Flight Delay Prediction

Mahesh Bharadwaj K

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. The aim of this project is to predict arrival delay of a flight after its departure using a two-stage machine learning model.If the flight is predicted to have an arrival delay, the delay in minutes is predicted.

1 Introduction

The Federal Aviation Authority (FAA) considers flights to be delayed if the arrival delay is greater than 15 minutes. Delays not only agonise 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 important to predict arrival delays accurately. This project aims to design a two stage model to classify and predict the arrival delay of flights based on the dataset 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 Dataset

The flight dataset contains flights from all airports in the USA from 2016 to 2017. The weather dataset 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 datasets 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 data points available in the processed dataset. 75% of these data points are designated as training data and the remaining 25% of the data points are designated as testing data.

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 and aims to classify flights as delayed or not. Flights which are delayed have the target variable *'ArrDel15'* = 1 and those which are not have the target variable *'ArrDel15'* = 0.

Models Used

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

Performance Metrics for Classifier Models

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 as Not Delayed
4. **FN:** False Negatives
   Flights Delayed classified incorrectly as Not Delayed

Metrics

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

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

Classifier Performance

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

Table 4: **Performance of Classifier Models**

Receiver Operation Characteristics of the Classifiers

Area under ROC curve 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

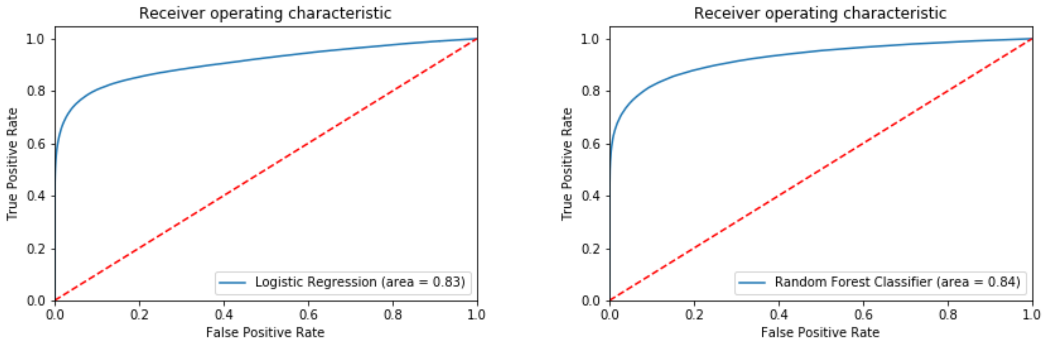

Figure 1: **Logistic Regression & Random Forest Classifier**

4 Class Imbalance In Dataset

The poor performance on class 1 relative to class 0 of the classification algorithms above is due to the inherent bias in the dataset towards non-delayed flights.
Out of the 18,51,115 data points in the dataset, only 3,87,948 data points are delayed. The distribution can be seen in figure 2.

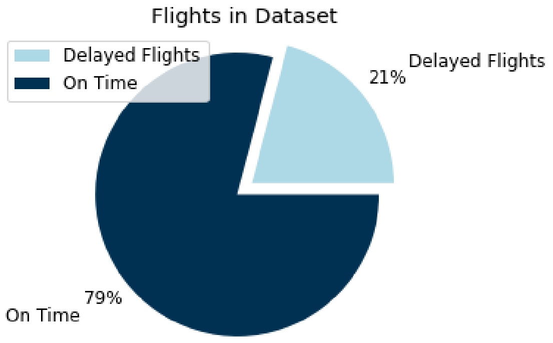

Figure 2: **Pie Chart of Dataset Class Distribution Before SMOTE**

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 over-sample the delayed flights. The dataset distribution after SMOTE is given in figure 3.

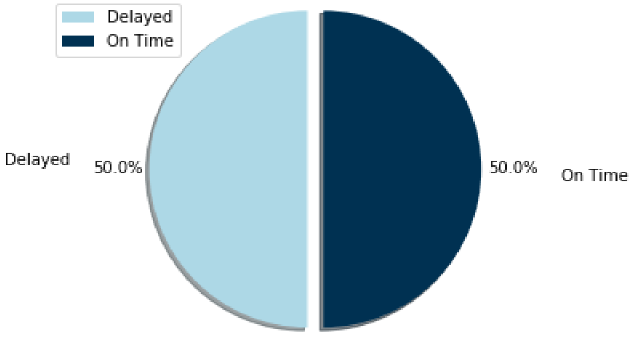

Figure 3: **Pie Chart of Dataset Distribution After SMOTE**

Classifier Performance After SMOTE

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

Table 5: **Performance of Classifier Models After SMOTE**

Receiver Operation Characteristics After SMOTE

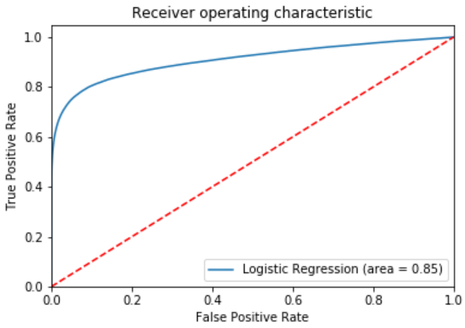

Figure 4: **Logistic Regression**

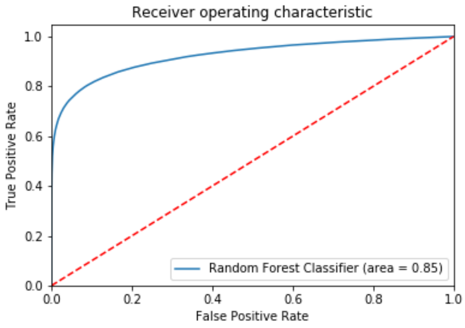

Figure 5: **Random Forest Classifier**

SMOTE oversampling improves the performance of Logistic Regression model by increasing the recall of class 1. Random Forest Classifier and Gradient Boosting Classifier do not show any improvements as they are built to work on imbalanced classes. F1 Score gives equal importance to precision and recall and hence, the classifier chosen was Random Forest Classifier having the highest F1 Score (0.78).

