2021

FORECASTING THE DAILY PERCENTAGE OF DELAYED FLIGHTS BASED ON THE NATIONAL WEATHER DATA

Parto Mahmoudi
The University Of Montana

FORECASTING THE DAILY PERCENTAGE OF DELAYED FLIGHTS BASED ON THE NATIONAL WEATHER DATA

By

PARTO MAHMOUDI

Master’s degree, University of Zanjan, Zanjan, Iran, 2010

Bachelor's degree, University of Zanjan, Zanjan, Iran, 2007

Thesis

presented in partial fulfillment of the requirements

for the degree of

Master’s degree

Data Science

The University of Montana

Missoula, MT

May 2021

Approved by:

Scott Whittenburg, Dean of The Graduate School

Graduate School

Dr. Javier Perez Alvaro, Chair

Department of Mathematical Sciences

Dr. Johnathan Bardsley

Department of Mathematical Sciences

Dr. Simona Stan

College of Business

Mahmoudi, Parto, M.S., Spring 2021 Data Science

Forecasting the Daily Percentage of Delayed Flights Based on the National Weather Data

Chairperson: Dr. Javier Perez Alvaro

Abstract Content

Flight delays cost airlines and affect passenger's satisfaction. In this research work, we predicted the daily percentage of delayed flights based on the national weather data using the multiple linear regression and the random forest models. We extracted the passenger flight on-time performance data from the Bureau of Transportation Statistics and the weather dataset from NOAA National Centers for Environmental Information for the years from 2015 to 2019. We used the flight dataset for Seattle airport as the origin. We predicted the daily percentage of delayed flights for the Seattle-originated flights based on the features such as weather conditions of the origin and its top 10 destination airports on the date of flight, weather features of the day before the flight for the origin, the number of daily flights from Seattle to these destinations, year, month, and day of week. We conducted the random forest model by training and rigorously hyper-parameter tuning. We measured the assessment of the fitted model with the evaluation metrics, such as mean absolute error, root mean squared error, and coefficient of determination scores. The random forest model with the evaluation scores of 2.68, 4.08, and 0.79, respectively, outperformed the multiple linear regression model to predict the daily percentage of delayed flights.

Contents

List of Figures

1. INTRODUCTION

In 2007, national flight delays cost the U.S. economy $31.2 billion [1]. According to the Bureau of Transportation Statistics (BTS), flights are delayed if they depart or arrive 15 or more minutes later than the scheduled departure or arrival time [2].

Weather condition is one of the major causes of flight delays [3] [4] [2]. As stated by the BTS, the other causes of delays in addition to extreme weather conditions are air carriers, national aviation system, late-arriving aircraft, and security [2].

We conducted this study to forecast the percentage of departure delays for the Seattle International Airport, using weather data, because Seattle International Airport is one of the busiest airports in the US. In 2020, the number of daily flights changed due to the Covid19. Thus, in this work, we used the data for the years 2015-2019.

Due to the large number of daily flights from Seattle airport to the different destinations, we had a large dataset. Hence, we limited data to the top 10 destinations of flights originated in Seattle, which includes about 45% of flights from Seattle. To predict the daily percentage of delayed departure flights, we extracted the flight and weather details regarding Seattle airport and its top 10 destinations. The top 10 airports covering Los Angeles International Airport (LAX), Denver International Airport (DEN), Anchorage Ted Stevens International Airport (ANC), Phoenix Airport (PHX), San Jose International Airport (SJC), Dallas/Fort Worth International Airport (DFW), San Francisco International Airport (SFO), Salt Lake City International Airport (SLC), McCarran International Airport (LAS), and Chicago International Airport (ORD).

In this research project, we employed two algorithms: A multiple linear regression model as a baseline, and a random forest regressor algorithm. We compared the models for a set of evaluation scores such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) to find the most efficient algorithm.

This thesis is designed as follows: In Section 2, we have done a literature review from several other papers; in Section 3, we have explained the data preprocessing and cleaning techniques involved and described the methodologies. In Sections 4-5, we have done a comparative study.

2. LITERATURE REVIEW

Shah et al. applied random forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) models to classify flight delays by more than 10 minutes based on features such as Origin, Destination, NASDelay, WeatherDelay, LateAircraftDelay, Month, etc. They collected the data from BTS for the year 2016. The researchers reported accuracy, recall, F1-score, and precision for their different models. In their study, the random forest machine learning model with an accuracy of approximately 92.013% and f1- Score 0.88 gave the best results [5].

Belcastro et al. focused on the arrival delay of a scheduled flight using weather data. Both flight information and weather conditions in their work were considered. The Airline On-Time Performance (AOTP) dataset was from RITA—Bureau of Transportation Statistics from January 2009 to December 2013. The Quality Controlled Local Climatological Data (QCLCD) was from the National Climatic Data Center. The flight features in this study were origin airport, destination airport, scheduled departure, arrival time. They also considered the weather conditions at the origin and the destination airport. They predicted arrival flight delays using random forest in MapReduce. With a delay threshold of 15 minutes, their model provided an accuracy of 74.2% and a recall of 71.8%, and with a delay threshold of 60 minutes, their model obtained an accuracy of 85.8%, and a recall of 86.9% on delayed flights [6].

Etani conducted a study to discover the correlation between flight data of a low-cost carrier in Japan (Peach Aviation) and weather data by implementing machine learning models. The researcher extracted the sea-level pressure data from March 2012 to December 2018 from Japan Meteorological Agency. The flight features such as departure date, arrival date, departure airport, and destination airport were from FLIGHTSTATS. To predict the arrival delays with and without weather data he applied the SVM, gradient boosting, random forest, decision tree, and adaboost classifiers. In his work, the random forest classifier model with weather data obtained the highest accuracy of 77% [7].

Ding implemented a multiple linear regression algorithm to predict flight delays for domestic airports in China from November 2015 to March 2016. He compared the prediction results with Naive Bayes and C4.5 based predictive models. The models were based on the route distance, the departure and arrival airports, the flight, and the weather features, etc. The researcher suggested that the multiple linear regression model with an accuracy of approximately 80% outperforms the other two models [8].

Elangovan et al. predicted the departure delays in minutes for the airline using logistic regression, random forest regression, and support vector regression models. In their research, the random forest model had the best performance compared to the other two models with the evaluation score of 0.083, 0.289, and 0.007 for MAE, MSE, and RMSE, respectively [9].

Manna et al. applied a gradient boosted decision tree regression model to predict flight departure and arrival delays. They collected flight on-time performance data from the U.S. department of transportation for the period April-October 2013. They used features such as day of week, carrier, origin airport ID, destination airport ID, CRSDepTime, DepDelay, CRSArrTime, and ArrDelay. Their model gained the highest R^2 of 92.3185% in case of arrival and 94.8523% in case of departure [10].

Kalliguddi et al. applied regression models like decision tree regressor, random forest regressor, and multiple linear regressor on the flight data for predicting both departure and arrival delays. The data set was from BTS to analyze the domestic flight activity from January to December 2016. In their research, they used departure delay, taxi in, taxi out, carrier delay, security delay, weather delay, late aircraft delay, distance, and national air system delay as the model variables. They realized that the random forest model had the best performance compare to the other two models based on the evaluation criteria. Also, they found that the significant factors for departure delay are late aircraft delay, carrier delay, weather delay, and NAS delay. In their work, the random forest model provided the R^2 of 0.94, which was significantly larger than the multiple linear regression model. They got a prediction error of 153.94, and the RMSE of 12.5 minutes for the random forest model, which was significantly lower than both applied models [11].

Ebenezer et al. developed a predictive system to classify flight delays based on weather data. The ensemble method, decision tree, and random forest implemented to the balanced data. Flight data set was from US Department of Transportation and weather data set was from Hourly Land-Based Weather Observations from NOAA. They used both flight and weather data sets to predict flight delays. They selected features such as month, year, day of week, carrier, day of month, origin and destination airport ID, departure delays, and arrival delays and canceled from flight data. From the weather data, they extracted the following features: year, adjusted month, adjusted delay, adjusted hour, time zone, visibility, dry bulb Fahrenheit, dry bulb Celsius, dew point Fahrenheit, dew point Celsius, relative humidity, and wind speed. In their study, the ensemble model based on accuracy, precision, and recall with the values 97.83%, 95.04%, and 95.34%, respectively had the best performance compared to the other models [12].

In another attempt to analyze the flight data, Kuhn et al. applied decision tree, logistic regression, and neural networks classifiers to predict if a given flight's arrival will be delayed or not. In that research, all three classifiers gave a test accuracy of approximately 91% and a ROC of 0.96%. The selected flight features in their study were information about flight, origin, destination, departure, flight-journey, arrival, diversion, and cancellation. They observed that the decision tree classifier performed slightly better at predicting on-time flights, whereas the neural network performs a little better at predicting delayed flights [13].

Chakrabarty et al. applied gradient boosting classifier to analyze and predict possible arrival delays of the flights. The model employed on flight information of US domestic flights operated by American Airlines. Data were from top 5 busiest airports of US for years 2015 and 2016. They used a 200% randomized SMOTE technique to reduce the imbalance between classes. In their article, there were two strategies to follow. In strategy one, they skipped the data imbalance removal step for data preprocessing, and in strategy two, they used SMOTE technique for data preprocessing. For both strategies, they did a grid search, and calculated the Mean score. They also created ROC curves and confusion matrix. The comparison showed that the data imbalance removal step resulted in a much better performance with a validation accuracy of 85.73% [14].

Many works have done on flight datasets of the US or other countries, and some of them have considered weather features. But to my knowledge, forecasting the daily percentage of delayed flights based on the weather data using machine learning algorithms such as random forest is innovative.

3. METHODOLOGY

In this section, we apply the multiple linear regression model and the random forest regressor algorithm on the flight dataset along with weather features for Seattle and its top 10 destinations to predict the daily percentage of delayed flights.

3.1. DATASETS

In the next sub-sections, we describe our flight and weather datasets.

