2015 Flight Delays and Cancellations

Nikita Ganvkar
Computer Science and Engineering
PES University
Bangalore

Ishaan Samant
Computer Science and Engineering
PES University
Bangalore

R.Darsini
Computer Science and Engineering
PES University
Bangalore

***Abstract*: This report primarily focuses on analyzing flight delays based on the flight information and delay data. The purpose of the analysis is to ultimately be able to predict the delays and look into the factors that affect delay. This is accomplished using various regression models, and ultimately settling on polynomial regression.**

Keywords: Delay, Airlines, Regression, EDA, Visualization

I. Introduction

In the airline industry, efficient schedule implementation is a necessity in order to establish a structured airline business. A major issue in this industry is recurrent flight delays which may occur due to unforeseen circumstances. Hence, incorporating flight delay prediction into the aviation industry is required to stay on top of the line and to keep up with the current air travel demand.

The Federal Aviation Administration (FAA) guidelines state that a flight is considered delayed when it exceeds its scheduled arrival/departure time by fifteen minutes. Some common reasons for flight cancellation/delays are: Fueling and flight maintenance, unexpected aircraft parts malfunctioning, security issues, extreme weather conditions, flight delay due to previous flight's delay and so on.

Repercussions due to Delay of Flights:

- Loss incurred by airlines: If an airline's plane occupies the tarmac for a considerable amount of time, they are charged a hefty penalty by the FAA.
- Loss incurred by passengers: Flight delays may cause disruption in the passengers' travel plans. Delays in flights may cause a passenger to miss their next connecting flight and hence may incur loss of money and time. Disappointment, vexation and even Air rage is witnessed at certain times.

The project's main goal is to be able to figure out the amount of delay caused based on basic flight details and schedule. The biggest challenge with this dataset is the size of the dataset and the amount of missing values that are present in this dataset. The initial approach that we are going for is to try out various models, one set for arrival delay and finally a polynomial regression model to predict departure delay . The library functions that we intend to use for this are pandas, numpy, scipy, sklearn and scikit-learn. We can try to implement different types of regression like Simple Linear Regression, Lasso/Ridge Regression, Polynomial Regression or any combination of these if we need to.

II.II. Literature overview

A. Paper [1] *:Detecting periodic patterns of arrival delay*

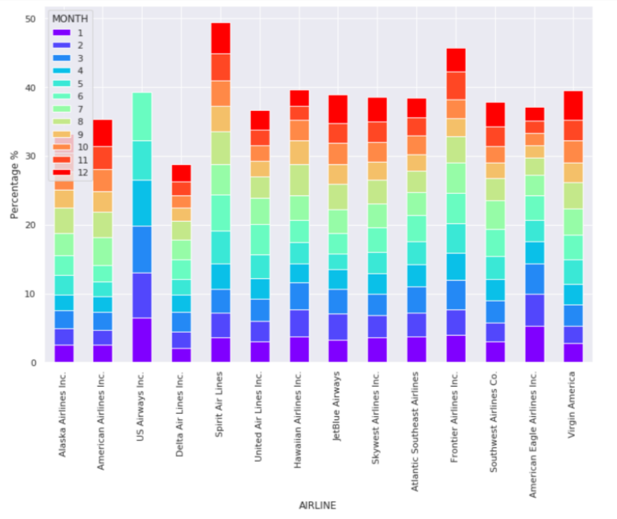
The study identifies the periodic patterns of arrival delay for the Orlando International Airport in 2001-2003. It looks into finding recurring delay patterns (and the frequencies of these patterns) and also detects the factors associated (such as weather) with these patterns This is done using a "two-stage approach", where mathematical frequency analysis and statistical analysis techniques are used.

*Assumption***:** By using their novel two-stage approach, they would be able to identify regularly repeating events in flight delay, and based on the delay, they can identify important variables too. The results may be able to show the flight delay trend and help airport authorities create constructive strategies to lessen flight delay.

