Airline Efficiency: Strategies of Flight Delay Forecasts

Team members: Xiyu Li, Xiaoya Huang, Elsie Ye, Wenpei Zhao, Yu Zhao

1. Business Background

Flight delays represent a pervasive issue within the aviation sector, significantly affecting passengers, the functioning and profits of airlines, and airport infrastructure as a whole. Accurate prediction of flight delays is paramount, serving as a catalyst for mitigating financial setbacks, enhancing operational efficiency, and elevating passenger experiences. To make accurate predictions of flight delays, it requires a detailed analysis of historical flight data and other factors such as weather conditions, which can significantly affect flight schedules. [1]

For airlines, predicting delays accurately can lead to better resource allocation, more efficient flight scheduling, and improved customer satisfaction. For passengers, it translates to more reliable travel plans and a better overall travel experience. These outcomes not only diminish potential losses for airlines but also contribute to enhanced profitability. By understanding the reasons behind delays and developing predictive models capable of forecasting these delays, we can identify patterns and propose actionable solutions to mitigate these delays and thus maximize benefits in the future.

2. Introduction

This research project is dedicated to the development of an flight delay forecast product, specifically tailored to the air route connecting Los Angeles International Airport (LAX) and McCarran International Airport in Las Vegas (LAS) and the air route joining John F. Kennedy International Airport (JFK) and LAX — routes distinguished by their high traffic and the prominence of all destinations for both business and leisure travel. Our analysis encompasses the integration of multiple datasets, combining extensive flight data from June 2018 to May 2023 with corresponding weather information for Los Angeles, Las Vegas, and New York during the same period. To facilitate a nuanced understanding and effective utilization of the data, a comprehensive data dictionary has been provided for both the flight and weather datasets to aid in understanding and utilizing the data effectively.

The primary goal of this project is to contribute to the reduction of operational expenses for airlines. This will be achieved by exploring the underlying reasons for various types of flight delays on the selected route and developing predictive models capable of forecasting these delays. Simultaneously, our project aims to empower passengers with valuable scheduling insights, enhancing predictability and efficiency in air travel.

This report delves into the development of model insights and translating them into practical product solutions. We meticulously outline the datasets utilized in our research, providing detailed descriptions of the application of several machine learning models to the prediction process. Our analysis encompasses a range of models, including but not limited to Linear Regression and Random Forest, showcasing the diversity of approaches employed to achieve accurate and reliable flight delay forecasts. [2]

3. Data Understanding

3.1 Dataset Collection

Our project has implemented a comprehensive data collection and processing approach in order to build a powerful predictive model for flights between Los Angeles and Las Vegas and between Los Angeles and New York. It is primarily built upon two crucial datasets: the Flight dataset and the Weather dataset. These datasets encompass comprehensive information related to flights between LAX and LAS, and flights between JFK and LAS. To ensure the reliability and accuracy of our prediction, we obtained monthly flight delay data from the U.S. Bureau of Transportation Statistics (BTS) [3] and daily weather data from Visual Crossing [4], a leading provider of real-time and historical weather information.

Our data collection process gathers historical flight data spanning from June 1st, 2018 to May 31th, 2023, covering features such as flight date, marketing airlines, departure times, arrival times, corresponding delay intervals, cancellation status, delay types, etc. Upon initial retrieval, our raw dataset comprised a staggering 103,237 flight records, featuring a total of 93 distinct attributes. Subsequently, through an extensive data preprocessing phase, we narrowed down the dataset to a comprehensive flight history database with 103,237 records specifically indicating the flight routes between LAX and LAS, and 95,407 records of flight routes between LAX and JFK, retaining 46 essential attributes.

We acquired daily weather data for the same time period, focusing on the weather in Los Angeles, Las Vegas and New York, capturing features like maximum and minimum temperature, wind direction, precipitation, visibility, and atmospheric conditions, which may have a significant impact on flight operations. The daily granularity of this weather data enabled us to take into account subtle changes in weather conditions that impact flight delays.

3.2 Datasets Description

The Flight Dataset

Among the noteworthy discoveries within this dataset, our analysis revealed the dataset contains 14 different operating airlines and 9 different marketing airlines. However, certain airlines are categorized under larger marketing airline companies. For instance, SkyWest Airlines with its airline code "OO," is categorized under all of Alaska Airlines, Delta Airlines, and United Airlines. It is crucial to understand the airline industry's dynamics before and after the prevalence of COVID-19. These distinctions are vital as they provide insights into market share and operational efficiency, both of which can impact passengers' travel preferences.

Further, we explored five distinct delay reasons, some of which were not entirely transparent. Notably, certain weather-related delays were classified under NAS (National Air System) delays, possibly encompassing unforeseen severe weather conditions which have been communicated directly to the airports by the national air system. This discovery sheds light on the multifaceted nature of flight delays. The classification of weather-related delays under NAS delays underscores the role of external factors in causing delays, emphasizing the need for advanced weather forecasting and contingency planning to minimize disruptions, and it leads to further analysis on the delay reasons.

We analyzed delay minutes by airline, duration, and month, categorizing the data for future use. We also examined missing values and outliers to ensure data integrity, crucial for accurate analyses and predictions.

The Weather Dataset

The Weather Data Set has highlighted crucial meteorological features. Wind speed, wind gust, precipitation, visibility, snow condition, severe risk condition and weather conditions icons have been believed to be pivotal elements affecting flight operations. These variables range from the intensity of wind forces to the amount of precipitation, the clarity of visibility, and the overall weather conditions.

According to the explanation provided by Visual Crossing, wind speed was measured at a standard height of 10 miles above the ground in an unobstructed area. [5] In Los Angeles, the maximum wind speed of 42.8 mph and a minimum of 1.3 mph. In Las Vegas, the maximum is 64.5 mph and the minimum is 7.4 mph. In New York, the maximum could reach 50.3 mph and the minimum is 8.0 mph.

Wind gust represents the highest wind speed measured within a short timeframe, typically less than 20 seconds. Los Angeles has a maximum value of 94.7 mph and a minimum of 0 mph; Las Vegas has a maximum of 98.0 mph and a minimum of 20.5 mph; while New York has a maximum of 109.1 mph and a minimum of 13.3 mph. Wind speed and wind gust are critical for flight safety. Excessive wind speeds and gusts can affect passenger comfort and safety. Pilots rely on this data to ensure smoother flights, enhancing the well-being of passengers.