5 Regression

Regression is the second stage of the two-stage model. 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.

Models Used

- Linear Regressor
- Gradient Boosting Regressor
- Extra Tree Regressor

Performance Metrics for Regression Models

$$\textbf{Mean Squared Error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$\textbf{Root Mean Square Error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$\textbf{Mean Absolute Error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$R^2 \textbf{ Score} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

Regression Performance

Regression Model	RMSE	MAE	R^2 Score
Linear Regression	16.43	11.20	0.926
Gradient Boosting Regression	15.49	10.52	0.934
Extra Tree Regression	15.16	10.47	0.937

Table 6: **Regression Performance**

R^2 Score is a measure of the ability of a model to predict the variances in the dataset accurately. The Regressor chosen was Extra Tree Regressor having R^2 Score (0.937) and RMSE (15.16).

6 Regression Analysis

The arrival delay ranged from 0 to 2142 minutes. Frequency Distribution plot (figure 6) of the arrival delay reveals that the maximum frequency is observed between 0 - 100 and 100 - 200. The dataset was split into ranges of arrival delay minutes and performance of Linear Regression model is studied in these ranges (table 7).

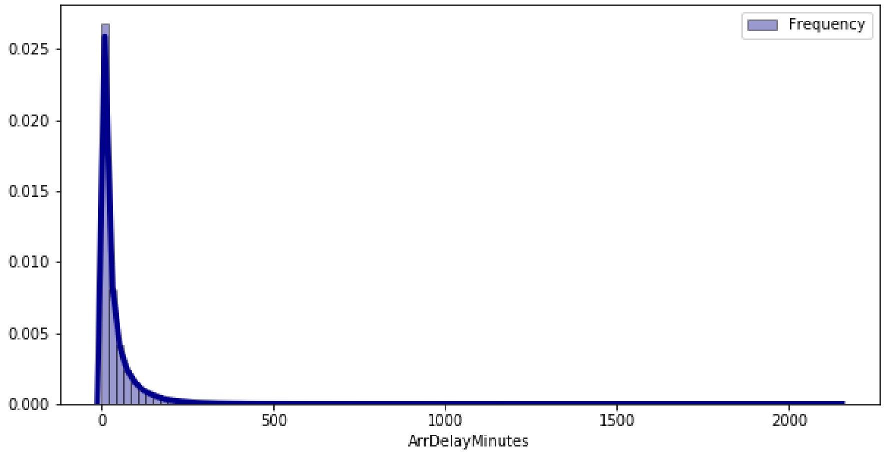

Figure 6: **Frequency Distribution of Arrival Delay**

Range	RMSE	MAE
0 - 100	14.51	10.44
100 - 200	27.50	17.74
200 - 500	32.34	20.15
500 - 1000	42.37	35.17
1000 - 2000	69.53	65.55

Table 7: **Range-wise Regression Analysis**

The results obtained are in line with the frequency distribution histogram. Most data points have *'ArrDelayMinutes'* ranging between 1 - 100 minutes and hence, MAE (10.43) and RMSE (14.51) are least in this range. As the range increases, the number of data points decreases and as result, the values of RMSE and MAE increase.

7 Pipelined Model

The Classifier chosen is Random Forest Classifier as it has the maximum F1 Score and area under ROC. The Regression Model chosen in Extra Tree Regressor as it has the highest R^2 Score. The Flow Chart given below (figure 7) is the representation of the two-stage machine learning model built.

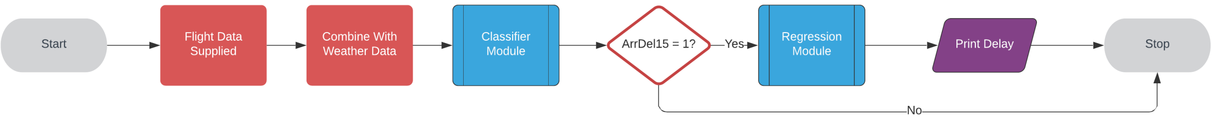

Figure 7: **Flowchart of Pipelined Model**

Performance of Pipelined Model

Metric	Value
Root Mean Squared Error	10.74
Mean Absolute Error	7.93
R^2 Score	0.937

Table 8: **Pipelined Model Performance**

8 Conclusion

The flight and weather data were combined into a single dataset for training the models. It was observed from the classifier that bias in dataset(towards non delayed flights) affected the performance of class 1 relative to class 0 in the classifier models chosen. This bias was overcome by oversampling the delayed flights in the dataset using SMOTE. The Classifier chosen was Random Forest Classifier having the highest F1 Score (0.78) & Accuracy (0.91).The Regression Model chosen was Extra Tree Regressor having the highest R^2 Score (0.937) & least RMSE (15.16) . The pipelined model was designed using the aforementioned Classifier and Regression Models and performed with a good accuracy.