Real time flight delay prediction

ABSTRACT

The objective of this project is to predict the flight delay based on the data of 15 airports (refer in Table 1) in the United States of America using a pipeline model. The project involves an extensive data acquisition and data preprocessing of flight and weather data. A pipeline consisting of a classifier and a regressor was modeled, built, and analyzed to predict the departure delay of flight. This model predicts the delay in two steps, in the first step the classifier makes a prediction about a possible delay and in the second step the regressor then goes on to quantify the delay. The accuracy of the classifier and regressor were found to be 86% and 92% respectively.

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

Table 1: Flight terminals Table

INTRODUCTION

Major cause of flight delays are bad weather conditions and technical problems with the aircrafts. Other factors such as air traffic congestion, security issues and late arrival of passengers can also contribute to flight delays.

Initially the Flights are classified as *delay* or *not delay* using a classifier model and then the fights which are classified to be delayed are subjected to a regressor which predicts the delay of flights in minutes. The best regression model is evaluated within every possible interval in which the model can predict and this process gives a comprehensive idea of where the model performs poorly and where the model excels.

1 DATASET

The dataset for predicting delay was created by merging of flight and weather data from year 2016 to 2017. Weather data contains hourly weather conditions of 15 airports which includes columns like the *windSpeedKmph, precipMM, visibility, arrTime, ArrDelayMinutes* and many other columns which are mentioned in Table 3. The flight dataset consist of columns which describes the flights departure and arrival details such as year, month, day, origin and destination which are mentioned in Table 2.

FlightData	Quarter	Year	Month	DayofMonth
DepTime	DepDel15	CRSDepTime	DepDelatMinutes	OriginAirportID
DestAirportID	ArrTime	CRSArrTime	ArrDel15	ArrDelayMinutes

Table 2: Flight Details

WindSpeedKmph	WindDirDegree	WeatherCode	precipMM	Visibility
Pressure	Cloudcover	DewPointF	WindGustKmph	tempF
WindChilliF	Humidity	date	time	airport

Table 3: Weather Details

2 FEATURE SELECTION

NaN values which were present in the combined dataset were cleaned and Columns like *FlightDate, Origin, Dest, time, date, airport, ArrTime, ArrDel15, ArrDelayMinutes, DepTime* were removed because these parameters were either redundant, not related to departure delay or would over-fit the model.

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1.

From the Heatmap (Figure 1) we can infer that *month and quarter* are positively correlated, *windspeedkmph and windGustKmph* are positively correlated and *WindChillF and tempF* are positively correlated. So, the columns *Quarter, WindGustKmph and WindChillF* were removed.

