**MGTA601 - Machine Learning for Business**

**Term Project: Final Report**

**Sarah Mansoor, Meet Kothari, Albert Yu**

**December 14, 2022**

**Table of Contents**

[**1 Business Problem and Motivation**](#_jmfmarm3qn3r) **3**

[**2 Dataset**](#_l0wn0ta6nxao) **3**

[**3 Analysis**](#_10rc299ouihk) **4**

[3.1 Exploratory Analysis](#_rwo9kzjkxfkn) 4

[3.2 Machine Learning Models](#_287eg9f3jlxm) 5

[3.2.1 Linear Regression](#_667ty0rcm0ww) 5

[3.2.2 Logistic Regression](#_fkiuttx2sdtp) 6

[3.2.3 Regression Tree](#_fktg4m18wi7l) 6

[3.2.4 Support Vector Machine (SVM)](#_9nxb1xotg85x) 6

[**4 Results**](#_4xxd1mpj2qjx) **6**

[4.1 Exploratory Analysis](#_csk89bxpnpex) 6

[4.2 Linear Regression](#_km4ks3w4cc1u) 7

[4.3 Logistic Regression](#_tt9ju5bcyvaa) 7

[4.4 Regression Tree](#_avo14xwk860n) 7

[4.5 Support Vector Machine](#_e9xw3uvsck4q) 7

[**5 Outcome and Recommendations**](#_9snlkig6x7bz) **7**

[**Appendix**](#_x9822voelhfj) **7**

# **1 Business Problem and Motivation**

Flight delays are a common occurrence in the world of air travel. They can be caused by a variety of factors, including weather, technical issues with the airplane, air traffic control issues, and more. Predicting flight delays is a significant challenge for the aviation industry, as it can have a major impact on the operations of airlines and the travel plans of passengers. For airlines, flight delays can lead to increased costs due to the need to provide accommodations and alternative transportation for affected passengers, as well as lost revenue from missed connections and cancelled flights. For passengers, flight delays can cause inconvenience and frustration, especially when they have important plans or tight schedules.

Developing a model that can accurately predict the likelihood and duration of flight delays, based on available data, can help airlines to better manage their operations and resources and can help passengers to make informed decisions about their travel plans. The motivation for solving this problem is to improve the efficiency and reliability of air travel, and to provide a better experience for passengers.

Additionally, governments and regulatory bodies may use flight delay predictions to monitor the performance of airlines and identify potential issues that need to be addressed. If an airline wants to reduce their flight delay timings to increase customers’ satisfaction and improve its competitiveness, it is essential for them to reduce their flight delays to less than the industry standards of 27 minutes. Thus, our central goal is to understand the flight delays of major American airlines from 2012 to 2017.

# **2 Dataset**

The dataset used for this flight delay prediction includes a variety of relevant information, such as flight details (such as the origin, destination, operating airline, and scheduled departure and arrival times), taxi information (such as taxi in and out times), wheels on and off information, and delay information. The variables "taxi in" and "taxi out" refer to the times when the airplane begins and ends its taxiing from the terminal to the runway, respectively. The "wheels on" and "wheels off" variables refer to the times when the airplane's landing gear is lowered and raised, respectively. These variables can be used to calculate the time the airplane spends on the ground at the airport, which can be an important factor in determining the likelihood of a flight delay. Delay information includes the following: carrier delays, which are delays caused by the airline (e.g. due to staffing or equipment issues); weather delays, which are delays caused by adverse weather conditions; NAS (National Airspace System) delays, which are delays caused by issues with the air traffic control system; security delays, which are delays caused by security concerns; and late aircraft delays, which are delays caused by the previous flight on an aircraft arriving late.

The data was carefully cleaned and preprocessed to ensure its quality and reliability. Additionally, the dataset is robust, meaning it is large and diverse enough to accurately capture the complex relationships between the various factors that can influence flight delays. To ensure the robustness of the data, the data includes a wide range of flights from different airlines, airports, and routes in the dataset. In this report we will narrow down our analysis to four major airlines in America: American Airline, Delta Airlines, Southwest Airlines, and United Airlines. We also focus on the following major states: New York, California, Florida, and Texas; and the following major cities: Miami, Orlando, New York City, Chicago, Dallas, Los Angeles, San Diego, and San Francisco.

Our data contained about 6.5% missing values which were removed, which is totally acceptable given that we have 102,515 rows even after removing the missing values. One risk within our dataset is that the data available is only for January of 2022. So, we will not be able to analyze the effects of different months and the trends over years in our prediction.

# **3 Analysis**

## **3.1 Exploratory Analysis**

Exploratory analysis is a type of data analysis that is used to uncover patterns, trends, and relationships in a dataset. In the context of flight delay prediction, exploratory analysis could be used to investigate the relationships between flight delays and various factors such as the day of the month, the day of the week, the airline, the state or city of origin or destination, and so on. This analysis could also be used to investigate the relative contributions of different types of delays (such as carrier delays, weather delays, NAS delays, security delays, and late aircraft delays) to the overall flight delay time.

