Flight Delay Prediction

Vedanth Subramaniam

Abstract

A flight delay occurs when an airplane takes off later than its scheduled time of departure and/or arrives later than its scheduled time of arrival. Various factors such as errors in management and trouble with the weather conditions attribute to delays. The project aims 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

Flying has become synonymous with delays, frustrating the well laid out plans of not just the passengers but also the airports and airlines involved. In the present world, knowing the delay beforehand would be an advantage as air traffic has increased significantly and this would help in solving scheduling problems. A flight covers multiple trips each day and thus a delay in arrival or departure in one place creates a cascading effect on the rest of them. Flight delays also result in major economic losses for all parties involved.

The weather conditions also play a role when it comes to departure or arrival delays. This project aims to predict the arrival delay of flights based on weather data of airports in the USA during 2016 and 2017 and the flight data of all flights during the same period. The first stage of the two-stage model is classification. Classification section involves the prediction of a flight as delayed or non delayed. The second stage of the two-stage model is regression. The regression section predicts the arrival delay in minutes of flights which are classified delayed.

2 Data Preprocessing

The Flight dataset contains information about the flights that flew within United States of America during 2016 and 2017. The airport codes that are considered are given in Table 1.

Table 1 : Airport Codes

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

The flight attributes that are considered are given in Table 2.

Table 2 : Flight Attributes

FlightDate	Quarter	Year
Month	DayofMonth	DepTime
DepDel15	CRSDepTime	DepDelayMinutes
OriginAirportID	DestAirportID	ArrTime
CRSArrTime	ArrDel15	ArrDelayMinutes

The weather dataset has the weather conditions of the airports on an hourly basis for 2016 and 2017. The weather attributes that are considered are given in Table 3.

Table 3 : Weather Attributes

WindSpeedKmph	WindDirDegree	WeatherCode
precipMM	Visiblity	Pressure
Cloudcover	DewPointF	WindGustKmph
tempF	WindChillF	Humidity
date	time	airports

The two datasets are merged on the basis of Airport, Date and Time and a final dataset which contains details concerning each flight and the weather conditions is obtained.

3 Classfication

Classification is the first stage of the two-stage machine learning model. The aim of this module is to predict whether a flight is delayed or not. According to the data, a flight is classified as delayed if the arrival delay is more than 15 minutes. Flights which are delayed have the target variable 'ArrDel15' = 1 and for those which are not delayed have the target variable 'ArrDel15' = 0.

The following classification models are considered:

- Logistic Regression
- Decision Tree Classifier
- Extra Trees Classifier
- Gradient Boosting Classifier

3.1 Classification Metrics

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

TP - True Positives TN - True Negatives
FP - False Positives FN - False Negatives

- True Positives: A true positive is an outcome where the model correctly predicts the positive class. Flights delayed classified correctly as delayed.
- True Negatives: A true negative is an outcome where the model correctly predicts the negative class. Flights not delayed classified correctly as non delayed.
- False Positives: A false positive is an outcome where the model incorrectly predicts the positive class. Flights on time classified incorrectly as delayed.
- False Negatives: A false negative is an outcome where the model incorrectly predicts the negative class. Flights delayed classified incorrectly as non delayed.

Recall is the number of delayed flights that are correctly classified as delayed. Recall is an important factor during prediction because our aim is to classify flights as 'delayed' as accurately as possible.

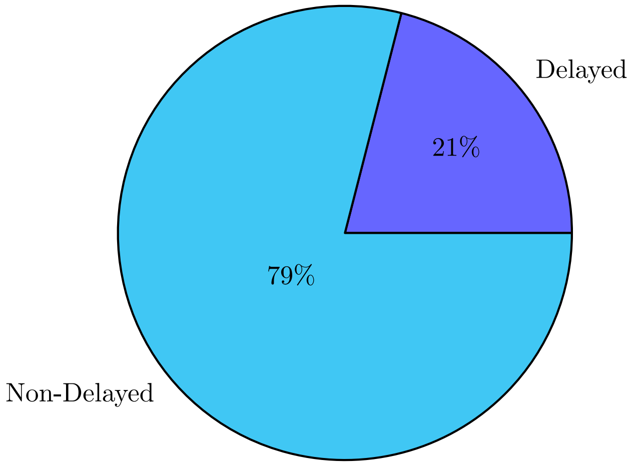
3.2 Classification Scores

Table 4 : Classification Scores

Model	Precision		Recall		F_1 Score		Accuracy
	0	1	0	1	0	1	
Logistic Regression	0.92	0.89	0.97	0.68	0.95	0.77	0.91
Gradient Boosting Classifier	0.92	0.89	0.98	0.68	0.95	0.77	0.91
ExtraTrees Classifier	0.92	0.85	0.97	0.68	0.94	0.76	0.91
Decision Tree	0.92	0.68	0.91	0.71	0.92	0.69	0.87