Approach: Various approaches were explored; transforming the dependent variable to its logarithm to develop a nonlinear regression, however, it did not improve the fit of the model. Other methods such as the analysis of variance (ANOVA) were able to find the variables that were significant statistically. In terms of periodic patterns the LSD(least significant difference) and Tukey's tests showed that on Thursday and Friday, average daily delay was notably higher than the rest of the weekdays, and similarly Tuesday and Saturday's delay was comparatively lower than the rest. Additionally, the tests also picked up seasonality patterns. Out of all the seasons, summer's average daily delay was particularly higher. This may be attributed to the fact that in Orlando, summer is the rainiest season.

*Main Claims***:** The paper showed that using frequency analysis, significant periodic patterns (daily, weekly and seasonal) are present in the arrival delay. By using statistical analysis it is seen that factors such as the flight distance, the season, the time and week of the day were significantly correlated with the arrival delay.

Since one of the objectives of this project is to make analysis on the basis of periodicity of time, this paper allows us to look into how that can be made possible, and to look out for the various factors that may influence the periodicity in the delay of the flight. Moreover, as the paper suggests, the findings and the approaches implemented in the study can be extended to other airports as well, including our own airline delay dataset.



B. The figure above shows the variation in the number of flights scheduled in each airline according to months.

C. Paper [2]*: Predicting flight delay based on multiple linear regression*

Highlights:

- The paper presents experiments performed on a reliable dataset of various airports and shows that the accuracy of the proposed model approximates 80%, which is arguably better than the C4.5 or the Naive-Bayes approaches.

- The paper states that this method of prediction is easier for calculation, and also can predict delays accurately.

Data Source: The dataset for this paper was taken from the website *www.umetrip.com*. The dataset consists of records of 119432 domestic flights from November 3, 2015 to March 5, 2016.

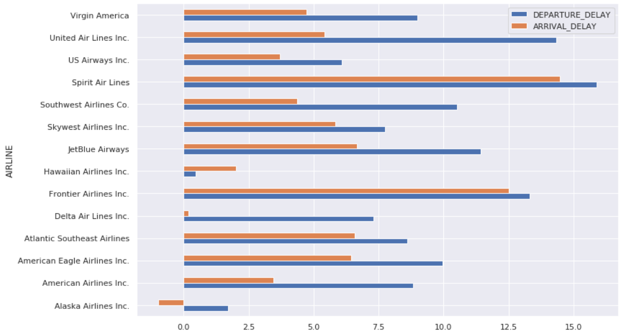
Problem Statement: To predict flight arrival delay at a given airport using Flight Departure time.

Approach:
In this paper, a dual prediction approach is used:

1) Regression (which gives an estimation of the arrival delay).
2) Classification, which gives a simple ‘YES/NO’ labeled prediction stating whether the arrival delay is more or less (as defined by the set threshold).

Implementation: The author has planned to predict arrival delay using departure delay. In order to predict whether a flight has been delayed or not, the author has proposed a model with two labels:
There is no flight delay is the delay time falls between [-INFINITY,30].
There is a delay if the delay time falls between [-INFINITY,30].

D. The figure below shows departure and arrival delay for each airline.



A multiple linear regression equation : $Y = B_1X_1+B_2X_2+B_3$ is used where the target variable Y is the arrival delay and the predictor variables are : X1 is departure delay and X2 is route distance.
The author has added a new field called departure delay by taking the difference between the planned departure time and the actual departure time. In the next phase, he has used the attribute departure delay along with a route-distance training model to find the values of B1, B2 and B3.

Conclusion: The problem statement was originally considered as a regression task followed by a categorical classification task and then a multiple linear regression model was used to predict the delay.

E. Paper [3]*: Flight delay prediction based on aviation big data and machine learning*

This study implements machine learning methods to predict flight delay based on aviation big data. The data is obtained from automatic dependent surveillance-broadcast (ADS-B) messages which in essence are sequential data periodically broadcasted by ongoing flights. The ADS-B messages are further integrated with attributes such as airport information, flight schedule and weather conditions (sourced from various sites mentioned in the paper).