Precipitation is recorded in inches. Los Angeles has a maximum of 69.6 inches, Las Vegas 20.1 in, and New York 71.6in. It is vital for resource allocation at airports. It informs decisions on runway closures and allocation of staff and equipment.

Visibility, the distance that can be seen in daylight expressed in miles. In Los Angeles, the minimum visibility is 3.5 miles; in Las Vegas, it is 11.0 miles; and in New York, it is 2.2 miles.

Snow conditions demonstrate the amount of new snow that has fallen in the time period measured in inches. Las Vegas has a maximum of 0.8 inches of snow, and New York has a maximum of 17.1 inches. Los Angeles, unsurprisingly, has a maximum of 0 inches. This nature is later specially adapted to our analysis, emphasizing the weather characteristics in the two routes.

Severe risk is a value between 0 and 100 representing the risk of convective storms such as thunderstorms, hail or tornadoes. Typically, a severe risk value less than 30 indicates a low risk, between 30 and 70 a moderate risk and above 70 a high risk. According to the data, Los Angeles and Las Vegas have low risk in these five years, while New York posts a strong tendency of having hazardous storms with a maximum value of 100.

Lastly, the icon feature provided an overview of the daily weather conditions, and in Los Angeles and Las Vegas, the datasets indicate a predominance of clear days over a total of 1,858 days; yet, New York has rainy days more than half of the time.

4. Data Preparation

4.1 Data Preprocessing

During the initial stage of our data processing pipeline, we undertook the process of data integration to combine weather and flight information. Subsequently, we executed data cleaning procedures aimed at eliminating disorderly or missing data, thereby upholding the integrity of our predictive models. The accumulated information underwent a transformation into a format comprehensible to our system, akin to the conversion into binary for optimal functionality. Following this data transformation, our focus shifted to feature selection, involving meticulous choices of the most pertinent details. Proper partitioning of these features facilitated rigorous testing of our forecasts, ensuring their accuracy and precision. The bias has been mitigated to the utmost extent within our system, thereby ensuring a fair and unbiased representation of outcomes.

We preprocessed the data, removing missing values and outliers, and integrated flight delay data with daily weather data. We reduced the feature count from 93 to 70 by eliminating 23 empty features and identified numerous missing values as shown in Figure 1.

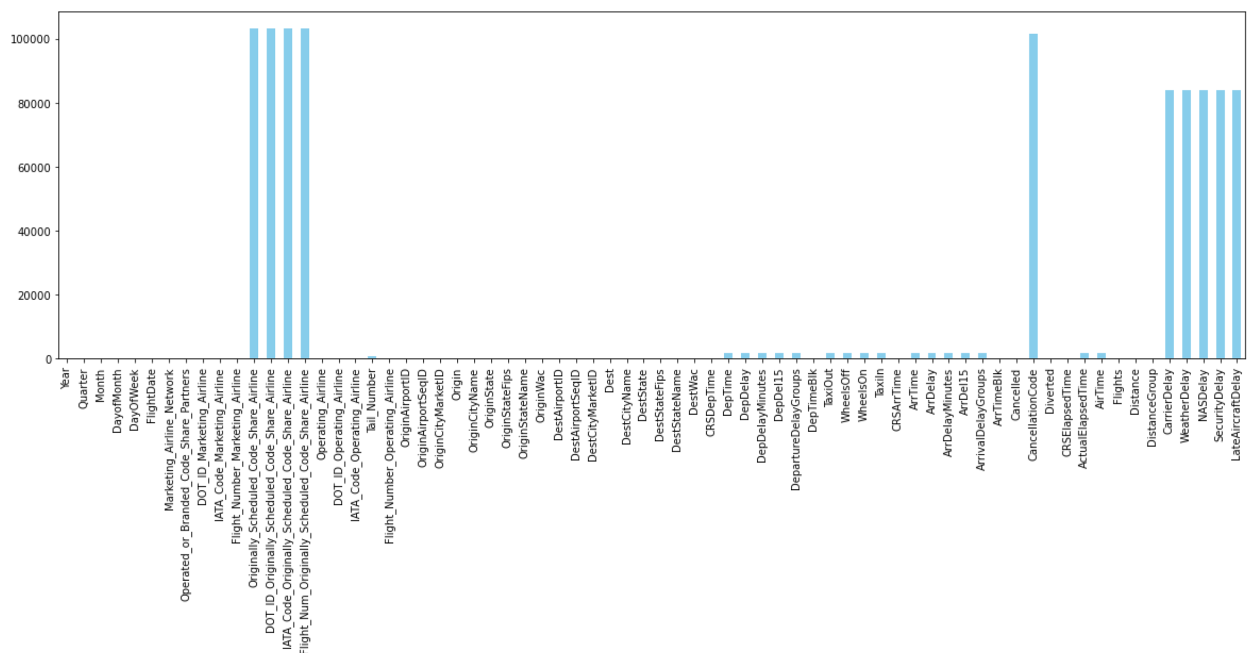

Figure 1

- Originally_Scheduled_Code_Share_Airline 103237
- DOT_ID_Originally_Scheduled_Code_Share_Airline 103237
- IATA_Code_Originally_Scheduled_Code_Share_Airline 103237
- Flight_Num_Originally_Scheduled_Code_Share_Airline 103237

The first 4 features listed above: The number of missing values equals the number of data. This means that there are no entries in these features, and the features are unhelpful for delay prediction and we removed them.

CancellationCode: Missing values in *CancellationCode* are reasonable. By finding the number of canceled flights, we found every canceled flight has its cancellation code. Otherwise, there are no entries.

Last 5 delay reason features: All missing values in these 5 features are in the same rows. Delayed flight without delay reason: By checking the count of canceled/not delayed flights, we found there are around 17000 delayed flights missing delay reasons. But all flights were delayed over 15 mins and only 1862 within 15 mins (total 21074 flights) have delay reasons stated. In our analysis, we consider only flights delayed over 15 minutes as delayed flights. The missing values are reasonable under the situation of small delay minutes.

All departure features have missing values at the same rows. All arrival features have missing values at the same rows, except there are 46 rows that have only ArrTime entries. Thus, we focus on *DepTime*, *ArrTime* and *ArrDelay*, instead of all features.

In 1816 flights that have no *WheelsOn* and *TaxiIn* entries, 1812 flights were canceled, 4 rows were removed. The 10 flights that have *TaxiOut* and *WheelsOff* but no *WheelsOn* and *TaxiIn* entries, were urgently canceled. *ActualElapsedTime* and *AirTime* cannot be estimated accurately. We decided to leave it because of their similar features as ArrTime.