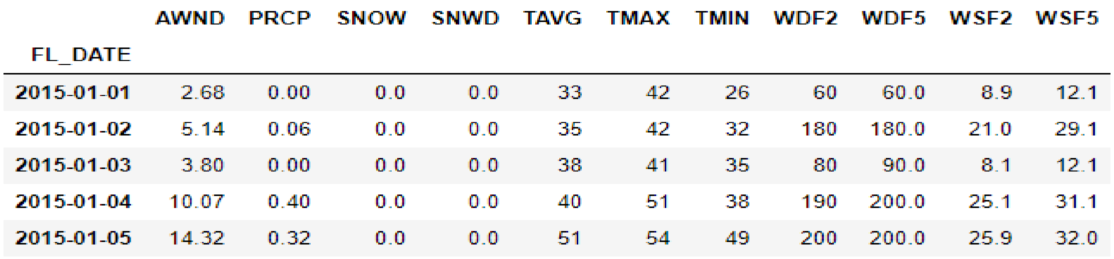
3.1.1. FLIGHT DATASET

For this study, we took the flight on-time performance data from the US Department of Transportation’s Bureau of Transportation Statistics (BTS) website. The dataset contained the flight information for the years between 2015 to 2019, regarding the flight dates, the quarters, the months, the day of month, the day of week, the origins, the destinations, the departure delay status (it assumes two values for the status of the flight, 0 and 1, where 0 means no departure delay and 1 indicates a departure delay) and an indicator for the canceled flights.

3.1.2. WEATHER DATASET

We collected the daily airport-based weather data for Seattle airport and its top 10 destinations for the years 2015-2019 from the National Centers for Environmental Information of NOAA website [15]. The desired features were the average wind speeds (m/s), the precipitations (mm), the snowfall (inch), the snow depth (inch), the minimum, average, and maximum temperatures (°F), the direction of fastest 5-second and 2-minute winds (degrees), and the fastest 5- second and 2-minute wind speeds (degrees). The first 5 rows of the weather data frame for Seattle looks like this:

Table 1: First 5 rows of Seattle weather data



	AWND	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	WDF2	WDF5	WSF2	WSF5
FL_DATE											
2015-01-01	2.68	0.00	0.0	0.0	33	42	26	60	60.0	8.9	12.1
2015-01-02	5.14	0.06	0.0	0.0	35	42	32	180	180.0	21.0	29.1
2015-01-03	3.80	0.00	0.0	0.0	38	41	35	80	90.0	8.1	12.1
2015-01-04	10.07	0.40	0.0	0.0	40	51	38	190	200.0	25.1	31.1
2015-01-05	14.32	0.32	0.0	0.0	51	54	49	200	200.0	25.9	32.0

In addition to considering the daily weather features, we added the weather for the day before the flight for Seattle to the flight data set.

3.2. DATA PREPROCESSING

Before training the models, in order to achieve the desired data frame, we preprocessed the data sets as follows.

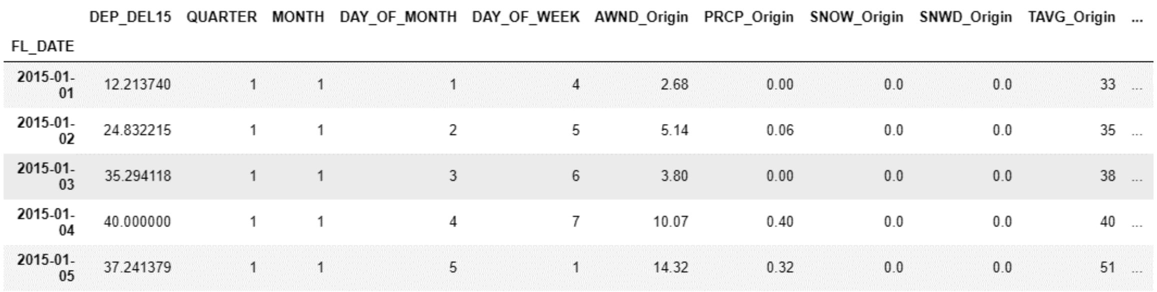
we concatenated the datasets row-wise for each month of 2015-2019 to achieve the complete flight delays data frame. To provide better results, we removed the canceled flights from the data frame, since their corresponding data values for the departure delay variable were reported as missing values. In order to achieve our goal, we filtered the dataset to include only Seattle originated flights. Also, to make this dataset more feasible for analysis, we used only about half of the data and we filtered the data to include the top 10 destination airports for Seattle originated flights. They were LAX, DEN, ANC, PHX, SJC, DFW, SFO, SLC, LAS, and ORD.

We counted the total number of daily flights to each destination, and it was added to the dataset as a variable. In addition, we calculated the target variable, the daily percentage of delayed flights for the flights originated in Seattle airport, as follows:

$$\text{Daily Percentage of Delayed Flights } = \frac{\text{Total number of daily delayed flights}}{\text{Total number of daily flights}} \times 100$$

We cleaned the weather datasets by removing the features with a large number of missing values. However, we replaced the remaining missing values with zeros.

To build the final data frame for this study, we joined all the aforementioned datasets together which contains 1826 days with 128 attributes. The attributes are the number of flights per day, the weather variables for the origin, the weather variables for the day before the flight for the origin, and the weather variables for the top 10 destinations, the quarters, the month, the day of month, and the day of week. The response variable is the daily percentage of delayed flights. The data was Shuffled and Split into train and test sets with 70% of data forming the training set and 30% of data forming the test set. A part of the final data frame that represents the first 5 rows and 10 columns looks like this:



FL_DATE	DEP_DEL15	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	AWND_Origin	PRCP_Origin	SNOW_Origin	SNWD_Origin	TAVG_Origin	...
2015-01-01	12.213740	1	1	1	4	2.68	0.00	0.0	0.0	33	...
2015-01-02	24.832215	1	1	2	5	5.14	0.06	0.0	0.0	35	...
2015-01-03	35.294118	1	1	3	6	3.80	0.00	0.0	0.0	38	...
2015-01-04	40.000000	1	1	4	7	10.07	0.40	0.0	0.0	40	...
2015-01-05	37.241379	1	1	5	1	14.32	0.32	0.0	0.0	51	...

3.3. EXPLANATORY DATA ANALYSIS

In this section, we visualize the relationship between the daily percentages of delayed flights as the response variable and a few features. **Error! Reference source not found.** r epresents the daily percentages of delayed flights over time between the years 2015 to 2019. The graph fluctuates mostly between 5 to 20 percent. Also, there are some spikes where their corresponding daily percentages of delayed flights are between 40 to 85%. These spikes have occurred for different reasons that one of them would be the poor weather conditions at those days. Generally, daily percentages of delayed flights have increased in 2019.

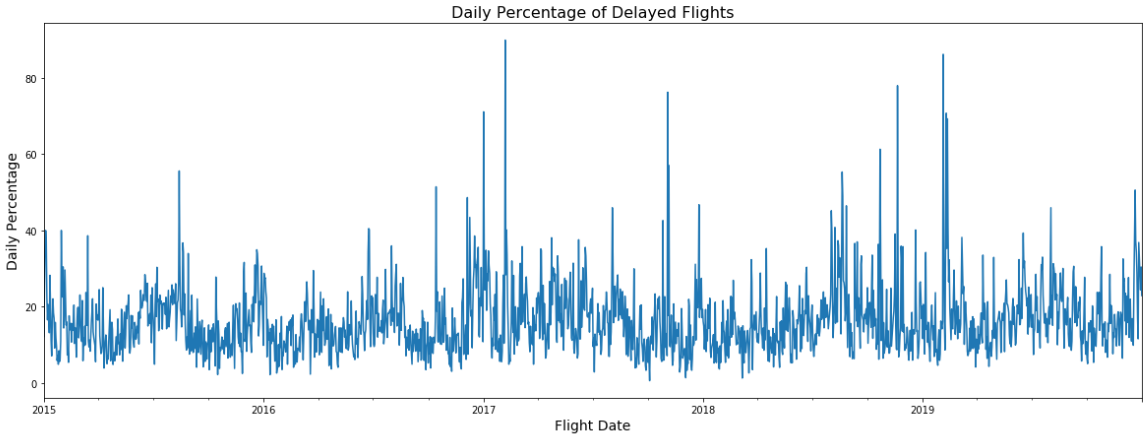


Figure 1: Daily percentages of delayed flights for years between 2015-2019.

The relationship between the daily percentage of delayed flights and air traffic in Seattle is illustrated in Figure 2 and Figure 3. According to these figures, climbing the volume of air travel continues results in an increase in the daily percentage of delayed flights.

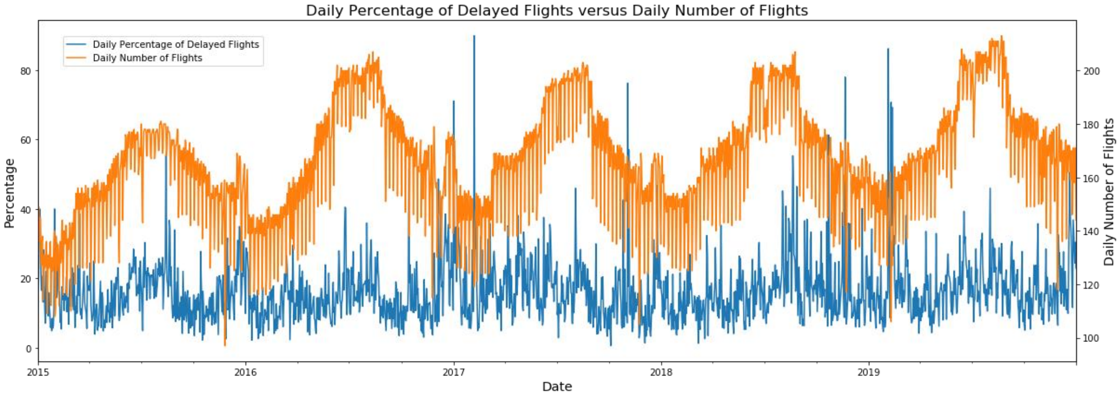


Figure 2: Daily percentage of delayed flights against daily volume of flights.

