higher ranges, we see that Southwest also experiences higher-than-industry average departure delays, with the difference more pronounced in the months of 6 and 7.

4.2.6 Relationship Between Busy Airports and Delays

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(southwest_df$DEPARTURE_DELAY, southwest_df$Freq)
```

```
## [1] 0.6049447
```

We now focus on other factors that might contribute to Southwest's delay performance being worse than industrial average. Initially, we had a hypothesis that flying from busy airports would lead to higher likelihood of delays. Focusing on Southwest's records only, we tried to correlate whether the volume of flights at the `ORIGIN_AIRPORT` would affect the `DEPARTURE_DELAY`. Our hypothesis was

supported by the correlation that we got, 0.6049, and could be explained where **busier airports tend to have to manage more flight and air traffic within limited number of runways**.

4.2.7 Summary of Key EDA Findings

In summary, we compared Southwest's departure delay performance to others in the industry and found that **the company is actually performing slightly worse than the industry averages**, across the year as well as within each month, even while being the industry leader in market share and flight volumes. By narrowing our focus onto departure delays, we also found that the month of the flight and the popularity of the airport where the flight is departing from are likely contributors to delays in departure.

5 Predictive Analysis on Flight Departure Delay

To assist Southwest Airlines in management of flight delays, we can predict the exact number of minutes of departure delay with regression or whether a flight will be delayed with classification. We believe that when a flight is delayed, the exact number of minutes may not be crucial; e.g. a flight delayed for 30 minutes is not much different from another delayed for 40 minutes. We explored Multi-linear Regression; the R-square value is only slightly over 0.2 with a root mean square error of over 23 minutes. The linear regression model did not perform well. As such, we proceed this problem as a binary classification problem.

5.1 Predictive Analysis Data Preprocessing

As there will be some features calculations based on route and origin airport information (refer to Section 5.1.2), a few outliers were excluded where the frequency of a route did not meet our minimum criteria of 1 flight per month (i.e. 12 flights in a year). This is so that the calculated values (e.g. estimates of on-time performance) can be more accurate and reflect more points of data.

5.1.1 Data splitting into train, validation and test sets

To have a fair evaluation of different models' performances later, we split the dataset into *train*, *validation* and *test* sets with a ratio of 60-20-20. The training set will be used to explore features and train models; the validation set will be used for hyperparameter tuning and final model selection; and the test set will only be used for the final model evaluation.

The final best model will be trained with the combination of the training and validation sets.

5.1.2 Feature Engineering

In addition to the preprocessing steps described in Section 4, more processing was required to prepare the dataset for model training. These include some feature selection and engineering which are explained below. As the ultimate purpose of the model is to predict departure delay on unseen data, the new features engineered will only be from the train set for feature and model selection, and the combination of the train and validation sets for final model training.

For feature selection, given that our model seeks to predict the possible *departure delay* of a new flight, we will not be able to use information related to the actual flight in or around the time of the actual departure, such as the duration of taxi-in, taxi-out and the actual flight. Therefore, we will

only use the **scheduled information** that are known before the day of the flight, including origin and destination airports.

Given that feature selection now leaves us with a few variables such as date of flight and origin/departure airports, the team engineered new features for the prediction models. Inspired by the positive positive correlation between the frequency of flights from an airport and the mean delay of the airport, we use the train set to come up with quantiles to group airports based on their frequency as origins. `origin_qtile` indicates which quantile, out of 10 quantiles, the origin of a flight is in, with `origin_qtile = 10` indicating that the flight's origin airport is in the 90-100th percentile in terms of origin airport frequency. Below shows the origin quantile of 3 airports.

airport	origin_qtile
ABQ	7
ALB	4
AMA	2

Next, we conduct feature extraction for on-time performance of a flight. We define the measure of on-time performance to be the proportion of flights that were delayed (more than 15 minutes) out of all known completed flights in the data. The `flight_num` and `tail_num` features can be used to aggregate the mean proportion of delayed flights in the train set. The new features, `plane_otp` and `flightnum_otp`, could be interpreted as the proportion of delayed flights per individual aircraft and flight number respectively. Any missing values due to the random split not capturing specific flight/tail numbers are imputed with the average on-time performance on the train set.