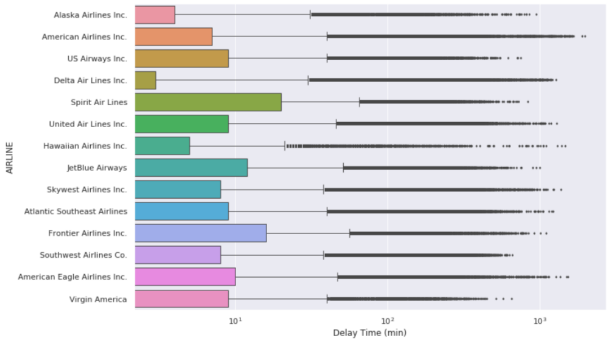
Approach: The paper incorporates three major models - LSTM (long short-term memory) neural network(NN), Random Forest Based Model, and a basic regression model, with the focus primarily on the first two models. LSTM is chosen to convey the potential of NNs in the field of aviation, while the other models are used to see how LSTM performs against these classic models. In terms of the model plan, the study covers all routes and airports obtained from ADS-B unlike conventional schemes that focus on a single route or single airport.

Implementation: Weather(w), time(t) and flight schedule(s) are allocated their own vectors for their respective attributes, and are included in the input vector x along with the flight date(d), ICAO identity number(n1), flight number(n2) and the route's traffic flow(f) in the form of x = [d; n1; n2; f;w; t; s]. As the main goal is to get acceptable accuracy results, the proposed model also attempts to fit in a hidden function for both regression and classification models. Ultimately, if the prediction is handled as a classification task, the outcome would be the predicted probability of each of the delay classes.

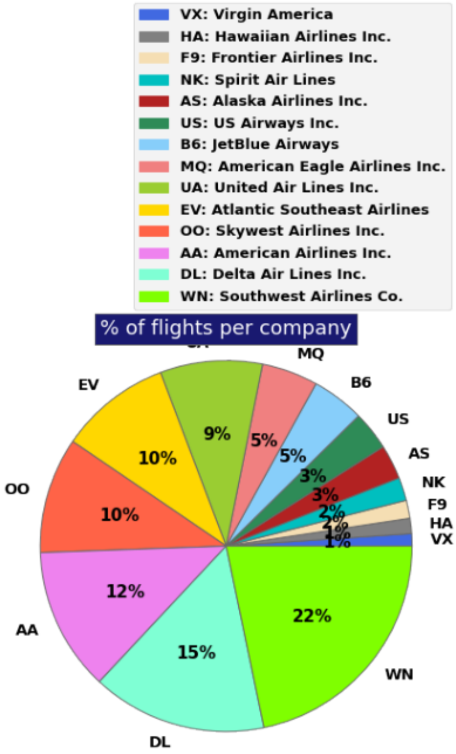
The LSTM plays different memory depths with three different network architectures - standard, added fully-connected and added dropout.. The Random Forest-based Classifier handles three types of tasks - binary classification, three categories and four categories classification. Finally the training and testing accuracy is computed for each instance.

Outcomes: Since the dataset is still not adequate enough for the LSTM NNs, overfitting occurs and thus its not an optimum solution. Nevertheless, the inclusion of a dropout layer in contrast to a fully connected layer did show better testing accuracy, hence it does have a future scope in this field. In hindsight however, the generalization ability of the random forest based method is much stronger and suited for the dataset (with a testing accuracy of 90.2%). Since the primary aim of the study was to introduce NNs into this field, it does show the potential of LSTM-based methods, with the claim that it can give a promising performance if the overfitting problem can be overcomed using a larger dataset (one or two more years worth of data according to the paper).