4.2 Exploratory Data Analysis

4.2.1 Correlation Analysis

We focused on a subset of key features to avoid data overload, particularly emphasizing airline delays, our analysis included features such as DepDelay, ArrDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay, DayOfWeek, CRSDepTime, CRSArrTime, Marketing_Airline_Network, Origin, and Dest, as illustrated in Figure 2. Correlations were observed, including a strong positive correlation (0.97) between Departure and Arrival Delays, indicating that delayed departures often lead to late arrivals. A 0.19 correlation between NASDelay and ArrDelay revealed a moderate impact on arrival times, while a 0.10 correlation with DepDelay suggested a slight influence on departure delays.

Unexpected findings included a neutral correlation (−0.03) between WeatherDelay and CarrierDelay, challenging the assumption that weather delays are less likely attributed to carriers. NAS delays, potentially influenced by factors such as traffic control or airport operations, exhibited a slight negative correlation (−0.12) with Late Aircraft Delay, which typically results from aircraft arriving late from a previous destination. Contrary to expectations, DayOfWeek demonstrated low correlations (around 0.01 or less) with most delay types, indicating a negligible impact of specific weekdays on delays.

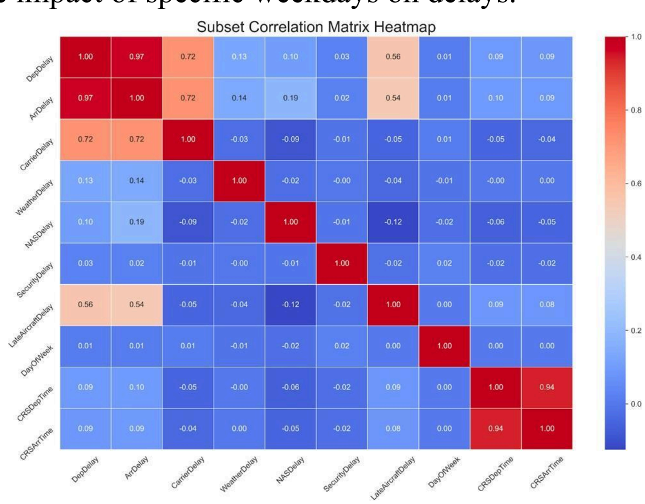

Figure 2

4.2.2 Visual Exploration on Important Features

'ArrDelayMinutes'

We started by examining the distribution of flight delays. Delays under 15 minutes, considered typical and insignificant for daily operations, were omitted from further analyses. Delays from 15 minutes to an hour were less common but notable at 13,394, while those lasting 1-3 hours totaled 4,959. Delays exceeding 3 hours, especially surpassing 6 hours, were infrequent. Addressing delays beyond 15 minutes is crucial to mitigate potential disruptions, as illustrated in Figure 3. Southwest Airlines (WN) stood out with the highest number of delays within 15 minutes, totaling 21,219, while Allegiant Air (G4) recorded the fewest with only 62 flights over five years. Our focus is now on uncovering reasons behind delays longer than 15 minutes.

Further examination involved charts representing the frequency of delayed flights for different airlines across distinct delay periods. Key takeaways include Airline 'AA' ranking among the top two for the most delays, 'WN' predominantly experiencing delays under 3 hours, and delays beyond 24 hours being notably infrequent.

Considering varying total flight numbers, a figure showing the percentage of total vs. delayed flights over 15 minutes for each airline was created as shown in Figure 4. Key findings from this visual data include 'WN' having the highest total flights, 'B6' having the highest delay ratio at 34.18%, 'G4' being the most punctual with the lowest delay ratio at 10.13%, and 'DL' and 'AA' facing operational challenges despite their scale.

In evaluating airline performance, both the absolute number of delays and the delay ratio should be considered. This information is crucial for potential passengers and stakeholders to make informed decisions.

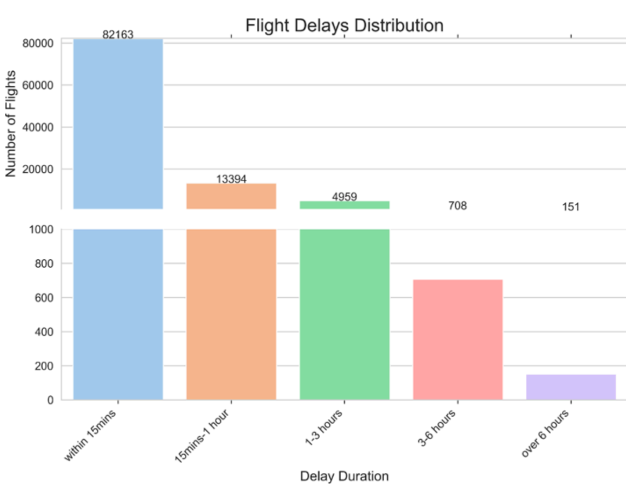

Figure 3

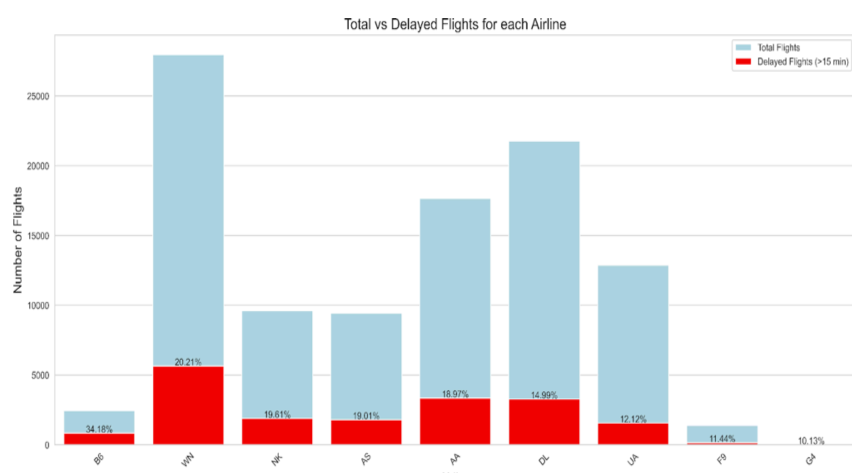

Figure 4

Airline Analysis

The original flight dataset had three airline attributes: Marketing airline, Operating airline, and originally operating airline. The "originally" attribute, intended for scheduled airline changes during flights, contained "NA" as no changes occurred in the selected time period. We removed this column.

