# Flight Delay Prediction

Vedanth Subramaniam

**Abstract**

A flight delay occurs when an airplane takes off and/or lands later than its scheduled time of departure or arrival.Management errors and weather conditions attribute to such delays.This project aims to predict the delay in arrival of flights using a two-stage machine learning model.

## 1 Introduction

Flying has become synonymous with delays, frustrating the well laid out plans of not just the flyers but also the airports and airlines involved.In the present world,knowing the delay beforehand would be a major benefit as the 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 of one place creates a cascading effect for the rest of them.Flight delays also result in major economic losses for all parties involved.

As a delay is usually caused with problems relating to departure and arrivals of flights, our aim is primarily to calculate the delay concerned with the arrivals.Since the weather also plays a role when it comes to departure/arrival delays, we take these factors into consideration and try to arrive at a solution.

## 2 Data Preprocessing

The Flight dataset contains information about the flights that flew within United States Of America during the years 2016 and 2017. The following airports are considered:

Airport Codes

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

Feature Selection has been applied to the given dataset so that only the features that are relevant are considered while we apply the respective models and thus redundancy is eliminated.

Flight Features

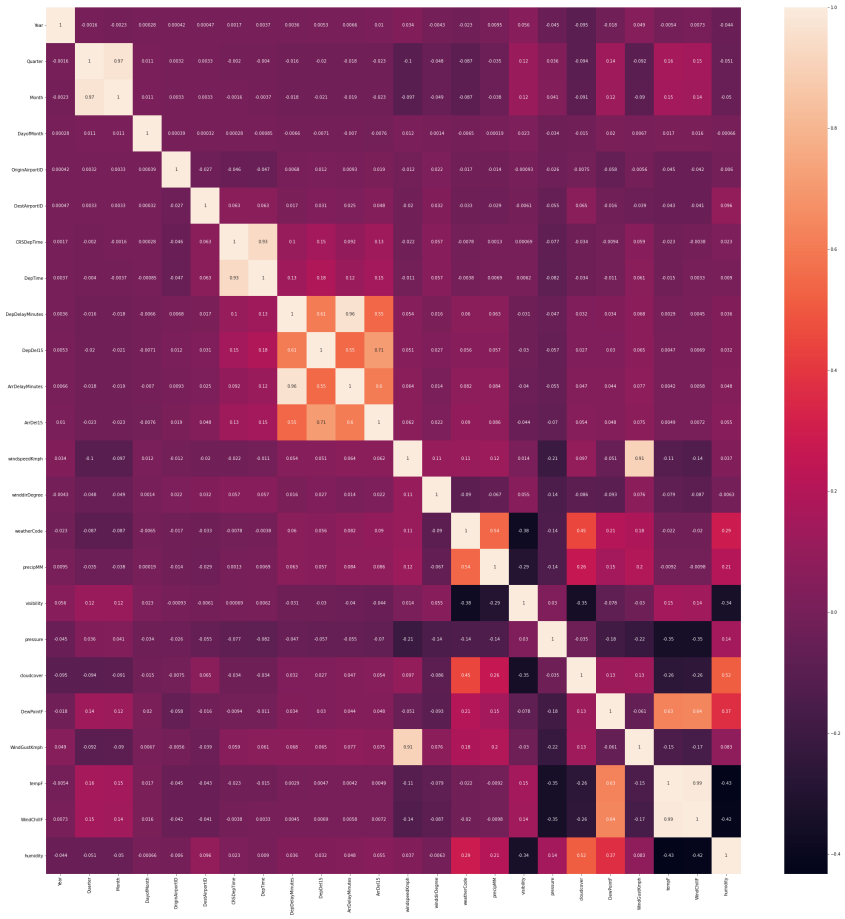
| | | |
|---|---|---|
| FlightDate | Quarter | Year |
| Month | DayofMonth | DepTime |
| DepDel15 | CRSDepTime | DepDelayMinutes |
| OriginAirportID | DestAirportID | ArrTime |
| CRSArrTime | ArrDel15 | ArrDelayMinutes |

The weather dataset is of type JavaScript Object Notation(JSON) and contains the weather conditions for each hour,each day and for a period of 12 months for the years between 2013 and 2017. Again, feature selection is carried and the following features are retained:

Weather Features

| | | |
|---|---|---|
| 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 with its respective weather report.



There is a correlation of 0.96 betweeen Departure delay and Arrival Delay, which means that a delayed take off would most likely result in a delayed arrival. However, since the correlation is not 1, a flight could pick up more speed on the air to avoid running late as much as possible or other factors at destination might affect the scheduled arrival.

# 3 Classfication

The aim of this module is to predict whether a flight will be delayed or not. The dataset is fed into the classifier and if the output is 1 then the flight is delayed and 0 if not. The 'ArrDel15' feature is used as the threshold involved in this classifying task.

## Classification Models

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

### 3.1 Metrics

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

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

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

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

Classifier Scores:

| Classifiers | Class | Precision | Recall | f1 | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 1 | 0.89 | 0.68 | 0.77 | 0.91 |
| | 0 | 0.92 | 0.97 | 0.95 | |
| Decision Tree | 1 | 0.68 | 0.71 | 0.69 | 0.87 |
| | 0 | 0.92 | 0.91 | 0.92 | |
| Extra Trees | 1 | 0.85 | 0.68 | 0.76 | 0.91 |
| | 0 | 0.92 | 0.97 | 0.94 | |
| Gradient Boosting | 1 | 0.89 | 0.68 | 0.77 | 0.91 |
| | 0 | 0.92 | 0.98 | 0.95 | |

### 3.2 Data Imbalance

The number of instances belonging to class 1 is very feeble compared to that of class 0. As a result of this imbalanced distribution, sampling techniques are used to tackle this.