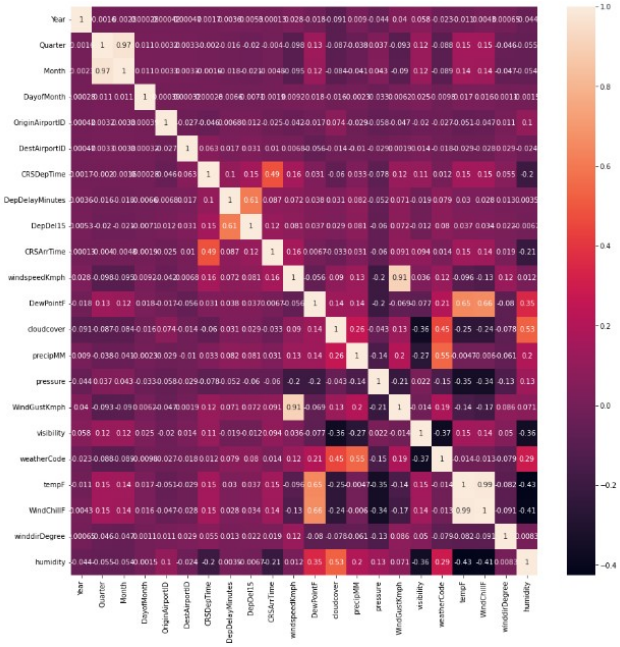


Figure 1: Correlation Heatmap

3 CLASSIFICATION

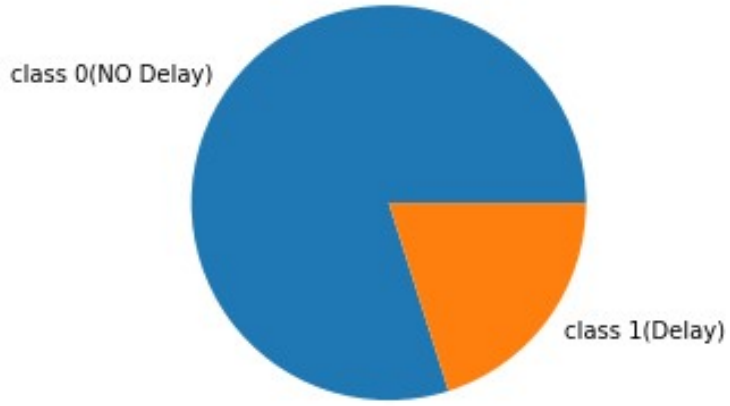
Classification is a process where the model segregates the datapoints into different classes. A Classification model was built to predict if a flight is delayed on departure or not. Departure delay which is lesser than 15 minutes is not considered. Models like Logistic Regression, DecisionTreeClassifier, RandomForestClassifier, XGBOOST, ExtraTrees Classifier were used and their performance was analysed using various metrics.

3.1 CLASS IMBALANCE

Class imbalance is a problem in which population of datapoints in one class is higher than other class. This imbalance in the dataset would affect the performance of the classifier as it tends to be more biased towards the majority class, causing bad classification of the minority class.

The majority class here is *Class 0* which indicates that the flight is delayed and the minority class is *Class 1* which indicates the flight is not delayed. From the Figure 2, we can infer that there is an massive class imbalance. Undersam-

pling and oversampling are two techniques which are used to address this class imbalance in a dataset. Elimination of class imbalance would help increase the performance of the classifier.



0 - 1479357
1 - 372079

Figure 2: Class imbalance pie chart

3.2 ANALYSIS METRICS

Various classification metrics like precision, recall, F1-score were used to compare different models.

TP (True Positive): A classification outcome where the model correctly predicts the positive class.

TN (True Negative): A classification outcome where the model correctly predicts the negative class.

FP (False Positive): A classification outcome where the model incorrectly predicts the positive class.

FN (False Negative): A classification outcome where the model incorrectly pre-

dicts the negative class.

Recall: Recall is a metric that measures the proportion of true positive instances that are correctly identified by a model.

Precision: Precision is a metric that measures the proportion of true positive instances out of all positive instances identified by a model.

Accuracy: Accuracy is the ratio of the number of correct predictions and the total number of predictions.

F1 score: F1 score is a harmonic mean of recall and precision that combines the two metrics to give an overall measure of model performance.

$$\text{Recall} = \cdot\frac{TP}{TP+FN}$$

$$\text{Precision} = \cdot\frac{TP}{TP+FP}$$

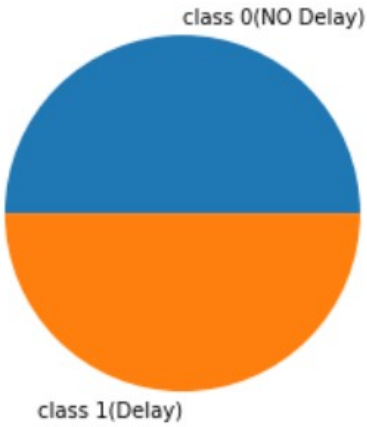
$$\text{Accuracy} = \cdot\frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{F1 Score} = 2 \cdot\frac{Precision \cdot Recall}{Precision+Recall}$$

3.3 UNDERSAMPLING

Undersampling is a technique to balance uneven datasets by deleting or removing samples(datapoints) from the majority class.

From Table 4 we can infer that XGBOOST was the best performing model and the least preforming model was Logistic regression. The F1 score is considered to determine the best classifier model as it is a metric which consolidates both precision and recall.



0 - 372079
1 - 372079

Figure 3: After Undersampling

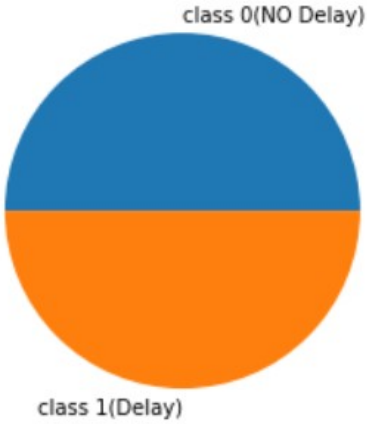
Classifier	Precision		Recall		F1-score		Accuracy
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	
Logistic regression	0.60	0.59	0.58	0.61	0.59	0.60	0.60
DecisionTree	0.59	0.59	0.59	0.59	0.59	0.59	0.59
XGBOOST	0.66	0.66	0.66	0.66	0.66	0.66	0.66
ExtraTrees	0.62	0.63	0.65	0.61	0.64	0.62	0.63
RandomForest	0.64	0.65	0.66	0.62	0.65	0.63	0.64