F. The figure below is a boxplot of delay time (min) vs airlines.



G. The figure below shows the percentage of carriers for each airline company



III.III. Exploratory data analysis

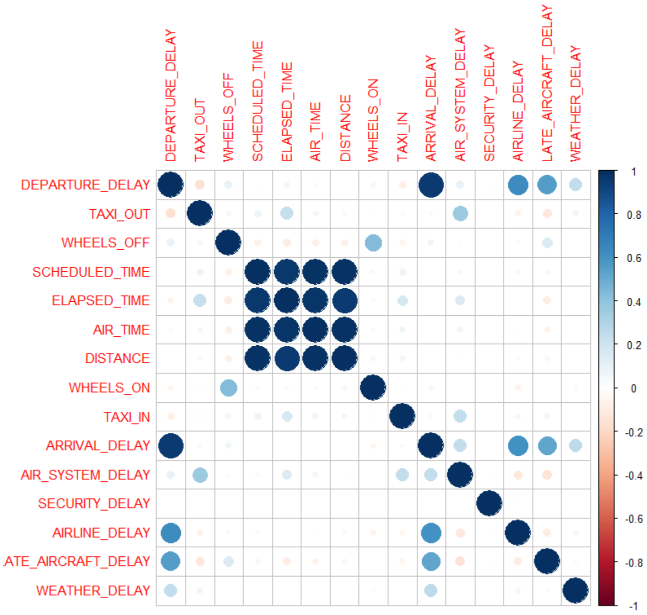
The entire dataset consists of the three csv files - 'airlines.csv', 'airports.csv', and 'flights.csv'. The airlines dataset contains the list of all the commercial airlines along with its IATA code (Eg: UA- United Airlines), which we will be using to refer to the corresponding airlines throughout the datasets. The 'airports.csv' contains the list of airports along with geographical details. The airports too are referred to with their own unique IATA codes.
"airports.csv" - Attributes: 7, Instances: 322
"airlines.csv" - Attributes: 2, Instances: 14

The primary dataset – flights.csv contains many details about the flights along with time- related attributes. The EDA will mainly focus on this dataset:
"flights.csv" - Attributes: 31, Instances: 5819079
The categorical variables (12) contain basic aircraft details such as the airlines, airports etc. and also if the particular flight has been cancelled with its reason.
The numerical values focus on the delay timings and what factors contribute how much to the delay.

Focussing on the flights dataset, we can find the correlation between the numerical variables available and infer that out of all the delay factors, weather delay, late aircraft delay (when a previous flight of the same aircraft arrives late, which causes the present flight to depart late) and air system delay (cumulatively refers delay due to airport operations) influence the departure and arrival delay the most.

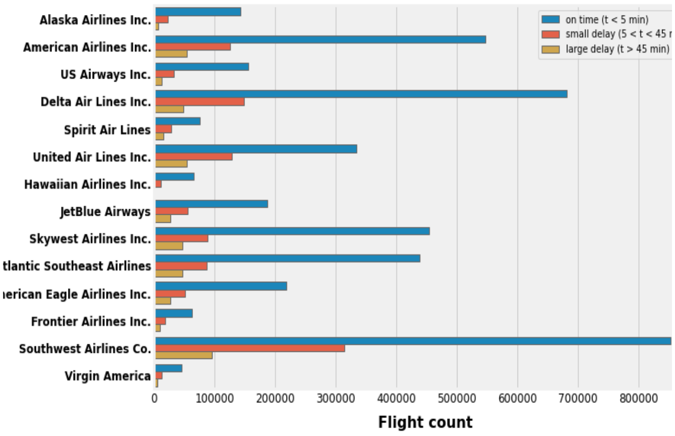
Since the data is consistent and does show normal behaviours or trends, data preprocessing techniques are not necessary in our case. However, due the high presence of null values, data cleaning would be done prior to visualizing the data and building the model.

.

