Flight Arrival Delay Prediction

Jerrick Gerald
Computer Science
Solarillion Foundation

Abstract. Flight Delays create a huge impact on airports and passengers. Prediction of flight delays is very crucial in aviation management. It not only creates inconvenience to passengers but also creates economic losses to airline companies. In this study, this paper presents the complete analysis of flights operating over 15 airports in the US and predicting possible arrival delay of flights using various classifiers and regression.

1 Introduction

A flight delay occurs when it departs or arrives airport later than the scheduled departure or arrival time. Flight delay has a negative impact mainly economic, for passengers, airlines, and airports. Passengers usually plan to travel many hours earlier for their appointments, increasing their trip costs, to ensure their arrival on time but due to delays of airlines, they may also lose trust in airline companies. So, to overcome this case an accurate prediction system is required. This model analyses flights covering 15 airports and gives arrival prediction i.e., the flight is delayed or not upon arrival.

2 DataSet:

- Flight Data: The data of all the flights that flew inside the US is provided for the year 2016 and 2017 in CSV format.
- Weather Data: The weather data consists of weather details of different airports in the USA between 2014 to 2017 in JSON format.

Flight Data :

FlightDate	Year	Quater
Dayofmonth	DepDel15	CRSDepTime
OriginAirportID	DestAirportID	ArrTime
Year	Month	DepDelayMinutes
CRSArrTime	ArrDel15	ArrDelayMinutes

Table 1: Flight Table

Weather Data:

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

Table 2: Weather Table

3 Data Preprocessing:

Flight data: The first preprocessing step takes place with the merging dataset of flight data of the year 2016 and 2017. The

flight data is filtered with the airport codes which are given below. The flights which are traveled between these places are considered in the dataset. The concatenated data is saved as Flight Data.

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

Weather Data: The second preprocessing step takes place with weather data. The JSON file is flattened and features are extracted for the prediction (features names are mentioned in Table2). These features are concatenated and saved as Weather Data in CSV format.
Final Dataset: The final step for preprocessing data is merging flight data and Weather data concerning 3 cases.

- Time
- Date
- Airport

This merged data is saved in the format of CSV and ready for classifier and regression.

3.1 Exploratory Data Analysis:

EDA is performed to investigate on a dataset with the help of statistics values and graphical representation. Some of the analysis are:

1. Dataset comprises of 1851436 observations and 30 characteristics.
2. Dataset has int, float, and object values.
3. Dataset columns have no null/missing values.
4. Correlation Matrix using Heatmap-lighter the colour higher the correlation values.

3.2 Feature Selection :

The main purpose of feature selection is to,

- Reduce complexity of the model.
- Improves accuracy of the model.
- Make the ML algorithm to train faster.

A correlation matrix in the form of a heatmap is shown in Fig 1, those features are taken for further prediction. The ground values are removed from the features.

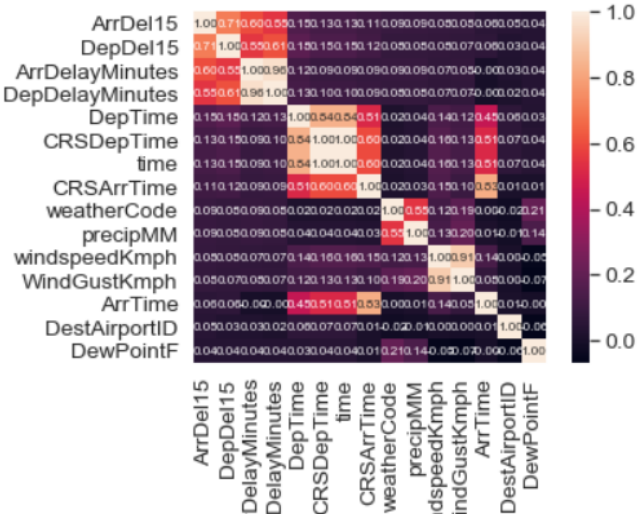

Figure 1: Correlation matrix for top 15 features with respect to ArrDel15.

3.3 Imbalanced Data and Removal:

A binary classification needs to be performed on the label, ArrDel15 which assumes binary values 0 and 1 where,

- 0 denotes no arrival delay flights.
- 1 indicates that there has been an Arrival delay of the concerned flight.

But, the number of instances with

- 0 has 1463378 samples.

- 1 has 388058 samples.