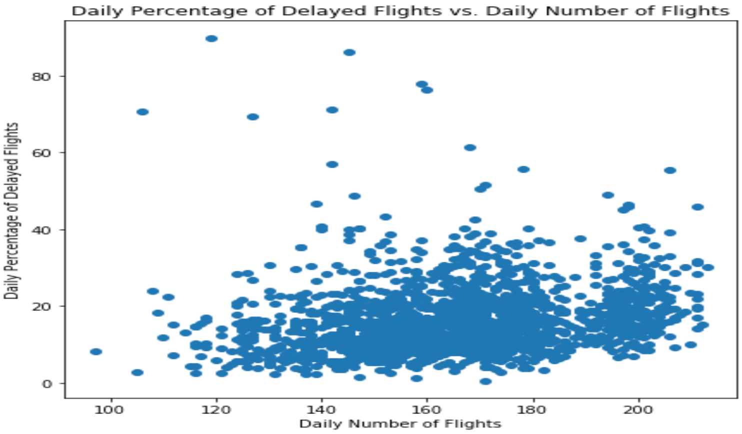


Figure 3: Relationship between daily percentage of delayed flights and daily number of flights.

To study the effect of variables, quarters, months, day of week, and day of month on the daily percentages of delayed flights, we plotted the following graphs. Figure 4 and Figure 5 show the distributions of daily percentage of delayed flights with respect to the quarters and months, respectively. Figure 4 interestingly indicates that the daily percentage of delayed flights were a little higher in summer on average. However, in winter and fall there are the most unusual percentages of delays which happened in February and November. The delay in these months can be due to the extreme weather conditions. In August and December, the distributions of the daily percentages of delayed flights are higher than other months that might be because of the large number of flights in these months since people travel more at those times for school and holidays, respectively.

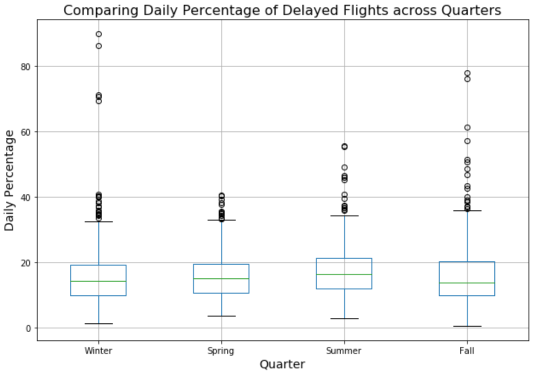


Figure 4: Comparing daily percentage of delayed flights across quarters.

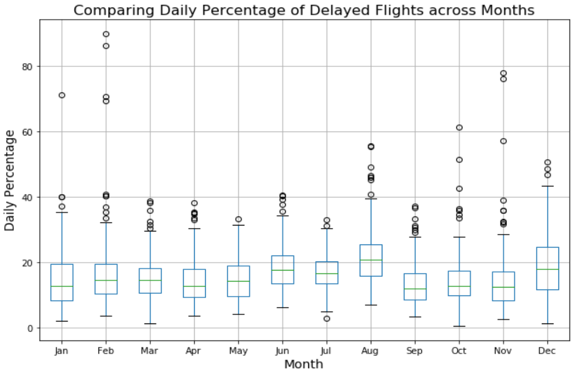


Figure 5: Comparing daily percentage of delayed flights across months.

Figure 6 and Figure 7 show the distributions of daily percentage of delayed flights with respect to the variables, day of month and day of week, respectively. These graphs indicate that the daily percentages of delayed flights are higher in 20th of each month and Friday of each week, on average.

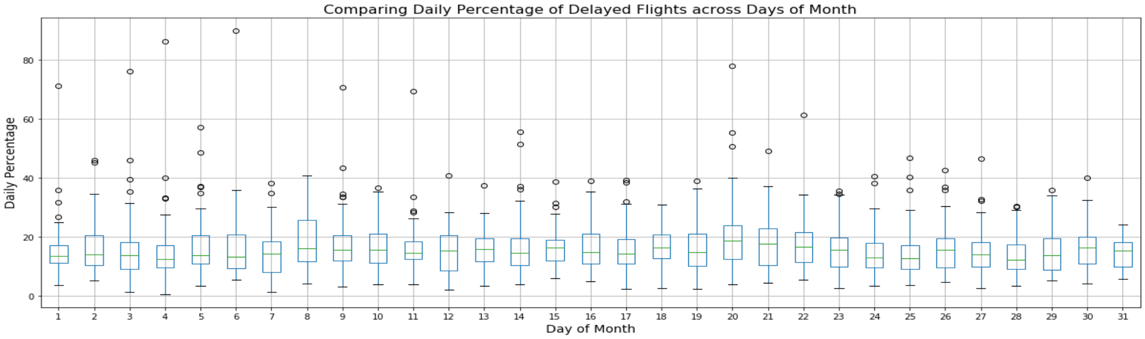


Figure 6: Comparing daily percentage of delayed flights across days of month.

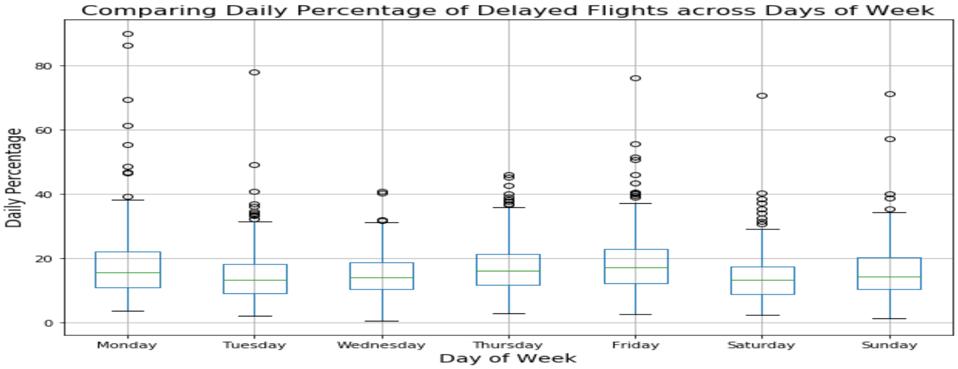


Figure 7: Comparing daily percentage of delayed flights across days of week.

Now to understand how weather features can affect the daily percentage of delayed flights and to see how these features can justify the spikes in the target variable, we visualize a few of these relationships in the following figures. Figure 8 is a graph with twin axes that shows the daily percentage of delayed flights (left axis) along with the minimum temperature in Seattle (right axis). The minimum temperature can explain some of those spikes in 2017 and 2019 especially when the minimum temperature drops under 30°F. Similarly, Figure 9 is a graph with twin axes that shows the daily percentage of delayed flights (left axis) along with the snowfall in Seattle (right axis). It indicates that some spikes in the graph of flight delays can be explained by the increase in the amount of snowfall. Hence, the snowfall followed by the temperature dropping in Seattle could explain some spikes. We can have a closer look at those spikes and the amount of snowfall in Figure 10 and Figure 11.

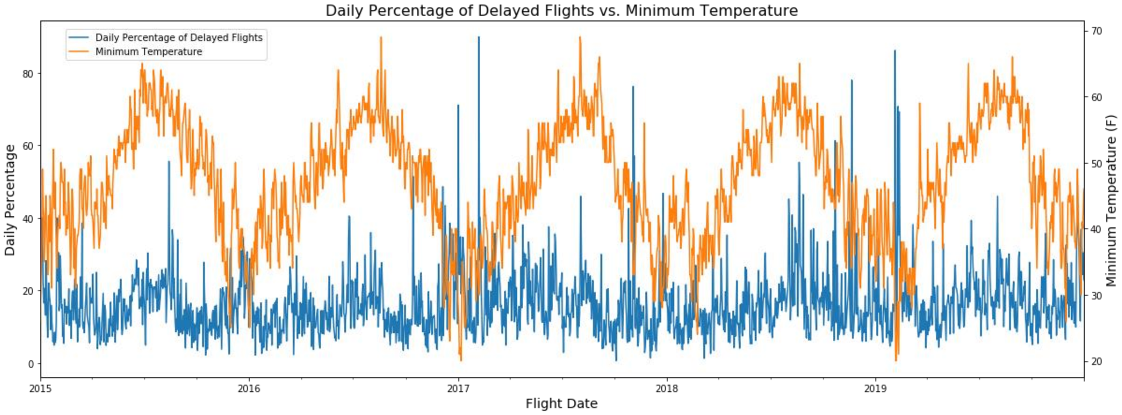


Figure 8: Daily percentage of delayed flights vs. minimum temperature.

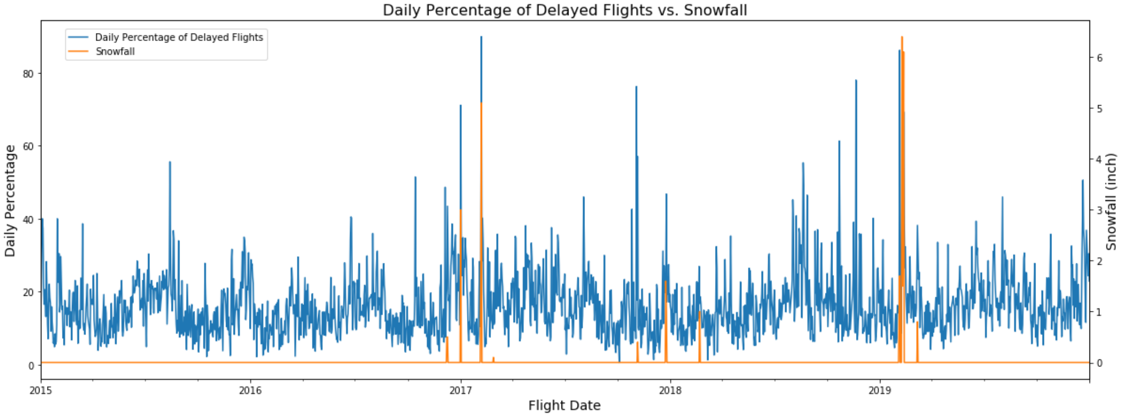


Figure 9: Daily percentage of delayed flights vs. snowfall.

