Flight Delay Prediction Using Flight and Weather Data

Shreyas Srinivasan

Abstract

A flight delay occurs if a flight arrives earlier or later than the scheduled time. This delay will be affected by factors such as weather conditions, air traffic, thunderstorm and other weather glitches. The aim of this project will be to build a two-stage predictive machine learning model to train a dataset and predict if there will be a delay in arrival time and by how much time will the delay be.

1 Introduction

Flight delay affects airlines, airports and passengers. Prediction of flight delay helps passengers and airlines reduce financial losses caused by delays and also be wary of possible accidents that can occur due to bad weather conditions. Hence it is essential to predict flight delays.

This project aims to build a two-stage model to predict whether a flight will be delayed or not. It also aims to predict the delay of flights based on Weather and Flight Data across 15 airports in USA in the years 2016 and 2017. Different classification and regression models are studied and compared in this project.

2 Dataset

The flight dataset contains data of all flights that flew in the USA during the years 2016 and 2017. The individual flight details which have their origin and destination in the 15 specified airports are used to get the Flight Dataset.

The weather data was a json file, and it was restructured into csv to get the Weather Dataset. Flight and Weather Datasets are merged twice based on, Departure Date, Absolute Departure Time and Departure Airport, and merged again based on, Arrival Date, Absolute Arrival Time and Arrival Airport, both seperately.

Table 1 shows the airport codes considered. The weather features considered are listed in Table 2, and the flight features considered are listed in Table 3.

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

Table 1: Chosen Airport Codes

WindSpeedKmph	WindDirDegree	WeatherCode	precipMM
Visibility	Pressure	Cloudcover	DewPointF
WindGustKmp	tempF	WindChillF	Humidity
date	time	airport	

Table 2: Weather Details

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

Table 3: Flight Details

3 Classification

It is essential to know if a flight will be delayed or not first before finding out by how much time will there be a delay. Hence the first stage is classification. In classification, we classify whether the flight is delayed or not by 15 minutes or more. The ground truth used for classification is ArrDel15 which tells us if Arrival Delay of flight is more than 15 minutes or not. If ArrDel15 = 1.0, the flight is delayed by 15 minutes or more. ArrDel15 = 0.0 for other cases.

3.1 Classification Metrics

- **TN:True Negative**

  The model has predicted correctly that a flight will not be delayed

- **TP: True Positive**

  The model has predicted wrongly that a flight will not be delayed

- **FN: False Negative**

  The model has predicted wrongly that a flight will not be delayed

- **FP:False Positive**

  The model has predicted wrongly that a flight will be delayed

- **Accuracy**

  Accuracy is the ratio of true results to the total number of results that are examined.

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

- **Precision**

  Ratio of true positives to total positives. Gives a proportion of predictive positives which are truly positive.

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

- **Recall**

  It tells us what proportion of actual positives are correctly classified.

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

- F_1 **Score**

  F_1 Score is the harmonic mean of precision and recall.

$$F_1 Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.2 Classifier Models

Different Classifier models are used and the best classifier is chosen based on F1 Score. The different classifier models used are Logistic Regressor, Gradient Boost Classifier, Extra Tree Classifier and Random Forest Classifier.

3.3 Classifier Perfomance

Algorithm	Precision		Recall		F_1 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	0.69	0.91	0.71	0.92	0.70	0.87
Gradient Boosting Classifier	0.92	0.90	0.98	0.70	0.95	0.79	0.92
ExtraTrees Classifier	0.89	0.57	0.89	0.57	0.89	0.57	0.82
Random Forest Classifier	0.92	0.89	0.98	0.70	0.95	0.78	0.92

