Flight Arrival Delay Prediction

Jerrick Gerald
Computer Science
Solarillion Foundation

Abstract. This paper presents the analysis of flights operating in 15 airports in the US and predicts the possible arrival delay of flights by applying classification and regression tasks. The classification tasks classify the flights into delayed and non-delayed flights. Among classifiers, the ExtraTree classifier was the best performing model. If the flights are predicted as delayed, we predict the arrival delay in minutes (Regression). Among various regressors, XGBoost Regressor was found to be the best performing regressor and recorded the least errors. The regression analysis has carried out using the XGBoost regressor to check the performance of the model on various ranges. Extra Tree classifier and XGBoost regressor have chosen to build the pipeline to test the entire model.

1 Introduction:

A flight delay occurs when a flight departs or arrives at the airport later than the scheduled time. With the rapid development of the aviation industry, flight delays have become an important problem for air transportation systems all over the world. Passengers usually plan to travel many hours earlier for their appointments by increasing their trip costs to ensure their arrival on time but due to delay, it creates a problematic situation for the passengers. On the other hand, airlines suffer extra crew costs, penalities associated with accommodating disrupted passengers, and aircraft rescheduling. So, to overcome this case an accurate prediction system is required. I proposed this model to analyze flight delays covering 15 airports and predicts delay on arrival.

2 Dataset:

- Flight Data: The data of the flights that flew inside the US is provided for the years 2016 and 2017 . (Refer Table1).
- Weather Data: The weather data consists of weather details of 15 different airports across USA between 2014 to 2017 (Refer Table2) .

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

Table 1: Flight Table

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

Table 2: Weather Table

3 Data Collection:

Flight data: The flight data contains the details of the flights from all airports in the US from 2016 and 2017. The flight data is filtered with the 15 airport codes (Refer Table3), where the flights which are traveled between these 15 places in the year 2016 and 2017 are considered in the dataset.

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

Table 3: Airport Codes.

Weather Data: The second step takes place with weather data. The necessary features are mentioned in Table2 which are considered for the years 2016 and 2017.

Final Dataset: The final step is merging flight data and Weather data concerning 6 features.

- Time - CRSDepTime.

- Date - FlightDate.
- Airport - Origin.

Now, we get the desired dataset for the study. It contains 1851436 observations and 30 characteristics.

3.1 Feature Selection:

The main purpose of feature selection is to,

- Reduce complexity of the model.
- Improve accuracy of the model.
- Allows the machine learning algorithm to train faster.

Features are selected based on the correlation matrix using a heatmap. A correlation heatmap uses colored cells, to show a 2-D correlation matrix between two discrete dimensions. Features are selected based on the top 15 features correlated to (Target Variable) ArrDel15 (Refer to Fig1). The darker shades in the chart represent negative correlation values and the lighter shades show the positive correlation values.

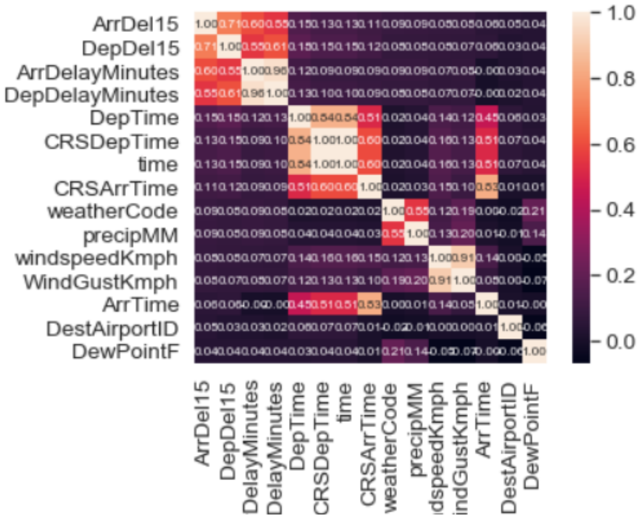

Figure 1: Correlation Heatmap for top 15 features with respect to ArrDel15.

3.2 Train Test Split:

The selected features are 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 classifiers considered are ,

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

4.1 Classification Metrics:

Precision, Recall and F1-score were selected as classification metrics for comparing the performance results of different models with respect to two classes.

