Analysis of Flight Delays

Authors

Amrutha Dokkidi

Ashutosh Rastogi

Tejas Prakash Agrawal

Vineeth Reddy Tati

Abstract

According to a study, flight delays cause approximately $40 billion loss to the United States civil aviation industry. As more and more people are opting for air travel instead of road trips to reach their destinations, it becomes imperative to maximize cost efficiency and customer satisfaction without any drop in operational efficiency. Flight delays are influenced by many events such as weather, security, airline scheduling, air traffic control, flight maintenance etc.

Data analytics can help immensely to predict flight delays and various factors making a significant contribution to those events. The objective of this project is to employ various supervised learning methods i.e., Linear Regression, Principal Component Regression, Ridge Regression and Lasso Regression which are parametric and Random Forest Regression and Support Vector Machine Regression which are non-parametric to predict the response variable Arrival Delay.

Furthermore, we used 10-Fold cross validation to validate our results. Root Mean Squared Error and R-Squared are the performance metrics considered for the model comparison. Principal Component Regression is found to have the least Test Root Mean Squared Error of 7.563 and R-Squared of 0.8725.

Introduction

For the past several years, air travel has become both more hectic and expensive, with tighter security regulations and longer waits. Yet, it is still the fastest and safest mode of transportation. The number of flights flown around the world by the airline industry including passenger and freight increased steadily since the start of the 21st century. Airlines have always had the need to use data analytics to help them analyze and develop models that can predict better both the customer and the outside events behavior.

Airlines, like any other industry, are focused on finding the best way to be more cost efficient, and the easiest way to achieve this is by avoiding unnecessary costs. Delayed aircraft are estimated to have cost the airlines several billion dollars in additional expenses. Delays also drive the need for extra gates and ground personnel and impose costs on airline customers in the form of lost productivity, wages, and goodwill. Data analytics play a very important role in helping airlines predict any event that might cut their profit.

The dataset presents a wide variety of problems and offers different ways to approach them but the main question to be answered is “How to avoid flight delays?”, and since there is no one easy solution, it gives space for us to analyze this data set in many ways. Knowing the causes and the effects for both the passengers and airlines for flight delays can help avoid costs for both the passengers and the airlines.

Data Description

Data Manipulation

The source of the dataset is 2015 Flight Delays and Cancellations from Kaggle which originally comes from the United States Department of Transportation's Bureau of Transportation Statistics. The original dataset has around 1 million rows and 31 columns.

The response variable is Arrival Delay which is in continuous format. The predictor variables are in both continuous and categorical format. of the continuous predictors are Scheduled Departure, Departure Delay, Taxi Out, Scheduled Time, Elapsed Time, Distance, Wheels On, Arrival Time, etc. Some of the categorical predictors are Day of Week, Airline, Flight Number, Origin Airport, Diverted, Cancelled, Cancellation Reason etc. The categorical variables are encoded by default.

The timeline of the original dataset was for the first three months of 2015 i.e., January, February, and March. Since the original dataset is of huge size, data visualization is done only for the month of February because it has the least number of days.

Missing values are replaced by 0 for continuous predictors, although it has different interpretation for each predictor. The continuous predictors are in time stamp and time duration format indicating 0 minutes and 12:00 am respectively when missing values are replaced. The categorical predictors are transformed into factor data type and missing values are replaced by 0.

Due to computation constraints, the dataset size is further reduced for predictive modeling. The first 14 days of February are considered for prediction. American Airlines (AA) and Delta Air Lines (DL) are chosen from the Airline column as they are the first two largest airlines in the United States passenger aviation industry respectively. Similarly, Dallas (DFW) and Atlanta (ATL) are chosen from the Origin Airport column as they are the focus hubs for the previously mentioned airlines respectively.

Correlation between each continuous predictor is calculated. It is employed to eliminate highly correlated continuous predictors. The cutoff chosen for elimination is 0.7. It is shown in the correlation plot.

The filtered dataset has around 15000 rows and 22 columns after manipulation.