Below shows the on-time performance of 3 aircrafts respectively.

TAIL_NUMBER	plane_otp
7819A	0.2028169
7820L	0.2631579
N200WN	0.1839378

Furthermore, we believe that the characteristics of the origin and destination airport may affect how the airlines are scheduled to depart and arrive. The EDA in Section 4 shows the popularity of an airport is correlated with delay times. Knowing that all the flights actually form a weighted network where airports are nodes, flights are edges and weights are frequency of routes, we can use the various node centrality measures to capture the different characteristics of airports. The specific node centrality measures used are:

1. Indegree: the number of incoming flights arriving at this airport
2. Outdegree: the number of flights departing from this airport
3. Betweenness: the degree of brokerage of an airport in this network
4. Closeness: the closeness of the airport to all other airports

Below shows the graph features of 3 airports.

ORIGIN_AIRPORT	outdegree	indegree	betweenness	closeness
SFO	106.28235	105.77647	0.0001630	0.5214724
BOS	75.16471	75.14118	0.0004555	0.5483871
LAS	524.90588	523.57647	0.1270750	0.7727273

As our dataset contains features that specifies causes of delays, such as the `WEATHER_DELAY` and `AIR_SYSTEM_DELAY`. We aggregate the `WEATHER_DELAY` by month and origin airports as the weather conditions in different time periods and locations are different. `AIR_SYSTEM_DELAY` is aggregated by origin airports as we are concerned about departure. `AIRLINE_DELAY` is irrelevant as we only focus on the Southwest Airlines only; `LATE_AIRCRAFT_DELAY` is taken into account by the `PLANE_OTP` variable. Below shows the mean weather delay of an airport in 3 different months in 3 different airports.

MONTH	ORIGIN_AIRPORT	MEAN_WEATHER_DELAY
1	ABQ	0.2174688
10	ABQ	0.1962775
11	ABQ	0.3617021

Besides the aforementioned features, there are some features that we explore but do not work well for the logistic regression. As people are more likely to travel during holidays and thus we include the U.S. public holidays in 2015 as a feature; however the feature does not improve our performance and we believe it is because the `MONTH` feature actually captures the holiday effect, both public and school holidays. The `is_weekend` feature does not contribute much as well. The `SECURITY_DELAY`, which is generated by aggregation of `SECURITY_DELAY` by origin airports, only improves the performance minimally due to a nearly 0 correlation with the `DEPARTURE_DELAY` variable.

5.2 Evaluation Metrics

The evaluation metrics we employ are Precision, Recall, F1-Score and Area under Recipient Operating Curve (AUC), with AUC being the main criteria for model selection.

5.3 Model Exploration

As regression models and tree-based models use different approaches in drawing the decision boundary, we believe that we should explore models under both types and select the better one. The two classification models we select to explore are Logistic Regression and Classification Tree.

Due to limited computational resources and the large number of observations in our dataset, we discover that a Random Forest model needs hours to train, even with fewer than 100 trees. Thus, we do not explore ensemble models.

5.3.1 Logistic Regression

The code chunk below builds the logistic regression model based on selected features in the training dataset. The feature 1 to 17 in `use_col` are the selected significant features.

```
lgr_model <- glm(DEP_DELAYED ~ ., data=train[use_col[1:17]], family=binomial)
```

5.3.2 Classification Tree

Due to the limited computational resources, we manually searched for the best `cp` value for the classification tree based on AUC, instead of using grid search with cross validation.

From the performances below, we can see that the best value for `cp` is 0.00001. Then we build the classification tree with the best value.

