Flight Delay Prediction

Sachin Krishan T

March 2020

Abstract

Flight delayed is when an airplane fails to adhere to schedule. Flight delays negatively affect airlines, airports and passengers. Multiple factors contribute to delay of flights. This project is particularly interested in the weather's influence on flight delay. The project aim to predict the delays of commercial airborne commutation. This project assesses multiple Machine learning algorithms to select the one that fits our necessity the best. The project will be performing Classification and Regression.

1 Introduction

Flight delays are an economic burden for airlines, airports and passengers. Given the uncertainty of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip costs, to ensure their arrival on time. On the other hand, airlines suffer penalties, fines and additional operation costs, such as crew and air crafts retention in airports. Furthermore, from the sustainability point of view, delays may also cause environmental damage by increasing fuel consumption and gas emissions. Moreover, flight delays result in an unpleasant memory for the passengers this affects the airlines as passengers desire reliability in flying.
This project utilises with Real time Weather data collected during 2016 and 2017 at 15 Airstations in USA. The project aims to build a model which will use incurred departure delay and forcasted weather data to predict the arrival delay of a flight. This is incrementally done over three modules. The first module is data pre-processing. The second module uses this processed data to classify if a flight will be delayed and pipelines it to the third module. The third module takes details of the delayed flights and predicts the delay in minutes.

2 Data Pre-processing

The flight data set was acquired in the .csv format which has information of flight schedules and actual flight travel timings and date in the United States of America. the flight data set holds the information for the years 2016 2017. Categorical values like the airport codes were encoded into numerical values using label encoding. The weather data set is a JSON file that contain weather data that is recorded periodically for every one hour for the years 2013-2017 for 15 airports.

The listed features were used from the Flight data set

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

Table 1: Features used from Flight data.

The listed features were used from the Weather data set

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

Table 2: Features used from Weather data.

The flight data set is pre-processed to drop irrelevant attributes. The weather dataset is reconstructed. The time attribute in the flight dataset is rounded off to the nearest hour and dropped all other flight details whose airports are not in the set of 15 airports given in the weather dataset. Both the data sets are merged based on the time, day, month, year and airport code attributes. After merging the dimension of the whole data set is (1780890, 46).

3 Classification

This module deal with the prediction of the possibility of a flight delay. The processed data is fed into machine learning models here to classify weather a flight will be delayed or not. The feature *ArrDel15* in the data set determines weather a flight is delayed at least 15 minutes. The project will use *ArrDel15* as the y-axis feature. The data set was split with a ratio of 80:20 to train and test the models. The algorithms used in this module are Logistic Regression, Decision Tree Classifier, Extra Trees Classifier, Gradient Boosting Classifier. Flights that are classified to be delayed by the model with the best performance criteria are pipe-lined into the next module.

3.1 Metrics Used

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

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

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

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

where TP - True Positives, TN - True Negatives, FP - False Positive, FN - False Negative.

F1-score is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases than the Accuracy Metric.

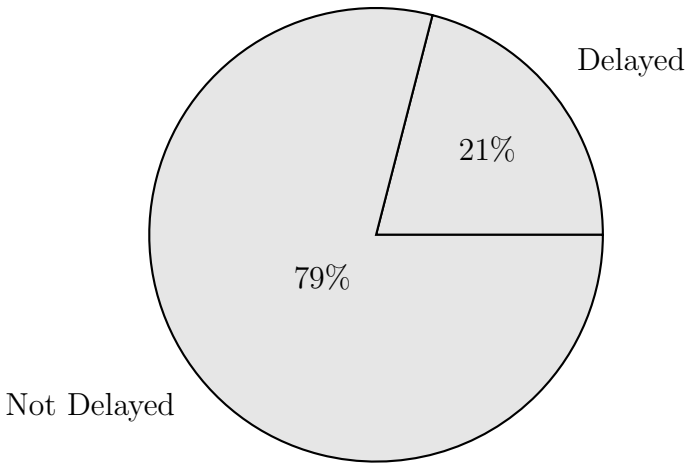
3.2 Performance Analysis

Table 3: Performance of Classifiers

Algorithms	precision		recall		f1-score		accuracy
	0	1	0	1	0	1	
Logistic Regression	0.92	0.89	0.98	0.68	0.95	0.77	0.92
Decision Tree Classifier	0.92	0.69	0.91	0.71	0.92	0.70	0.87
Extra Trees Classifier	0.94	0.79	0.95	0.75	0.94	0.77	0.91
Gradient Boosting Classifier	0.92	0.89	0.98	0.69	0.95	0.78	0.91

3.3 Data Set Imbalance

On inspection of the performance of the two labels we can see that one label performs better than the other. To further investigate this we plot the label distribution. From this we can understand that there's a skew in the in the data set. To cure the dataset of this imbalance in hopes of bettering the models' performace we sample the data. Now we are left with two choices to under-sample or over-sample the data set. We experiment with both and check which gives us better performance. For under-Sampling we use the Tomek links. For over-sampling we use Synthetic Minority Over-sampling Technique (SMOTE).



3.4 Performance analysis after sampling

Table 4: Performance of Classifiers after Sampling

Sampling Algorithm	Regression Algorithm	precision		recall		f1-score		accuracy
		0	1	0	1	0	1	
SMOTE	Logistic Regression	0.94	0.74	0.93	0.78	0.93	0.76	0.90
	Decision Tree Classifier	0.92	0.68	0.91	0.71	0.92	0.69	0.87
	Extra Trees Classifier	0.94	0.77	0.94	0.77	0.94	0.77	0.90
	Gradient Boosting Classifier	0.93	0.83	0.96	0.73	0.95	0.78	0.92
Tomek Links	Logistic Regression	0.92	0.88	0.97	0.69	0.95	0.77	0.92
	Decision Tree Classifier	0.93	0.67	0.91	0.72	0.92	0.70	0.87
	Extra Trees Classifier	0.94	0.78	0.94	0.76	0.94	0.77	0.91
	Gradient Boosting Classifier	0.92	0.88	0.98	0.69	0.95	0.78	0.92

We use F1-score as the deciding performance metric because accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes as in the above case. Of the trained models Gradient Boosting Classifier with sampled using SMOTE has the best performance. Hence we will be pipe-lining the output from this model.