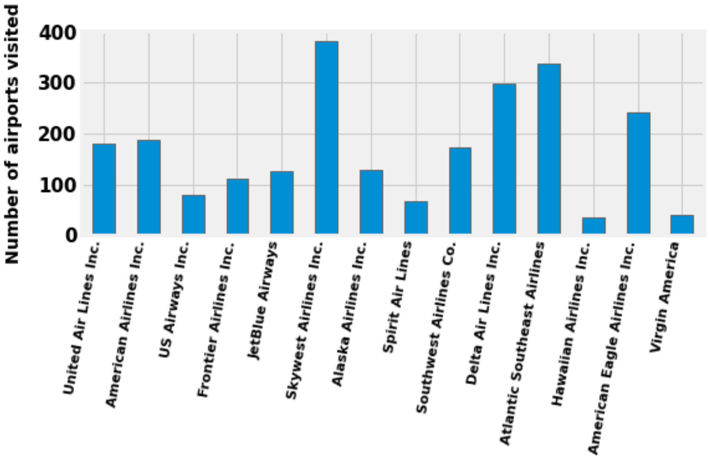


IV.IV. Visualization

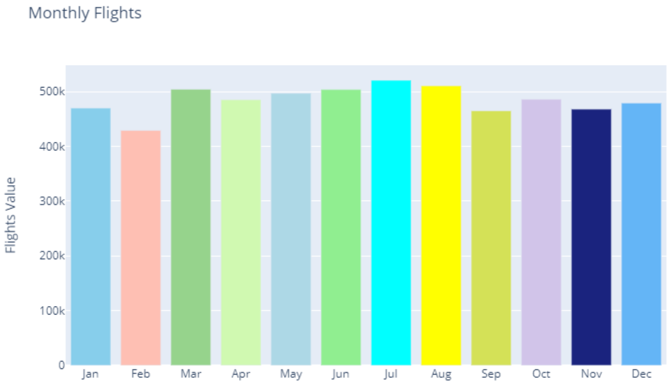
A. a) The figure below gives the frequency of the flight delays in three ranges: lesser than five minutes, between five and forty-five minutes and more than forty-five minutes . It can be seen that for any type of airline, flight delays greater than forty-five minutes are very few. For instance, SkyWest Airlines, have delays > 45 minutes by approximately 30% lesser than with reference to the delays in the range of five to forty-five minutes. SouthWest Airlines seems to be doing better since delays > 45 minutes are four times less frequent than the flight delays of range 5 minutes < t (time)< 45 minutes.



B. b) The figure below shows the number of destination airports for each airline

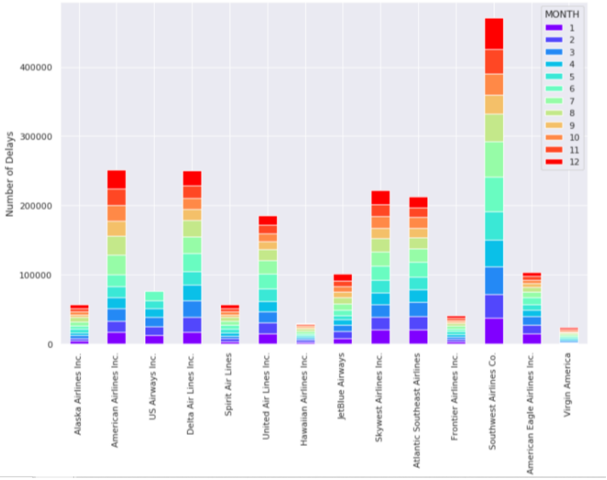


C. c) The figure below shows the number of flights for each airline month.



D.

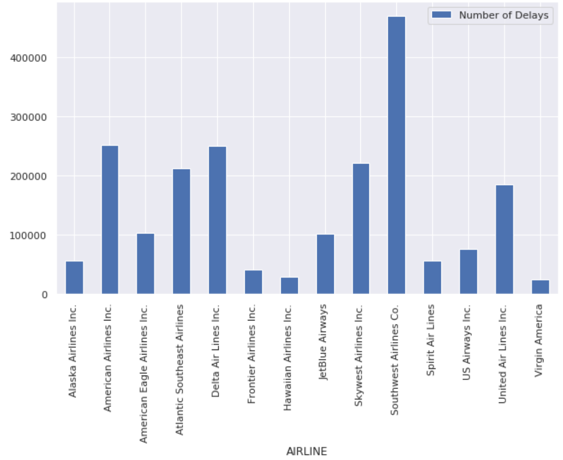
E. d) The figure below shows the variation in the number of delays in each airline according to months.



