FLIGHT DELAY PREDICTION

S Sowmiya Sree

College Of Engineering, Guindy

ABSTRACT

The main goal of the project is to predict the arrival delay of the flights across 15 stations in the US. A two stage model was built which classifies delays in the former stage and predicts the duration of delays in the latter. Strong and robust predictive models were built through ensemble techniques and regularization. The predictive models were analyzed with reliable metrics and the results were tabulated. The best analyzed classifier had 89% recall score and the best regressor had 93.5% r^2 score.

INTRODUCTION

A delay is defined as any flight which arrives at least 15 minutes later than the scheduled arrival time. Data shows that an average of 2950 flights is delayed for more than 15 minutes every day. Factors like air traffic, weather conditions, bird strikes, mechanical problems, and many more are known to impact the delay of the flights. In 2019, the leading cause of flight delays was late aircraft arrival which accounted for 39.7 percent of the total delay minutes. The flight delays impact the passengers, airline operations and the economy as well. With the increase in air transport demand, the subject of flight delays has drawn much research interest in the past two decades.

The raw data of flight and weather features were processed into a single data file as explained in [**1**]. This was followed by building predictive models. The binary classification was achieved through the classifiers as elaborated in [**2**]. The feature selection in [**2.1**] visualizes the multiple correlations and the highly correlated features were chosen for classification. The models were analyzed with metrics which are described in [**2.4**]. Further, the train data's imbalanced class was dealt in [**2.5**]. The binary classification of the delays was followed by predicting the duration of the delays in [**3**]. The regularization techniques discussed in [**3.2.1**] were used to build robust models. The regression metrics are elaborated in [**3.3**] and the best concluded regressor was analyzed in [**3.2**].

1 DATA PREPROCESSING

The combined data of flight and weather details contributed to predicting the arrival delay of flights. Table 1 shows the 15 station codes whose data were considered.

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

Table 1 : Airport station codes.

The flight data comprised of several observations for 2 years namely 2016 and 2017. Table 2 shows the flight details that were considered.

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

Table 2 : Flight details

The weather data files that corresponded to 2016 and 2017 were chosen. Table 3 presents the weather features chosen for the prediction.

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

Table 3 : Weather features

The processed flight and weather data were merged into a single data file which had **18,51,436** observations. This data file was then used as a source for the upcoming classification and regression tasks.

2 CLASSIFICATION

The classification algorithms were used to predict whether the arrival of flights would be delayed or not after their departure. This classification was achieved through stages as explained below.

2.1 FEATURE SELECTION

Feature selection is the process of selecting only those features that contribute most to the prediction. Considering the irrelevant features results in poor performance of the model. The Heatmap was used to visualize the correlations between all the features as shown in Figure 1.

Heatmap is a two-dimensional plot which contains values represented by various shades of the same colour. The darker shades of the chart represent higher values(correlation) than the lighter shade. This is denoted by the gradient scale to its right.

The **Univariate Selection Method** was used to identify the top correlated features. The choice of the number of features depends on the shape of the data file. Here, top 16 features were chosen which are shown in figure 2. The **SelectKbest** is a class under univariate feature selection that scores the top correlated features with a statistical function. **ANOVA**(Analysis Of VAriance) tests the relationship between a categorical(target variable) and a numeric variable(features) by testing the differences between their means. This test produces a p-value to determine whether the relationship is significant or not. This statistical tool was used in SelectKbest.