Table 4: Classifier Performance

4 Data Imbalance Problem

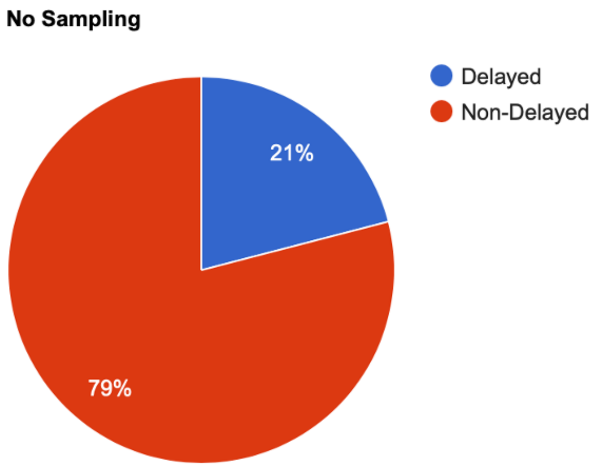

Figure 1: Dataset Distribution Before Sampling

In the above classification algorithms, the Class 1 (delayed flights) performance is weaker than Class 0 (non delayed flights). This is due to more number of non Delayed flight data points present in the dataset, as shown in Fig 1. Table 4 shows the classifier performance before overcoming the bias.

This bias in the dataset can be overcome by applying Oversampling or Undersampling Techniques like

Random Under Sampler

This technique involves randomly duplicating data from the majority class and adding it to the training data.

Near Miss

In this technique, we eliminate majority class examples by checking if there are instances of two different classes that are very close to each other in the feature space. We remove the instances of the majority class to increase the space between the two classes.

Random Over Sampler

This technique involves randomly duplicating data from the minority class and adding it to the training data.

Synthetic Memory Oversampling Technique(SMOTE)

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.

4.1 Undersampling

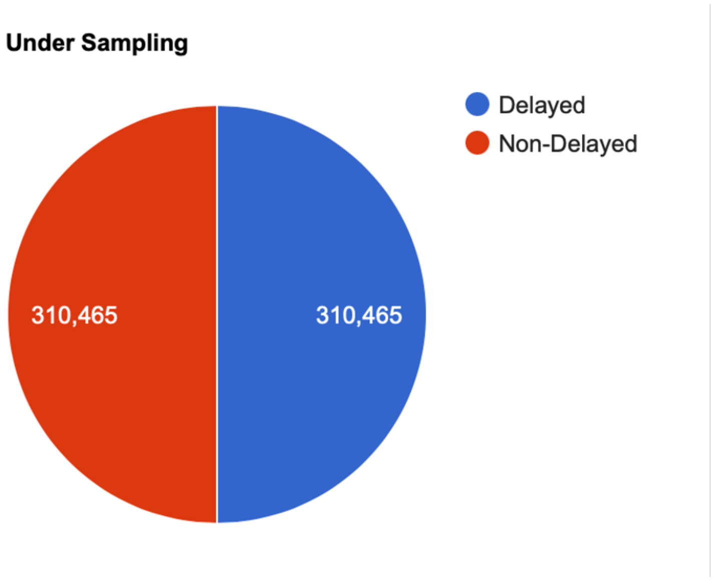

Figure 2: Dataset Distribution After Undersampling

Algorithm	Precision		Recall		F_1 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	0.94	0.51	0.79	0.81	0.86	0.62	0.80
Gradient Boosting Classifier	0.95	0.73	0.92	0.81	0.93	0.76	0.90
ExtraTrees Classifier	0.90	0.39	0.70	0.71	0.79	0.50	0.71
Random Forest Classifier	0.95	0.72	0.92	0.81	0.93	0.76	0.89

Table 5: Classifier Performance using Random Under Sampler

Algorithm	Precision		Recall		F_1 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	0.94	0.51	0.79	0.81	0.86	0.62	0.80
Gradient Boosting Classifier	0.90	0.39	0.70	0.71	0.79	0.50	0.71
ExtraTrees Classifier	0.95	0.73	0.92	0.81	0.93	0.76	0.90
Random Forest Classifier	0.95	0.72	0.92	0.81	0.93	0.76	0.89