Hence, the dataset is highly imbalanced and the imbalance is visually shown in Fig 2. To reduce the imbalance dataset three sampling methods are performed.

- RandomOverSampler(Oversampler): Aims to increases the number of minority class (label1) members in the training set.
- NearMiss(UnderSampler): Aims to reduce the number of majorities (label 0) samples to balance the class distribution.
- SMOTETomek(Under-OverSampler): Combine Over and Under-sampling.

3.4 Data Spliting:

The dataset is splitted into Training and Testing Sets.

- Training-80%:
  - The actual dataset that we use to train the model.The model observes and learn from this data.
- Testing-20%:
  - The sample of data used to evaluation of a final model fit on the training dataset.

4 Classification:

The purpose of classification is to find whether the flight arrived is delayed or not.The imbalanced data is balanced and gives different classifier results.The classifier considered are ,

- Logistic Regression.
- Decision Tree Classifier.
- Extra Tree Classifier.
- XGBoost Classifier.

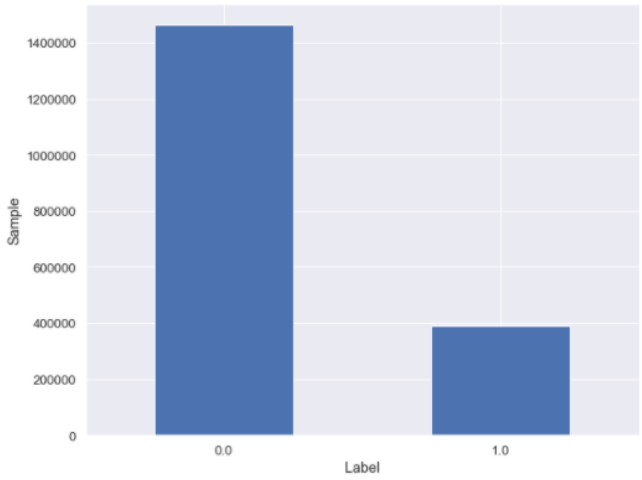

Figure 2: Bar plot of imbalance data between two labels.

Classifier	Precision		Recall		F1-Score	
	0	1	0	1	0	1
Logistic	0.94	0.50	0.78	0.83	0.85	0.62
DecisionTree	0.93	0.32	0.52	0.86	0.66	0.47
ExtraTree	0.95	0.36	0.59	0.88	0.72	0.51
XGBoost	0.94	0.35	0.57	0.87	0.71	0.50

Table 3: UnderSampling-Classifier Scores

Classifier	Precision		Recall		F1-Score	
	0	1	0	1	0	1
Logistic	0.94	0.73	0.98	0.78	0.93	0.75
DecisionTree	0.92	0.67	0.91	0.68	0.91	0.68
ExtraTree	0.93	0.84	0.96	0.71	0.94	0.77
XGBoost	0.94	0.73	0.92	0.79	0.93	0.76

Table 4: OverSampling-Classifier Scores

Classifier	Precision		Recall		F1-Score	
	0	1	0	1	0	1
Logistic	0.94	0.73	0.983	0.78	0.93	0.75
DecisionTree	0.92	0.66	0.91	0.69	0.91	0.68
ExtraTree	0.93	0.81	0.95	0.74	0.94	0.77
XGBoost	0.92	0.88	0.98	0.75	0.95	0.77

Table 5: SMOTETomek-Classifier Scores

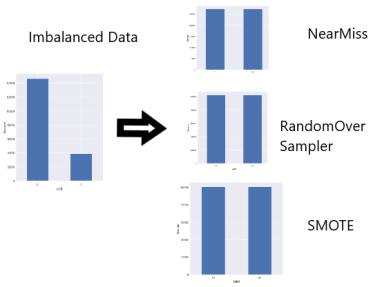

Figure 3: Reduction in Imbalance Data.

Data can be manipulated either by oversampling or undersampling or by both. Classification results were better for SMOTETomek than Undersample and Oversample data. SMOTETomek was considered for sampling as it gave a better result than the other samplers.

- SMOTETomek-Table5.
- RandomOverSampler-Table4.
- NearMiss-Table3.