4 Regression

In this Module we use regression module which predicts continuous value in the previous model we predicted categorical value. The algorithms explored in this section are Linear regression, Decision Tree Regressor, Extra Tree Regression and Gradient Boosting regression. Like the previous model we have the train and test split ratio as 80:20. Only delayed flight records are used in training regression model. The performance of the models are given below

4.1 Metrics used

$$\text{Mean absolute error } \mathbf{MAE} = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t|$$

$$\text{Mean square error } \mathbf{MSE} = \frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2$$

$$\text{Root mean squared error } \mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$$

4.2 Performance observation

Table 5: Models trained using delayed flights

Algorithms	MAE	MSE	RMSE
Linear Regression	14.516	398.233	19.955
Decision Tree Regressor	16.469	573.034	23.938
Gradient Boosting Regressor	11.827	294.816	17.170
Extra Trees Regressor	11.673	277.735	16.665

5 Regression testing

To assess the performace of our best regressor model, Extra Trees Regressor, in practical scenarios the project plot and visualises the distribution of the number of delayed flights based on their delay duration. The project then splits the test dataset into 4 classes of delay less than 100 minutes, 100 - 500 minutes, 500 - 1000 minutes and greater than 1000 minutes. The regressor model's performance in the specified delay intervals are assessed using the separated test dataset. The performance of the model in these intervals is given in the table below

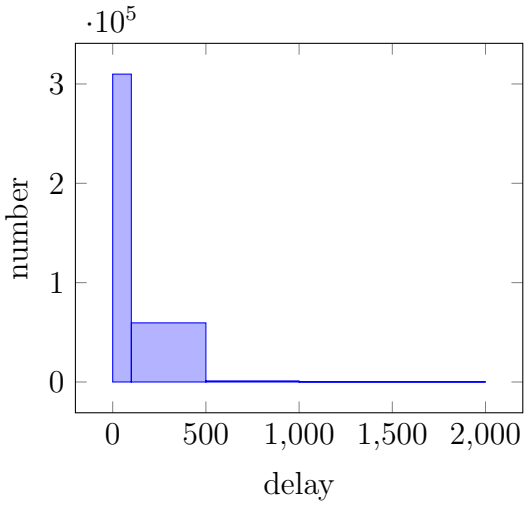
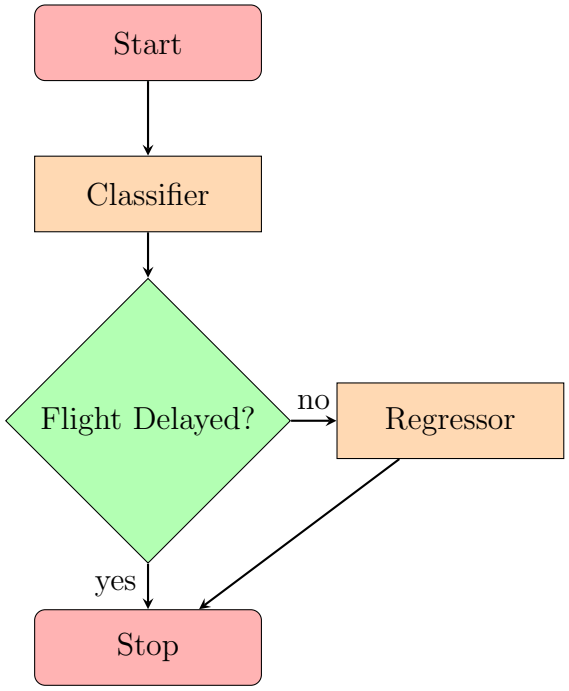


Table 6: regressor performance in different delay intervals

Arrival Delay Range in Minutes (x)	MAE	MSE	RMSE
x≤ 100	2.567	38.323	6.190
100<x≤ 500	5.103	177.038	13.305
500<x≤ 1000	4.891	187.912	13.708
1000<x	8.415	942.225	30.695

6 Pipelining

The project uses the best classifier which is pre-trained to predict if the flight is delayed or not. And then pass the flights classified as delayed to the best regressor which is also pre-trained to find the minutes by which the flight is delayed. The flowchart of this architecture is given below. The performance of the regressor in this architecture is **MAE(4.699), MSE(118.045), RMSE(10.864)**.



7 Conclusion

The flight and weather data sets were formatted, prepossessed and merged into one data frame. Examining the data set with visualization we observed that the data set was imbalanced. We re-sampled the data set to reduce the imbalance. however sampling did not help improve the performance our models. Out of the classifiers explored Gradient Boosting Classifier had the best performance score with a f1-score of 0.78 for the 1 class. Among the explored regressors Extra Tree Regressor performed best with MAE(11.673), MSE(277.735), RMSE(16.665) scores. In regression testing section we see that the model performs significantly better for delays less than 100 minutes. looking at the distribution graph. we can infer that this is because there is a higher density of flights in this interval. Hence, training the model better in this interval. we can also observe that the performance of the model in an interval correlate with the density of the flights in the corresponding interval. The regressor model performs better when pipelined with flights from the classifier.