Flight Delay Prediction

Rishi Vardhan K

November 27, 2019

Abstract

Flight delay is when an airline flight takes off and/or lands later than its scheduled time. Several factors affect flight delay with respect to weather parameters and management errors. The project works with real time weather data on a two stage predictive machine learning engine that predicts the chance of a flight to be delayed and the amount of delay.

1 Introduction

Flight delays are an important factor to consider whilst flight operation. There are factors that are under the direct control of the carrier, such as aircraft turnarounds between flights, passenger punctuality, technical and crew performance, etc. There are also even more factors that are outside of the airline's control, such as weather, air traffic control, security, airport conditions, etc. The reality is such that so long as airplanes continue flying, flight delays will be a part of the experience. According to the Bureau of Statistics, about 20% of all flights are delayed by 15 minutes or more.

As flight delays are caused by either of factors affecting departure or arrival, the project works on dealing with predicting and calculating delay pertaining to arrival of flight. The project is divided into three modules each with different sets of operations. Module 1 focuses on Data Pre-processing, Module 2 on Classification and Module 3 on Regression. Each Module works incrementally on previous module's output.

2 Data Pre-Processing - Module 1

The first module operates to process the data to fit under the required problem statement. The flight data is obtained from the Bureau of Transportation Statistics between years 2016-2017. The corresponding weather data for the years are also collected. Each data-set is filtered to contain features with maximum relevance to the given problem. The following are the list of airport codes considered for the model. The data is processed so as to work with only the arrival and departure of flights under these airports.

Table 1: Airport Codes

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

The following are the features considered from real-time weather data-set.

Table 2: Weather Features

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

The following are the features considered from real-time flight performance data-set.

Table 3: Flight Features

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

The given flight data and weather data are compared using date, time and airport of weather data and similar arrival features of flight data as merging parameters to merge the two data-sets. The output set is further made void of duplicate features so as to avoid redundancy in data. The categorical values are label encoded to fit the necessary machine learning algorithms. The final data-set comprises of 23 features in total.

3 Classification - Module 2

Module 2 works on predicting the chance of flight delay, i.e classifying flights to be either delayed or not. The processed data-set from module 1 is made use of in the current module. The standing criteria for a flight not to be delayed is set under a threshold of 15 minutes. The feature **ArrDel15** denotes whether a flight has been delayed or not under our criteria. It contains values of Class 1 and Class 0 where Class 1 indicates flight delay and Class 0 denotes no flight delay. The train set and test set are split in a ratio of 80:20. The predicted chance of flight delay as obtained from the best model of this module is pipe-lined to the successive model for delay minutes prediction.

3.1 Metrics Used

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

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

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

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

where TP - True Positives, TN - True Negatives, FP - False Positive, FN - False Negative.

Since f1-scores are a measure of both precision and recall, they are suitable for deciding on the best model.

3.2 Observations

The following are the scores of various Classifiers considered.

Table 4: Classifier Scores

Classifiers	Class	Precision	Recall	f1	Accuracy
Decision Tree	1	0.68	0.71	0.92	0.86
	0	0.92	0.91	0.92	
Weighted Average		0.87	0.87	0.87	
Extra Trees	1	0.83	0.71	0.76	0.90
	0	0.93	0.96	0.94	
Weighted Average		0.91	0.91	0.91	
Gradient Boosting	1	0.89	0.68	0.77	0.91
	0	0.92	0.98	0.95	
Weighted Average		0.92	0.92	0.91	

There stays a difference in values between Class 1 and Class 0 scores. The difference is attributed to the fact that data-set has less number of examples for delayed flights.

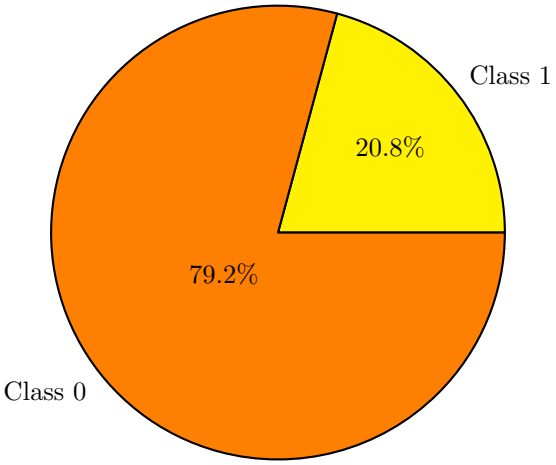

Figure 1: Class Distribution

This imbalance causes the classifier to not work properly due to less/more number of records supporting each case. This problem is solved by the inclusion of sampling methods. Sampling is done to solve data imbalance by creating or deleting records to compensate for the improper class distribution. Over all the observation is such that the model is bad at predicting flight delay.