Table 6: Classifier Performance using Near Miss

4.2 Oversampling

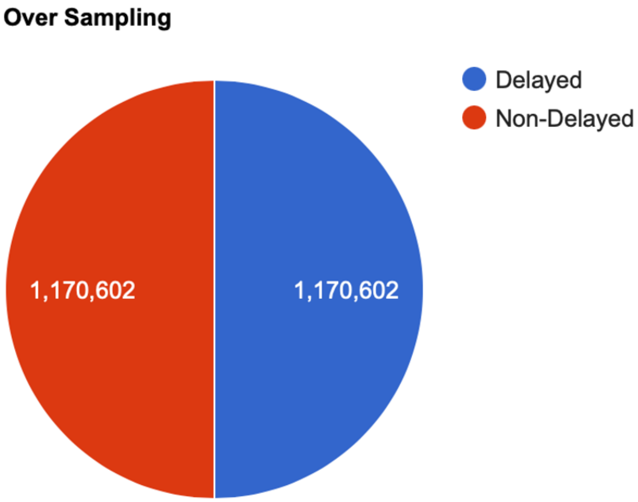

Figure 3: Dataset Distribution After Oversampling

Algorithm	Precision		Recall		F_1 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	0.69	0.92	0.70	0.92	0.70	0.87
Gradient Boosting Classifier	0.95	0.73	0.92	0.81	0.93	0.77	0.90
ExtraTrees Classifier	0.89	0.60	0.89	0.59	0.89	0.59	0.83
Random Forest Classifier	0.93	0.83	0.96	0.74	0.95	0.78	0.91

Table 7: Classifier Performance using Random Over Sampler

Algorithm	Precision		Recall		F_1 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	0.71	0.91	0.69	0.87
Gradient Boosting Classifier	0.93	0.87	0.97	0.71	0.95	0.78	0.92
Extra Tree Classifier	0.89	0.50	0.84	0.62	0.86	0.55	0.79
Random Forest Classifier	0.94	0.80	0.95	0.77	0.94	0.78	0.91

Table 8: Classifier Performance using SMOTE

From Tables 5,6,7 and 8, it can be concluded that Gradient Boost Classifier using SMOTE offers best perfomance due to high F1 Score for both Classes 1 and 0.

5 Regression

After finding out if a flight is delayed or not, it is important to know the delay time. In regression, delay period (in minutes) of delayed flights is obtained using different regression models. Delayed flights are those with ArrDel15 = 1.0. Our ground truth for regression will be ArrDelayMinutes, which is the time difference (in minutes) between the CRSArrTime (scheduled Arrival Time) and ArrTime (actual Arrival Time).

5.1 Regression Metrics

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

The following notations stand for :
$\bar{Y}$: Mean Value Of Y
$\hat{Y}$: Predicted Value Of Y
N: Number of Data Points

- **Mean Absolute Error**

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

- **Mean Square Error**

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

- **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**

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

5.2 Regression Models

Different Regressor models are used and the best regressor is chosen based on R2 score, RMSE and MAE. The different regressor models used are Linear Regressor, Extra Tree Regressor, Gradient Boost Regressor and Random Forest Regressor.

5.3 Regression Perfomance

Regression Model	RMSE	MAE	R^2 Score
Linear Regressor	19.79	14.49	0.93
Extra Trees Regressor	16.55	11.64	0.95
Random Forest Regressor	16.53	11.64	0.95
Gradient Boosting Regressor	16.83	11.60	0.95

Table 9: Performance of The Regressors

From Table 9, it can be concluded that the best regressor is the Random Forest Regressor as it has the lowest MAE and RMSE.

6 Regression Analysis

The arrival delay time of flights ranges from 0.0 to 2142.0 minutes. Regression analysis is done to check performance of the best regressor, obtained from Table 4, on smaller distributions of time periods and form an analysis based on the same.

