Flight Delay Prediction

Vishal N

Abstract: Develop a two stage predictive machine learning model that predicts the delay of flights from the given datasets. The model is built to predict the arrival delay to a good measure. Classification is done to classify flights as 'late arrival' or otherwise. Data of the flights that were delayed is passed to regression algorithms to predict the arrival delay of flights in minutes. The best classifiers and regressors are combined after a comparative study.

1 Introduction

In the recent past, the world has seen a rather drastic inflation in the airlines industry owing to it's speed, comfort and sometimes, even luxury. About 933.1 million and 965.0 million passengers were reported to have used the services provided by the US airlines industry in the years 2016 and 2017, respectively. It is evident that there is a 3.4% relative increase in the passenger count from the year 2016 to 2017.

This study however suggests that these numbers are expected to increase over the following years. Moreover, the economic impact of flight delays for domestic flights in the US is estimated to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy. The losses incurred by the US airlines are over nine figures. Here, a simple machine learning machine learning model can be used in order to optimize flight operations and minimize the financial loss due to delays, thus raising GDP of the aviation industry.

In the US, a flight is considered delayed if it is 15 minutes late. The entries for whether or not a flight is delayed was included in the data set. Several classification techniques including Logistic Regression, Decision Tree Classification, Extra Trees Classification, XGBoost (eXtreme Gradient Boost) Classification were used to classify the flights as delayed or otherwise. The Delay minutes (if any) for every flight, was also included in the data set. Using this and various Regression techniques, the delay (in minutes) was predicted for every flight delayed at arrival. The key point is to reduce the error in the classified and predicted values so as to render the best possible machine learning model.

2 Dataset

The primary goal is to combine the two data sets after cleaning the data by removing missing entries. The input to the model is the matrix of features like Origin, Destination, Flight Date, Arrival Time, Departure Time etc. Using this,

Table 1. Table of airports in the US to be considered

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

the data is pre-processed by considering only those flights which are within the airports in the United States, mentioned in the Tab. 1.

The other input to the model is the weather data set of the above 15 airports for the years 2016 and 2017. The features available in the data set are precipitation, humidity, temperature etc.

The two data sets used are Flight Data and Weather Data. The flight data set has 109 columns of which only those 17 columns given in Tab. 2 are considered. The weather data set has 25 columns of which only those 12 columns given in Tab. 3 are considered.

Table 2. Table of Flight features to be considered

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

Table 3. Table of Weather features to be considered

windspeedkmph	DewPoint	cloudcover	precipitation
pressure	windgustkmph	visibility	weathercode
temperatures	humidity	date	time

The flight data set consists of records of various stations for 12 months in the years 2016 and 2017. The records are included if and only if both the origin and destination airports fall under those 15 airports mentioned above. The weather data set consists of many **JSON** files which store weather data for each month. Every JSON file consists of hourly data which means that the weather data was recorded and stored for every hour. The JSON files which had data for the years 2016 and 2017 are extracted and restructured as a CSV file. The strategy used here to merge the two data lies in the fact that both the data sets feature **date, time and airport** columns. Using these columns, merging was done.

Fig. 1 illustrates the distribution of the arrival delay in minutes over the days of each month. The figure is a boxplot confined in a range of 0 and 50 arrival delay minutes. It can be seen that the delays lie towards the lower end of the

range, meaning that the arrival delays were for the most part within 0 and 12 mins. From this, it can be inferred that most flights were not delayed.