- Class 0-Flight not delayed.
- Class 1-Flight delayed.

Precision:

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

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

Recall:

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

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

F1-Score:

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

$$\text{F}_1 - \text{Score} = \frac{2 \times (Precision) \times (Recall)}{Recall + Precision} \tag{3}$$

- TN-True Negative: Flights delayed, but classified as not delayed.
- TP-True Positive: Flights delayed and classified as delayed.
- FP-False Positive: On Time flights are classified as delayed.
- FN-False Negative: Flights delayed but classified as non delayed.

Classifier	Precision		Recall		F1-Score	
	0	1	0	1	0	1
Logistic	0.94	0.73	0.93	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: Classifiers Scores

F1-Score and recall were considered to find the best model. F1-score gives equal importance to recall (FN) and precision (FP). The recall is given more preference than precision because of two drawbacks, in which the model gives false results. The reasons are,

- Case 1: The model predicts the flight as not delayed, but the flight is actually delayed (FN).
- Case 2: The model predicts the flights as delayed, but the flight is not delayed (FP).

The first case is more trouble for passengers than the second one. So, the recall of the class (class1) delayed flights to be high. From Table 4 we can see that the ExtraTreeClassifier has the highest F1-Score of 0.77 and recall of 0.71.

5 Imbalanced Data and Removal:

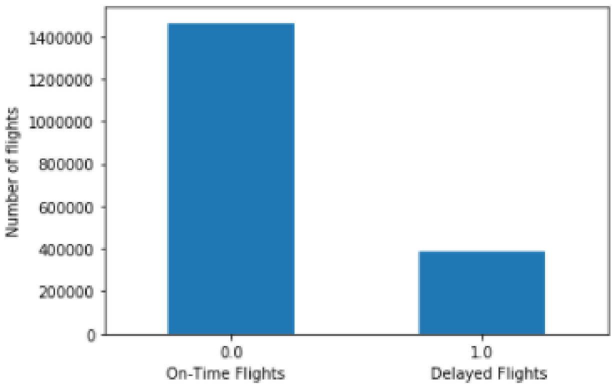

Figure 2: Bar plot of imbalance data between two class.

A dataset is said to be imbalanced when the number of observations per class is not equally distributed among the training dataset.
There has been a heavy class imbalance in our dataset where there is more number of flights which are not delayed and less number of flights that are delayed.

- Class 0 - On Time flights - 1463378 flights.
- Class 1 - Delayed flights - 388058 flights.

This shows that the data is highly imbalanced and the bar plot is visually shown in Fig2.
To reduce the imbalance dataset three sampling methods are performed. The samplers are,

- RandomOverSampler.
- NearMiss.
- SMOTETomek.

Random OverSampler:
Random OverSampler method is an **oversampling** technique. It aims to increases the number of minority class members in the training set.

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 5: Random OverSampler-Classifier Scores

Near Miss:
Near Miss is an **Undersampling** technique. It aims to reduce the number of majorities samples to balance the class distribution.

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 6: NearMiss-Classifier Scores

SMOTETomek:
SMOTETomek sampler is a combination of both **Under and Over Sampling** techniques.

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	0.66	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 7: SMOTETomek-Classifier Scores

ExtraTree classifier after SMOTETomek was the best performing classifier because it had a good F1-score and recall for the minority class, whereas other classifiers which were skewed towards the majority class.

6 Regression:

Regression is used to find the arrival time of the delayed flights which are extracted from the classifier. Three regressors are used to perform regression . The regressors used are,

- Linear Regression.
- Extra Tree Regression.
- XGBoost Regression.

7 Regression Metrics:

RMSE and MAE were selected as performance measures for comparing the prediction results of different models. Both RMSE and MAE gives the error in the predicted value. So the lesser the error, the better the model is. Along with that R-Squared value is also considered to find the goodness of fit.

7.1 Mean Absolute Error:

MAE is the sum of absolute differences between our target and predicted variables

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}i| \tag{4}$$

7.2 Root Mean Square Error:

The square root of the average of squared differences between prediction and actual observation.