For example, an exploratory analysis could show that flight delays are more common on certain days of the week (such as on Mondays and Fridays). It could also show that certain airlines are more likely to experience delays than others, or that certain states or cities have a higher incidence of delays due to weather or other factors. By uncovering these patterns and relationships, exploratory analysis can help airlines and passengers better understand the factors that contribute to flight delays and make more informed decisions about their travel plans.

## **3.2 Machine Learning Models**

There are several different types of machine learning models that can be used for flight delay prediction. The ones we will use include linear regression, logistic regression, regression tree, and support vector machine (SVM).

### **3.2.1 Linear Regression**

Linear regression is a statistical method that can be used to model the relationship between a dependent variable and one or more independent variables. In the context of flight delay prediction, a linear regression model could be used to analyze the relationship between flight delay times and various factors that may influence them, such as the distance of the flight, the time of day, the airline, and so on. By fitting a linear regression model to this data, it would be possible to make predictions about the expected flight delay time for a given set of input variables.

### **3.2.2 Logistic Regression**

Logistic regression is a type of regression analysis that is used to predict binary outcomes, in our case this is whether a flight will be delayed more than 3 hours or not. This model can be used to analyze the relationship between flight delays and various factors that may influence them, such as the airline or the weather conditions.

### **3.2.3 Regression Tree**

Regression tree is a type of decision tree that can be used to model the relationship between a dependent variable and one or more independent variables. This model can be used to make predictions about the expected flight delay time by dividing the data into different segments based on the values of the input variables.

### **3.2.4 Support Vector Machine (SVM)**

Support vector machine (SVM) is a type of machine learning algorithm that can be used for classification and regression tasks. In the context of flight delay prediction, an SVM model could be trained on a dataset of flight delay data to learn the relationship between the input variables (such as the distance of the flight or the arrival time) and the target variable (flight delay time). This model could then be used to make predictions about the expected flight delay time for a given set of input variables.

# **4 Results**

## **4.1 Exploratory Analysis (Sarah)**

To start off our exploratory analysis, we will look at the distribution of the variables that we are interested in: departure delay, taxi in and out times, flight elapsed time, distance, and various types of delay (carrier, NAS, security, weather, late aircraft, and arrival delay). We calculated summary statistics for each factor such as mean, first, second and third quartile, minimum, and maximum for each variable to get a better understanding of their characteristics (See Appendix 6.1.1). The mean delay time is 65 minutes, with a maximum of 2512 minutes and a minimum of -27 minutes (suggesting the flight left earlier than the departure time). A flight is typically delayed by an hour and we will further explore the factors that may influence this time. The mean for taxi in times is 10.6 minutes whereas the mean for taxi out times is 23.32 minutes. There could be several reasons for the difference in mean taxi in and taxi out times. One possibility is that the airport infrastructure is not designed to efficiently handle the movement of planes from the gate to the runway and vice versa. This could lead to longer taxi out times compared to taxi in times. The mean for the distance traveled is 815 miles and in the United States the mean is likely to vary depending on the specific routes and destinations involved. The United States is a large country with many major cities located far apart, so it is not surprising that the mean distance traveled by flights within the country is relatively high. Arrival delay time has the highest mean of 69 minutes, carrier delay has the second highest mean of 31 minutes and late aircraft delay has the third highest mean of 23 minutes. Arrival delay time seems to be the delay that makes up most of the delay time as we saw the mean for that was 65 minutes. Arrival delay time can be influenced by factors that are outside the control of the airline, such as congestion at the destination airport or delays in the baggage handling process. All of these factors can contribute to a higher mean arrival delay time. It is important to note that the specific reasons for high arrival delay times can vary from case to case and further analysis would be needed to determine the exact causes.

When we look at the amount of flight delay time by day of the month, we can see the increased number of flights during the middle of the week, as well as other factors such as weather conditions and airport congestion (See Appendix 6.1.2). However, without further information and data, it is not possible to accurately determine the reasons for the observed pattern in flight delay times. It is important to carefully analyze all relevant data and consider a variety of potential factors in order to accurately understand and predict flight delays which we will do using machine learning models.

Next, if we look at flight delay time by day of week, we can see that mondays tend to have higher delay times (See Appendix 6.1.3). It is possible that the higher flight delay times on Mondays could be due to increased travel on that day due to people traveling back from trips on mondays or people traveling for business and work related events.

American airlines have the highest time for flight delay and United airlines has second highest time for flight delay time (See Appendix 6.1.4). Which could be due to the high volume of flights that these airlines operate, which can lead to more opportunities for delays. Additionally, these airlines may have a larger number of flights to and from busy airports, which can also contribute to delays. Southwest airlines have the lowest flight delay time which could be due to efficiency of the airline's operations, the routes it operates, and the types of aircraft it uses. Southwest Airlines has a reputation for being one of the most efficient airlines in the industry, with a focus on on-time performance and customer satisfaction. Southwest Airlines also primarily uses Boeing 737 aircraft, which are known for their reliability and efficiency.