The various over-sampling and under-sampling techniques are applied to each classifier to find the best fitting technique. The results are tabulated as follows.

Table 5: Decision Tree

Samplers	Class	Precision	Recall	f1	Accuracy
SMOTE	1	0.67	0.70	0.68	0.86
	0	0.92	0.91	0.91	
Weighted Average		0.87	0.86	0.87	
ADASYN	1	0.36	0.40	0.38	0.72
	0	0.84	0.82	0.83	
Weighted Average		0.74	0.73	0.73	
Random Over Sampler	1	0.69	0.70	0.69	0.87
	0	0.92	0.92	0.92	
Weighted Average		0.87	0.87	0.87	
Near Miss	1	0.23	0.67	0.35	0.47
	0	0.83	0.42	0.56	
Weighted Average		0.70	0.47	0.51	
Random Under Sampler	1	0.50	0.80	0.62	0.79
	0	0.94	0.79	0.86	
Weighted Average		0.85	0.79	0.81	

On observation from Table 5 we find that Random Over-Sampler and SMOTE have best values for f1-scores. The other techniques fail to compare with them and show poor performances in prediction.

Table 6: Extra Trees

Samplers	Class	Precision	Recall	f1	Accuracy
SMOTE	1	0.80	0.73	0.76	0.90
	0	0.93	0.95	0.94	
Weighted Average		0.90	0.91	0.90	
ADASYN	1	0.79	0.73	0.76	0.90
	0	0.93	0.95	0.94	
Weighted Average		0.90	0.90	0.90	
Random Over Sampler	1	0.83	0.70	0.76	0.90
	0	0.92	0.96	0.94	
Weighted Average		0.90	0.91	0.90	
Near Miss	1	0.41	0.84	0.55	0.71
	0	0.94	0.68	0.79	
Weighted Average		0.83	0.72	0.74	
Random Under Sampler	1	0.67	0.81	0.73	0.87
	0	0.95	0.89	0.92	
Weighted Average		0.89	0.88	0.88	

The same trend observed in Table 5 arises in Table 6 as Random Over Sampler and SMOTE are found to show good performances in prediction of delays compared to others.

Table 7: Gradient Boosting

Samplers	Class	Precision	Recall	f1	Accuracy
SMOTE	1	0.83	0.73	0.78	0.91
	0	0.93	0.96	0.95	
Weighted Average		0.91	0.91	0.9	
ADASYN	1	0.80	0.75	0.77	0.90
	0	0.93	0.95	0.94	
Weighted Average		0.91	0.91	0.91	
Random Over Sampler	1	0.73	0.79	0.76	0.89
	0	0.94	0.92	0.93	
Weighted Average		0.90	0.90	0.90	
Near Miss	1	0.45	0.83	0.58	0.74
	0	0.94	0.73	0.82	
Weighted Average		0.84	0.75	0.77	
Random Under Sampler	1	0.73	0.79	0.76	0.89
	0	0.94	0.92	0.93	
Weighted Average		0.90	0.89	0.90	

The similar trend observed in Tabel 5 and Table 6 arises here in Table 7 as Random Over Sampler and SMOTE Classifiers rank better than others in prediction.

Here since sampled and non sampled data have the same best f1-scores, sampling tends to be non-effective. The conclusion from applying different sampling techniques would be that the different techniques tend to be less effective for the current data-set as they result in non reliable output.

The Gradient Boosting SMOTE sampled model is concluded to predict the chance of flight delay to a better extent than the other models. The train and test sets are re-partitioned with training sets to be flights in the year 2016 and testing sets to be flights in the year 2017. The results of test set as predicted by the Gradient Boosting using SMOTE model is pipe-lined to the next module for prediction of flight delay minutes.

4 Regression - Module 3

The last module of the project deals with predicting the amount of delay occurred, in case of flight delay. The module implements a regression model to predict the necessary output. Since only delayed flights are to be considered, the rest flight details are removed from the set. The set is partitioned into train and test sets in a ratio of 80:20. The module also makes use of the prediction from previous module of chance of flight delay for the year 2017 and predicts flight delay minutes.