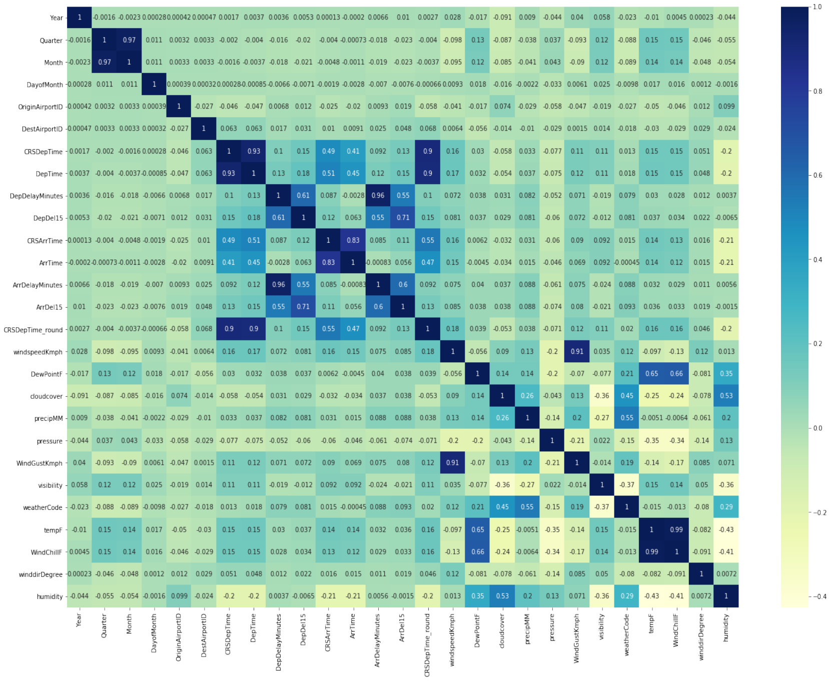
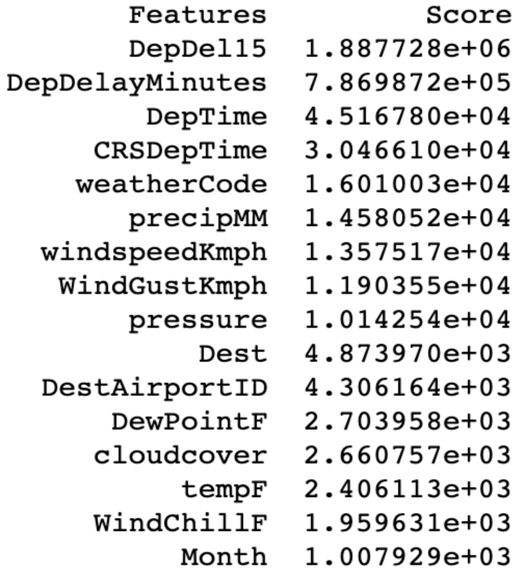


Figure 1 : Heatmap showing the multi-collinearity between the features.



```
       Features         Score
       DepDel15  1.887728e+06
DepDelayMinutes  7.869872e+05
        DepTime  4.516780e+04
     CRSDepTime  3.046610e+04
    weatherCode  1.601003e+04
       precipMM  1.458052e+04
  windspeedKmph  1.357517e+04
   WindGustKmph  1.190355e+04
       pressure  1.014254e+04
           Dest  4.873970e+03
  DestAirportID  4.306164e+03
      DewPointF  2.703958e+03
     cloudcover  2.660757e+03
          tempF  2.406113e+03
     WindChillF  1.959631e+03
          Month  1.007929e+03
```

Figure 2 : Top 16 correlated features for classification

2.2 TRAIN & TEST SPLIT AND LABEL ENCODING

The data with selected features were split into the train and test sets with a random state. 80% of the data was considered for the train set and the rest 20% for test set.

Label Encoding refers to converting the labels into a numeric form to convert it into the machine-readable form. It was done to convert non-numerical columns like station codes and flight dates to numbers.

2.3 CLASSIFIERS

Classifiers like Logistic regression and Decision trees were built and their performances were analyzed. Ensemble techniques were also used since they produce stronger predictive models from weak ones. The two main categories of ensemble techniques namely Boosting and Bagging were implemented.

1. **Bagging** - A Bagging classifier is an ensemble approach that fits base classifiers(Decision trees) on each random subsets of the original dataset(known as Bootstraps). It then aggregates their predictions to form the final prediction.
2. **Boosting** - Boosting is an iterative technique that adjusts the weight of an observation based on the last classification. This technique focuses on the wrongly predicted observation in every stage until best possible prediction is made.