## **4.2 Linear Regression (Meet)**

According to the analysis of a linear regression model, flights from Dallas TX, Los Angeles CA, New York NY, Orlando FL, and San Francisco CA were found to be the most significant in terms of P-value (See Appendix ). This is likely because these cities are home to some of the busiest airports in the United States and have a high frequency of flights. United Airlines and Southwest Airlines were also identified as significant factors in flight delays. The 18th of each month was found to be the day with the most delays which could be due to increased travel demand, severe weather conditions, and increased airport congestion. It is possible that the 18th falls at a time when many people are planning to travel for the weekend, leading to an increase in flights and potential delays. It is important to note that the specific reasons for the higher number of delays on the 18th of each month can vary and further analysis would be needed to determine the exact causes. Additionally, factors such as taxi in, taxi out, wheels off, wheels on, distance of the flight, and the types of delays mentioned in our explanatory analysis were found to be significant in the model. The model had a high R-squared value of 0.985, indicating that it was highly accurate in predicting flight delays.

## **4.3 Logistic Regression (Albert)**

In the logistics regression model, we used a step function to identify the significant features that increase the likelihood of a flight being delayed by more than 3 hours. This time frame was chosen because airlines are often required to compensate passengers if the flight is delayed by more than 3 hours. By analyzing these significant features, we can better understand which factors are most likely to lead to long flight delays and how they can be addressed to improve flight performance. This information can be useful for airlines and airport operators to manage their operations and for passengers to plan their travel accordingly. It can also help identify potential issues and inefficiencies in the aviation industry and guide efforts to address them. These features were found to be the operating airline Delta, taxi in and out times, carrier delay, weather delay, NAS delay, and late aircraft delay. In summary, flights from Delta Airline are more likely to be delayed. Additionally, having any of the identified types of delays increases the likelihood of a flight being delayed. The model's null deviance had a value of 3122.77 with 8303 degrees of freedom, and the residual deviance had a value of 135.58 with 8294 degrees of freedom. The Chi-Square test was applied to the model, resulting in a p-value that was much less than 0.05. This indicates that the model is highly effective in predicting the probability of a flight being delayed.

## **4.4 Regression Tree (Albert)**

In the regression tree model, each terminal node shows the predicted delay time for flights in that node, along with the number of observations from the original dataset that belong to that node. The regression tree model partitions the data into 13 parts, and the expected delay time for each part is estimated by the average delay time in that part. For example, according to the appendix B.5, there were 3% of the flights from the dataset with carrier delay less than 371 minutes and late aircraft delay between 116 and 205 minutes, and their average delay was 178 minutes.

The regression tree model can be used to identify which factors are most influential in determining flight delay times. For example, the model may show that flights with a long distance or flights that operate during busy periods at the airport are more likely to be delayed. This information can be useful for airlines and airport operators to identify potential issues and take steps to address them. For example, an airline may decide to adjust its flight schedules or use faster aircraft on long-distance flights to reduce the likelihood of delays.

Additionally, the regression tree model can help to identify patterns and trends in flight delay times. For example, the model may show that flights from certain cities or operated by certain airlines are more likely to be delayed at certain times of day or on certain days of the week. This information can be used to develop strategies for mitigating delays and improving the overall performance of the aviation industry.

Overall, the regression tree model is a powerful tool for understanding and predicting flight delay times. By applying this model, airlines and airport operators can make informed decisions to improve the performance of their operations and provide a better experience for passengers.

## **4.5 Support Vector Machine (Meet)**

In the SVM model, the following independent factors were used to predict flight delay time: taxi in and out times, wheels on and off times, arrival time, arrival delay, distance, carrier delay, weather delay, NAS delay, security delay, and late aircraft delay. The accuracy of this model was around 93.64% with sensitivity at 93.59% and specificity at 1. This was achieved by changing the threshold from 0.5 to 0.45, which increased the accuracy of the model without significantly impacting the overall accuracy. When the threshold was changed to 0.3, the accuracy and sensitivity improved even further, reaching 94.17% and 94.09%, respectively. The specificity remained unchanged. These results indicate that changing the threshold can improve the accuracy and sensitivity of the SVM model, and further optimization may be possible by adjusting the threshold value. The appendix B.6 provides more detailed information on the model's performance.

# **5 Outcome and Recommendations**

### Based on the results of the above models, it is clear that several factors can affect flight delay times, including the origin and destination cities, the operating airline, the day and time of the flight, and various other factors such as weather conditions, airport congestion, and the distance of the flight. By analyzing these factors, the models were able to make predictions about the likelihood of flight delays with a high degree of accuracy.

### One of the key recommendations that can be derived from these models is the importance of addressing the factors that are most likely to lead to flight delays. For example, airlines operating in cities with high volumes of air traffic or at airports with congestion may need to implement strategies to improve their on-time performance. Airlines may need to consider the specific routes and times of day that are most susceptible to delays and take steps to mitigate potential issues.