4.1 Metrics Used

$$\text{Root mean squared error } \mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

$$\text{Mean absolute error } \mathbf{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

$$\text{R Squared } \mathbf{R2} = \frac{TotalVariation - ExplainedVariation}{TotalVariation}$$

4.2 Analysis

4.2.1 Box plot: Arrival Delay Minutes

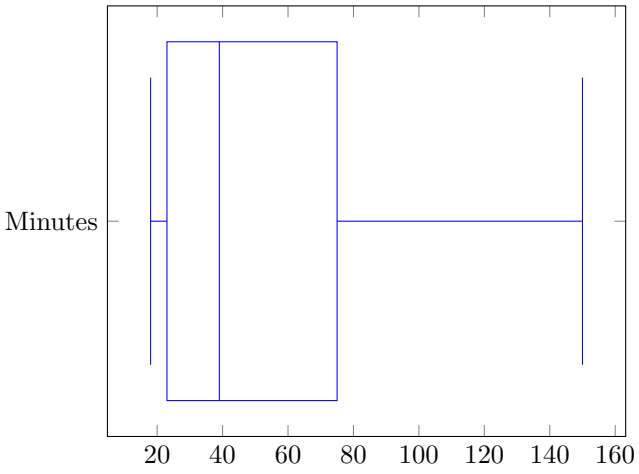

Figure 2: Box Plot of Distribution of Arrival Delay Minutes

From the box plot distribution we understand that the Inter Quartile range lies between 23 minutes and 75 minutes. This observation provides the fact that the data-set has delay minutes values more prominently between the range of 23 and 75 minutes. In a practical approach, the model is trained to predict values primarily lying between the Inter Quartile Range.

4.2.2 Feature Selection : Embedded Methods

Embedded methods are iterative in a sense that takes care of each iteration of the model training process and carefully extract those features which contribute the most to the training for a particular iteration. Regularization methods are the most commonly used embedded methods which penalize a feature given a coefficient threshold. Here we will do feature selection using Lasso regularization. If the feature is irrelevant, lasso penalizes it's coefficient and make it 0. Hence the features with coefficient = 0 are removed and the rest are taken.

Performing Lasso Regression the model picked 9 features to be important and discarded 13 other features to be not important. The feature importance plot is as shown below

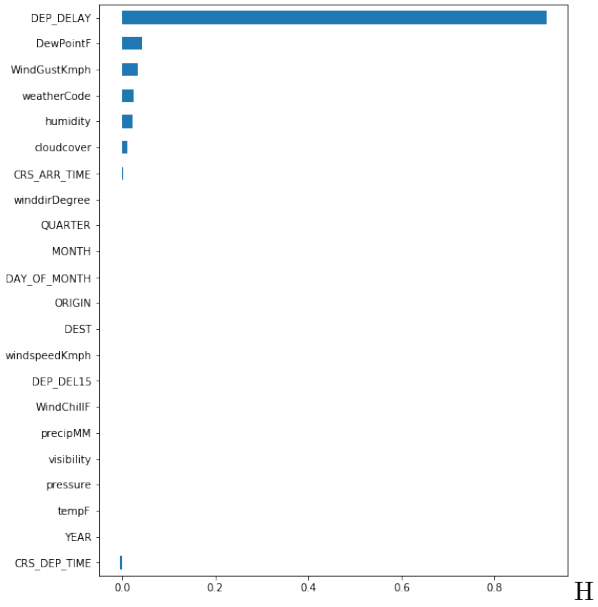

Figure 3: Feature Importance Using Lasso Model

From plot we observe that **Departure delay Minutes** feature denoting the number of minutes delayed during departure, **Dew Point** feature denoting the temperature to which air must be cooled to become saturated with water vapor during arrival and **Wind Gust speed** feature denoting the sudden increase of wind speed during arrival are the features that contribute the most in predicting **Arrival Delay Minutes**.

4.2.3 Correlation : SNS Heat Map

Filter method is used to find relevance between features using correlation matrix by implementing Pearson correlation. The correlation matrix heat map for the data-set is as follows.