Chart 1 summarizes the number of flights per airline and shows the relationship between operating and marketing airlines. For example, while the marketing airline may be AS (Alaska Airlines), some flights are operated by OO, and others by QX. Focusing on the operating airline for its direct impact on flight delays, we observed uneven data distribution among airlines. Some had many flights, while others had few (min: 79).

	Marketing_Airline_Network	Operating_Airline	n
1	AA	AA	17662
2	AS	AS	5479
3	AS	OO	2625
4	AS	QX	1340
5	B6	B6	2455
6	DL	CP	3998
7	DL	DL	9764
8	DL	OO	8019
9	F9	F9	1372
10	G4	G4	79
11	NK	NK	9603
12	UA	OO	3079
13	UA	UA	9792
14	WN	WN	27970

Chart 1

Figure 5 illustrates changes in monthly flights for each operating airline over time between 2020 and 2021, revealing inconsistencies. Nearly all airlines experienced a significant reduction in the number of flights, with the exception of OO Airlines. Airlines opened or closed routes, suggesting potential challenges in predicting delays using both airline and time-related variables simultaneously due to variable correlations.

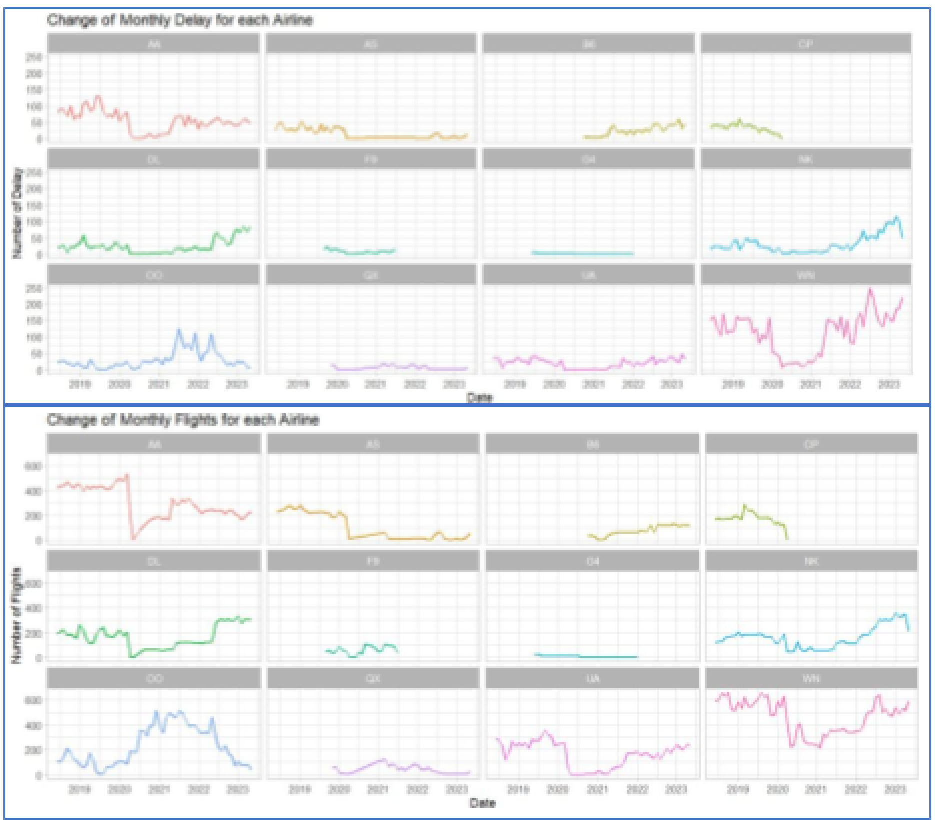

Figure 5

Weather Analysis

We combined the flightALL dataset with weather data based on timestamps and departure locations. Using scatter plots in Figure 6, we explored the relationship between Departure delay minutes and weather variables like WindSpeed, gusts, precipitation, etc. Surprisingly, no consistent relationships were found, encountering outliers like a 900-minute delay despite clear skies recorded for that day.