In addition to the recommendations mentioned above, there are several other steps that airlines and airport operators can take to minimize flight delays. Some of these steps include:

* Investing in technology and infrastructure to improve air traffic management and reduce congestion at airports.
* Working with meteorologists and weather forecasters to develop strategies for dealing with adverse weather conditions.
* Collaborating with other airlines and airport operators to share information and resources and coordinate operations to reduce delays.
* Providing passengers with timely and accurate information about flight delays and any potential compensation they may be entitled to.
* Training pilots and other aviation professionals to handle potential issues that can cause delays, such as technical problems with the aircraft or issues with air traffic control.

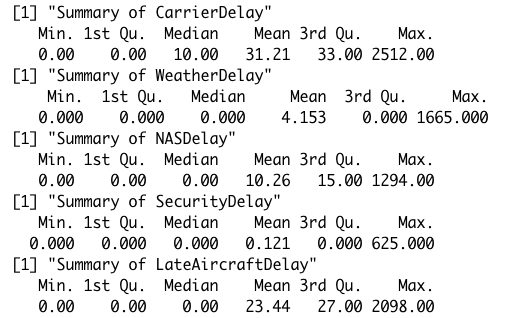
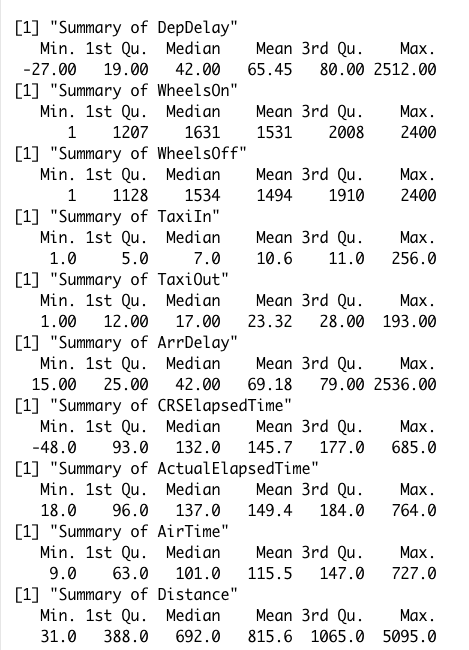
By implementing these and other measures, airlines and airport operators can help to reduce the number of flight delays and improve the overall performance of the aviation industry. This can benefit passengers by providing them with a better travel experience, and it can also help airlines and airports to save time and money and improve their reputations.

### Overall, the results of these models highlight the potential value of using machine learning techniques to predict and manage flight delays. By applying these techniques, airlines and airport operators can better understand the factors that affect flight performance and take action to improve it. This can help to enhance the customer experience, improve operational efficiency, and reduce the costs and disruptions associated with flight delays.