Data Visualization

Correlation: Scheduled Departure, Departure Time, Departure Delay, Taxi Out, Wheels Off, Scheduled Time, Elapsed Time, Air Time, Distance, Wheels On, Taxi In, Scheduled Arrival and Arrival Time are the predictors which are highly correlated with correlation coefficient greater than 0.7. Hence, a combination of predictors with the least correlation are employed for predictive modeling.

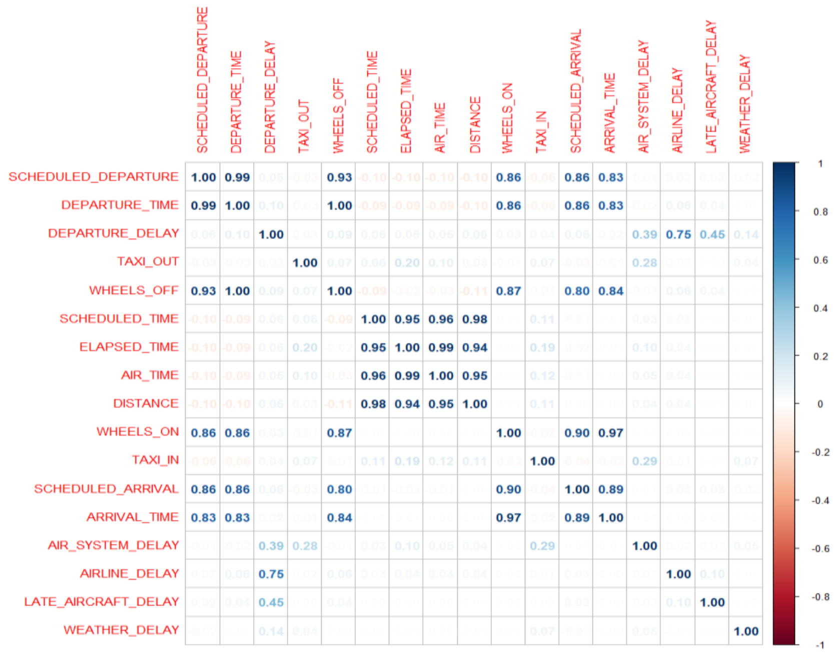

Figure 1. Correlation Plot

Delayed Arrival Time in Minutes by Airline: American Airlines (AA) and Delta Air Lines (DL) have the highest number of delayed flights in February. Similarly, American Airlines (AA) and Hawaiian Airlines (HA) have the flights with the longest delayed times. JetBlue Airways (B6) and Frontier Airlines (F9) have the highest average arrival delay.