Table 4: Classification with undersampling

3.4 OVERSAMPLING

Oversampling is a method for dealing with class imbalance, which occurs when one class in the dataset has a considerably low number of data points compared to the other classes. In order to fix the imbalance, It increases the number of instances in the minority class by replicating them.

SMOTE (Synthetic Minority Over-sampling Technique) is a popular oversampling method. Synthetic instances of the minority class are created. As a result, Random Forest Classifier model was found to be the best performing model and SMOTE gave better results when compared with undersampling (from table 5).



0 - 1479357
1 - 1479357

Figure 4: After Oversampling

4 REGRESSION

Regression finds the statistical relationship between dependent and independent variables, regression models are used to predict any continuous values using the statistical relationship between dependent and independent variables.

Classifier	Precision		Recall		F1-score		Accuracy
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	
Logistic regression	0.60	0.60	0.60	0.61	0.60	0.61	0.60
DecisionTree	0.80	0.80	0.79	0.80	0.80	0.80	0.80
XGBOOST	0.74	0.81	0.83	0.71	0.79	0.76	0.77
ExtraTrees	0.85	0.87	0.88	0.84	0.86	0.86	0.86
RandomForest	0.84	0.88	0.89	0.83	0.86	0.85	0.86

Table 5: classification with oversampling

4.1 ANALYSIS METRICS

Mean Absolute Error (MAE):
The mean absolute error is a metric used to evaluate the accuracy of a model's predictions. It's calculated by summing up the absolute differences between the predicted and actual values.

Mean Squared Error (MSE):
Mean Squared error is a metric which calculate average of the squared difference between the actual value and prediction value.

R-squared (R2):
R2 is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables.

	Linear Regression	XGBOOST	ExtraTree	RandomForest
MAE	36.94	11.73	12.20	12.66
MSE	4375.45	395.63	379.35	415.46
R2	0.1583	0.92	0.927	0.92

Table 6: Regression results

From the Table 6 we can infer that ExtraTrees Regressor is the best model. XGBOOST, RandomForest and ExtraTrees had similar R2 scores but Extra-trees Regressor had the lowest mean squared error among the three and finally, Linear Regression was the worst performing model.

5 PIPELINE

The pipeline consists of a classifier and a regressor where the classifier initially classifies the datapoints in the dataset into delayed and not delayed classes. Then the regressor is used to predict the delay in minutes if the datapoint is classified as *delay* by the classifier.
From Table 5 and Table 6, the best classification and regression models were found to be RandomForest classifer and ExtraTree regressor respectively. Using these two a pipeline was built.

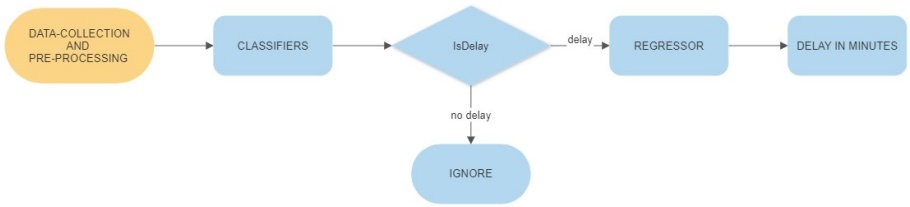


Figure 5: Pipeline flowchart

6 REGRESSION ANALYSIS

The predicted delays are divided into 5 different intervals, then the best regressor model is used to make delay prediction specifically for that interval. This gives an idea where the model performs better and also to analyse in which intervals the model would perform poorly and would have lower accuracy.

Figure 6 (Boxplot) shows that most of the datapoints are in the range 15 to 100. We can infer from the plot that the datapoints are scarce above 2000. Since the datapoints are less in this interval, the accuracy is very high.

From Table 7, we can infer that the Interval 100-200 had the worst R2 score and 200-500 had the best R2 score, hence regression model performs well at

interval 200-500 and performs poorly at interval 100-200. 15-100 interval had the most number of datapoints and the R2 value of the model at the interval is 0.89 which is very good for the number of datapoints in the interval.