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

7.3 R^2 :

R-squared is a measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{6}$$

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

Regressors	R2	RMSE	MAE
Linear	92.15	20.04	14.7
ExtraTrees	93.67	17.5	12.28
XGBoost	94.47	16.28	11.64

Table 8: Regression Scores

From Table 8, we can see that the XGBoost Regressor has minimum RMSE (16.28 mins) and MAE (11.64 mins) and highest r2 score of 94% when compared to other regressors.

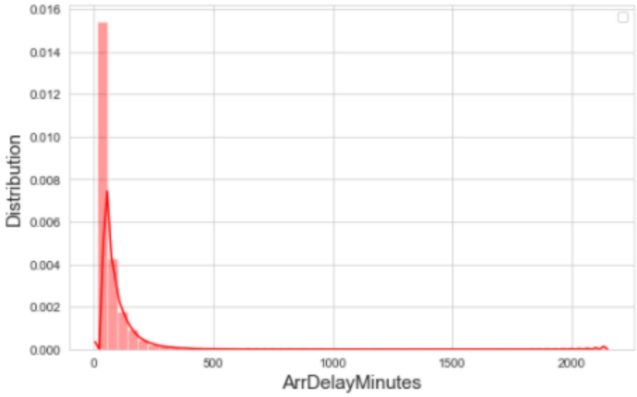

Figure 3: Dist plot for regression analysis

RANGE	MAE	RMSE
15-200	11.28	15.97
200-500	15.14	21.55
500-800	16.04	23.63
800-1971	19.5	33.12
2028-2142	144	144

Table 9: Regression Analysis for XGBoost

7.4 Regression Analysis:

The Regression analysis was carried out to check the performance of the best-concluded model i.e., XGBoost Regressor across various ranges of the target column. It can be concluded from the distplot (Figure 3) that most of the delays fall below 200 minutes. The column was split into 5 categories and the errors in each of them were analyzed as shown in Table 9. The range 15-200 with most of the data points recorded the least MAE and RMSE values which were 11.3 and 15.9 respectively. These low errors indicate the good predictions of the regressor. It is also observed that when the range increases the number of data points decreases and as a result, the values of RMSE and MAE increase. Hence, it can be concluded that the model performs better in the range 15-200 when compared to data in the range 2028-2142.

8 Pipeline:

The purpose of the pipeline is to test the entire model. The dataset was preprocessed and sampled for classification. ExtraTree classifier was chosen for the classification task. From classification, the delayed flights are considered to perform the regression task. XGBoost Regressor was chosen for the regression model. Fig 4 shows the structure of the pipeline. The performance of the pipeline model is shown in Table 10 and Table 11.

Regressors	R2	RMSE	MAE
XGBRegressor	94	16.89	11.81

Table 10: Pipline Scores for XGBoost Regressor

Classifier	Precision		Recall		F1-Score	
	0	1	0	1	0	1
ExtraTreesClassifier	0.93	0.81	0.96	0.73	0.94	0.77

Table 11: Pipline Scores for ExtraTree Classifier

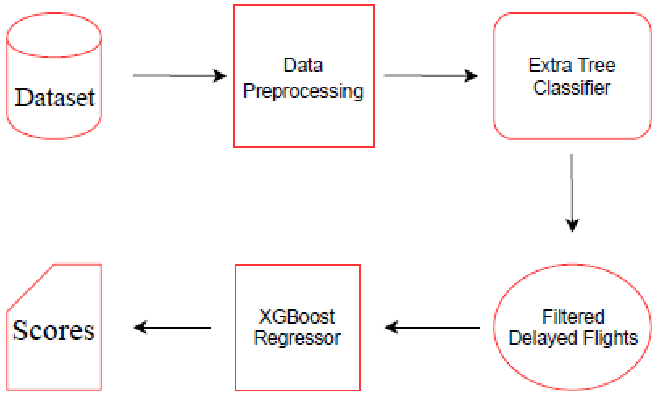

Figure 4: Pipeline Structure.

9 Conclusion:

This paper proposed the machine learning system which was able to predict the flight delays with good accuracy and minimal errors. ExtraTrees classifies the flights into delayed or not. It has good scores of F1-score and recall for minority class, after SMOTETomek. XGBRegressor predicts the arrival time of flights. It performs very well with an accuracy of 94% and has minimum errors than the other regressors. Regression analysis is carried out with the best regressor XGBoost Regressor. The pipeline model was deployed with the best classifier and regressor and performed well with good scores.