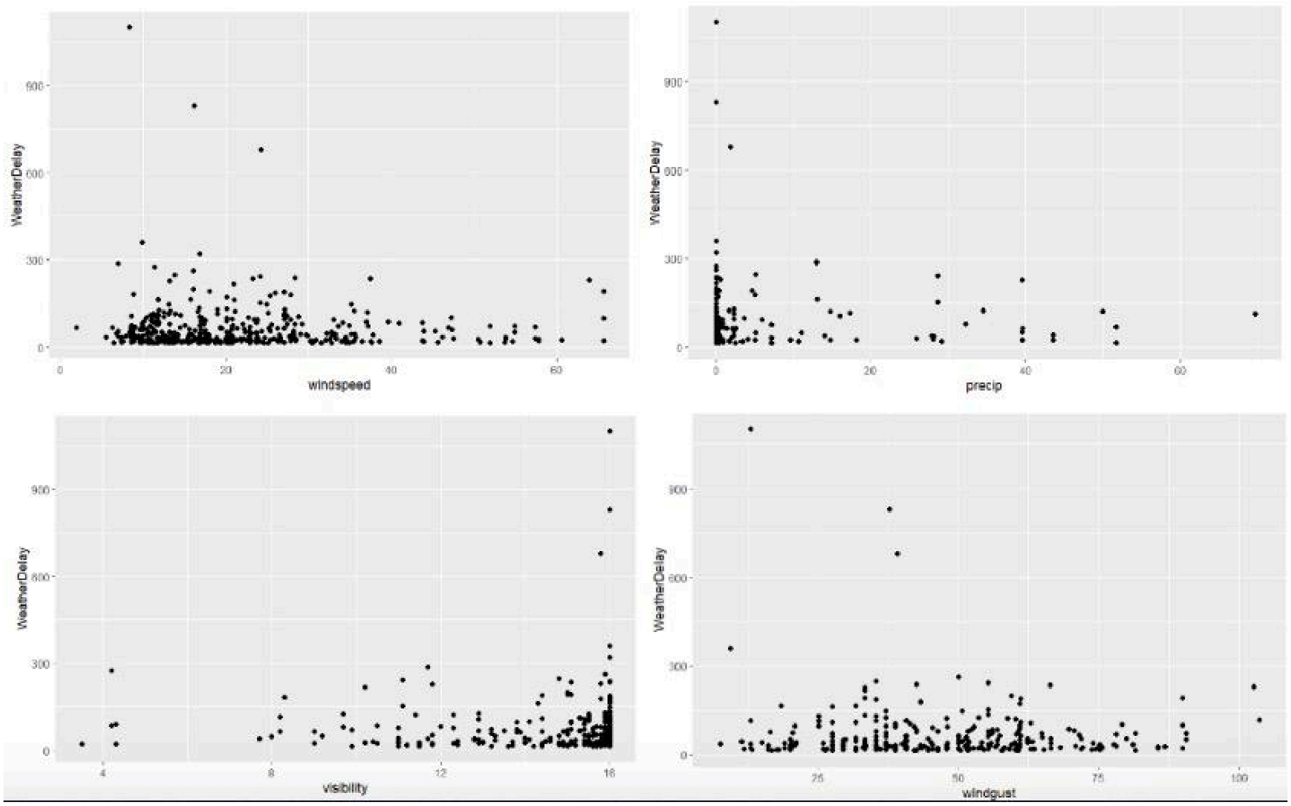

Figure 6

It's crucial to note that extreme weather events are relatively rare in the Los Angeles to Las Vegas route, while the route from Los Angeles to New York reveals a stronger correlation between weather and flight delays as shown in Figure 7.

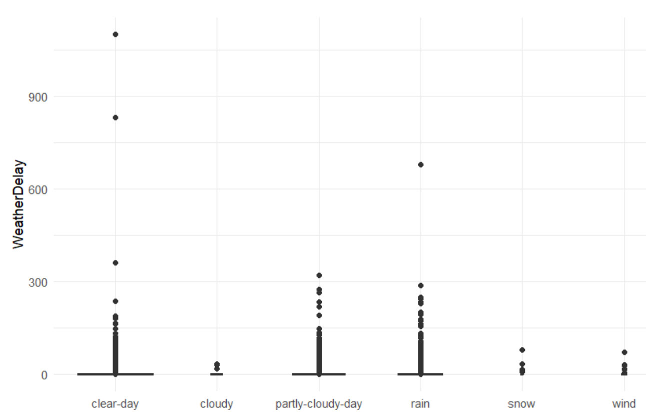

Figure 7

5. Modeling And Evaluations

5.1 Logistic Regression Model

A Logistic Regression Model is a statistical method we use to predict the probability of a flight being delayed. It works by analyzing past data, like weather conditions, previous delays, and airport traffic. The model takes this information and calculates the odds of a delay. Think of it like a sophisticated odds-maker that looks at all the factors and then makes a bet on whether your flight will leave on time or not.[6] In our report, we'll explain how this model uses the data we have to make these predictions and how accurate it has been in forecasting delays for airlines.

We conducted a comprehensive exploration of various feature combinations. Based on their significant impact on the model's predictive performance, three key predictors are selected in the final model: "TimeBlk," "Icon," and "Operating Airline". The modeling coefficients are shown in the below Figure.

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -1.60522    0.01960 -81.903  < 2e-16 ***
Operating_AirlineDL      0.21817    0.02105  10.364  < 2e-16 ***
Operating_AirlineOO     -0.11872    0.03221  -3.686 0.000228 ***
Operating_AirlineUA     -0.19857    0.03482  -5.704 1.17e-08 ***
Operating_AirlineWN      0.62105    0.02183  28.446  < 2e-16 ***
TimeBlkMorning          -0.70735    0.01925 -36.750  < 2e-16 ***
TimeBlkNight             0.09557    0.02010   4.755 1.99e-06 ***
iconcloudy               0.19688    0.06582   2.991 0.002780 ** 
iconpartly-cloudy-day    0.09869    0.01969   5.011 5.40e-07 ***
iconrain                 0.35649    0.02104  16.942  < 2e-16 ***
iconsnow                 0.58160    0.07653   7.600 2.96e-14 ***
iconwind                 0.24441    0.16781   1.456 0.145269    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8

Regarding "Operating Airline," the positive coefficient of "Operating_AirlineDL," 0.21817, suggests an increased likelihood of delays associated with this airline; . Conversely, "Operating_AirlineOO" and "Operating_AirlineUA" exhibit negative coefficients, indicating a decreased likelihood of delays, while "Operating_AirlineWN" shows a positive coefficient, implying a higher likelihood of delays.

Analysis of the "TimeBlk" predictor reveals that flights in the morning (TimeBlkMorning) are associated with a lower likelihood of delays, as indicated by the negative coefficient of -0.70735. Conversely, flights during the night (TimeBlkNight) exhibit a positive coefficient, 0.09557 , suggesting a higher likelihood of delays during this time.

Furthermore, the impact of weather conditions, represented by different icons, plays a significant role in predicting delays. Specific icon types, such as cloudy, partly-cloudy-day, rain, snow, and wind, are associated with varying degrees of increased likelihood of delays.

In terms of model performance, both null and residual deviances, 101193 and 97787, respectively, suggest that the model significantly improves upon a null model. The Akaike Information Criterion (AIC), with a value of 97811, indicates a reasonable trade-off between model fit and complexity.

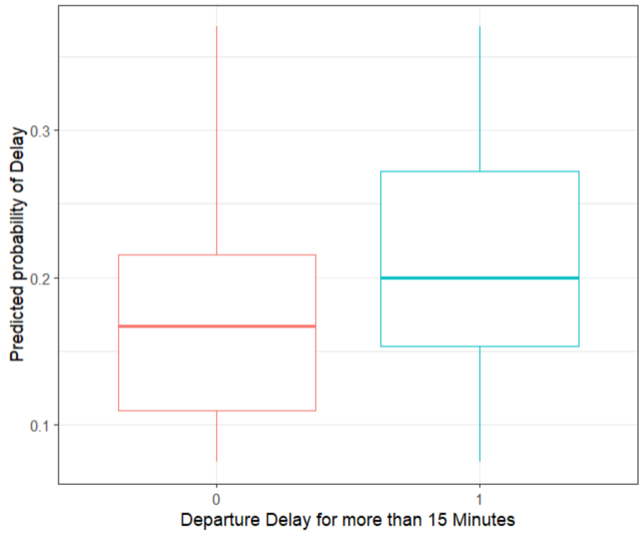

Figure 9

Figure 9 is a box plot illustrating the relationship between our predicted probability of flight delays and the actual on-time information of the flights. Despite the two boxplots having a similar range, a noticeable distinction exists in the medians of each group.

