Flight Arrival Delay Prediction

Jerrick Gerald
Computer Science
Solarillion Foundation

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

1 Introduction

A flight delay occurs when it departs or arrive 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 on airline companies.So,inorder to overcome this case a accurate prediction system is required.This model analyse 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 Data

3 Preprocessing:

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

is filtered with the airport codes which is given below .The flights which is travelled between these places is 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 preprocess step takes place with weather data.The json file is flatten 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 with respect to 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 in order to investigate on dataset with the help of statictics 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 - It is necessary to remove correlated variables to improve the model.

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

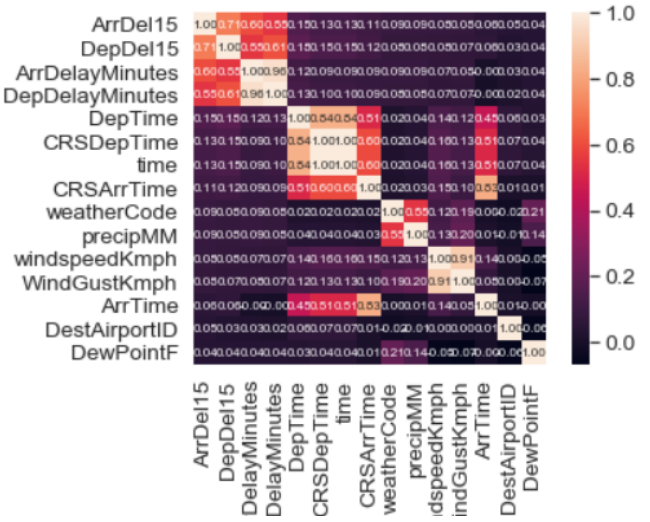

Figure 1: Correlation matrix for top 15 features with respect to ArrDel15.

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 Arrival delay of the concerned flight.

But,the number of instances with

- label 0 have 1463378 sample.
- label 1 have 388058 sample.

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

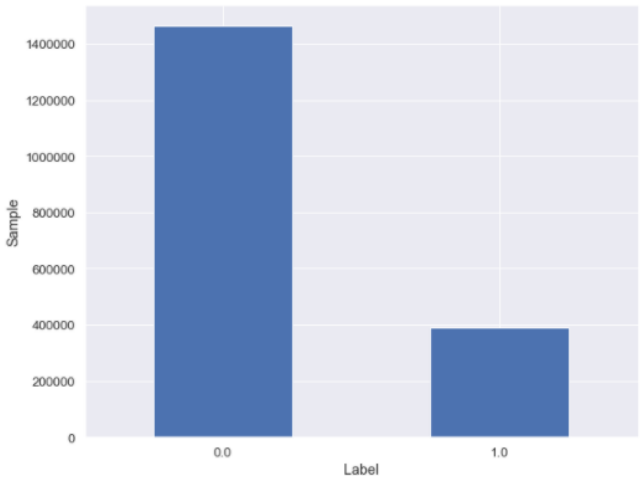

Figure 2: Bar plot of imbalance data between two labels.

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

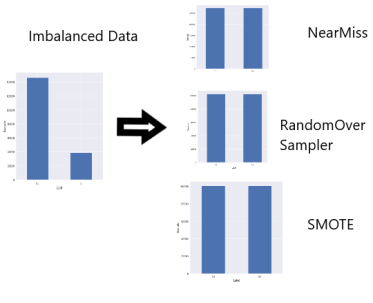

Figure 3: Reduction in Imbalance Data.

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.

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

Data can be manipulated either by oversampling or undersampling or by both. Classification results were better for SMOTETomek than Undersample and

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	0.69	0.95	0.77

Table 5: SMOTETomek-Classifier Scores

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:

Recall is given preference than precision because of two drawbacks, in which the model gives false results.The reason 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 to the passengers than the second one.So,the recall of the class is needed i.e,(class1) delayed flights to be high. XGBClassifier after SMOTETomek performs very well than other classifiers.It has good 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:

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

6.2 Mean Absolute Error:

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

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.So,lesser the error better the model is.
From the Table 4,XGBoost Regressor have RMSE 16.28 and MAE 16.82. RMSE is given important,because RMSE has the benefit of penalizing large errors.It avoids the use of taking the absolute value.XGBoost Regressor performs well and has less error when compare to other models in the Table.

6.5 Regression Analysis:

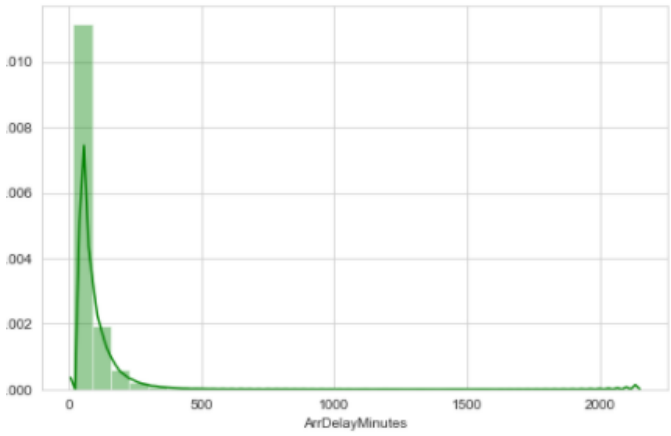

Figure 4: Dist plot for regression analysis

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 (250-300) is very high and gradually decreases at the end i.e., the delay of flights are more between that period and decreases gradually over the time.

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

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 XGBoost Classifier performs very well with recall 0.98 for majority class and has 0.70 for minority class,F1-score of 0.95 for majority class and 0.78 for minority class over SMOTE-Tomek Sampling.
Among Regression XGBRegressor performs very well with 94% accuracy and has MAE of 16.82 mins and RMSE 16.28mins.
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.