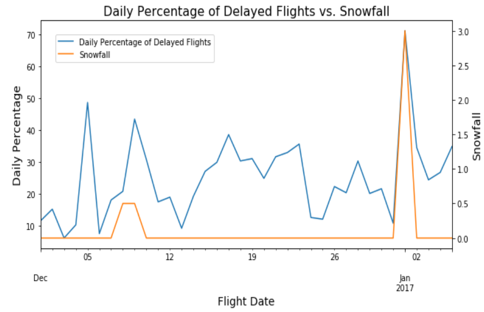


Figure 10: Daily percentage of delayed flights vs. snowfall for December and January 2017.

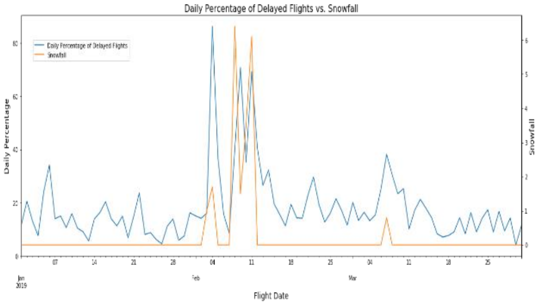


Figure 11: Daily percentage of delayed flights vs. snowfall from January to March 2019.

The effect of the amount of precipitation in Seattle on the daily percentage of delayed flights is illustrated in Figure 12. This figure shows that increasing in the amount of precipitation in Seattle is followed by increasing in the daily percentage of delayed flights at some days. The effects of the high amount of precipitation in some destinations on the daily percentage of delayed flights are illustrated in Figure 13 to Figure 17. These figures show that the higher amounts of precipitation in some destinations may cause an increase in the daily percentage of delayed flights at some days.

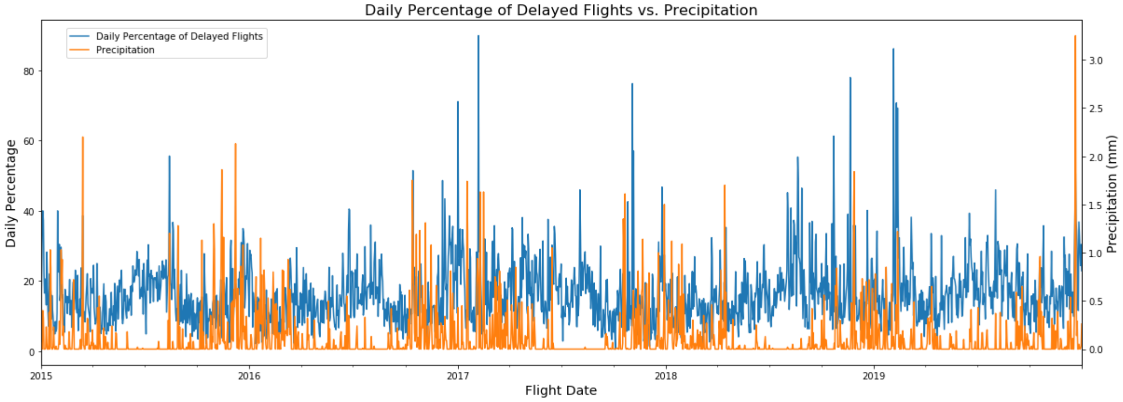


Figure 12: Daily percentage of delayed flights vs. precipitation in Seattle.

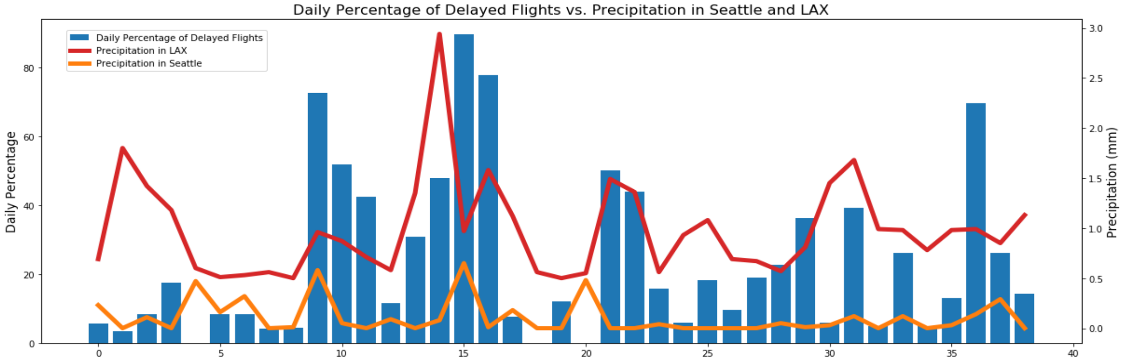


Figure 13: Daily percentage of delayed flights vs. precipitation for those days that the amount of precipitation in Los Angeles International Airport >=0 .5mm.

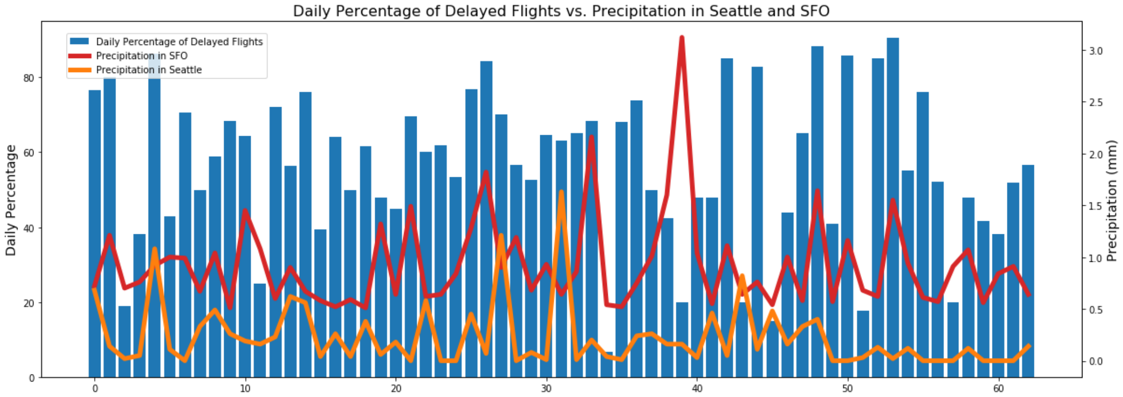


Figure 14: Daily percentage of delayed flights vs. precipitation for those days that the amount of precipitation in San Francisco International Airport >=0 .5mm.

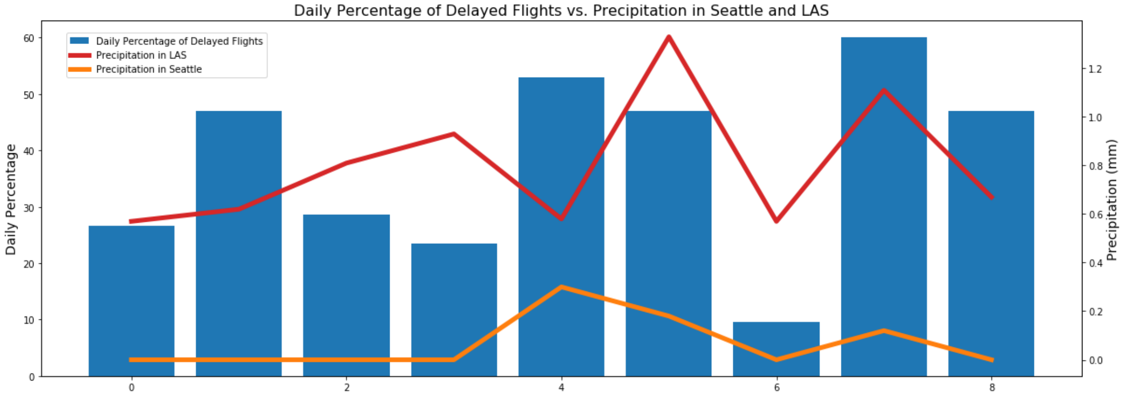


Figure 15: Daily percentage of delayed flights vs. precipitation for those days that the amount of precipitation in McCarran International Airport >=0 .5mm.

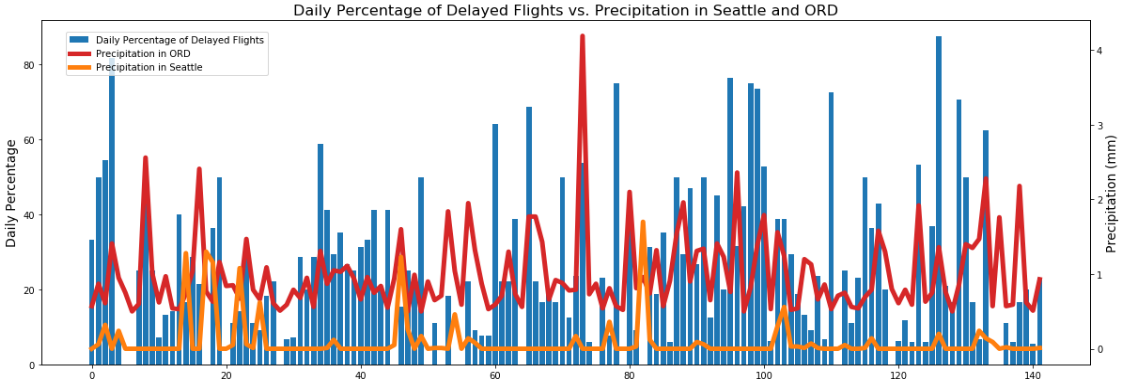


Figure 16: Daily percentage of delayed flights vs. precipitation for those days that the amount of precipitation in Chicago O'hare International Airport >=0 .5mm.