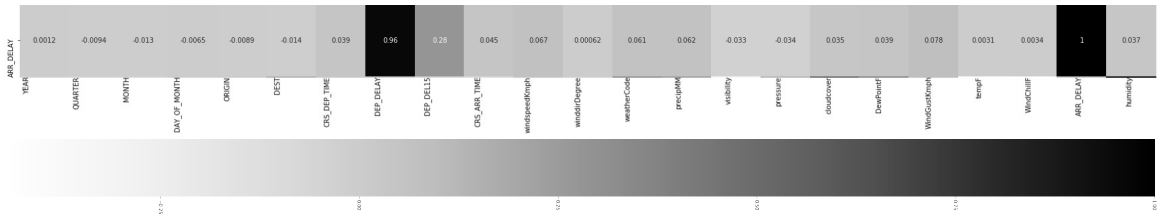

Figure 4: SNS Heat Map

From the heat map we observe the following features to be of maximum relevance to **Arrival Delay minutes**.

Departure delay minutes feature denoting the number of minutes delayed during departure. **Departure delay** feature indicating whether a flight has been delayed or not and **Wind Speed** denoting the speed of wind in Kmph during arrival.

4.3 Tabulated Results

The results of different models that predicted flight delay minutes for flight delays from actual data-set are tabulated as follows.

Table 8: Regression model scores - Actual Delays

Model	MAE	RMSE	R2
Linear Regression	12.34	17.77	0.94
XGBoost	11.63	16.88	0.94
Extra Trees	11.85	16.92	0.94
Decision Trees	16.59	24.13	0.88

The results of different models that predicted flight delay minutes for flight delays predicted by module 2 are tabulated as follows.

Table 9: Regression model scores - Predicted Delays

Model	MAE	RMSE	R2
Linear Regression	10.78	15.11	0.88
XGBoost	10.55	14.81	0.89
Extra Trees	10.37	14.48	0.89
Decision Trees	14.86	20.66	0.79

From the table 8 and table 9 we observe that the models predict values with less errors from pipe-line output obtained from previous model than the original data-set.

4.4 Post Regression Analysis

4.4.1 Box plot: Predicted Arrival Delay Minutes

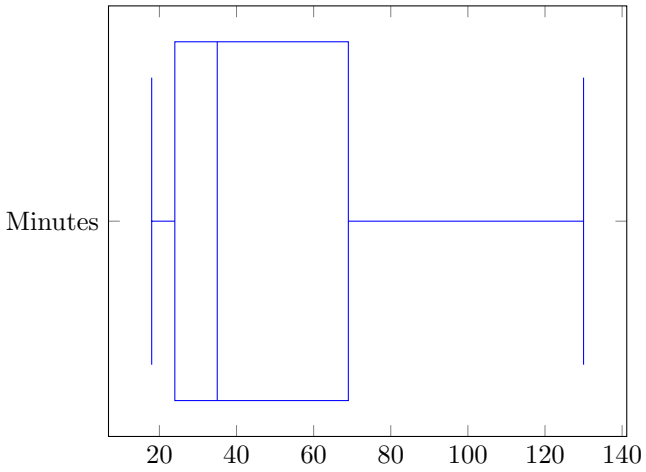

Figure 5: Box Plot of Distribution of Predicted Arrival Delay Minutes

From the plot we observe that the Inter Quartile Range lies between 24 minutes and 69 minutes. The range is maintained the same to an extent as in before prediction with consideration of Figure 2. Thereby the trained model predicts delay minutes primarily in the Inter Quartlie Range.

4.4.2 Feature Importance

The Feature Importance concept from the tree models post regression is able to perform the ranking of features based on importance of features to the machine learning model. Upon ranking and observation, we infer that the **Scheduled departure time** denoting the scheduled time of departure of flight, **Departure delay** indicating whether a flight is delayed during departure or not and **Departure delay minutes** denoting the number of minutes flight has been delayed during departure are the set of features that act as most important according to Feature Importance.

5 Conclusion

The first stage deals with classifying flights to be delayed or not. Due to data-imbalance, sampling techniques are employed to compensate for the improper data distribution. Sampling does not however solve the situation and is ineffective in predicting flight delay for the given data-set. Hence by consideration of better scores among all other classifiers implemented, the **Gradient Boosting Classifier** sampled under SMOTE proved better with an accuracy of 91 %. The follow up module worked on predicting the amount of delay for delayed flights by running different regression algorithms. Among the various regressors considered, the **XGBoost Regressor** resulted in the most efficient output with minimum MAE (11.63) and RMSE (16.88) score. The two-stage predictive machine learning model is implemented on basis of above conclusion of techniques to work efficiently for the given problem statement.