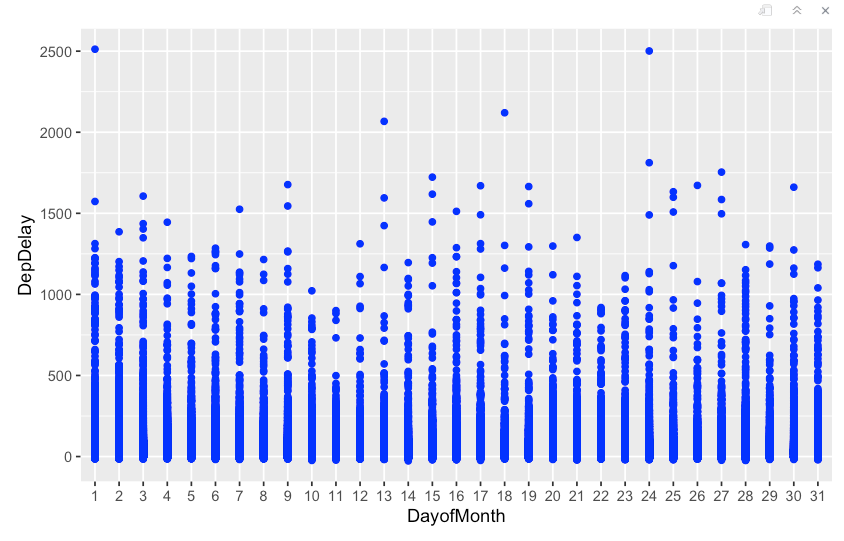
# **6 Appendix**

**6.1 Exploratory Analysis**

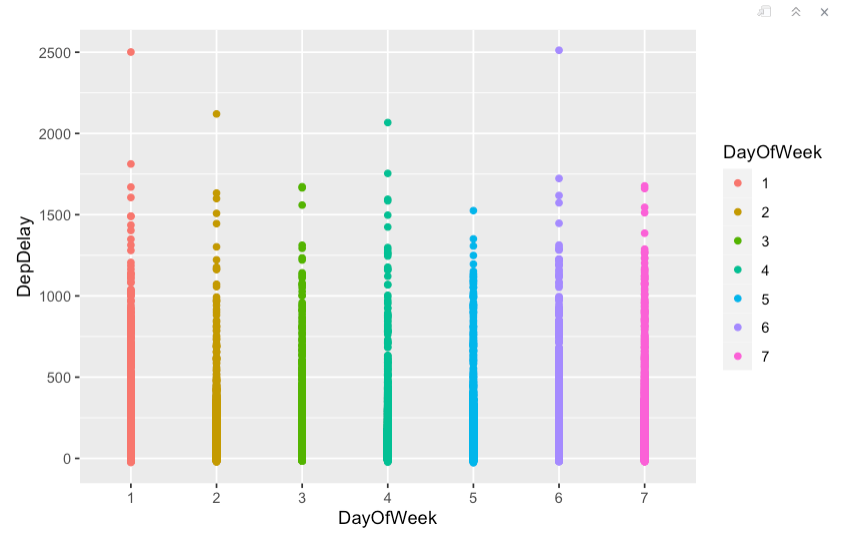
6.1.1



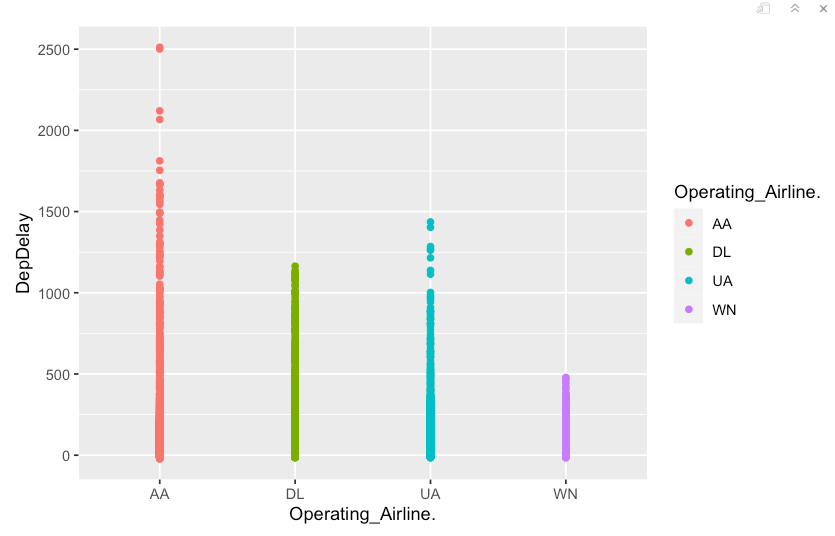
6.1.2



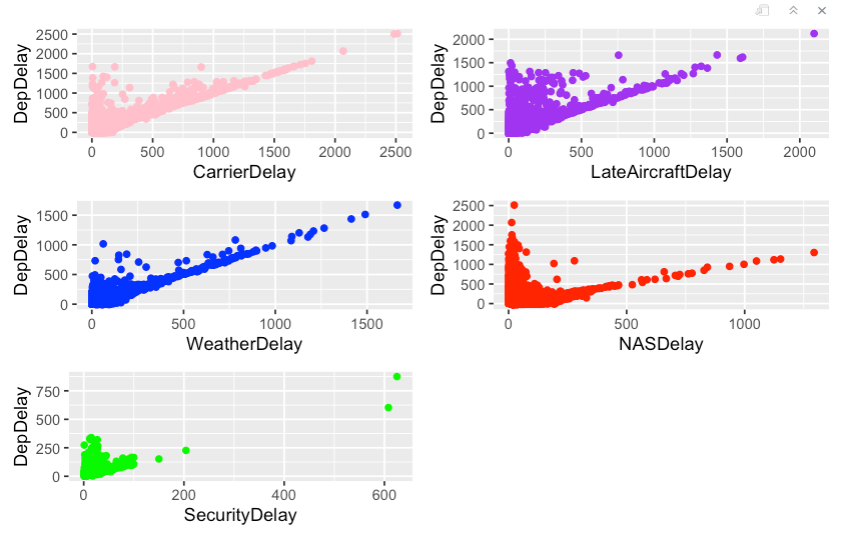
Appendix A.3