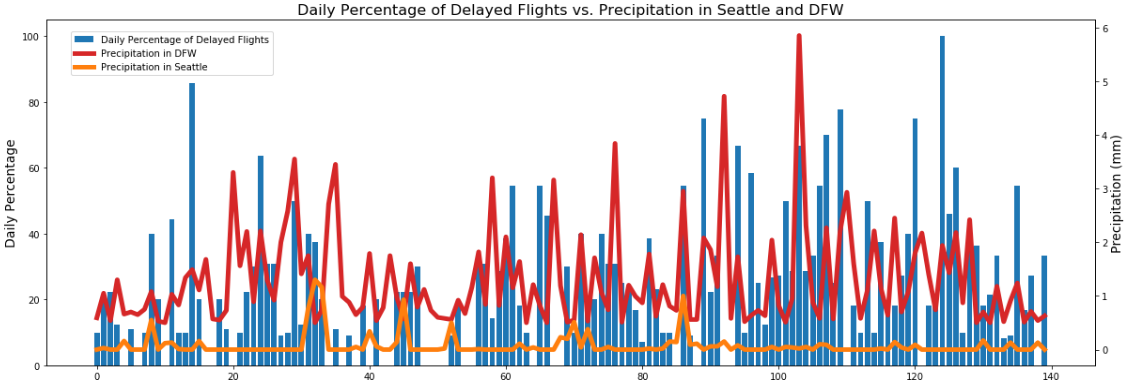


Figure 17: Daily percentage of delayed flights vs. precipitation for those days that the amount of precipitation in Dallas/Fort Worth International Airport >=0 .5mm.

In order to compare the changes in the daily percentage of delayed flights against the higher amounts of the fastest 2-minute wind speed we draw Figure 18. As figure shows, some spikes in the daily percentage of delayed flights can be explained by increasing in the wind speed.

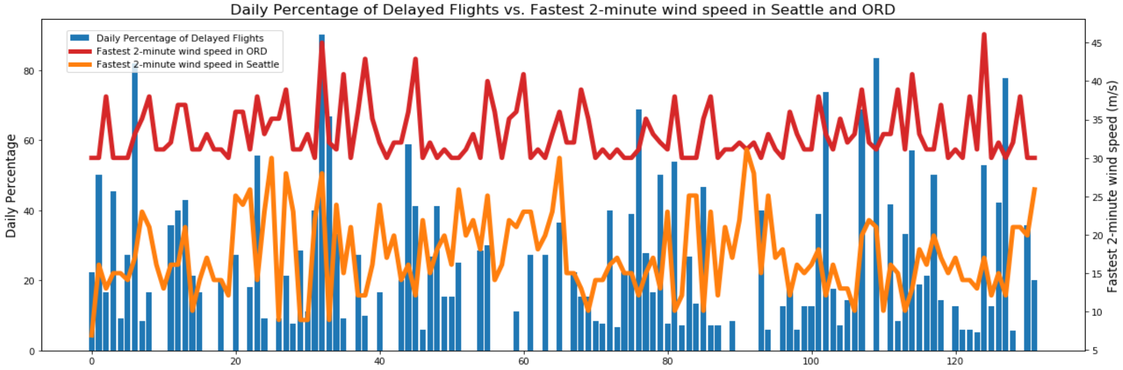


Figure 18: Daily percentage of delayed flights vs. fastest 2- minutes wind for those days that the amount of fastest 2-minutes wind in Chicago O'Hare International Airport >=30m/s

Figure 19 and Figure 20 show the daily percentage of delayed flights to the destinations of Salt Lake City and Denver International Airports vs. snowfall in Seattle and SLC and DEN, respectively, for those days that the amount of snowfall in those cities was greater or equal to 0.5 inches. These graphs show that increasing in the amount of snowfall in the destinations could explain some high percentages of delayed flights (spikes) in Seattle.

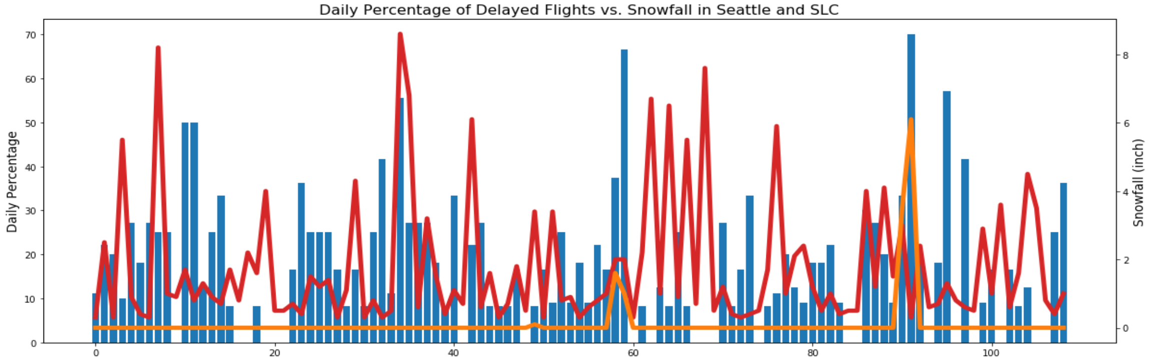


Figure 19: Daily percentage of delayed flights vs snowfall for those days that the amount of snowfall in Salt Lake City International Airport >=0.5inch.

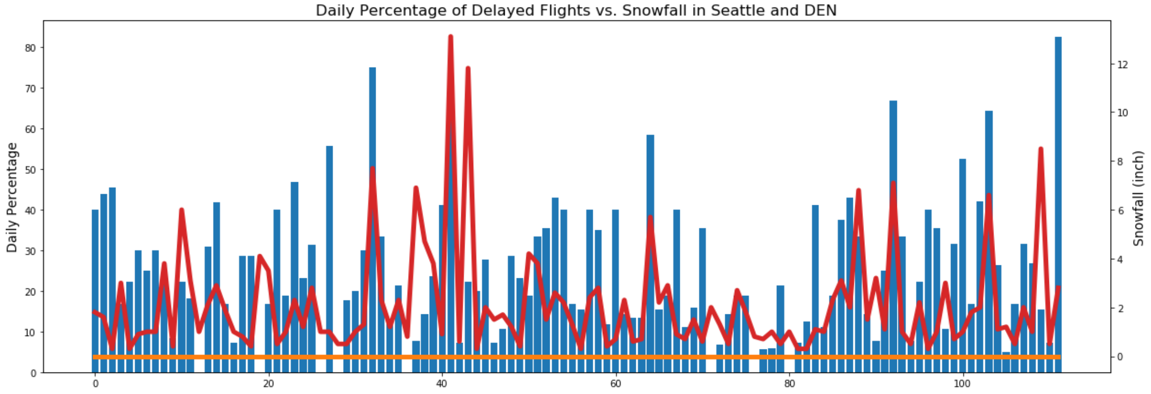


Figure 20: Daily percentage of delayed flights vs snowfall for those days that the amount of snowfall in Denver International Airport >=0.5inch.

After the explanatory analysis, we fitted linear regression and random forest models to predict the response, daily percentage of delayed flights, based on the explanatory variables.

3.4. MODELS

In this study, we used multiple linear regression, and random forest regressor algorithms to predict the daily percentage of delayed flights.

3.4.1. LINEAR REGRESSION

Linear regression is a method to model the association between the response variable (y) and the explanatory variables (X).

For more than one independent variable, the method is called multiple linear regression as given in Equation (1). The goal is to learn the coefficients of the linear equation to describe the relationship between the explanatory variables and the response variable, which can then be used to predict the targets that were not in the training dataset.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon, \qquad i = 1, \dots, n \qquad (1)$$

where y_i is the response variable, x_j is the explanatory variable, β_0 is the y intercept, β_j for $j = 1, \dots, k$ is the coefficient for each explanatory variable, ϵ is the model error.

In linear regression, the aim is finding the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ that minimizes $\parallel \boldsymbol{y} - X\boldsymbol{\beta} \parallel_2^2$. It is well known, that $\boldsymbol{\beta}$ satisfies the so- called normal equations.

$$X^T X\boldsymbol{\beta} = X^T \boldsymbol{y},$$

where $\boldsymbol{y}$ is the $n \times 1$ response vector and X is the matrix of observations.

3.4.2. RANDOM FOREST REGRESSION

The random forest algorithm [16] is a type of ensemble machine learning algorithm called bootstrap aggregation or bagging. An ensemble technique combines the predictions from various machine learning algorithms together to make more accurate predictions than the individual predictions. In random forest method, the samples for each tree are selected by bootstrapping from the training dataset. Bootstrap refers to random sampling with replacement. Then a decision tree grows from the bootstrap samples. At each node of the tree some features are selected randomly without replacement. In random forest regression, the splitting criterion of the node is Mean Square Error (MSE) (2) or Mean Absolute Error (MAE) (3). Thus, the node is split using the feature that provides the least MSE or MAE, which leads to tree growth. These steps are repeated k times. Each model makes a prediction. The final prediction is calculated as the average of the predictions over all decision trees. So, the random forest is called a bagging model because it makes each weak model run in parallel, then at the end aggregates the predictions to make more accurate predictions than any individual models.

A random forest usually outperforms an individual decision tree because of randomness which helps to reduce the model's variance. In the context of decision tree regression, the

MSE is often referred to as within node variance, which is why the splitting criterion is also better known as variance reduction. The random forest is also less sensitive to outliers.

$$MSE(t) = \frac{1}{N_t} \sum_{i \in \mathrm{D}_t} (y^{(i)} - \hat{y}_t)^2, \tag{2}$$

$$MAE(t) = \frac{1}{N_t} \sum_{i \in \mathrm{D}_t} \left| y^{(i)} - \hat{y}_t \right|, \tag{3}$$

where N_t is the number of training examples at node t, D_t is the training subset at node t, $y^{(i)}$ is the true target value, and $\hat{y}_t$ is the predicted target value (sample mean):

$$\hat{y}_t = \frac{1}{N_t} \sum_{i \in \mathrm{D}_t} y^{(i)}.$$

A random forest model involves a few to many decision trees and Figure 21 shows one of these trees. In order to make each decision tree in a random forest algorithm, different splits take place at different points all over the training samples. These splits are assessed using a cost function. The split with the minimum cost value is chosen as the threshold. The feature that obtains the least cost function compared to other features in the training samples is placed in the first node (root). All the features and all the possible split points are assessed, and the best split point is selected each time. By defining a minimum count on the number of training cases given to each leaf node, the splitting process knows when to stop splitting as it works its way down the tree. If the sample size at the node is less than some minimum, then the split process stops, and the node is taken as a final leaf node. In this model, the count is set to 2 samples. As shown in Figure 21, there are 1154 samples in the root node because, for the training of each tree of the random forest, the model uses a random subset of the data points with replacement. The predictor *SNWD_origin* at the threshold of 0.6 obtained MSE = 71.729, and the predicted response value using the samples at the first node is their average (15.995).