Based on these findings, we established a threshold of 0.2: predictions above 0.2 are classified as delays, while those below 0.2 are deemed on-time. Transforming the results into binary outcomes allows us to compute the True Positive Rate (TPR) and True Negative Rate (TNR) using the following formulas in Figure 10.

$$\text{TPR (Sensitivity)} = \frac{\text{True Positives}}{\text{Actual Positives}}$$

$$\text{TNR (Specificity)} = \frac{\text{True Negatives}}{\text{Actual Negatives}}$$

Figure 10

The computed rates are as follows: If the flight is actually delayed, our model has a 46% chance to predict the delay (TPR); if it is not delayed, our model has a 72% chance to make the correct prediction (TNR).

- True Positives: 2235
- True Negatives: 16040
- False Positives: 6276
- False Negatives: 2572
- True Positive Rate (Sensitivity or Recall): 0.46
- True Negative Rate (Specificity): 0.72

5.2 Random Forests Model

A Random Forests Model is an ensemble learning method that's really good for predicting flight delays. Imagine it as a team of decision-making trees where each tree has a vote in determining if a flight will be late. These trees look at different samples of our data, like weather reports, mechanical issues, or air traffic, to make their decisions. They each come up with their own prediction, and then the most common outcome gets picked as the final answer. In our report, we'll break down how this 'forest' of multiple decision trees leads to a more robust and reliable prediction about flight delays compared to using a single decision tree. The primary objective of our analysis was to predict flight delay minutes, quantified as DepartureDelayGroups with each group representing a 15-minute interval.

To address the challenge of predicting DepartureDelayGroups for flights, our team developed two distinct Random Forest models tailored to the JFK—LAX and LAS—LAX routes. The necessity for separate models arises from the substantial dataset sizes and the unique feature sets pertinent to each route.

The differing climatic conditions between New York City and Southern California necessitated the use of route-specific weather features. For instance, the "preciptype" variable in the dataset for the LAS and LAX routes exclusively contained "rain," whereas the JFK dataset comprised a mix of "freezingrain," "snow," "ice," and "rain," with some entries combining multiple conditions such as "freezingrain, snow, ice" and "rain, snow."

Additionally, the diversity in operating airlines across these routes influenced the one-hot encoding process, resulting in distinct feature sets for each model.

- ☐ LAS—LAX Route Features
  The model for this route incorporated weather-related variables such as "precip," "precipprop," "precipcover," "snow," "snowdepth," "windgust," and "windspeed." Visibility metrics and temporal features like 'Month', "DayofMonth," "DayOfWeek," and "CRSDepTime" were also included. Airline-specific attributes were captured through one-hot encoded "Operating_Airline" and "DepTimeBlk."

- ☐ JFK—LAX Route Features
  This route's model integrated the aforementioned features along with an additional feature 'ice', derived from the "preciptype" to capture the more complex weather patterns affecting New York City.

For the route of JFK—LAX, the Random Forest model's accuracy for predicting the 'DepartureDelayGroups' is approximately 58.12%. The Overall (weighted) evaluation is: F1-score = 0.50, Precision = 0.46, Recall = 0.58.

Looking at the classification report, the model performs best at predicting the '-1.0' class with a precision of 0.64 and a recall of 0.89, indicating a high true positive rate for this class. The F1-score for this class is also the highest at 0.75, reflecting the balanced precision and recall for this class. However, the model does not perform well on some other classes, especially those with fewer instances (lower support).

The top 5 features are: (1) CRSDepTime: 0.171; (2) windspeed: 0.119; (3) windgust: 0.113 ; (4) DayofMonth: 0.107; (5) Month: 0.076

These features represent the variables that contribute most significantly to the predictions of the Random Forest model. The departure time (CRSDepTime) is the most influential feature, followed by weather-related features such as wind speed and wind gust, which are likely related to the potential for delays.

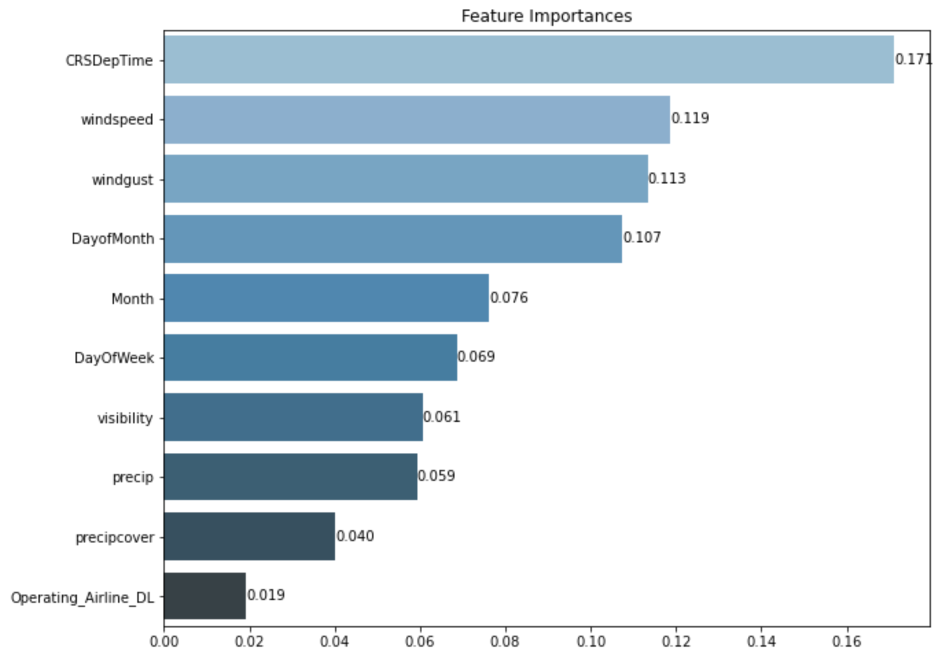

Figure 11

```
Accuracy: 0.5812123360510457
Classification Report:
              precision    recall  f1-score   support

        -2.0       0.00      0.00      0.00        73
        -1.0       0.64      0.89      0.75     17417
         0.0       0.28      0.14      0.18      5885
         1.0       0.09      0.02      0.03      1743
         2.0       0.05      0.01      0.02       837
         3.0       0.04      0.01      0.02       504
         4.0       0.03      0.01      0.01       371
         5.0       0.05      0.01      0.02       254
         6.0       0.00      0.00      0.00       183
         7.0       0.07      0.01      0.02       163
         8.0       0.04      0.01      0.01       139
         9.0       0.00      0.00      0.00        88
        10.0       0.00      0.00      0.00        95
        11.0       0.00      0.00      0.00        69
        12.0       0.08      0.02      0.03       389

    accuracy                           0.58     28210
   macro avg       0.09      0.08      0.07     28210
weighted avg       0.46      0.58      0.50     28210
```