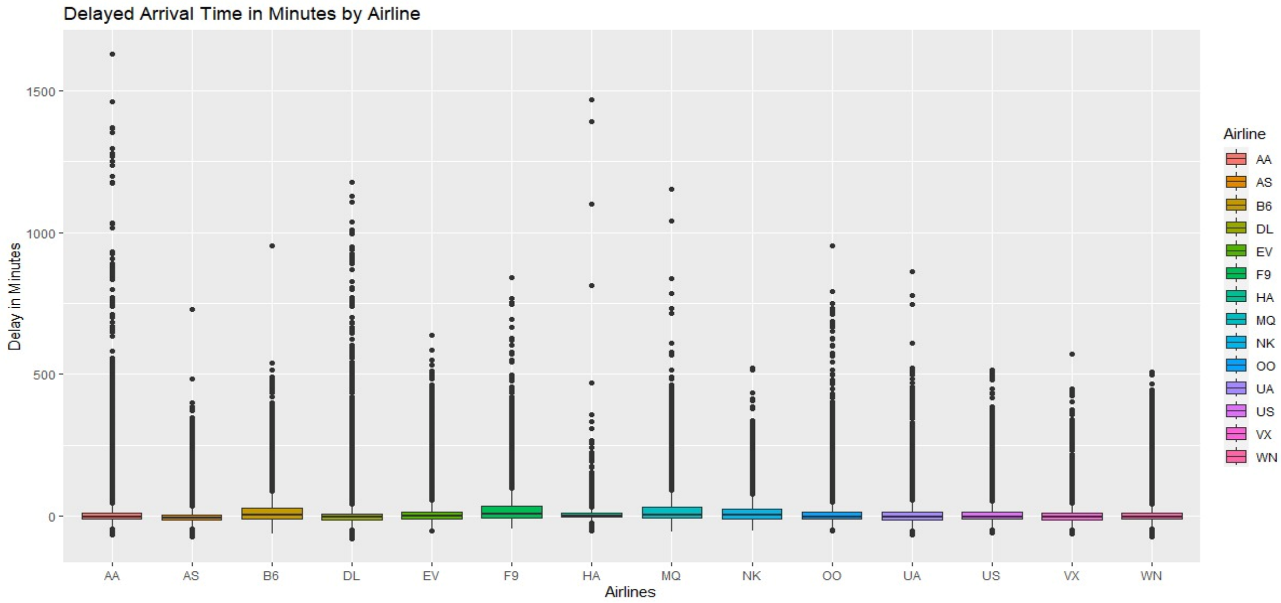

Figure 2. Delayed Arrival Time in Minutes by Airline

Delayed Arrival Time in Minutes by Day of the Week: Monday and Saturday have the flights with the longest delay time since weekday and weekend start on these days. Similarly, Tuesday and Friday have the shortest delay time.

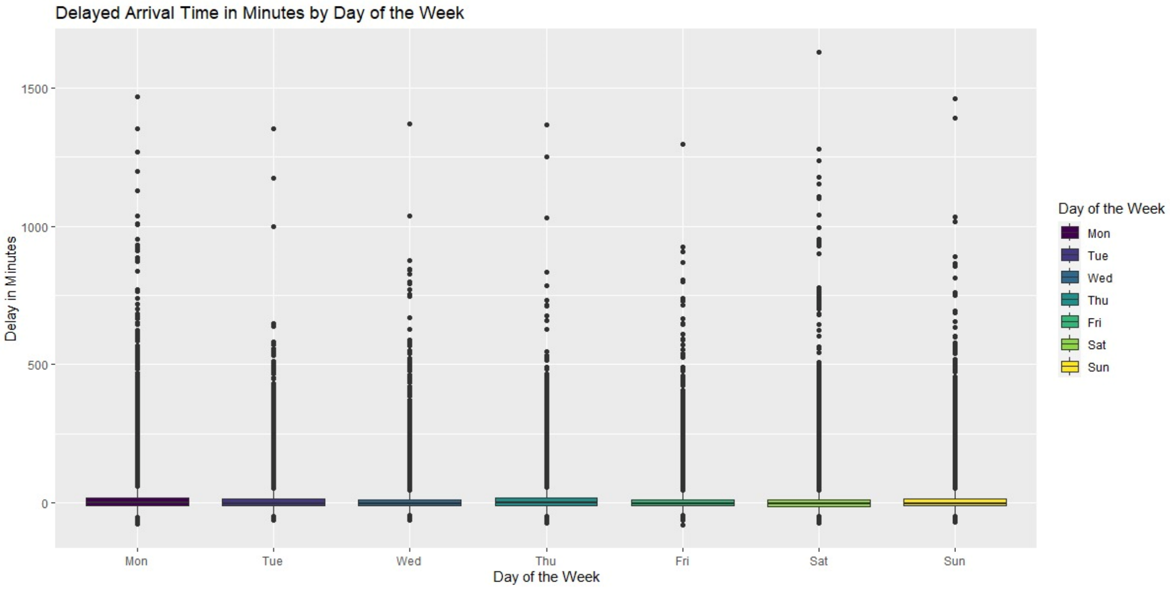

Figure 3. Delayed Arrival Time in Minutes by Day of the Week

Average Delay in Minutes Related by Part of the Day: Flights have the highest average delay during the Night. Surprisingly, the average delay time is negative during the Morning which indicates that the flights are leaving early.