Bagging techniques like **RandomForest Classifier** & **ExtraTrees Classifier** and Boosting technique like **XGBoost** were analyzed.

2.4 CLASSIFICATION METRICS

The classifiers were analyzed with classification reports. The report shows the main classification metrics like precision, recall, and f1-score for each class. The two classes of the target column include **Class 0** which indicates that the flight is not delayed and **Class 1** which denotes the delay of the flight. The metrics are calculated by using true and false positives, true and false negatives.

1. **True positive**(TP) - an outcome where the model correctly predicts the positive class. In exact, the delayed observations are predicted as delayed by the classifier as well.
2. **False positive**(FP) - an outcome where the model incorrectly predicts the positive class. In exact, the flights which are not delayed are predicted to be delayed by the classifier.
3. **True negative**(TN) - an outcome where the model correctly predicts the negative class. In exact, the non-delayed flights are predicted as non-delayed by the classifier as well.
4. **False negative**(FN) - an outcome where the model incorrectly predicts the negative class. In exact, the delayed flights are predicted as non-delayed by the models.

Precision(positive predictive value) is the ability of each of the classifiers not to label a non-delayed flight as delayed.

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

Recall(sensitivity) is the ability of the classifiers to correctly predict all the delayed flights.

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

F1 score can be interpreted as a weighted average of precision and recall. This score takes both false positives and false negatives into account.

$$f1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

Classifiers	**Precision**		**Recall**		**f1-score**	
	Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Logistic regression	0.92	0.89	0.98	0.68	0.95	0.77
Decision Trees	0.92	0.67	0.91	0.70	0.91	0.68
Random Forest	0.92	0.88	0.98	0.68	0.95	0.77
Extra trees	0.93	**0.85**	0.97	**0.72**	0.95	**0.78**
XGBoost	0.92	0.90	0.98	0.68	0.95	0.77

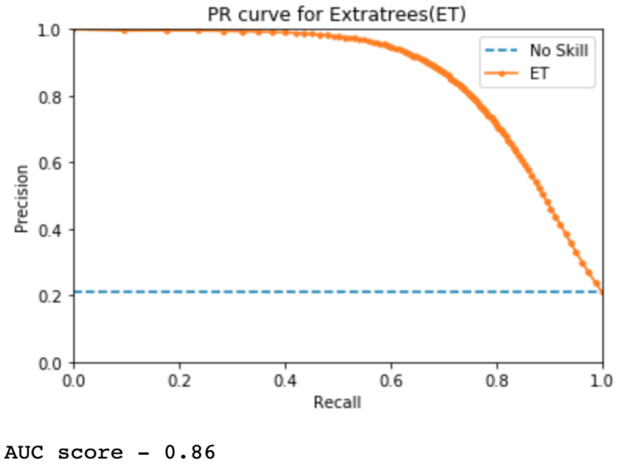
Table 4 : Performance of the classifiers.

The accuracy scores of class 1 were chosen to compare the models since predicting the delay is the main purpose. The recall metric is considered more important when compared to the other metrics as misclassifying the delayed flights as non-delayed deteriorates the model's performance more than the error caused predicting non-delayed flights as delayed. Extra trees had the highest recall score(0.72).

Precision-recall(PR) curve was used to analyze the performance of extratrees for the following reasons.

1. The ultimate aim is to predict the delays. The true negative values(non-delays) are not as important as the delays. Both precision and recall doesn't take true negatives into account.
2. The train data had imbalanced classes(elaborated in **2.5**). Precision and recall make it possible to assess the performance of a classifier on the minority class.

The no-skill classifier in figure 3, indicates the poor performance. It is the precision score which is proportional to the positive observations(delays). Due to class imbalance, the no-skill classifier is plotted at 0.2 precision. The AUC(Area under Curve) quantifies the performance of the model. The classifier had 85% of AUC.