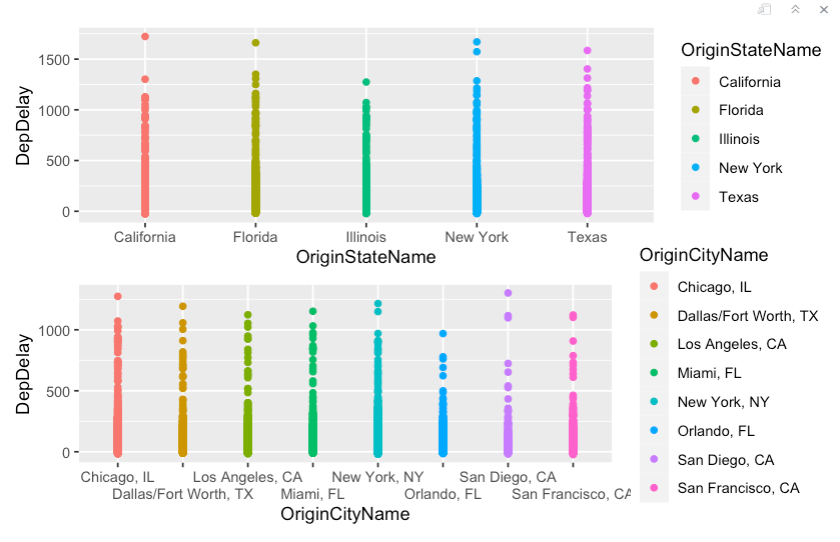
Appendix A.4



Appendix A.5

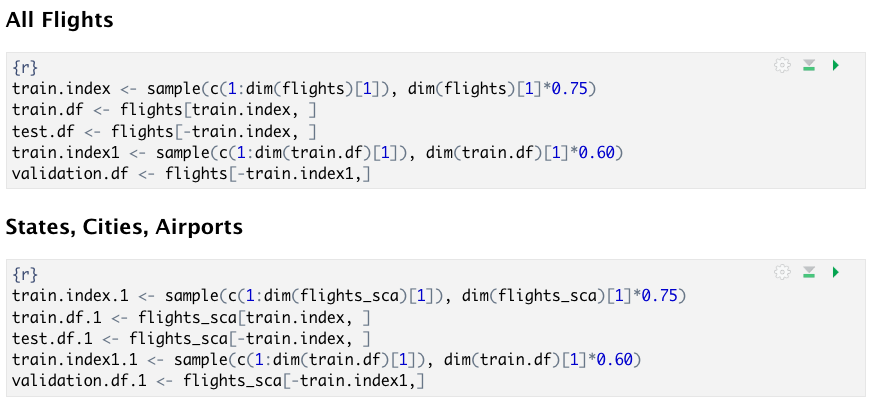


Appendix A.6

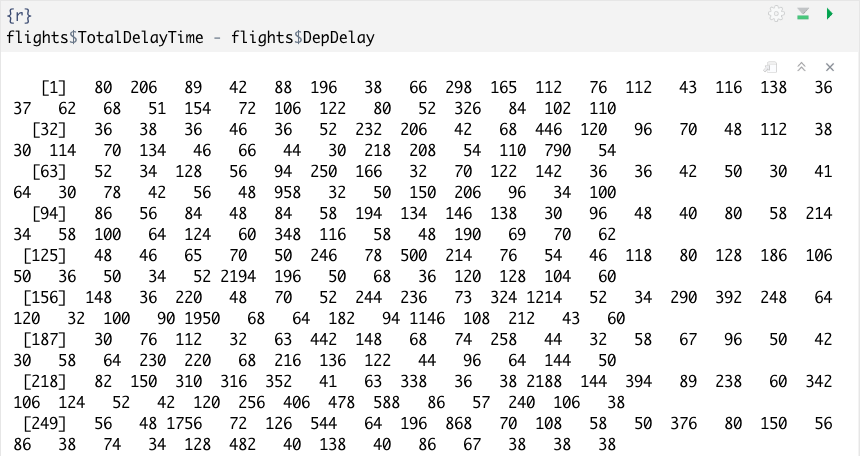


Appendix B: Models

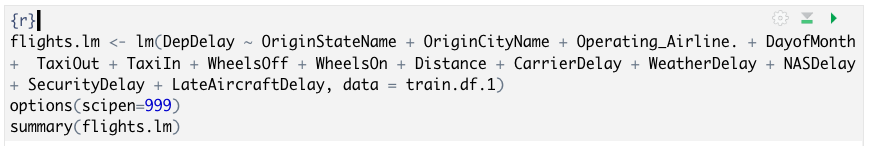