4 Sampling

Figure 2 : Class Distribution for Delayed Flights



From Figure 2, it can be seen that the number of instances belonging to 'Delayed' is very sparse compared to that of 'Non Delayed'. As a result of this imbalanced distribution, sampling is used to make it a balanced dataset.

- **Synthetic Minority Oversampling Technique(SMOTE)**

  It is one of the few common Oversampling technique. In this technique, the new instances are generated by randomly selecting one or more of the k-nearest neighbors for each instance in the feature space in the minority class.

- **Random Under Sampler**

  In this technique, samples from the majority class are randomly removed, to get an even distribution. This technique may discard useful or important samples.

Table 5 : Classification Results with SMOTE

Model	Precision		Recall		F_1 Score		Accuracy
	0	1	0	1	0	1	
Logistic Regression	0.94	0.74	0.92	0.78	0.93	0.76	0.89
Gradient Boosting Classifier	0.93	0.82	0.96	0.73	0.94	0.77	0.91
ExtraTrees Classifier	0.93	0.82	0.96	0.72	0.94	0.76	0.91
Decision Tree	0.92	0.68	0.91	0.71	0.91	0.69	0.87

Table 6 : Classification Results with Random Undersampler

Model	Precision		Recall		F_1 Score		Accuracy
	0	1	0	1	0	1	
Logistic Regression	0.94	0.72	0.93	0.77	0.93	0.76	0.90
Gradient Boosting Classifier	0.94	0.73	0.92	0.79	0.93	0.76	0.89
ExtraTrees Classifier	0.95	0.64	0.87	0.83	0.91	0.72	0.86
Decision Tree	0.94	0.50	0.79	0.80	0.86	0.62	0.80

For each classifier, SMOTE and Random Undersampling techniques are applied. From Table 5, the SMOTE with Gradient Boosting Classifier has the best recall(0.73) and precision(0.82) for class 1.

5 Regression

Regression is the second-stage of the model. Delayed flights are given as inputs to regressors. The Arrival Delay is predicted.

Regression Models

- Linear Regression
- Decision Tree Regressor
- Extra Trees Regressor
- Gradient Boosting Regressor

Regression Metrics

To evaluate the regressor models, we use the following metrics.

- **Mean Absolute Error**

$$Mean\ Absolute\ Error(MAE) = \frac{1}{N}\sum_{i=1}^{N} \mid Y_i - \hat{Y}_i \mid$$

- **Root Mean Square Error**

$$Root\ Mean\ Square\ Error(RMSE) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$$

- R^2 **Score**

$$R^2 Score = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

$\bar{Y}$: Mean Value Of Y
$\hat{Y}$: Predicted Value Of Y
N: Number of Data Points

Table 7 : Regression Scores

Model	**MAE**	**RMSE**	**R2**
Linear Regression	12.16	17.53	0.94
Decision Trees	16.65	24.15	0.88
Extra Trees	11.85	16.88	0.94
Gradient Boosting	11.65	16.86	0.94

R2 Score indicates how close the data is fit to the regression line. Higher the R2 value better the model fits the data. Low MAE value indicates better performance of the model. Lower the RMSE value better the fit.

Gradient Boosting Regressor had the best performance with a high R2 Score(0.94) and low RMSE(16.86) and MAE(11.65) values.

6 Pipeline

Under the classification models, it is observed that the SMOTE sampling performed with Gradient Boosting Classifier has the best scores. A pipeline is constructed with Gradient Boosting Classifier and Gradient Boosting Regressor. The 2016 data is used as the training set and the 2017 data is used as the test set for the pipeline model.