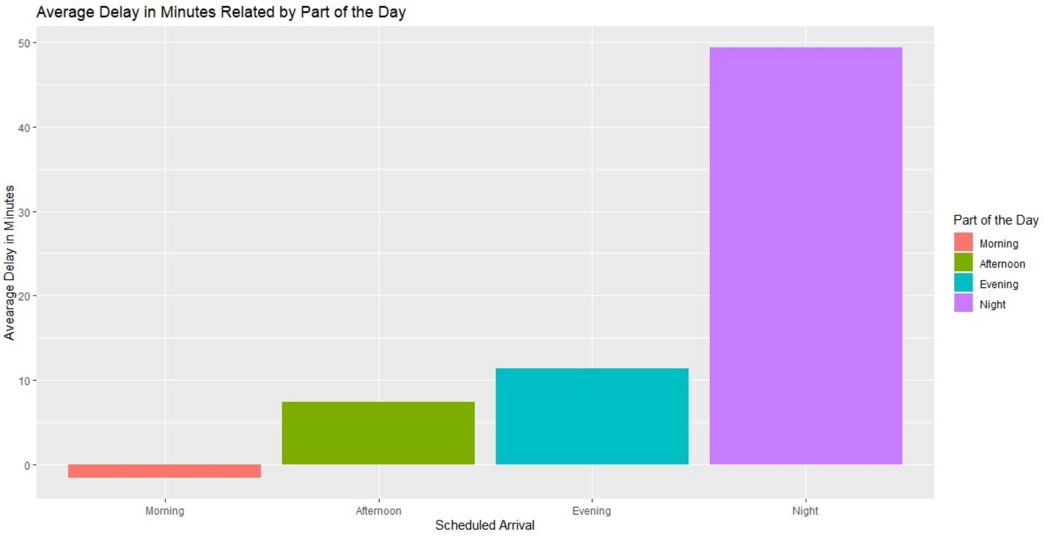

Figure 4. Average Delay in Minutes Related by Part of the Day

Methodology

Predictive Modeling

Linear Regression: Linear Regression without cross validation is performed.

Principal Component Regression: Principal Component Regression is tuned. Optimal number of principal components obtained are 16.

Ridge Regression: Ridge Regression is tuned with 10-fold cross validation. Optimal lambda obtained is 1.5872.

Lasso Regression: Lasso Regression is tuned with 10-fold cross validation. Optimal lambda obtained is 0.0375.

Random Forest Regression: Random Forest is built with 1000 trees and 5 predictors are sampled at each split.

Support Vector Machine Regression: Support Vector Machine with linear kernel is tuned. Optimal cost obtained is 5.573.

Performance Comparison

Validation Method: 10-fold cross validation is chosen for resampling the dataset.

Train-Test Split: The dataset is randomly split into 80:20 ratio for training and testing the predictive models respectively.

Performance Metrics: Root Mean Squared Error and R-Squared are considered as the performance metrics.

Results

Principal Component Regression has the least Test Root Mean Squared Error of 7.563 and the highest R-Squared of 0.8725.

Model	Test Root Mean Squared Error	Test R-Squared
Linear Regression	8.430	0.8658
Principal Component Regression	7.563	0.8725
Ridge Regression	8.533	0.8661
Lasso Regression	8.491	0.8643
Random Forest	9.298	0.8512
Support Vector Machines	9.573	0.8467

Conclusion

- Predictors having significant contribution to the response Arrival Delay are Day of Week, Wheels On, Taxi In, Distance, Air System Delay, Airline Delay, Late Aircraft Delay, Security Delay and Weather Delay.
- Linear Models i.e., Linear, Principal Component, Ridge and Lasso Regression have slightly better prediction performance compared to Random Forest and Support Vector Machine Regression.
- Dataset manipulation has a significant influence on the prediction performance.
- Non-parametric models may have performed better with other manipulation criteria.

GitHub Repository

Contributions

Contributions	**Members**
Data Manipulation	Amrutha Dokkidi, Tejas Prakash Agrawal
Data Visualization	Ashutosh Rastogi, Vineeth Reddy Tati
Predictive Modeling	Tejas Prakash Agrawal, Vineeth Reddy Tati
Performance Comparison	Ashutosh Rastogi, Vineeth Reddy Tati