Various over-sampling and under-sampling techniques are applied to each classifier to figure out the best fitting one. The results are tabulated as follows

Logistic Regression

| Samplers | Class | Precision | Recall | f1 | Accuracy |
|---|---|---|---|---|---|
| SMOTE | 1 | 0.73 | 0.78 | 0.76 | 0.89 |
| | 0 | 0.94 | 0.92 | 0.93 | |
| Random Over Sampler | 1 | 0.73 | 0.78 | 0.76 | 0.89 |
| | 0 | 0.94 | 0.93 | 0.93 | |
| Near Miss | 1 | 0.56 | 0.80 | 0.66 | 0.83 |
| | 0 | 0.94 | 0.83 | 0.86 | |
| Random Under Sampler | 1 | 0.74 | 0.77 | 0.76 | 0.90 |
| | 0 | 0.94 | 0.92 | 0.93 | |

Decision Trees

| Samplers | Class | Precision | Recall | f1 | Accuracy |
|---|---|---|---|---|---|
| SMOTE | 1 | 0.68 | 0.70 | 0.69 | 0.86 |
| | 0 | 0.92 | 0.91 | 0.91 | |
| Random Over Sampler | 1 | 0.69 | 0.70 | 0.69 | 0.87 |
| | 0 | 0.92 | 0.92 | 0.92 | |
| Near Miss | 1 | 0.36 | 0.83 | 0.50 | 0.66 |
| | 0 | 0.93 | 0.62 | 0.74 | |
| Random Under Sampler | 1 | 0.50 | 0.80 | 0.62 | 0.79 |
| | 0 | 0.94 | 0.79 | 0.86 | |

Extra Trees Classifier

| Samplers | Class | Precision | Recall | f1 | Accuracy |
|---|---|---|---|---|---|
| SMOTE | 1 | 0.82 | 0.72 | 0.77 | 0.91 |
| | 0 | 0.93 | 0.96 | 0.94 | |
| Random Over Sampler | 1 | 0.86 | 0.68 | 0.75 | 0.90 |
| | 0 | 0.92 | 0.97 | 0.94 | |
| Near Miss | 1 | 0.35 | 0.86 | 0.50 | 0.64 |
| | 0 | 0.94 | 0.58 | 0.72 | |
| Random Under Sampler | 1 | 0.64 | 0.83 | 0.72 | 0.86 |
| | 0 | 0.95 | 0.88 | 0.91 | |

Gradient Boosting

| Samplers | Class | Precision | Recall | f1 | Accuracy |
|---|---|---|---|---|---|
| SMOTE | 1 | 0.82 | 0.73 | 0.77 | 0.91 |
| | 0 | 0.93 | 0.96 | 0.94 | |
| Random Over Sampler | 1 | 0.73 | 0.79 | 0.75 | 0.89 |
| | 0 | 0.94 | 0.92 | 0.93 | |
| Near Miss | 1 | 0.47 | 0.82 | 0.60 | 0.77 |
| | 0 | 0.94 | 0.75 | 0.83 | |
| Random Under Sampler | 1 | 0.73 | 0.79 | 0.76 | 0.89 |
| | 0 | 0.94 | 0.92 | 0.93 | |

Among the four different sampling techniques which were tested out, only SMOTE(Synthetic Minority Oversampling Technique) and Random Over Sampling functions better than the rest.

However, both the sampled and non sampled methods almost yielded the same scores so it is possible to conclude that sampling has not been effective in tackling the data imbalance problem.

# 4 Regression

Regression is used to predict the delay in minutes. Tuples which are delayed are fed as input and the output is predicted based on Machine Learning models.

## 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**

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

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 |

# 5 Pipeline

Under the classification models, it was observed that the SMOTE sampling performed under Gradient Boosting Classifier yielded the best result. As a result, the output of this model is given as the input for best regressor model, again Gradient Boosting Regressor and thus a pipeline is constructed. Unlike training the entire dataset for this model, the 2016 data is used to train the model and the 2017 data is used as the test set.

Pipeline - Classifier

| Classifiers | Class | Precision | Recall | f1 | Accuracy |
|---|---|---|---|---|---|
| Gradient Boosting | 1 | 0.81 | 0.73 | 0.77 | 0.91 |
| | 0 | 0.93 | 0.95 | 0.94 | |

Pipeline - Regression

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

# 6 Regression Analysis

In general, the arrival delay(in minutes) for flights lies betweeen 0 and 2150. But for flights classified as delayed, the range is 15-2150.

| ArrivalDelayMinutes | RMSE | MAE | R Square |
|---|---|---|---|
| 15 - 300 | 16.63 | 11.55 | 0.89 |
| 300 - 550 | 29.4 | 19.2 | 0.77 |
| 550 - 800 | 27.95 | 17.69 | 0.84 |
| 800 - 1200 | 21.22 | 16.83 | 0.95 |
| 1200 - 2150 | 91.96 | 50.23 | 0.91 |

Frequency Distribution

# 7 Conclusion

Even though the accuracy obtained by the classifier model was fairly decent, it is still considered to have performed poorly because of data imbalance. The prediction of class 1 was poor when compared to that of class 0 and it is mainly due to the lack of training instances with label 1. Even various sampling techniques could not solve this problem, although the recall of class 1 did increase. Under regression modelling, Gradient Boost Regressor yielded the best results among the other models that were used. Thus, a two stage machine learning model has been constructed to tackle the given problem statement.