cp value	AUC on validation set
0.00005	0.6930096
0.00001	0.709647
0.000001	0.6878494

```
southwestTree = rpart(DEP_DELAYED ~ ., data=train[use_col],
                      control=rpart.control(cp=0.00001))
```

5.4 Model Evaluation and Selection

For the baseline model, we classify all data points in the validation set to be 0, the majority class.

model	score
Baseline	0.5000000
Logistic Regression	0.7179081
Classification Tree	0.7131200

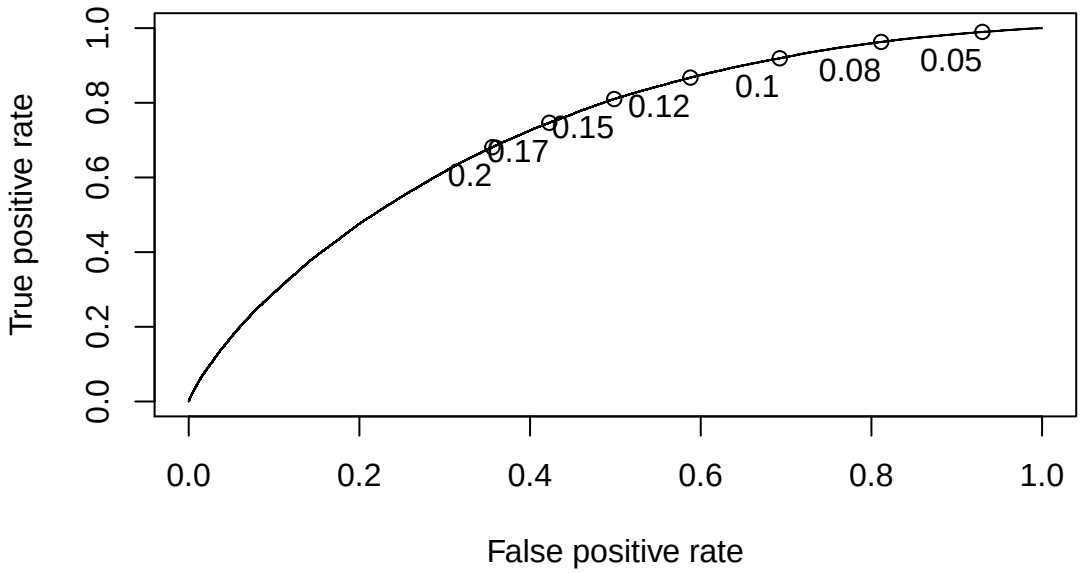
As logistic regression performs better on the validation set and requires much less computational resources to train and predict on new data, we select logistic regression as our final model.

5.5 Final Model Building and Testing

We use the combined train-validation set to build all significant features in the previous Logistic Regression model and train a new one as the final model on the combined set.

```
## [1] "Logistic Regression AUC on test data: 0.7196"
```

Below shows the Receiver Operating Curve of the model performance on test data. We use a threshold range of 0.05 to 0.20 with a step 0.025. From the ROC, we observe that the increase in the False Positive Rate gives a lower increase in the True Positive Rate. As such, the optimal threshold we choose is 0.15. With the threshold, we obtain the confusion matrix (left table) and other performance metrics of our model (right table).



	pred_pos	pred_neg
actual_pos	40925	9596
actual_neg	98677	99226

	score
Precision	0.2931548
Recall	0.8100592
F1	0.4305108

6 Discussion, Limitations and Future Extension

6.1 Business Values to Southwest Airlines

Using the classification model built in Section 5.5, Southwest Airlines can predict which of the future scheduled flights are likely to face delays of more than 15 minutes. We believe that Southwest can apply the output of this model to their business operations planning in the following ways:

1) Southwest can use the prediction output to **better manage customer expectations of delays**, such as by informing passengers ahead of time e.g. having a disclaimer/warning that is displayed to passengers at the time of flight browsing/booking, stating that a particular flight may be susceptible to delays. A management of expectations could help to **reduce dissatisfaction levels** in passengers as such delays will not come as an unpleasant surprise to them, and passengers may have already made plans taking into account the likelihood of delays, thus reducing the disutility they experience. This ultimately helps the company **reduce risks of unhappy customers and damages to its reputation** as a reliable carrier.