5 Classification Metrics:

5.1 Precision:

Precision is defined as the proportion of correctly predicted positive observations of the total predicted positive observations.

$$\text{Precision} = \frac{TN + TP}{TN + TP + FP + FN} \quad (1)$$

5.2 Recall:

Recall or Sensitivity is defined as the fraction of correctly identified positives.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

5.3 F1-Score:

F1-Score is the harmonic mean (HM) of Precision and Recall.

$$\text{F1} - \text{Score} = \frac{2 * (Precision) * (Recall)}{Recall + Precision} \quad (3)$$

- TN-True Negative
- TP-True Positive
- FP-False Positive
- FN-False Negative

5.4 DrawBacks:

The recall is given preference than precision because of two drawbacks, in which the model gives false results. The reasons are,

- The flight is delayed but appears not to be delayed.
- The flight is not delayed but appears to be delayed

The first case is more trouble for passengers than the second one. So, the recall of the class needs to be high i.e,(class1) delayed flights to be high. ExtraTreeClassifier after SMOTETomek is considered as it has high recall and F1-score.

6 Regression:

Regression is used to find the arrival delay time. The delayed flights are extracted from the classifier, the same number of features are used to train the model. The regression used are,

- Linear Regression.
- Extra Tree Regressor.
- Lasso Regression.
- XGBoost Regression.

6.1 Regression Metrics:

6.2 Mean Absolute Error:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \quad (4)$$

Regression	R2	RMSE	MAE
Linear	92.15	20.04	14.7
Lasso	92.15	20.04	14.7
ExtraTrees	93.67	17.5	12.28
XGBoost	94.47	16.28	16.82

Table 6: Regression Scores

6.3 Root Mean Square Error:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \qquad (5)$$

6.4 R2 Score:

$$\text{R2} = 1 - \frac{\sum y - \hat{y}}{\sum y - \bar{y}} \qquad (6)$$

$\hat{y}$- predicted delay duration
y(bar)- Mean Value of y
N-Total number of samples in the dataset.

RMSE and MAE gives the error for the predicted values when compared to actual values. So, the lesser the error better the model is.
RMSE is given important because RMSE has the benefit of penalizing large errors. It avoids the use of taking the absolute value.
From Table 6, we can see that XGBoost Regressor has RMSE 16.28 and MAE 16.82 is less than the actual delay where as the actual delay is very high.

6.5 Regression Analysis:

RANGE	MAE	RMSE
0-550	11.62	16.70
550-1097	17.29	24.03
1150-1649	30.62	39.48
1711-2142	85.5	90.12

Table 7: Regression Analysis for XGBoost

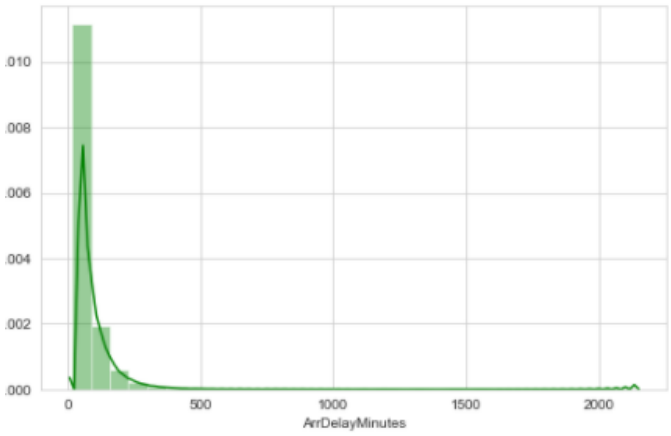

Figure 4: Dist plot for regression analysis

Regression analysis is performed for the XGBoost Regressor as it performs very well when compared to other models. The regression analysis is splitted into 4 ranges and the analysis is carried out. From the distplot(fig3) we can see that the tip between 0 and 300 is very high and gradually decreases at the end i.e., the flight delays are more between that period and decrease gradually over time.

7 Conclusion:

This paper proposed the machine learning system which was able to predict the flight delays with a good accuracy and minimal errors.Among classifiers ExtraTreeClassifier classifies the data into delayed or not.It has recall of 0.98 for majority class and has 0.74 for minority class,F1-score of 0.95 for majority class and 0.77 for minority class over SMOTETomek Sampling.
Among Regression XGBRegressor performs very well with 94% accuracy and has MAE of 16.82 mins and RMSE 16.28mins with actual delays.
The regression analysis is also carried with the best regressor(XGBoost) and the errors are minimum and shows that the delays

predicted are similar to actual delays.