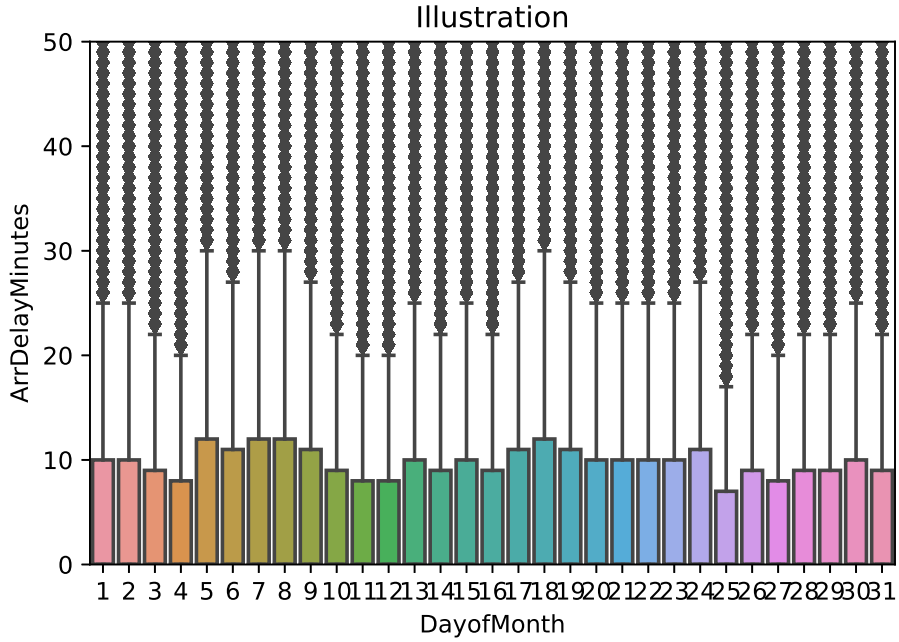

Fig. 1. Illustration of the Arrival Delay Minutes in the dataset

3 Classification

The pre-processed data is split into training and validation sets in the ratio (75 : 25). In the data, the column *'ArrDel15'* is kept as the label. It has values '0.0' and '1.0' which refer to the flight as *'not a delayed flight'* and *'delayed flight'* respectively. Keeping the column *'ArrDel15'* as the ground truth value, multiple classification algorithms are modeled to train based on the ground truth values and classify the flights as delayed or otherwise. As for the training of the models, the features in the Tab. 2 are used. Using these features, the model is able to classify the flights as delayed.

Different classifiers are used and tested namely

- Logistic Regression
- Decision Tree Classification
- Extra Trees Classification
- XGBoost Classification

To view the results of all the classifiers, refer to the figure. (see Tab. 4).

Table 4. Arrival Delay Classification

Algorithm	**Precision**		**Recall**		**F1-Score**		**Accuracy**
	0	1	0	1	0	1	
Logistic Regression	0.94	0.79	0.95	0.76	0.94	0.77	0.91
Decision Tree Classification	0.93	0.79	0.95	0.74	0.94	0.76	0.90
Extra Trees Classification	0.93	0.79	0.95	0.74	0.94	0.76	0.90
XGBoost Classification	0.94	0.79	0.95	0.75	0.94	0.77	0.91

4 Data Imbalance

Imbalanced data typically refers to any classification problem where the number of observations per class is not equally distributed; often data scientists encounter a large amount of data/observations for one class (referred to as the majority class), and much fewer observations for one or more other classes (referred to as the minority classes).

Data imbalance usually reflects an unequal distribution of classes within a dataset. The major reason the classifiers did not perform well owes to the data imbalance present in the dataset provided. Fig. 2 illustrates a pie chart to show the ratio of the number of flights that were classified as not delayed to that of delayed.

There are several solutions to eradicate or rather escape from data imbalance. A tool called Sampling helps in removing data points of the majority class (Under-Sampling ex: NearMiss, TomekLinks) or adding general points of the minority class (Over-Sampling ex: SMOTE, ADASYN) and reduces the bias and trains the model more effectively thereby, increasing the accuracy and other metrics. Here SMOTE is used to oversample the minority class and makes an equal distribution between the two classes, as seen in Fig. 3

Bias to the number of flights that were classified as not delayed to that of delayed