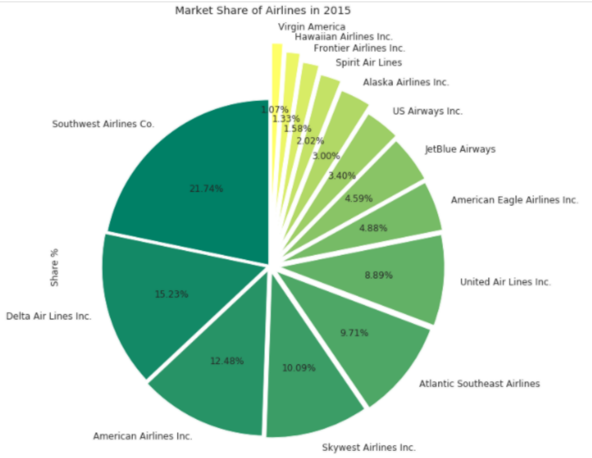
F.

G.

H. e) The figure below shows the number of times a flight of a particular airline company got delayed.



f) The figure below shows the market share of each airline in 2015.



V. Proposed Solution

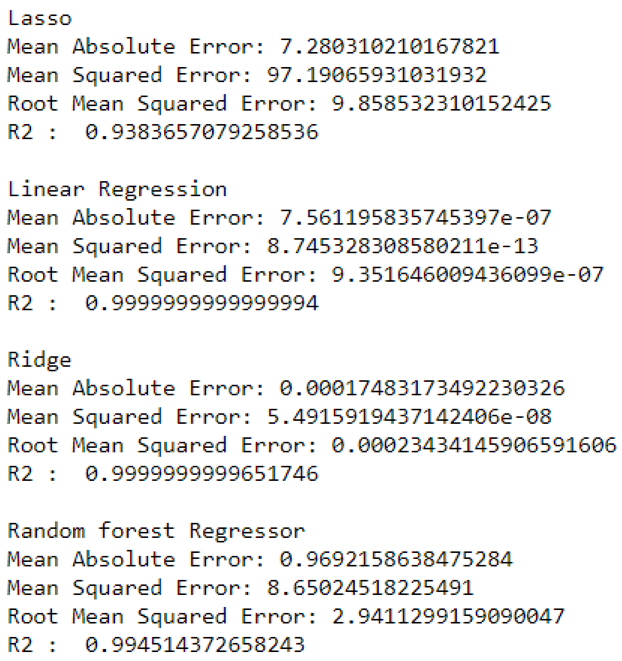
Before implementing the models, several factors of the dataset have to be pre-processed to make it operable. In the dataset, there were certain columns that contained a lot of null values, rendering them unusable (such as air security delay). Due to the lack of contribution from such variables, these attributes were dropped from the model. Another issue which was encountered in the dataset was the formats of certain values - modification had to be done to the time-related attributes to be made suitable for the model. The attributes such as *Airport* in the regression model encoded using a label encoder in order to incorporate it into the model.

In order to predict delay in flights, one of the two predictor variables that can be used to build the model are - Arrival Delay and Departure Delay. The preliminary models would use Arrival Delay to be predicted, and according to the results, a model or a combination of models can be used to predict the departure delay. The models being tested are regression models and their variants - Linear Regression, Lasso Regression and Ridge Regression. As the Random Forest model is a popular method used in the aviation sector to find out delays, it's framework is also implemented.

On computation of the first set of models, we have decided to predict departure delay by implementing ridge regression to incorporate with the polynomial regression model to provide an overall better fit. With this scheme, we have implemented the models of the project.