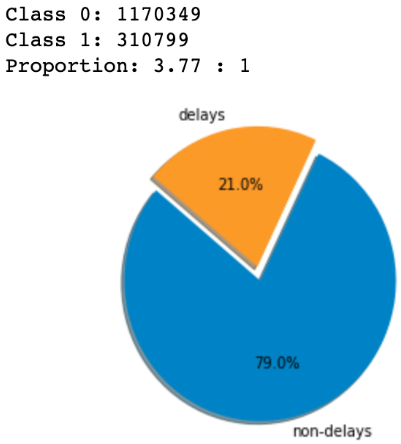


```
AUC score - 0.86
```

Figure 3 : PR curve for the best analyzed classifier.

2.5 CLASS IMBALANCE

Class Imbalance is the problem of classification when there is an unequal distribution of classes in the training dataset. The count of both the classes(classes 0 and 1) and their proportion is shown in figure 4. Such imbalanced classes mislead the predictive algorithms. Sampling techniques were used to balance the classes.



```
Class 0: 1170349
Class 1: 310799
Proportion: 3.77 : 1
```

Figure 4 : Distribution of classes.

2.5.1 UNDERSAMPLING

Undersampling refers to a group of techniques designed to balance the class distribution by reducing the count of the majority class. **Random Undersampling** was implemented on the training dataset. The technique randomly chooses observations from class 0(non-delayed flights) such that both the classes become proportionate.

Classifiers	**Precision**		**Recall**		**f1-score**	
	Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Logistic regression	0.94	0.74	0.93	0.78	0.93	0.76
Decision Trees	0.94	0.49	0.79	0.80	0.85	0.61
Random Forest	0.94	0.70	0.91	0.79	0.93	0.74
Extra trees	0.95	0.69	0.91	**0.81**	0.93	0.75
XGBoost	0.94	0.73	0.92	0.79	0.93	0.75

Table 5 : Random Undersampler on the train set.

2.5.2 OVERSAMPLING

Oversampling does the same job as undersampling to balance the class distribution but differs by the fact that it creates duplicate observations of the minority class. **Random Oversampling** was applied to the dataset. It randomly replicates the data points of delayed flights to equal the count of non-delayed flights.

Classifiers	**Precision**		**Recall**		**f1-score**	
	Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Logistic regression	0.94	0.74	0.93	0.78	0.93	0.76
Decision Trees	0.92	0.69	0.92	0.70	0.92	0.70
Random Forest	0.93	0.83	0.96	0.72	0.94	0.77
Extra trees	0.93	0.81	0.96	0.74	0.94	0.77
XGBoost	0.94	0.73	0.92	0.78	0.93	0.76

Table 6 : Random Oversampler on the train set.

2.5.3 SMOTE

SMOTE(Synthetic Minority Oversampling Technique) is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.

Classifiers	**Precision**		**Recall**		**f1-score**	
	Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Logistic regression	0.94	0.74	0.93	0.78	0.93	0.76
Decision Trees	0.92	0.65	0.90	0.70	0.91	0.67
Random Forest	0.93	0.83	0.96	0.72	0.94	0.77
Extra trees	0.94	0.78	0.94	0.76	0.94	0.77
XGBoost	0.93	0.79	0.95	0.75	0.94	0.77

Table 7 : SMOTE on the train set.

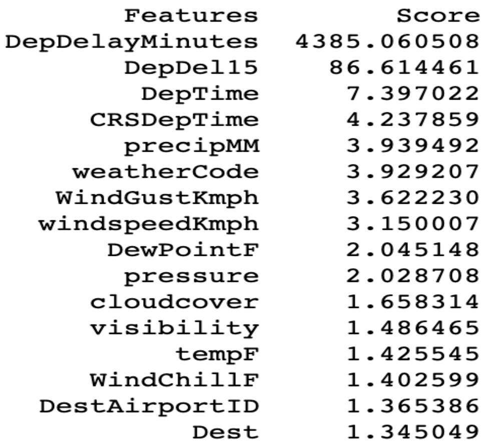
The random undersampling had higher recall scores for class 1. Among the five classifiers, Extra trees had the highest recall score of 81%. Hence, the Extra trees classifier with undersampled train data was analyzed to make reliable predictions.

