

Flight Delay Prediction

Mahesh Bharadwaj K

20th March, 2020

Abstract

Flights are said to be delayed when they arrive later than the scheduled arrival time. The delay can be due to several factors including but not limited to adverse weather conditions. This project aims to design a two staged machine learning model to predict if a flight will be delayed upon arrival due to weather conditions before departure and if so, the arrival delay in minutes

1 Introduction

The Federal Aviation Authority [FAA] considers flights to be delayed if the arrival delay is greater than 15 minutes. Delays not only cause agony to the travellers but also have a *domino effect* on the background work involving allocation of gates at the airport, ground crew, wastage of food to name a few. This causes *losses of billion of dollars* to the airline and hence, it is of importance to predict if a flight will be delayed upon arrival and if so, the actual delay in minutes must be predicted with a high degree of accuracy. **This project aims to design a two stage model to classify and predict the arrival delay of flights** based on the data set selected which contains flight data of all flights in USA from 2016 to 2017 and weather data pertaining to 15 selected airports during the same time frame. In this project, the performance of various classification and regression models is studied and compared.

2 Data Set

The flight data set contains flights from all airports in the USA from 2016 to 2017. The weather data set contains the weather data obtained from 15 airports in the years 2016 - 2017. The flights for which the weather data is available are selected and the corresponding weather data is appended to the flight data. The flight and weather data sets are merged based on: *Departure Airport, Departure Time, and Date*.

The Airports for which weather data is available are listed in table 1. The various weather data points considered are given in table 2. The various Flight Performance Metrics considered are given in table 3. There are 18,51,115 flights available in final the processed data set. 75% of these flights are designated to the training set and the remaining 25% of the flights are designated to the test set.

ATL	CLT	DEN	DFW	EWR
IAH	JFK	LAS	LAX	MCO
MIA	ORD	PHX	SEA	SFO

Table 1: **Airport Codes of the Chosen Airports**

WindSpeedKmph	WindDirDegree	WeatherCode	PrecipMM
Visibility	Pressure	CloudCover	DewPointF
WindGustKmph	tempF	WindChillF	Humidity
Date	time	SnowFallCM	

Table 2: **Weather Data Points Considered**

FlightDate	Quarter	Year	Month
DayofMonth	DepTime	DepDel15	CRSDepTime
DepDelayMinutes	OriginAirport	DestAirport	

Table 3: **Flight Performance Metrics Considered**

3 Classification

The classifier is the first stage of the two stage model mentioned and aims to classify flights as delayed or not. Flights which are delayed have the 'ArrDel15' = 1 and those which are not have 'ArrDel15' = 0.

3.1 Models Used:

- Logistic Regression
 - Decision Tree Classifier
 - Gradient Boosting Classifier
 - Random Forest Classifier
-

3.2 Performance Metrics for Classifier Models

3.2.1 Terms used:

1. **TP:** True Positives:
Flights Delayed Classified correctly as Delayed
2. **FP:** False Positives:
Flights On Time Classified as Delayed
3. **TN:** True Negatives:
Flights Not Delayed correctly classified
4. **FN:** False Negatives:
Flights Delayed classified incorrectly as not Not Delayed

3.2.2 Metrics

$$\text{Accuracy} = \frac{TP + TN}{\text{TotalNo.of Samples}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3 Classifier Performance

Classification Model	Performance Metric						
	Accuracy	Precision		Recall		F1 Score	
		0	1	0	1	0	1
Logistic Regression	0.92	0.92	0.89	0.98	0.68	0.95	0.77
Decision Tree Classifier	0.87	0.92	0.68	0.91	0.71	0.92	0.69
Gradient Boosting Classifier	0.92	0.92	0.90	0.98	0.69	0.95	0.78
Random Forest Classifier	0.92	0.92	0.88	0.98	0.70	0.95	0.78

Table 4: Performance of Classifier Models

3.4 Receiver Operation Characteristics of the Classifiers

Area under ROC is an indicator of the performance of the classifier model trained. Higher the area, better the performance of the model

- If the **Area** ≈ 0.5 , the model is unable to distinguish between the classes (represented by red line)
- If the **Area** ≈ 0.0 , the model is predicting the exact opposite of the actual classes
- If the **Area** ≈ 1.0 , the model is predicting the classes accurately

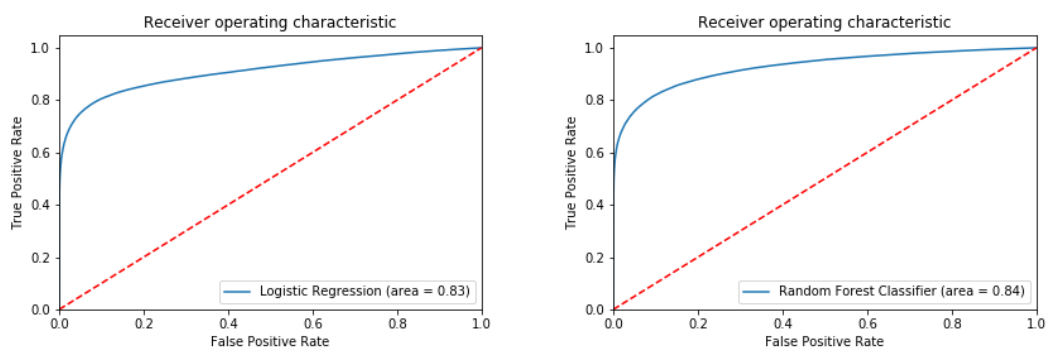


Figure 1: Logistic Regression & Random Forest Classifier

4 Class Imbalance In Data Set

The poor performance of the classification algorithms above is due to the **inherent bias** in the data set towards non-delayed flights.