To use this tree for prediction, we should pass a sample across the tree. At first, the value of the predictor *SNWD_origin* as the root of the tree should be compared with the threshold (0.6), and all samples that have a feature value less than or equal to 0.6, fall into the left child node otherwise they fall into the right child node. If so, the next predictor is *Total_DEP_DEL15* with the threshold of 163.5. Again, if the value of this predictor for samples at this node is less than or equal to 163.5 the samples fall into the left child node, if not they fall into the right child node. Similarly, the model moves samples through the tree by considering the thresholds at each node and make a decision to move to the left or to the right. As a result, we conclude the estimate for the daily percentage of departure delays.

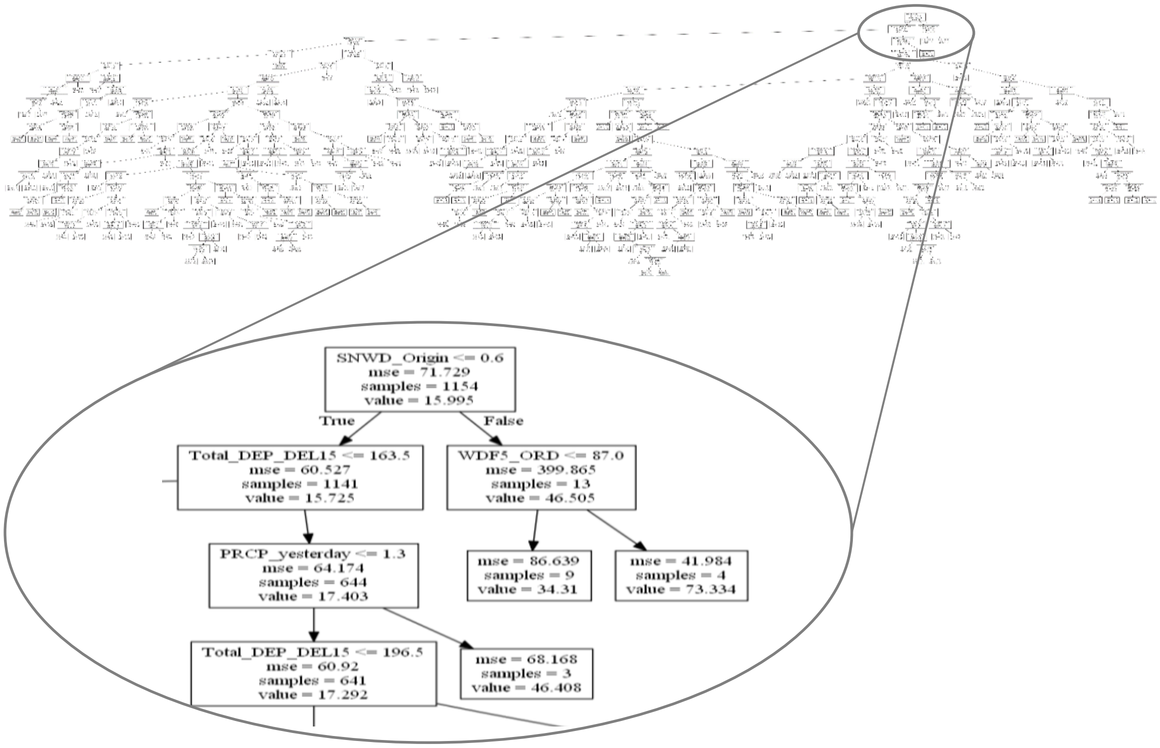


Figure 21: The first tree of the random forest.

3.4.2.1. FEATURE IMPORTANCE

In order to measure the usefulness of all the explanatory variables in the entire random forest model, we can take a look at the relative importance of the variables. Feature importance represents how much a particular variable is important for the predictions and is calculated as the decrease in the node MSE weighted by the probability of reaching that node. To calculate the node probability, we need to divide the number of samples that reach the node, by the total number of samples. When these values are higher for a feature it shows that the feature is more important.

In a decision tree regressor, the node importance is calculated using MSE, and by assuming only two child nodes (binary tree) it is calculated as follows:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)},$$

where ni_j is the importance of the node j, w_j is the weighted number of the samples reaching the node j, C_j is the MSE value of the node j, $left(j)$ indicates the child node from the left split on the node j, $right(j)$ indicates the child node from right split on the node j. The importance for each feature on a decision tree is then calculated as:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k},$$

where $fi_i$is the importance of the feature i, ni_j is the importance of the node j.

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$norm\ fi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}.$$

The final feature importance, at the random forest level, is its average over all the trees. The sum of the feature importance value on each tree is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} norm\ fi_{ij}}{T},$$

where $RFfi_i$ is the importance of the feature i calculated from all trees in the random forest model, $norm\ fi_{ij}$ is the normalized feature importance for i in tree j, and T is the total number of trees.

3.5. EVALUATING THE MODELS

3.6. MODEL METRICS

In this work, we used and reported MAE, RMSE, and R^2 scores to assess the model performances.

3.6.1. MEAN ABSOULUTE ERROR

Mean absolute error (MAE) is one of the measures for quantifying the quality of machine learning models. It is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i |, \tag{4}$$

where y_i and $\hat{y}_i$ are true value and prediction, respectively, and n is the number of observations.

3.6.2. ROOT MEAN SQUARE ERROR

Another measure for quantifying the quality of machine learning models is root mean square error (RMSE) and is calculated as:

$$RMSE(t) = \sqrt{\frac{1}{n}\sum_{i\in 1}^{n} (y_i - \hat{y}_i)^2}, \tag{5}$$

where y_i and $\hat{y}_i$ are true value and prediction, respectively, and n is the number of observations.

3.6.3. COEFFICIENT OF DETERMINATION

Coefficient of determination (R^2) shows how much the variability of the response variable is explained by the predictors or the model and is given as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{6}$$

where y_i, $\hat{y}_i$, and $\bar{y}$ are true value, prediction, and mean of observations, respectively.

3.7. GRID SEARCH

Machine learning models have hyperparameters that need to be set to modify the model to the dataset. Hyperparameters control training process of a model and are used to help estimate model parameters. They cannot be estimated from the dataset and must be set manually and tuned. Grid search is a technique to find the best set of performing hyperparameters and combinations of interacting hyperparameters for a given dataset. Using the values provided by the grid search result in the best performance of a model. In this study, we used the GridSearchCV function of the sklearn library for processing the grid search.

To perform a grid search for a given model, we defined a dictionary, where the dictionary keys were the model hyper parameters, and the dictionary values were discrete values, or a distribution of values for the hyper parameters.

we implemented a cross-validation object in the grid search to evaluate the model performance for each combination of hyperparameters. Cross-validation technique was repeated multiple times and the mean result across all runs was reported. For this process, the data split into K number of subsets, called folds, we then iteratively fit the model K times, and each time $K - 1$ subsets used to train the model and the other subset used for evaluating. The final results for the scores are the average of the scores calculated for each fold.

The score for regression models is a negative error metric, such as a negative version of the mean absolute error as given in Equation (7). Thus, maximizing the negative score and making that closer to zero leads to a better result and less prediction error by the model. Lastly, we used the best set of hyperparameters values corresponding to the least score and then fit the model on the data.

4. RESULTS

For the data preprocessing and model development we used open source libraries such as NumPy [17], Pandas [18], and Scikit-Learn [19] in **Python** programming language [20].

First, we applied the multiple linear regression model on the data to predict the daily percentage of delayed flights that is visualized in Figure 22. In viewing this figure, the predicted values using the regression model (red line) follow the main trend of the daily percentages of delayed flights (blue line) and would not be able to forecast the spikes except a few of them. The fit assessments of this model resulted in the MAE, the RMSE, and the R^2 scores of 5.23, 7.22, and 0.34, respectively, reported in Table 2. The coefficient of determination of 0.34 demonstrates that the regression model can explain only 34 percent of the variations of the target variable.

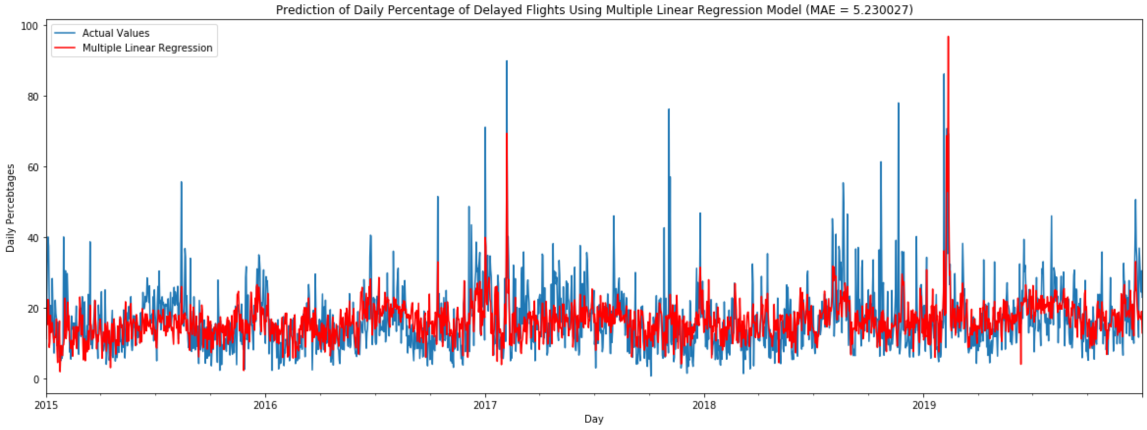


Figure 22:Prediction of the daily percentage of delayed flights using multiple linear regression model.