3 REGRESSION

The binary classification of the flight delay was followed by predicting the duration of delay, for which regression algorithms were used. The best analyzed classifier(Extra trees) was used to predict the binary classification on the entire data. The classifier's predictions were pipelined to the regressors.

3.1 FEATURE SELECTION AND TRAIN & TEST SPLIT

The univariate selection method was used to select the top 16 correlated features and their scores are shown in figure 5. Like Classification, 80% of the data was split into the train set, and 20% to test set with a random state.



```
         Features        Score
 DepDelayMinutes  4385.060508
        DepDel15    86.614461
         DepTime     7.397022
      CRSDepTime     4.237859
        precipMM     3.939492
     weatherCode     3.929207
     WindGustKmph     3.622230
   windspeedKmph     3.150007
       DewPointF     2.045148
        pressure     2.028708
      cloudcover     1.658314
      visibility     1.486465
           tempF     1.425545
      WindChillF     1.402599
   DestAirportID     1.365386
            Dest     1.345049
```

Figure 5 : Top 16 correlated features for regression

3.2 REGRESSORS

Initially, **multiple linear regression** was implemented with the selected features followed by the regularization techniques.

3.2.1 REGULARIZATION

This is a form of optimization, that regularizes or shrinks the coefficient estimates towards zero. The ultimate goal is to reduce the cost function. The cost function quantifies the error between predicted values and expected values. The Residual Sum of Squares(RSS) is the most commonly used cost function.

$$\hat{y_i} \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{4}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{5}$$

n - total number of observations

p - number of features considered for prediction

x - features

y_i - actual delay duration

$\hat{y_i}$- predicted delay duration

β_0 - constant

β_{1-p} - coefficients

RIDGE REGRESSION

The ridge regression penalizes the cost function by L2 regularization. In exact, it modifies the cost function by adding the penalty equivalent to the square of the magnitude of coefficients along with the tuning parameter(α or λ)

$$Total\ cost\ function = RSS + \lambda \sum_{i=1}^{p} \beta_i^2 \tag{6}$$

LASSO REGRESSION

The lasso regression works very much similar to that of ridge regression differing by the fact that it performs L1 regularization. It penalizes the cost function by adding the sum of the absolute value of coefficients.

$$Total\ cost\ function = RSS + \lambda \sum_{i=1}^{p} |\beta_i| \tag{7}$$

Further, the ensemble techniques were also implemented to predict the duration of delays. **ExtraTrees** and **XGBoost** were implemented.

3.3 REGRESSION METRICS

The performance of the regressors was analyzed with RMSE(Root Mean Squared Error), MAE(Mean Absolute Error) and R^2 as shown in table 8.

MAE

MAE measures the average magnitude of the errors in a set of predictions. It's the average over the test sample of the absolute differences between predicted delay and the actual delay.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{8}$$

RMSE

RMSE is an absolute measure of fit.RMSE is a quadratic scoring rule that measures the average magnitude of the error. It's the square root of the average of squared differences between predicted delay and the actual delay.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{9}$$

R^2

R^2 is a goodness of fit measure for regression models. It is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{10}$$

where, y_i is the actual delay, $\hat{y}_i$ is the predicted delay and $\overline{y_i}$ is the mean of the actual delays in equations 8, 9 and 10.

	Linear Regression	Ridge regression	Lasso regression	ExtraTrees	**XGBoost**
MAE	12.93	12.93	13.22	12.89	**12.35**
RMSE	18.11	18.11	18.28	17.94	**17.52**
R^2	93.05%	93.05%	92.92%	93.18%	**93.5%**

Table 8 : Perfomance of the regressors

XGBoost had the highest R^2 value of **93.5%**. Before concluding the goodness of fit, the residual values(here MAE and RMSE) were also analyzed. Residuals are the differences between the actual delay and the predicted delay. Lesser the residuals, better the performance of the model. XGBoost had the least residual scores, **12.35** MAE and **17.52** RMSE.