Figure 12

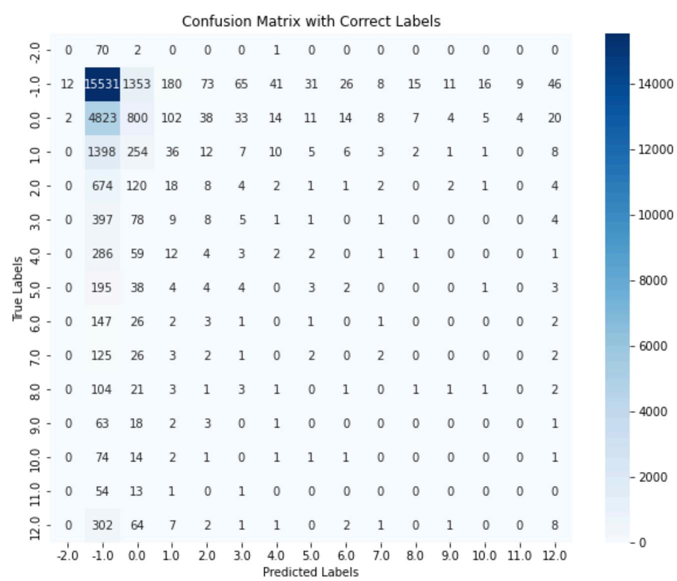

Figure 13

For the route of LAS—LAX, the random forest model's accuracy for predicting the 'DepartureDelayGroups' is approximately 57.61%. The Overall (weighted) evaluation is: F1-score = 0.51, Precision = 0.48, Recall = 0.58.

Results for both models are quite similar, the model shows its highest efficacy in predicting the '-1.0' class, achieving a precision of 0.65, a recall of 0.89, and a F1-score of 0.75.

The top 5 features, according to their importance in influencing the model's predictions, are: (1) CRSDepTime: 0.178; (2) windspeed: 0.139; (3) DayofMonth: 0.126; (4) visibility: 0.107; (5) Month: 0.083

The most critical factor in predicting delays is the scheduled departure time (CRSDepTime), followed by windspeed and DayofMonth, instead of windgust comparing to the previous model.

Additionally, for both models, the confusion matrix indicates that the model is quite good at predicting early departures (class '-1.0').

```
Classification Report:
              precision    recall  f1-score   support

        -2.0       0.00      0.00      0.00        48
        -1.0       0.65      0.89      0.75     17743
         0.0       0.36      0.24      0.29      6881
         1.0       0.14      0.05      0.07      2155
         2.0       0.08      0.02      0.03      1105
         3.0       0.08      0.02      0.03       729
         4.0       0.04      0.01      0.02       486
         5.0       0.04      0.01      0.02       300
         6.0       0.06      0.01      0.01       245
         7.0       0.00      0.00      0.00       162
         8.0       0.00      0.00      0.00       130
         9.0       0.06      0.01      0.02       100
        10.0       0.00      0.00      0.00        57
        11.0       0.00      0.00      0.00        48
        12.0       0.08      0.02      0.03       238

    accuracy                           0.58     30427
   macro avg       0.11      0.08      0.08     30427
weighted avg       0.48      0.58      0.51     30427
```

Figure 14

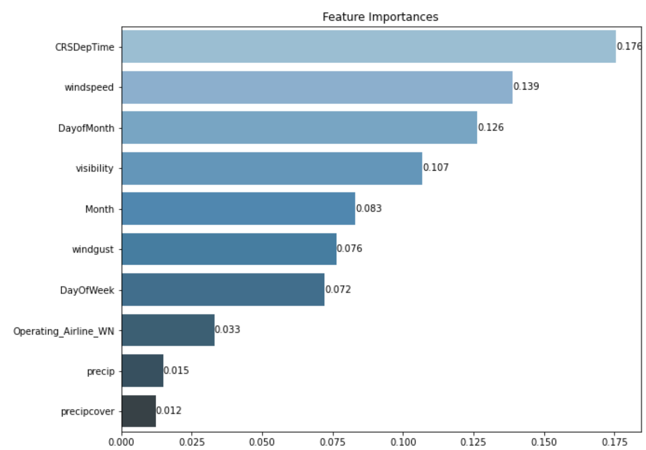

Figure 15

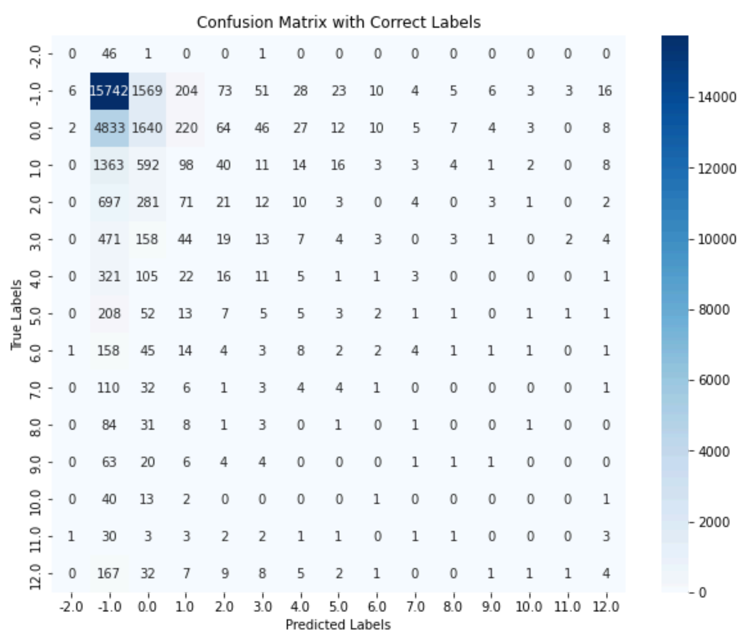

Figure 16

5.3 Key Findings of Modeling

When it comes to predicting flights that are on schedule or even ahead of time, both models perform notably well. They're much better at figuring out when a flight will leave on time (which we categorize as DepartureDelayGroup 0) or early (DepartureDelayGroup -1), rather than when there will be delays. Key elements like the planned takeoff time, wind speed, and how clear the sky is, play a big role in this prediction process. These factors highlight just how much flight times are at the mercy of scheduling and the weather.