Appendix B.1

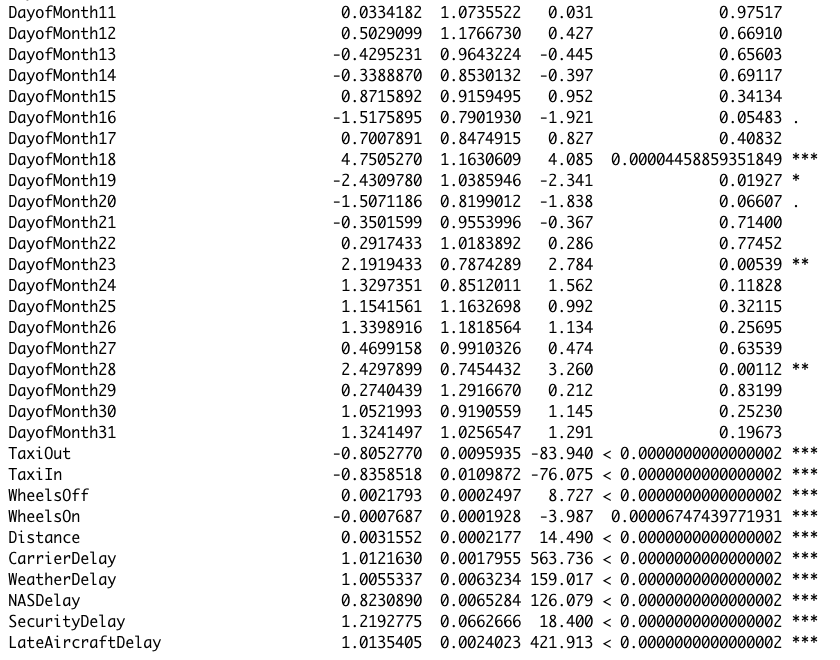
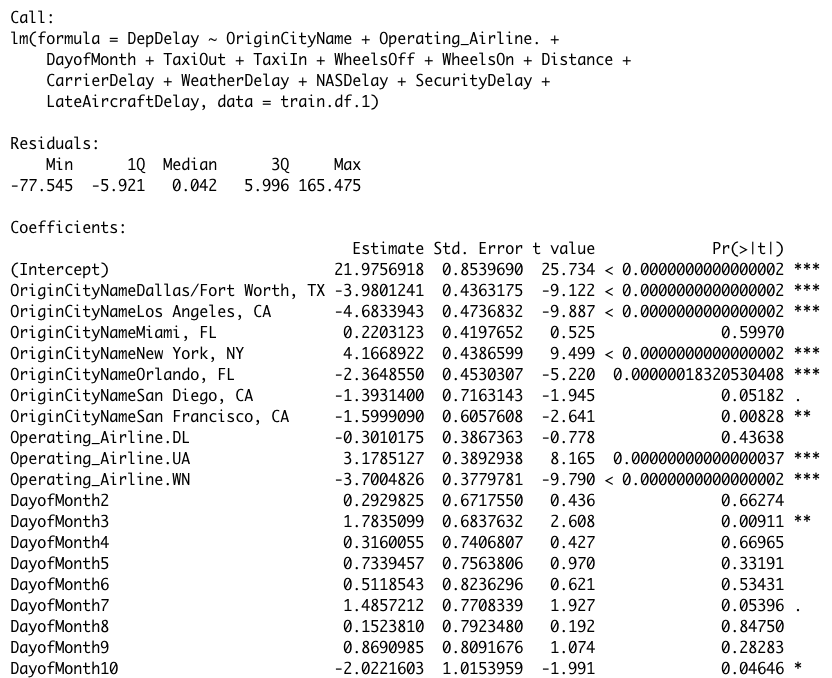
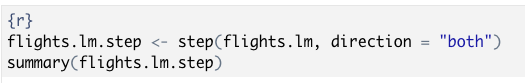


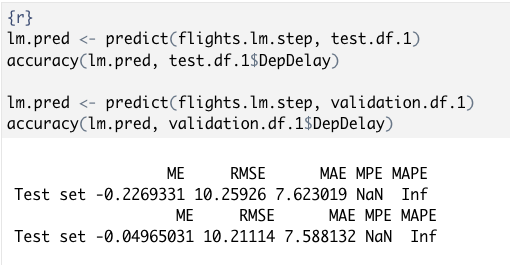
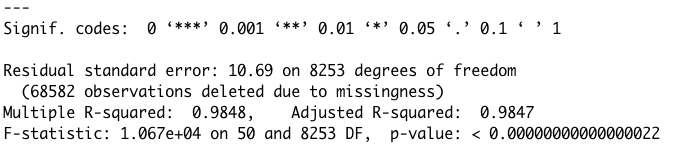
Appendix B.2