ArrivalDelayMinutes	No Of Flights	RMSE	MAE	R2 Score
0 - 100	323550	13.15	9.89	0.64
100 - 200	48954	15.81	11.98	0.66
200 - 500	14232	20.39	14.61	0.91
500 - 1000	1128	22.72	16.34	0.97
1000 - 2000	175	49.04	29.72	0.96

Table 10: Frequency Distribution And Range Wise Regressor Scores Of The Flights

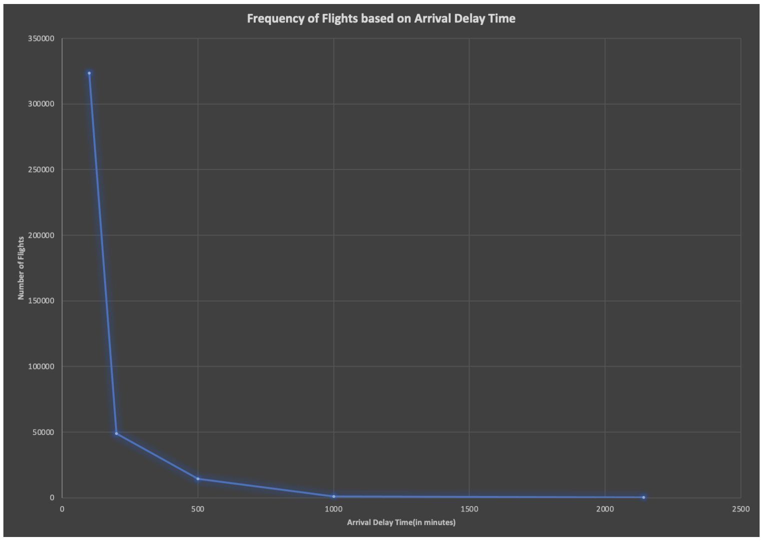

Figure 4: Frequency Distribution of Flight Delay in minutes

Low RMSE value indicates that predicted data is close to original data and low MAE value indicates the performance of the regressor. High R2 Score indicates how well the predicted data has been fit among original data. Based on the same, it can be inferred from Table 10 that best performance has been shown in the 200-500 range as it has the lowest RMSE and MAE Score.

7 Pipeline

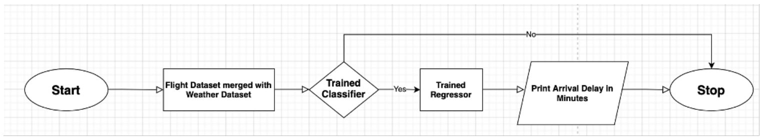
No
Start
Flight Dataset merged with Weather Dataset
Trained Classifier
Yes
Trained Regressor
Print Arrival Delay in Minutes
Stop

Figure 5: Pipelining Process

In pipelining we use our best classifier and regressor to classify and regress our data respectively. We first classify our data using our best classifier which is Gradient Boost Classifier using SMOTE to classify flights based on whether they're delayed or not by 15 minutes.

Then we regress our data using our best regressor, Random Forest Regressor, for the delayed flights. After regression, we check our R2 Score, Root Mean Square Error and Mean Absolute Error.

Metric	Value
MAE	11.64
RMSE	16.53
$R^2 Score$	0.95

Table 11: Performance of the Pipeline model

8 Conclusion

Classification models were used to classify the flights as delayed or non delayed. The classifier performance was observed, which showed the poor performance of Class 1 with respect to Class 0. The poor performance was due to more number of non-delayed flight data points being present in the dataset. This imbalanced data was overcome by applying SMOTE. After using SMOTE, the recall values of Class 1 increased. Gradient Boost Classifier was chosen for the pipeline model as it had the highest F_1 Score.

Regression models were used to predict the arrival delay in minutes for those flights classified as delayed. Random Forest Regressor was chosen for the pipeline model as it had a high R^2 Score, low RMSE and low MAE values. Regression Analysis was done to check perfomance of the best regressor in smaller datasets from the dataset based on delay in Arrival Delay, ArrDelayMinutes. The pipeline model with the selected classifier and regressor performed with reasonable accuracy.

Thereby a two-stage predictive machine learning model has been successfully built which predicted if a flight will be delayed or not and by how much time will there be a delay if it exists.

With this information, passengers and airlines can significantly reduce losses caused by delays and also help in prevention of accidents caused due to bad weather conditions.