The interval 15-100 and 100-200 have most datapoints but when compared with other intervals they have lesser R2 values. The intervals 500-1000 and 1000-2000 have similar R2 values but 500-1000 has more number of datapoints when compaired to 1000-2000.

Range	MAE	MSE	R2	Count
15-100	2.183	57.15	0.898	270782
100-200	3.562	162.62	0.816	43203
200-500	4.173	190.14	0.958	12713
500-1000	6.429	107.52	0.9540	1041
1000-2000	7.842	591.10	0.9666	146
2000+	0.0	0.0	1.0	2

Table 7: Regression results

7 CONCLUSION

Flight and Weather data set were preprocessed and features with high correlation were removed from the datasets. A new consolidated dataset was made by merging the weather and flight dataset using common columns which were present in both datasets.

The class imbalance in the consolidated dataset is then removed by using SMOTE an oversamling technique which gave better results when compared to random undersampling.

Further, A pipeline was built using **Random Forest Classifier** which was the better performing classifier with a **F1 score** of **0.86**. Then departure delay for the flights which are classified as delay are predicted using regression models. Extra trees regressor was the better performing regressor with **R2 value** of **0.92**. Furthermore, the ExtraTrees Regressor was analysed at every interval

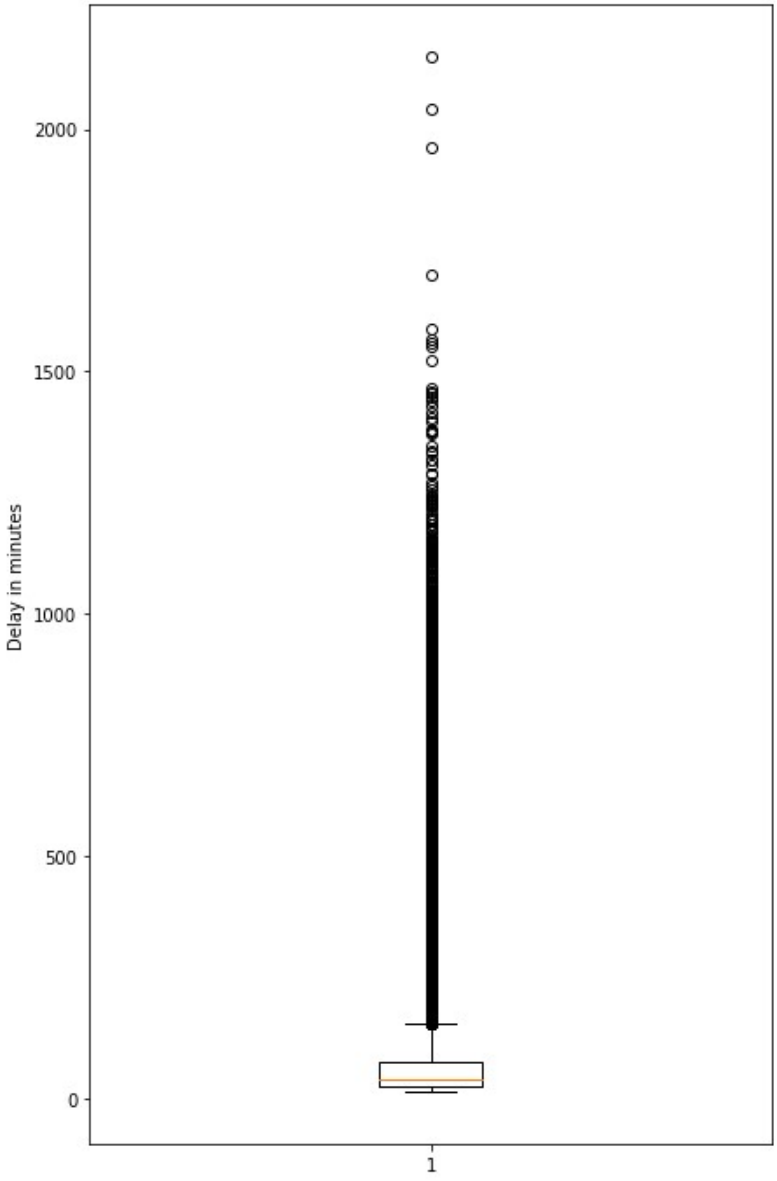


Figure 6: Delay distribution

in regression analysis which helps to evaluate the performance of the model in each delay intervals.