Out of the 18,51,115 Flights in the data set, only 3,87,948 entries are delayed. The can be seen in figure 2.

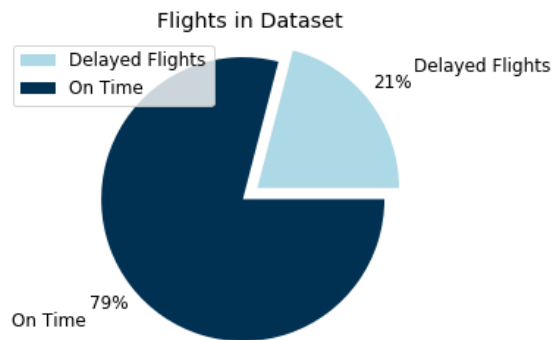


Figure 2: Pie Chart of Data Set Class Distribution Before SMOTE

To overcome this bias, there are two options

- **Under Sampling**

The Majority Class is under sampled to ensure there is an even distribution.

- **Over Sampling**

The Minority Class is over sampled to ensure there is an even distribution.

Here, Synthetic Minority Oversampling TEchnique referred to as **SMOTE** has been employed to oversample the non-delayed flights. The data set distribution after SMOTE is given in figure 3.

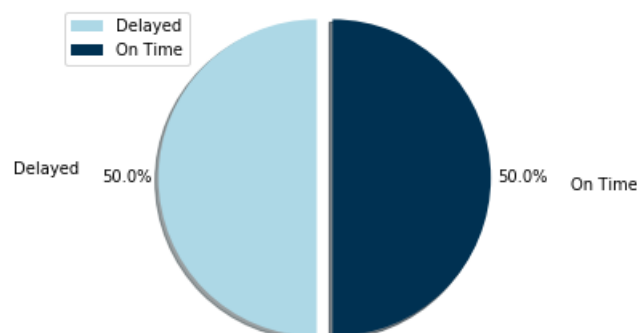


Figure 3: Pie Chart of Data Set Distribution After SMOTE

4.1 Classifier Performance After SMOTE

Classification Model	Performance Metric						
	Accuracy	Precision		Recall		F1 Score	
		0	1	0	1	0	1
Logistic Regression	0.90	0.94	0.74	0.93	0.78	0.93	0.76
Decision Tree Classifier	0.87	0.92	0.67	0.91	0.70	0.91	0.69
Gradient Boosting Classifier	0.91	0.93	0.85	0.97	0.72	0.95	0.78
Random Forest Classifier	0.91	0.93	0.83	0.96	0.73	0.95	0.78

Table 5: Performance of Classifier Models After SMOTE

4.2 Receiver Operation Characteristics After SMOTE

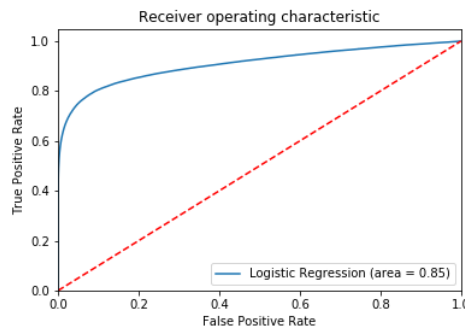


Figure 4: Logistic Regression After SMOTE

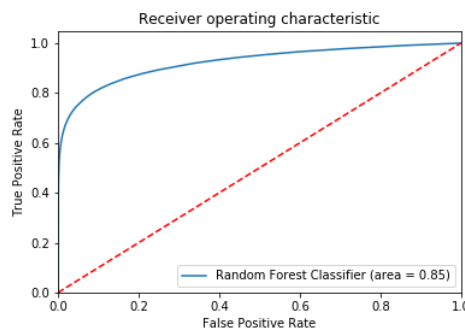


Figure 5: Random Forest Classifier After SMOTE

5 Regression

Regression is the second stage of the two stage model in which the Arrival Delay in minutes is predicted if the flight is classified as **Delayed** by the classifier. The flights having '**ArrDelayMinutes**' > 0 are used to train the regression model. The results of the various models used are located in table 6