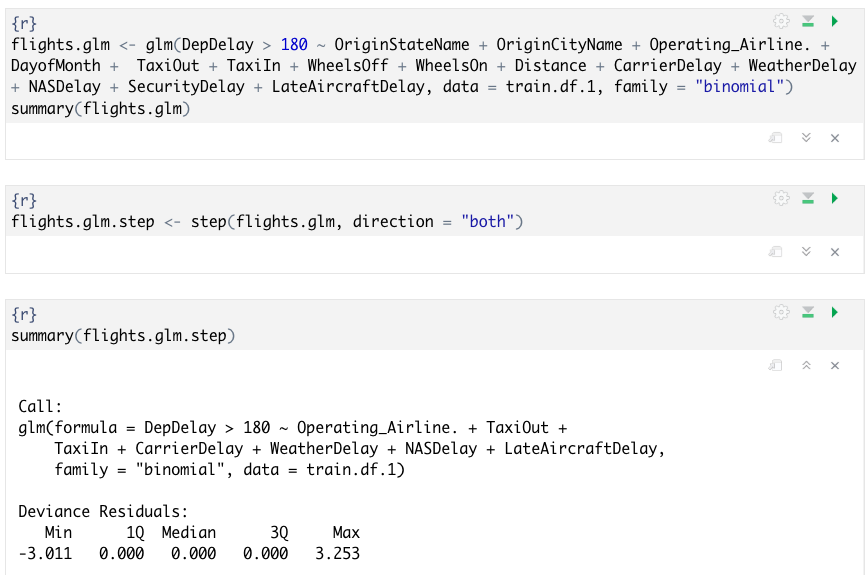
Appendix B.3

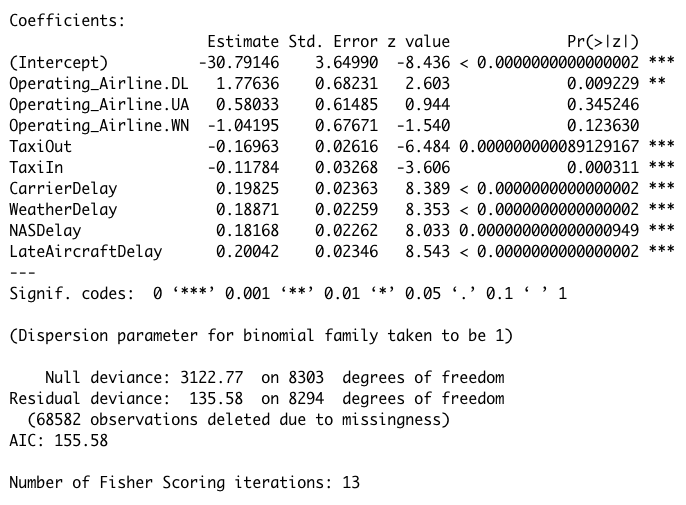




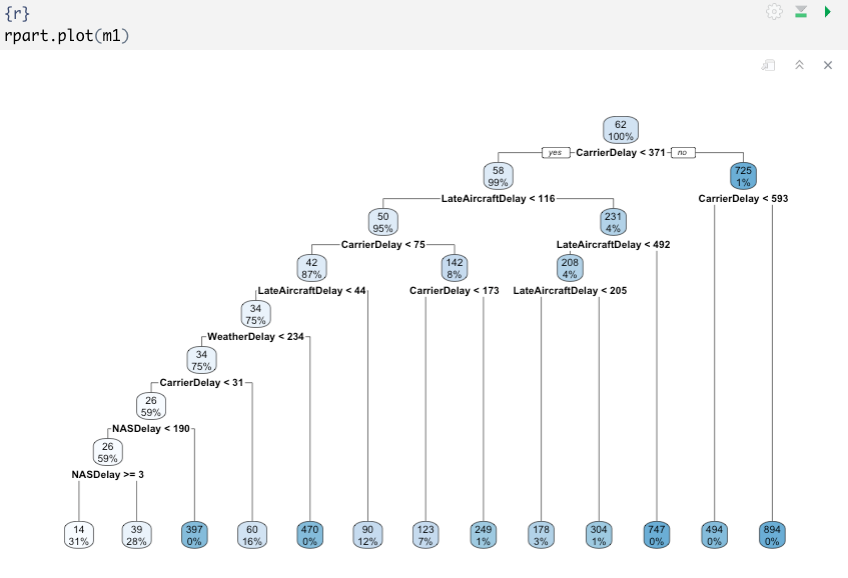


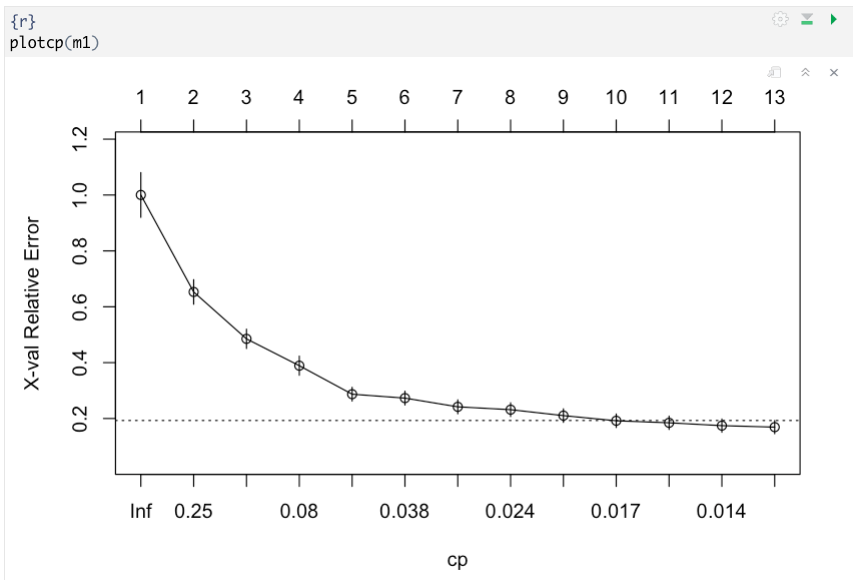
Appendix B.4





Appendix B.5





Appendix B.6 - SVM