However, there's a hitch when we expect these models to forecast longer hold-ups. Their effectiveness dips for predicting more substantial delays, falling off as we look at DepartureDelayGroups ranging from 1 to 12. This suggests that while the models are sharp at catching flights likely to be on the dot or quicker, they struggle to nail down the lengthier delays with lower performance metrics for DepartureDelayGroups 1 through 12.

6. Deployment to Business Product

6.1 Cost Reduction

The Cost Calculation Methodology we used is referred to by Professor Andrew that each minute of flight delay results in an estimated airline loss of approximately $75.
The financial implications are assessed based on the following formula:

$$\textit{Original Cost} = \textit{Delay minutes} \times \textit{\$75/minutes}$$

Before actual prediction took place, several adjustments were made, and have been reflected on the original formula as below.

- ☐ Cost when predicted delay underestimated (i.e. actual delay minutes exceed predicted delay minutes)
  $$\textit{Cost UnEst} = (\textit{Actual Delay Minutes - Predicted Delay Minutes}) \times \textit{\$75/minute}$$
- ☐ Cost when predicted delay overestimated:
  $$\textit{Cost OvEst} = (\textit{Predicted Delay Minutes - Actual Delay Minutes}) \times \textit{\$75/minute}$$

Thus we have:

$$\textit{Total Cost} = \textit{Cost UnEst} + \textit{Cost OvEst} = |\textit{Predicted Delay Minutes - Actual Delay Minutes}| \times \textit{\$75/minute}$$

In our financial analysis and predictive modeling, we have adopted a convention that treats any flight that departs earlier than scheduled, signified by negative delay minutes, as an on-time departure. Consequently, both actual and predicted delay minutes that are negative are reassigned a value of zero. This convention aligns with industry practices where early departures are typically not penalized and are considered equivalent to on-time performance.

In accordance with our predictive model's framework, a delay is defined as a flight departure that exceeds a 15-minute threshold. Consequently, our cost analysis incorporates the following adjustments:

No-Cost Threshold Adjustment: Actual Delay Minutes that are less than or equal to 15 minutes are assigned a cost impact of $0.

Predictive Cost Adjustment: Consistent with the above threshold, if the absolute difference between Predicted Delay Minutes and Actual Delay Minutes is equal to or less than 15 minutes, the cost impact is again set at $0. This adjustment aligns with the operational tolerance for flights departing within 15 minutes of the predicted time frame, either earlier or later.

For the LAX—LAS route, the test Set Size is 30427 with the original cost of $25,679,550. After prediction, the cost estimation is $19,506,375. Thus the money saved is $6,173,175 and the money saved per flight is 202.88 (Est). Based on the 1721 monthly flights on average, the estimated monthly saving for this route is $349,164.70.

For the JFK—LAX route, the test Set Size is 28210 with the original cost of $26,904,900. After prediction, the cost estimation is $21,403,125. Thus the money saved is $5,501,775 and the money saved per flight is $195.03. Based on the 1590 monthly flights on average, the estimated monthly saving for this route is $310,096.50.

The operational expenses associated with an application can vary significantly, ranging from several hundred dollars monthly for a modest and uncomplicated application to tens of thousands of dollars per month for a more intricate APP with an extensive user base.

6.2 Key Findings of Cost Estimation

The economic advantages stemming from the implementation of the delay prediction model are considerable. The monthly cost savings for the scrutinized routes significantly surpass the projected operational expenses associated with our APP. It is strongly recommended that the airline industry evaluates the adoption of this model to streamline operational processes and augment customer contentment by improving the management of flight delays.

It is imperative to underscore that our present approach to calculating benefits presupposes that precise delay predictions will lead to a reduction in costs from 75 to 0. Nevertheless, in practical application, further data may be necessary to determine the actual magnitude of these cost savings.

7. Conclusion

While the model's precision in forecasting longer delays needs improvement, the financial analysis reveals that even with the current accuracy level, significant cost savings are realized. The model effectively reduces the cost associated with flight delays, specifically within the crucial 15-minute early or on-time window, which represents a majority of the flight operations.

By focusing on the model's current strengths and the associated financial benefits, we can confidently advocate for its integration into operational workflows. Concurrently, we recommend ongoing model refinement to extend its predictive accuracy across all delay durations, thereby further enhancing cost-saving potentials.

8. Challenges And Next Steps

We encountered several challenges throughout our project, particularly in dealing with a vast dataset comprising 33 million records and over 140 features. The extensive preprocessing efforts required during the initial phase were substantial.

One major limitation we faced was the restricted access to operational and performance data from airport and airline companies. This data deficiency resulted in the loss of valuable insights into certain features, such as the aircraft's condition and the passenger count for each flight, which are closely tied to Carrier delay. Acquiring more detailed data in these areas would significantly enhance the accuracy of our model, thereby fostering greater economic savings.

In future work, we will need support in obtaining access to comprehensive operational and performance data from airports and airlines. This would greatly contribute to the refinement and improvement of our predictive model. Access to additional details on aircraft conditions and passenger counts would enable us to capture more nuanced patterns and better understand the factors influencing flight delays. Your collaboration in acquiring such data could substantially elevate the effectiveness and economic impact of our predictive modeling efforts.

References

[1] Passy, Jacob. “Flight Delays and Other Uncertainties Slow Business Travel's Rebound - WSJ.” *The Wall Street Journal*, 18 October 2023, https://www.wsj.com/lifestyle/travel/business-flights-delays-lounges-11ae091c. Accessed 20 December 2023.

[2] Truong, Dothang, et al. “Applications of Business Analytics in Predicting Flight On-time Performance in a Complex and Dynamic System.” . - *YouTube*, 24 June 2019, https://www.jstor.org/stable/10.5325/transportationj.57.1.0024. Accessed 20 December 2023.

[3] “Browse Statistical Products and Data.” *Bureau of Transportation Statistics*, 21 November 2023, https://www.bts.gov/browse-statistical-products-and-data. Accessed 20 December 2023.

[4] “Historical Weather Data & Weather Forecast Data.” *Visual Crossing*, https://www.visualcrossing.com/weather-data. Accessed 20 December 2023.

[5] “Weather Data Documentation” *Visual Crossing*, March 2023, https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/. Accessed 20 December 2023.

[6] McSwine, Damien. “.,.” *., - YouTube*, 7 November 2023, https://www.sciencedirect.com/topics/computer-science/logistic-regression. Accessed 20 December 2023.