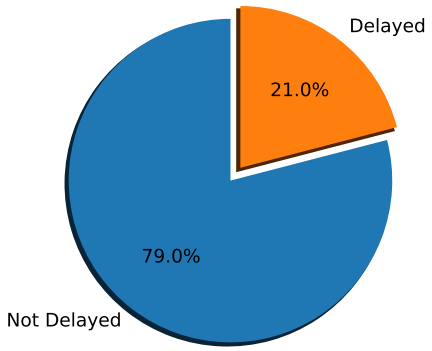

Fig. 2. Arrival Delay Classification before SMOTE

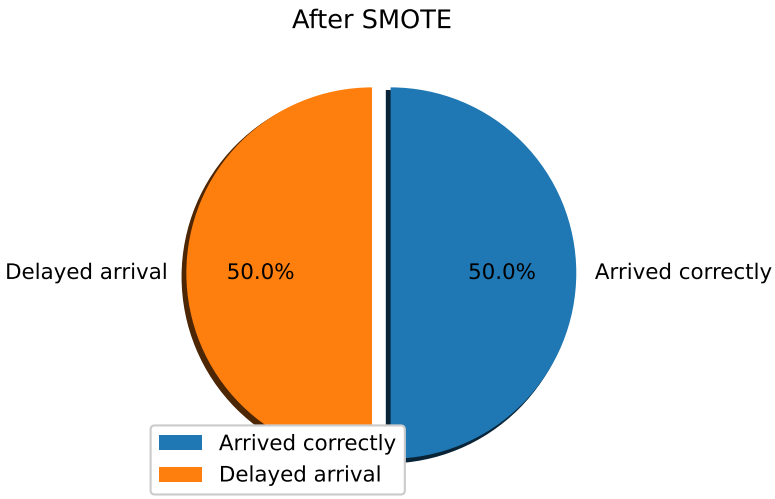

Fig. 3. Arrival Delay Classification after SMOTE

5 Regression

The pre-processed data is split into training and validation sets in the ratio (75 : 25). *'ArrDelayMinutes'* is the feature containing the arrival delay in minutes. Keeping the column *'ArrDelayMinutes'* as the ground truth value, multiple regression algorithms are trained to predict the delay in minutes of the flights based on the features given in Tab. 3 .

Different regressors are trained and tested namely

- Linear Regression
- Decision Tree Regression
- Extra Trees Regression
- XGBoost Regression

To view the results of all the regressors, refer to Tab. 5).

Table 5. Arrival Delay Minutes Prediction

Algorithm	R^2	**MAE (in mins)**	**RMSE (in mins)**
Linear Regression	0.938145	12.373360	17.935895
Decision Tree Regression	0.940375	12.120626	17.609614
Extra Trees Regression	0.902246	16.630704	22.547872
XGBoost Regression	0.941817	11.952818	17.395398

6 Regression Testing

It can be inferred from from Fig. 4 that the dataset has less delays (in mins). Further analysis was done to visualize how the distribution of the delay(in mins) values affects the model. For this, the Linear Regression model was considered. The model was trained on the training set while it was tested within each interval of the validation set. Tab. 6 shows the tabulation and metrics of the model trained under each interval. The metrics varied because the regressor was tested on different intervals and hence it accurately predicted only those intervals which the regressor is more familiar with. As expected, the trend shown in the metrics of the interval happen to decrease in accuracy because of the distribution of values.

Table 6. Regression Analysis

Interval (in mins)	**MAE (in mins)**	**RMSE (in mins)**
0 - 100	11.0677	14.785
100 - 200	18.3251	27.5173
200 - 1000	20.1862	32.3781
1000 - 2000	26.9474	31.9939