2) Prediction of flights more predisposed to delays also allows Southwest to be **more proactive** and pay more attention to such flights and **devise strategies to alleviate delay risks**. This includes possible incentive schemes to encourage earlier check-ins for passengers, or to schedule more buffer time for preflight activities to ensure internal factors do not contribute to delays. Undeniably, there will still be some external factors that are not within the control of the airline that might cause delays. However, by ensuring internal processes are in place, the airline can ensure that delays are not exacerbated.

6.2 Business Values to Airline Industry

The predictive model can serve as a proof-of-concept for other airlines too. Such prediction models can be used for **better financial planning and forecasting**, especially with regulatory requirements that provision passengers with a right to compensation for different severity of delays, or if the airline has such compensation policies as well. Thus by predicting likely delays, the company can **better estimate compensation amounts** that are expected to be paid out.

6.3 Limitation

One limitation of the current analysis is that the 2015 data is slightly outdated and it would be more relevant to perform the analysis against more recent data. Secondly, there are some inadequacies in the data, as the dataset is only limited to 2015 data. As such, some of the flight statistics features (e.g. on-time performance) were engineered on training data records. Ideally, these flight statistics should be calculated based on historical data and validated/tested against a designated "future" point-in-time (e.g. use 2010-2014 data to predict delays in 2015).

Lastly, the final predictive model's performance is still not excellent with the F1-score less than 0.5. The high recall is at the expense of a low precision, implying that there may be many false positives in the model predictions. This might cause an overestimation of compensation amounts (as in the aforementioned use-case of financial planning). Depending on the costs incurred to reduce risk of delays for flights with predicted delays compared to the costs of managing non-predicted delays, low precision may be a concern if the costs for the former is larger than the latter.

6.4 Future Extension

Understanding the potential risk of our model with high recall but low precision, One way to overcome this could be to **use this initial model as an early indicator of flights at risk of delays**. Subsequently, a separate model with more features and points of information could be developed that can serve as a re-assessor as to the likelihood of a delay when more information is available closer to the day of the scheduled flight (e.g. weather, passenger demographics, etc.).

Other future extensions of this predictive analysis could be to **develop a regression model to predict departure delay in minutes**. This could provide more useful information to airlines especially in forecasting flight delay compensation, since delay compensation policies often involve different tiers depending on severity of delay.

7 Conclusion

In summary, through data exploration, we found that flights from busy airports might be more susceptible to delays and that the month in which a flight happens contributes to the likelihood of delays (be it due to seasonality or peak holiday periods). With this knowledge, a predictive model with high recall was created to predict which flights were more likely to be delayed. This can serve as a planning tool for airlines to better pre-empt the occurrence of a delay and thus be able to better manage customer expectations, put in place measures to reduce likelihood or severity of delay, or budget for required compensation to affected passengers.

8 References

Airlines for America (2020). U.S. Passenger Carrier Delay Cos - Airlines For America. Retrieved November 2, 2020, from https://www.airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/

Bureau of Transportation Statistics (2017, December 12). 2015 U.S.-Based Airline Traffic Data. Bureau of Transportation Statistics. Retrieved November 2, 2020, from https://www.bts.gov/newsroom/2015-us-based-airline-traffic-data

CAPA (2010, November 11). Late flights cost US economy USD33 billion. CAPA. Retrieved November 2, 2020, from https://centreforaviation.com/analysis/reports/the-cost-of-delays-late-flights-cost-us-economy-usd33-billion-39215

IBISWorld (n.d.). Domestic Airlines in the US - Market Size. Retrieved November 2, 2020, from https://www.ibisworld.com/industry-statistics/market-size/domestic-airlines-united-states/