5.1 Models Used:

- Linear Regression
 - Gradient Boosting Regression
 - Extra Tree Regression
-

5.2 Performance Metrics for Regression Models

$$\text{Mean Squared Error [MSE]} = \sum_{i=1}^n \frac{(\text{Predicted}_i - \text{Actual}_i)^2}{n}$$

$$\text{Root Mean Square Error [RMSE]} = \sqrt{\sum_{i=1}^n \frac{(\text{Predicted}_i - \text{Actual}_i)^2}{n}}$$

$$\text{Mean Absolute Error [MAE]} = \sum_{i=1}^n \frac{|\text{Predicted}_i - \text{Actual}_i|}{n}$$

$$R^2 \text{ Score} = 1 - \frac{\sum_{i=1}^n (\text{Predicted}_i - \text{Actual}_i)^2}{\sum_{i=1}^n (\text{Predicted}_i - y_{\text{mean}})^2}$$

5.3 Regression Performance

Regression Model	MSE	RMSE	MAE	R^2 Score
Linear Regression	271.87	16.49	11.24	0.927
Gradient Boosting Regression	225.32	15.01	10.18	0.940
Extra Tree Regression	233.52	15.28	10.52	0.937

Table 6: Regression Performance

6 Regression Analysis

The maximum delay in the data set was 2142 and the minimum delay was 0. Frequency Distribution plot (figure 6) of the arrival delay reveals that the maximum frequency is observed between 0 - 100 and 100 - 200. The data set was split into ranges of arrival delay minutes and performance is studied in these ranges (table 7).

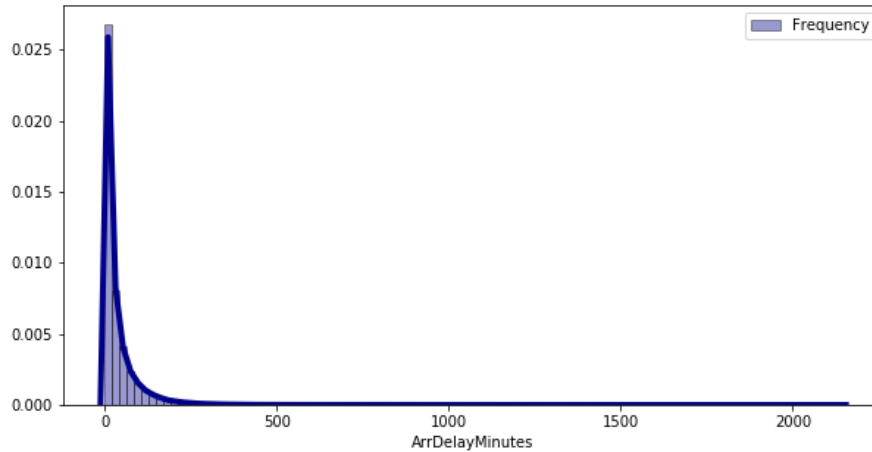


Figure 6: Frequency Distribution of Arrival Delay

RANGE	MSE	RMSE	MAE	R^2 Score
0 - 100	209.05	14.46	8.82	0.944
100 - 200	209.83	14.48	9.89	0.941
200 - 500	205.21	14.32	9.71	0.942
500 - 1000	221.72	14.89	9.82	0.919
1000 - 2000	220.01	14.83	9.71	0.901

Table 7: Range-wise Regression Analysis

The results obtained are in line with the frequency distribution histogram. Maximum R^2 is obtained in the range 0 - 100 which has the highest frequency and as the range increases, the R^2 decreases.

7 Pipelined Model

The Classifier chosen is Random Forest Classifier as it has the maximum F1 Score and are under ROC.

The Regression Model chosen in Gradient Boosting Regressor[XGBoost] as it has the highest R^2 Score.

The Flow Chart given below (figure 7) is the representation of the given two stage machine learning model to predict if a flight will be delayed upon arrival and if so, the actual delay in minutes.

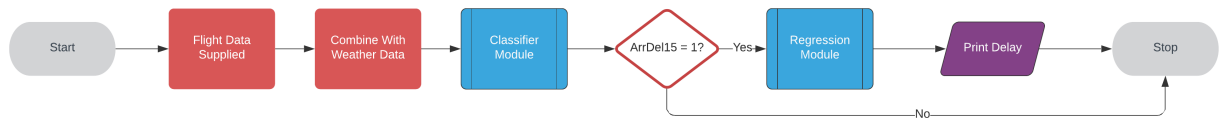


Figure 7: Flow Chart of Pipelined Model

Performance of Pipe-lined Model

Metric	Value
Mean Squared Error	164.01
Root Mean Squared Error	12.80
Mean Absolute Error	10.67
R^2 Score	0.911

Table 8: Pipe-lined Model Performance

8 Conclusion

The flight and weather data were combined into a single data set for training the models. It was observed from the classifier that bias in data set (towards non delayed flights) affected the performance of the classifier models chosen. This bias was overcome by oversampling the delayed flights in the data set using SMOTE. The *Classifier* chosen was **Random Forest Classifier** having the highest F1 Score of 0.78 . The *Regression Model* chosen was **Gradient Boosting Regressor** having the highest R^2 Score of 0.94. The pipe-lined model was designed using the aforementioned Classifier and Regression Modules and its performance serves as an indicator of the system performance.