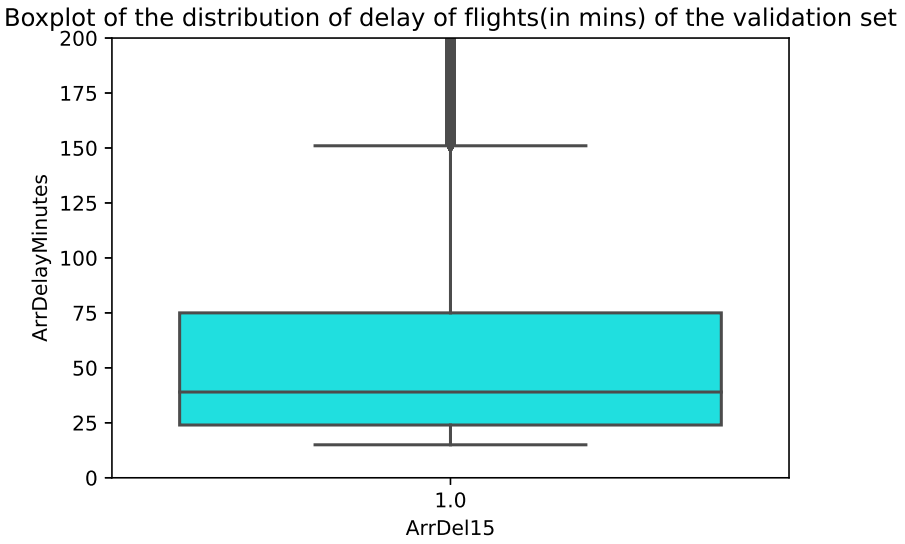

Fig. 4. Arrival Delay Minutes Prediction

7 Pipelining

The merged dataset comprises of both flight and weather data. The trained classifier and regressor which were dumped as a pickle format are loaded. The flights in the validation set that are classified by the classifier as delayed, are passed to the regressor as a dataframe. The regressor, thus predicts the output of arrival delay in minutes. A clear view of the pipeline model is given as a flowchart in Tab. 7.

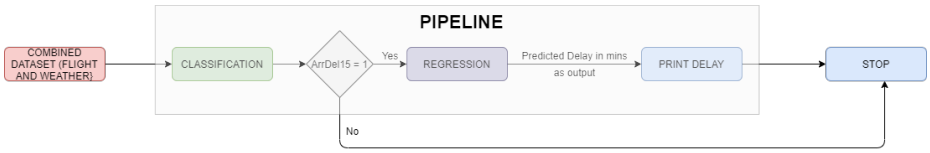

Fig. 5. Pipeline model flowchart

XGBoost Classifier and XGBoost Regressor have been used for modeling the pipeline algorithm. The reason for choosing this classifier and regressor is because of it's metrics. The two models are selected due to their high f1- scores and low Mean Absolute error. Since XGBoost showcased the best results in comparison with the other models, XGBoost was considered. (Refer Tab. 4 and Tab. 5)

Table 7. Pipeline Results

R^2	**MAE (in mins)**	**RMSE (in mins)**
0.9443	13.8259	18.9208

8 Conclusion

The flight and weather data were preprocessed and merged into a single dataset. This was used to train the classification and regression models. From the results of the classifier, it was discerned that there was an intrinsic inclination in the data towards non delayed flights and thus the classification models performed poorly in determining the flights that were actually delayed. Thus came a necessity for sampling since the dataset was imbalanced. It was possible to free the dataset from bias by incorporating oversampling using SMOTE. For the pipeline model, XGBoost Classifier and XGBoost Regressor was used for obvious reasons. (Refer Fig. 2 and Fig. 4). In the pipeline model, the classifier and the regressor were trained beforehand and the classifier classified outputs using the validation set and among those flights classified as delayed by the classifier, the regressor was tested on those flights and predicted the delay. Here it can be seen how the regressor compensates for the errors made by the classifier.

This two staged predictive model is a machine-learning algorithm built for the purpose of rendering accurate results. Aviation industries continually tend to lose billions of dollars per year due to delays in both passenger and shipment airlines. Accurate delay predictions will help in minimizing the delays and optimizing operations in the industry, thereby increasing the gross domestic product of the country.