VI. Experimental results

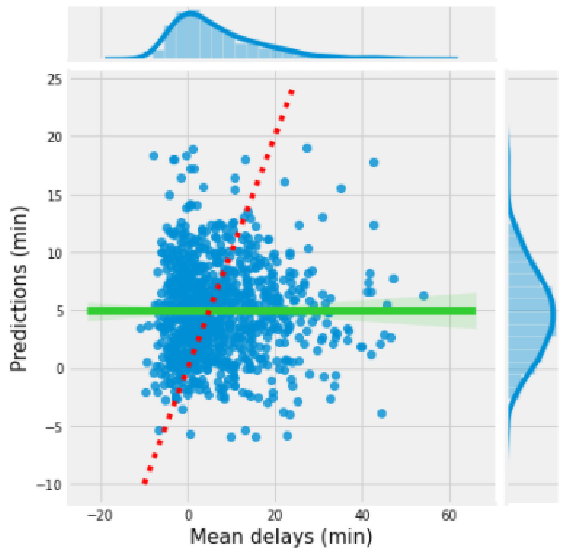
Phase 1: Implementation of Linear Regression and their variants - Lasso and Ridge, along with Random Forest Regressor.



```
Lasso
Mean Absolute Error: 7.280310210167821
Mean Squared Error: 97.19065931031932
Root Mean Squared Error: 9.858532310152425
R2 :  0.9383657079258536

Linear Regression
Mean Absolute Error: 7.561195835745397e-07
Mean Squared Error: 8.745328308580211e-13
Root Mean Squared Error: 9.351646009436099e-07
R2 :  0.9999999999999994

Ridge
Mean Absolute Error: 0.00017483173492230326
Mean Squared Error: 5.4915919437142406e-08
Root Mean Squared Error: 0.00023434145906591606
R2 :  0.9999999999651746

Random forest Regressor
Mean Absolute Error: 0.9692158638475284
Mean Squared Error: 8.65024518225491
Root Mean Squared Error: 2.9411299159090047
R2 :  0.994514372658243
```

The phase 2 model uses ridge regression instead of lasso regression since feature selection is not needed as the number of variables in our model is not a high number of variables, hence ridge regression is appropriate.

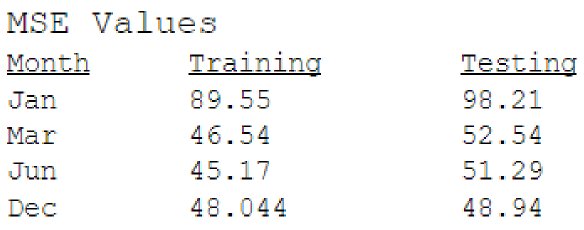
Phase 2: Polynomial Regression, including ridge regression.



Before implementing the model, a few extra tweaks have to be done to the dataset. After the label encoding done on attribute Airport (done in the initial pre-processing), these values are combined into a matrix with one hot encoding where each variable contains M labels with M columns and filled with 0 or 1 depending on the correspondence with the particular airports.

Finally the input variables used were the departure time, arrival time, origin airport, average departure delay and the average weekday with all of these values grouped by for a certain destination airport. Due to the small size of the dataset there was a high chance that to get a desirable accuracy, the polynomial regression model would have to overfit, thus to prevent overfitting we added a penalization criteria, the criteria that we chose was ridge regression.

To find the optimal polynomial order and the optimal value of alpha we ran the model on a training set for all the combinations of alpha and n and calculated the MSE. The range of values of n we took was 1 to 3 and the range of the alpha values we took was 0 to 20 but in steps of 2. Finally, we evaluated the model using the test data.


```
MSE Values
Month       Training        Testing
Jan         89.55           98.21
Mar         46.54           52.54
Jun         45.17           51.29
Dec         48.044          48.94
```

VI. Conclusions

Using basic attributes extracted from the flight information, preliminary analysis can be done to visually represent the effect of factors on the departure delay. This can be further explored and validated using regression models; however, choosing the right one greatly affects the prediction accuracy as shown by the models done.

Originally we thought that factors such as departure time, arrival time, original airport and average departure delay would affect the delay timings and after doing some intensive analysis, we can conclude the same.

Contributions:

- Nikita G: Report and Model
- Ishaan Samant: Data Pre-processing and Model
- R.Darsini: Data Visualization and Model

VII. References

[1] Abdel-Aty, M., Lee, C., Bai, Y., Li, X. and Michalak, M., 2007. Detecting periodic patterns of arrival delay. Journal of Air Transport Management, 13(6), pp.355-361.

[2] Ding, Y., 2017, April. Predicting flight delay based on multiple linear regression. In IOP Conference Series: Earth and Environmental Science (Vol. 81, No. 1, pp. 1-7).

[3] Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z. and Zhao, D., 2019. Flight delay prediction based on aviation big data and machine learning. IEEE Transactions on Vehicular Technology, 69(1), pp.140-