REGRESSION ANALYSIS

The target column was split into 5 categories. The residuals were analyzed in all the categories to ensure better predictions.

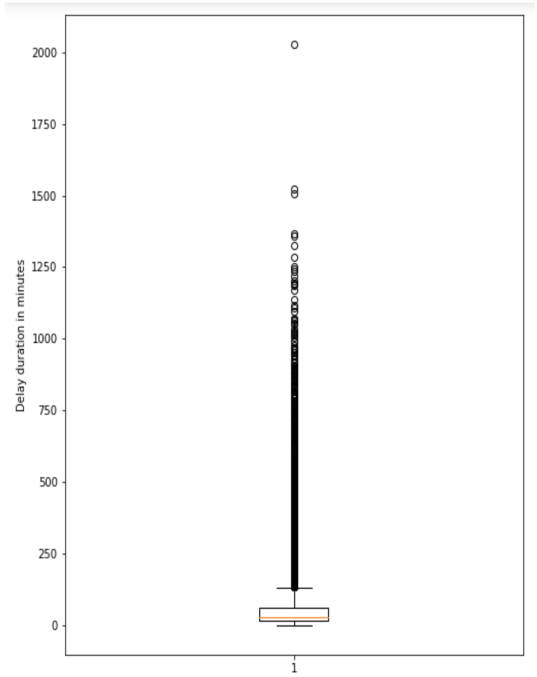


Figure 6 : Box plot to visualise the distribution of data points in the target variable.

The box plot shows that most of the duration of delays is less than 250 minutes. The data points above the upper whisker, indicate the uneven distribution of data points on the column. Table 10 shows the residuals in all 5 categories. The number of observations in each of the categories is given in paranthesis. As observed from the boxplot, majority of the delays(94,396) fall below 250 minutes. The residuals in this range were very low. In exact, 12.21 MAE and 17.12 RMSE is very much less than 250. This indicates that the predicted delays are very much closer to the observed delays.

	0 - 250(94,396)	250 - 500(1483)	500 - 750(159)	750 - 1250(97)	1250 - 2100(8)
MAE	12.21	14.68	18.89	18.45	84.82
RMSE	17.12	21.71	27.92	27.09	101.57

Table 10 : Performance of XGBoost on the 5 categories.

4 INFERENCE

The arrival delay was predicted for the entire dataset using XGBoost. Figure 7 shows the total delay of flights(predicted) for all 15 stations. **SFO(San Francisco)** was analyzed to be the busiest airport in the two years, 2016 and 2017.

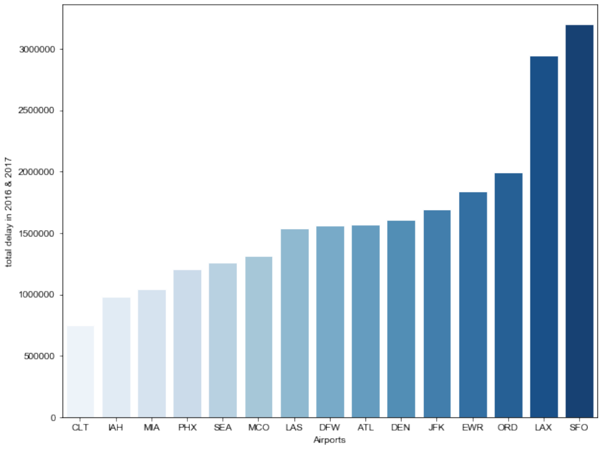


Figure 7 : total delay of airports.

CONCLUSION

The arrival delay of the flights was predicted to the best possible accuracy with supervised machine learning algorithms. The Bagging classifier, **Extratrees** predicted better with highest recall score on the required class i.e., Class 1. Also, the model's prediction was reliable with undersampled train data. The **XGBoost** regressor predicted the duration of the delay much better than the regularized models like ridge and lasso for this data. The regressor worked very well on the unevenly distributed target feature with very less residuals.