We split data to the training and testing sets in ratios of 70% and 30%, respectively. These partitions were invariant across various runs of the random forest model; we used training set to train the model and then to learn the hyperparameters for the random forest model; and we used the testing set only for examining the final evaluation metrics especially for checking for the overpredictions or the underpredictions.

We tuned the random forest regressor with a grid search for obtaining the best set of hyper-parameters. The grid search obtained the following values: 300 trees (n_estimators) with the maximum depth level (max_depth) of 40 for each tree. The maximum numbers of features (max_features) that the model is allowed to try in an individual tree were 50. The lowest numbers of samples needed to split an internal node (min_samples_split) were 10. The counts of samples that were needed as a minimum quantity to create a leaf (min_samples_leaf) were 2.

Lastly, we applied the random forest model on the data to predict the daily percentage of delayed flights which is visualized in Figure 23. In viewing this figure, the predicted values using the random forest (black line) follow the main trend of the daily percentages of delayed flights (blue line). The fit assessments of this model resulted in the MAE, the RMSE, and the R^2 scores of 2.68, 4.08, and 0.79, respectively, reported in Table 2. Note that the coefficient of determination of 0.79 demonstrates that the random forest model can explain 79 percent of the variations in the target variable.

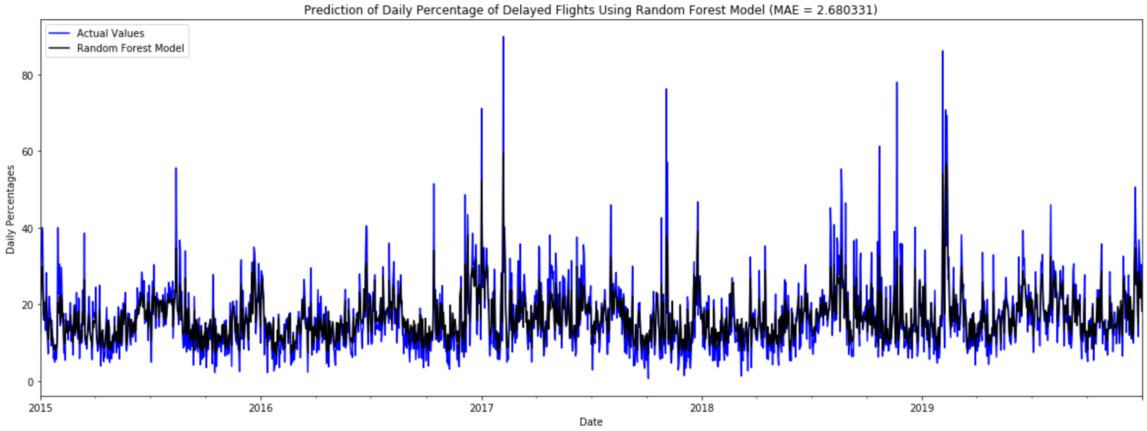


Figure 23: Prediction of the daily percentage of delayed flights using random forest model.

To compare the regression model with the random forest model, the predicted values obtained from both models are visualized in Figure 24. Since the data is large, the comparison using one graph is difficult. Hence, the predictions for both models for each year from 2015 to 2019 plotted in Figure 25 to Figure 29, separately. In viewing these figures, the predicted values using regression model (red line) follow the main trend of the daily percentages of delayed flights (blue line), however, the model would not be able to predict the spikes. In compare to the regression model, the predictions using the random forest algorithm (black line) made better predictions especially in predicting more spikes. According to Table 2 for the model’s assessment scores, the random forest technique outperformed the regression model.

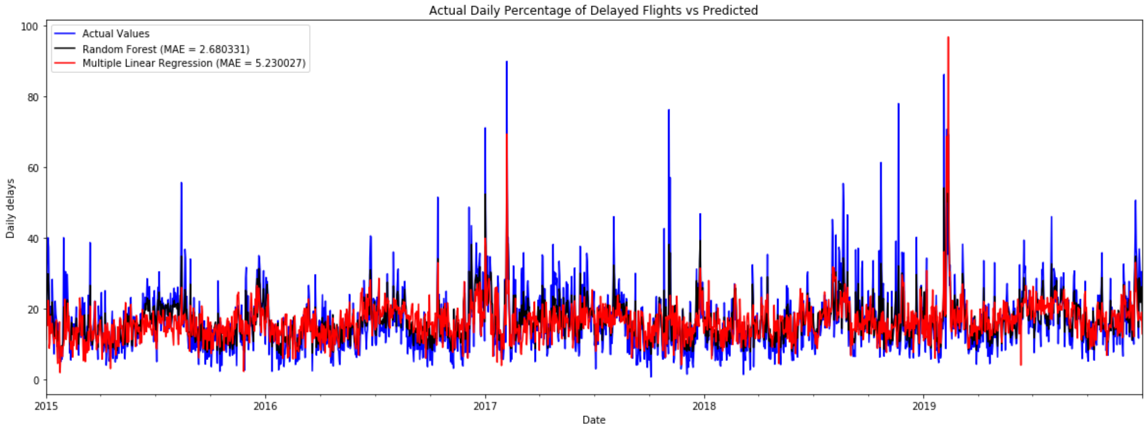


Figure 24: Prediction of daily percentage of delayed flights using regression and random forest model.

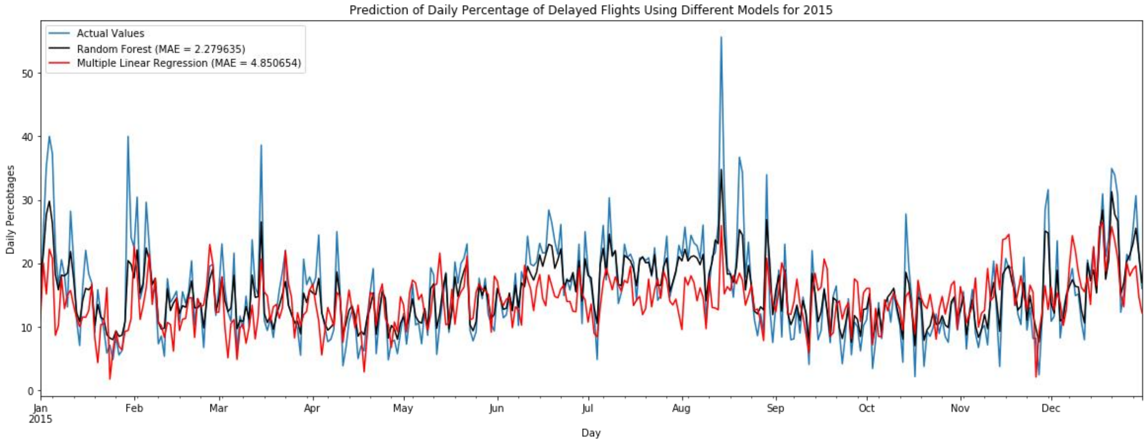


Figure 25: Prediction of daily percentage of delayed flights using regression and random forest model for 2015.

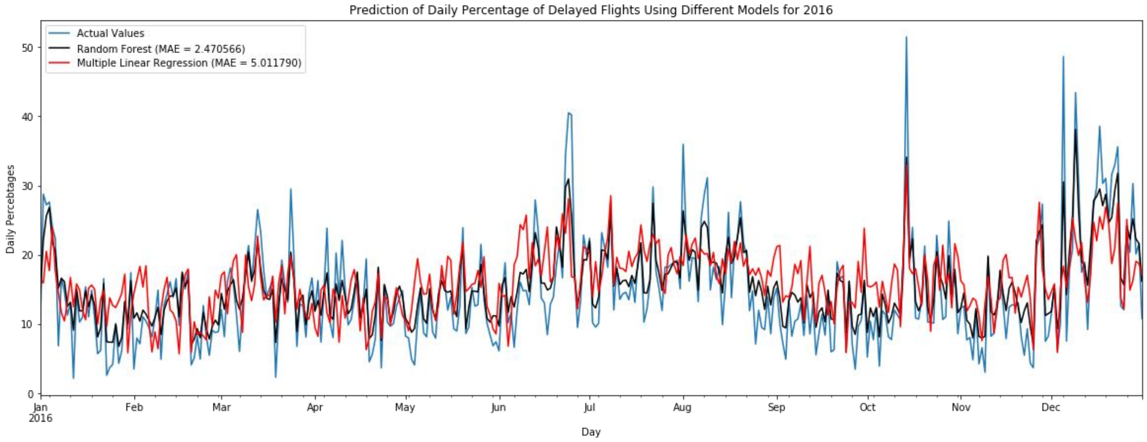


Figure 26: Prediction of daily percentage of delayed flights using regression and random forest model for 2016.

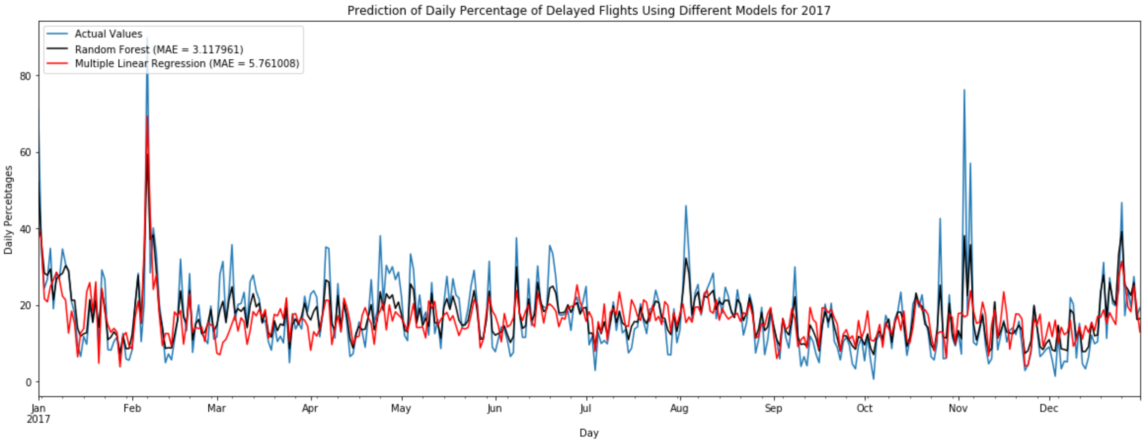


Figure 27: Prediction of daily percentage of delayed flights using regression and random forest model for 2017.