Figure 3 : Pipeline Model

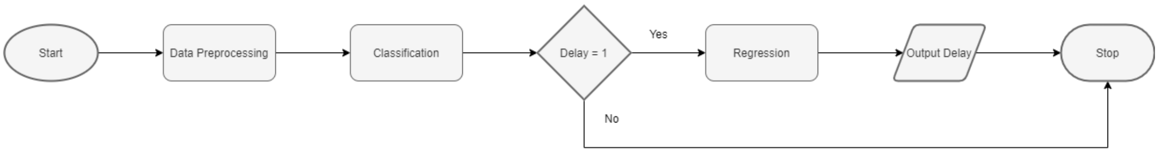


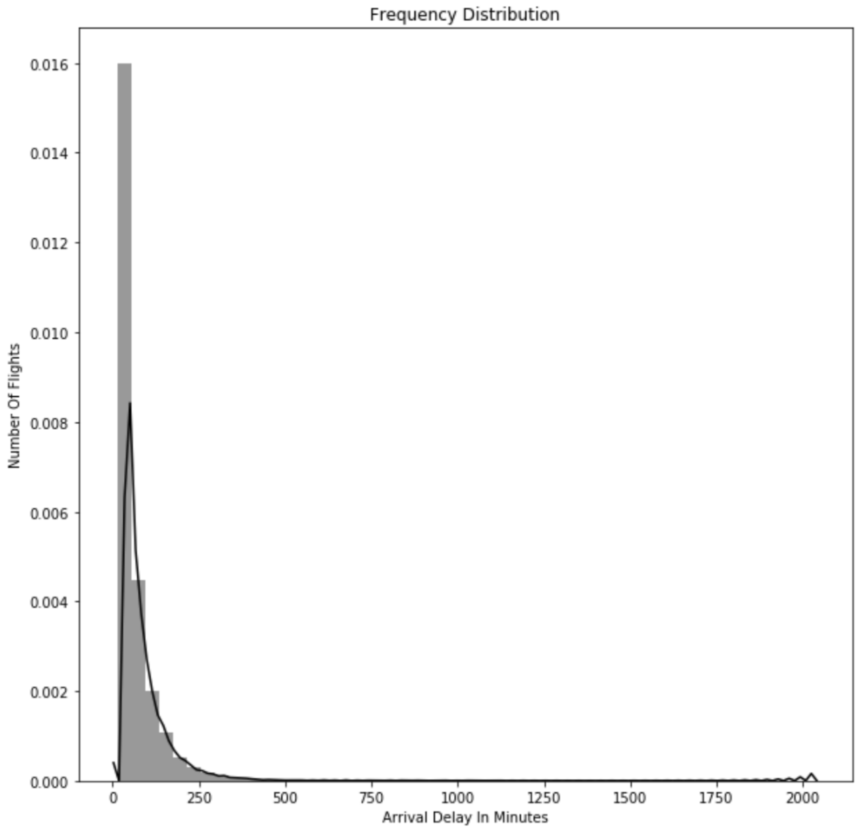
Table 8 : Pipeline - Regression

Model	**MAE**	**RMSE**	**R2**
Gradient Boosting	12.53	17.61	0.95

7 Regression Analysis

In order to evaluate the performance of the model under different ranges of arrival delays, the test set is split into ranges based on the amount of arrival delay. The results are given in Table 9.

Figure 4 : Frequency Distribution



The y-axis is the probability density function for the kernel density estimation.

Table 9 : Frequency Distribution of Arrival Delays

Range	RMSE	MAE
15 - 300	16.63	11.55
300 - 550	29.4	19.2
550 - 800	27.95	17.69
800 - 1200	21.22	16.83
1200 - 2150	91.96	50.23

The regressor best performs between the range 800 minutes and 1200 minutes with an RMSE of 21.22 and MAE of 16.83. This is because the error compared to the actual values is very insignificant. A majority of fights have a delay between 15 minutes and 300 minutes. As the range increases, the number of data points decreases and as a result, the values of RMSE and MAE increases.

8 Conclusion

The flight dataset and the weather dataset were merged into a single dataset and only the necessary features were retained. Due to data imbalance, the performance of the classifier on flights classified as delayed was lower when compared to flights classified as non-delayed, as a result sampling was performed and SMOTE with Gradient Boosting Classifier had the best performance. Regression models were applied to predict the arrival delay of flights which were classified delayed. Under regression models, the Gradient Boosting Regressor had the best performance. Pipeline was constructed with Gradient Boosting Classifier and Gradient Boosting Regressor and regression analysis was performed on the dataset.