# Flight Delay Prediction

Rishi Vardhan K

November 20, 2019

**Abstract**

**A 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.

| 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.

| 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.

| 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 columns 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 standing criteria for a flight not to be delayed is set under a threshold of 15 minutes. The feature **ArrDel15** contains values of Class 1 or Class 0 depending on whether a flight is delayed based on our criteria. Class 1 marking flight delay and Class 0 the negation.
The train set and test set are split in a ratio of 80:20.

## 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.

## 3.2 Observations

The following are the scores of various Classifiers considered.

| 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.
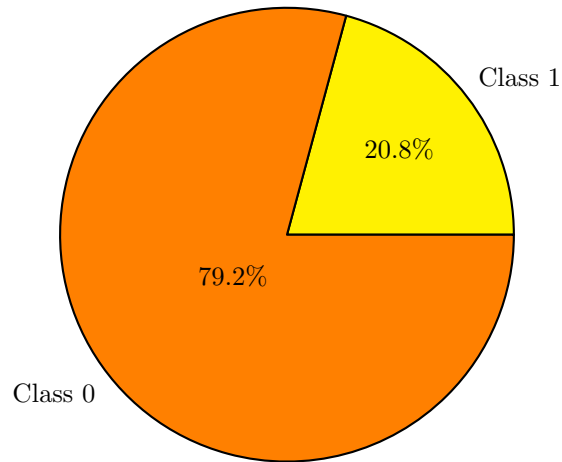


**Class Distribution**

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.

**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 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.

**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 arises as Random Over Sampler and SMOTE are found to show good performances in prediction of delays compared to others.

**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 |  |

A similar trend observed in previous cases arises here.

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

# 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.

## 4.1 Analysis: Arrival Delay Minutes
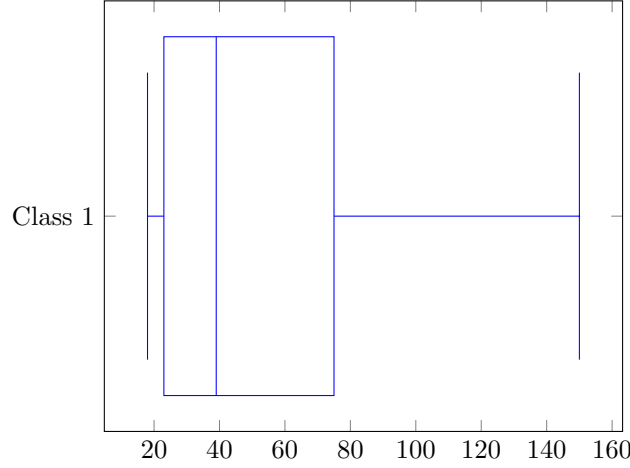


**Box Plot: Distribution of Arrival Delay Minutes**

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 Metrics Used

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

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

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

## 4.3 Tabulated Results

The results of different models are tabulated as follows.

| 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 |

## 4.4 Feature Importance

The Feature Importance concept 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 leads to improper output with limited changes. 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.