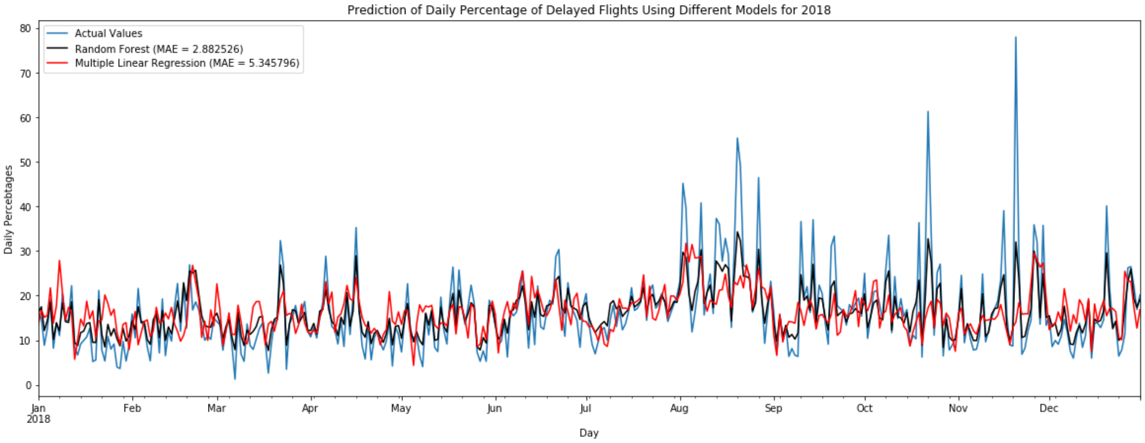


Figure 28: Prediction of daily percentage of delayed flights using regression and random forest model for 2018.

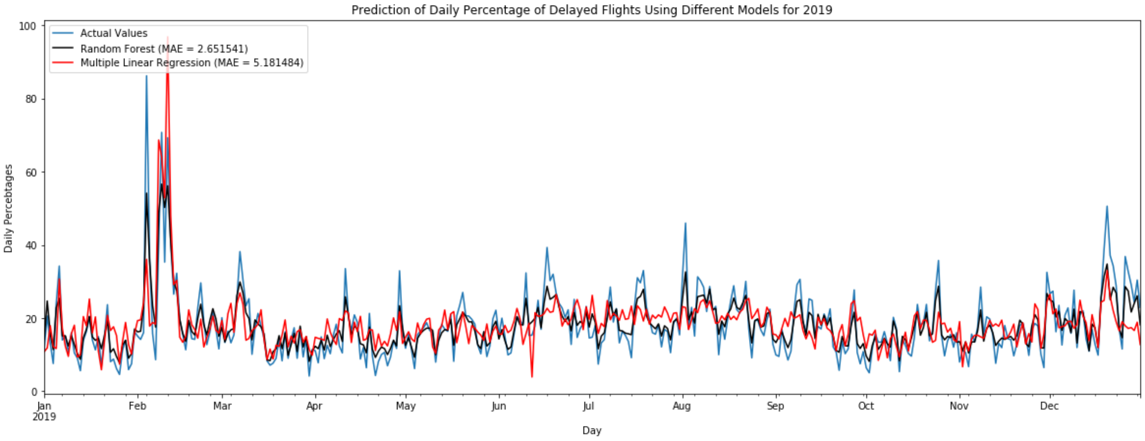


Figure 29: Prediction of daily percentage of delayed flights using regression and random forest model for 2019.

The top 50 important predictors for the random forest model illustrated in Figure 30. It shows that the departure flight features such as quarter or month of a year are the most important variables in the random forest model. The second most important set of predictors are the weather conditions at the flight date or even on the day before the flight for the origin. Additionally, the weather conditions for the destinations are other important features in the random forest model.

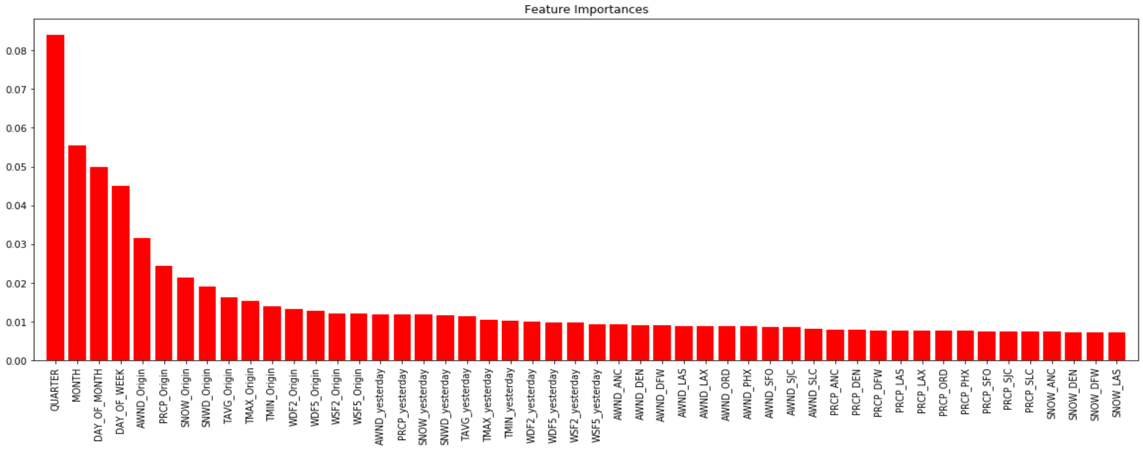


Figure 30: Feature importance for the random forest model

Table 2: Evaluation metrics for predictions

	Multiple Linear Regression	Random Forest
MAE	5.23	2.68
RMSE	7.22	4.08
R^2	0.34	0.79

5. CONCLUSION

In this study, we trained the models on the flight and weather datasets. The random forest model outperformed the multiple linear regression model according to the results of Table 2, and can be used to make a prediction of the daily percentage of delayed departure flights at the airports. Furthermore, some features were more important for training the random forest model, such as quarters, months, days of week, variables of the weather data for the origin at the date of flight and for a day before departure, variables of the weather dataset for the destination airports such as average wind speed and precipitation. Interestingly, the number of flights per day did not contribute much to increase the likelihood of the daily percentage of delayed flight.

A finer grid search might tune the model hyperparameters better; and considering other predictors besides the weather features can improve the prediction accuracy. Also, building some other models such as gradient boosting algorithm and neural network can help to find the best performing model to predict the daily percentage of delayed flight.

6. REFERENCES

[1] M. Ball, C. Barnhart, M. Dresner, M. Hansen, K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani and B. Zou, "Total Delay Impact Study," November, 2010.

[2] "Breau of Transportation Statistics," [Online]. Available: http://www.transtats.bts.gov/.

[3] S. S. Allan, J. A. Beesley, J. E. Evans and S. G. Gaddy, "Analysis of Delay Causality at Newark International Airport," *Lincoln Laboratory Massachusetts Institute of Technology Tech*, 2001.

[4] S. Oza, S. Sharma, H. Sangoi, R. Rutuja and V. C. Kotak, "Flight Delay Prediction System Using Weighted Multiple Linear Regression," *International Journal Of Engineering And Computer Science,* vol. 4, no. 4, pp. 11668-11677, 2015.

[5] D. Shah, A. Lodaria, D. Jain and L. D'Mello, "Airline Delay Prediction using Machine Learning and Deep Learning Techniques," *International Journal of Recent Technology and Engineering,* vol. 9, no. 2277-3878, pp. 1049-1054, 2020.

[6] L. Belcastro, F. Marozzo, D. Talia and P. Trunfio, "Using Scalable Data Mining for Predicting Flight Delays," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, July 2016..

[7] N. Etani, "Development of a Predictive Model for On-time Arrival Flight of Airliner by Discovering Correlation Between Flight and Weather Data," *Journal of Big Data,* vol. 6(85), 2019.

[8] Y. Ding, "Predicting Flight Delay Based on Multiple Linear Regression," *IOP Conf. Series: Earth and Environmental Science,* vol. 81, no. 1, p. 012198, 2017.

[9] T. Elangovan, M. Raheem and A. Arshad, "Predictive Analytics for Airline Departure," *Journal of Critical Reviews,* vol. 7, pp. 4034-4039, 2020.

[10] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta and S. Barman, "A Statistical Approach to Predict Flight Delay using Gradient Boosted Decision Tree," *International Conference on Computational Intelligence in Data Science(ICCIDS),* 2017.

[11] A. M. Kalliguddi and A. K. Leboulluec, "Predictive Modeling of Aircraft Flight Delay," *Universal Journal of Management,* vol. 5(10), pp. 485-491, 2017.

[12] K. Ebenezer and K. Brahmaji Rao, " Machine Learning Aproach to Predict Flight Delays," *International Journal of Computer Sciences and Engineering,* vol. 6, pp. 231-234, 2018.

[13] N. Kuhn and N. Jamadagni, "Application of Machine Learning Algorithms to Predict Flight Arrival Delays," 2017.

[14] N. Chakrabarty, "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines," in *9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON),* 2019.

[15] "National Centers for Environmental Information," [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/search.

[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, p. 5–32, 2001.

[17] T. E. Oliphant, A guide to NumPy,, vol. 1, Trelgol Publishing USA, 2006.

[18] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, pp. 51-56, 2010.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, vol. 12, pp. 2825--2830, 2011.

[20] G. a. D. J. F. L. Van Rossum, Python tutorial, Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.