Flight Delay Prediction

Sujay Sathya

July 2020

Abstract

A flight is said to be delayed when a particular flight does not arrive or depart from an airport in it's scheduled time. We have used a two-step process using two common machine learning tasks such as regression and classification. We try and classify whether a given flight will be delayed or not and then use regression techniques to find out the precise delay.

1 Introduction

There is an increasing number of flights getting delayed every day due to various reasons ranging from weather to technical difficulties. This includes both passenger and cargo flights and this delay is costing a lot of money to companies and passengers. The purpose of this project is to analyse these different conditions and to make a prediction whether a given flight might be delayed or not and how long it has been delayed by. We have predicted arrival delay considering the weather parameters at the departure airport

2 Dataset

The dataset used in this project contains information about various Airports in the USA between the years 2016 and 2017 and has relevant and important Flight details and weather details about when each flight took off and landed. 15 airports in the US has been selected for this project and only the flights flying from and to these 15 airports will be taken into consideration. All the models used in this paper were trained on eighteen hundred thousand data points. The dataset also had a significantly high number of weather parameters to choose from. For this project we have chosen 15 weather parameters.Table 1 represents the 15 airports, Table 2 represents the 15 weather parameters and Table 3 represents the 15 flight parameters.

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

Table 1: The 15 airports

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

Table 2: Eeather parameters

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

Table 3: Flight parameters

3 Classification

We have used various models to classify a flight as either delayed or not delayed. The model tries to predict the value of the column ArrDel15 using the pre-existing values of all the other relevant flight and weather data, if ArrDel15 is predicted as 1, then there is a delay, if it is predicted as 0, then there is no delay.

3.1 Models used

The following models were tested and their respective performances were noted:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGB Classifier
- SVM Classifier
- Artificial Neural Network

3.2 Performance metrics

$Precision = \frac{T.p}{T.p+F.p}$

$Recall = \frac{T.p}{T.p+F.n}$

$Accuracy = \frac{T.p+T.n}{T.p+F.p+F.n+T.n}$

$F1Score = \frac{2*(Recall*Precision)}{(Recall+Precision)}$

- True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this flight is delayed and predicted class also predicts the flight as delayed.
- True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class value indicates that this flight is not delayed and predicted class also predicts that this flight is not delayed.
- False Positives (FP) – When the actual class is no and the predicted class is yes. E.g. if actual class value indicates that this flight is not delayed but predicted class predicts that this flight is delayed.

- False Negatives (FN) – When the actual class is yes but the predicted class in no. E.g. if actual class value indicates that this flight is delayed and predicted class predicts that this flight is not delayed.

3.3 Results

Model	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	0.92	0.71	0.92	0.70	0.87
Random Forest Classifier	0.93	0.88	0.98	0.70	0.95	0.78	0.92
XGB Classifier	0.92	0.90	0.98	0.69	0.95	0.78	0.92
SVM Classifier	0.94	0.78	0.94	0.75	0.94	0.77	0.91
Artificial Neural Network	0.94	0.74	0.93	0.78	0.93	0.76	0.90

Table 4: Classifier Results

4 Class Imbalance

There is a huge difference between the number of data points of class 0 and class 1 (Figure .1), the model will have a slight bias towards class 0 and this will affect the output of the classifier, to solve this issue , we have used the random over sampling which tries to match the number of data points of both the classes by oversampling the minority class (Figure .2).This is done by going through all the data points of the minority class and using this data to generate similar data points which would belong to the same class. Under sampling is not used because doing so would decrease the number of data points of the majority class thereby reducing the overall number of data points which will later reflect on the accuracy of the model.

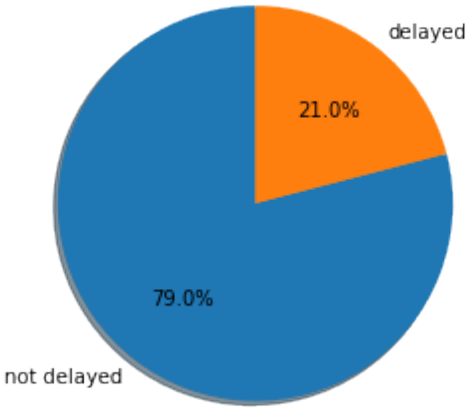

Figure 1: Pie chart representing the imbalance in class

As we can see in the pie chart there is a clear class imbalance. We have tried to remove this by oversampling.

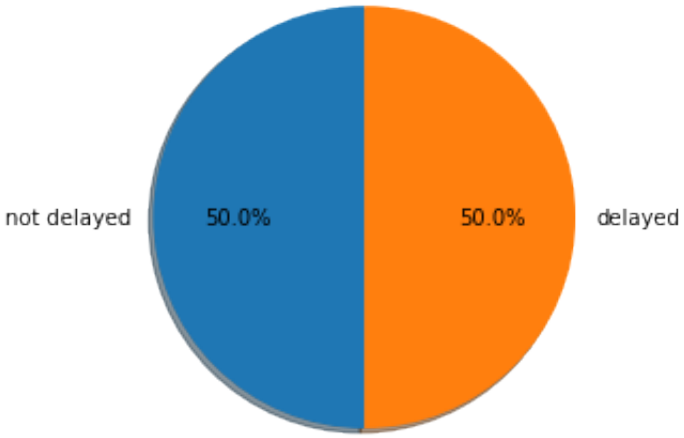

Figure 2: Pie chart representing the balance in class

As we can see, the class imbalance shown earlier has now ceased to exist because of oversampling the data,under sampling the data would yield the same distribution but the number of data points in each class will be significantly lower.

Model	Precision		Recall		F1		Accuracy
	0	1	0	1	0	1	
Logistic Regression	0.94	0.73	0.92	0.78	0.93	0.76	0.89
Decision Tree Classifier	0.92	0.69	0.92	0.70	0.92	0.70	0.87
Random Forest Classifier	0.93	0.83	0.96	0.74	0.95	0.78	0.91
XGB Classifier	0.95	0.74	0.92	0.80	0.93	0.77	0.90
SVM Classifier	0.94	0.74	0.94	0.95	0.94	0.77	0.91
Artificial Neural Network	0.95	0.65	0.88	0.82	0.91	0.72	0.87

Table 5: Classification results after oversampling

4.1 Classification Verdict

The model chosen to be the classifier is the XGB Classifier because of a high accuracy of 0.91 and a good recall of 0.80. Recall has been chosen as an important factor for picking the model because we need to classify as many delayed flights as "delayed" properly. Boosting methods have worked well here because of the extremely large number of data points. They help in avoiding over fitting while also allowing model to achieve a lower bias.

5 Regression

Regression is the second stage of the two-stage model. The Arrival Delay in minutes is predicted by the Regressor. Data points used to train the Regressor needs to have 'ArrDelayMinutes'greater than 0, so basically only the delayed flights are used to train the regressor.

5.1 Models Used

The following models were tested and their respective performances were noted:

- Linear Regression
- Random Forest Regression
- XGB Regression
- Artificial Neural Network
- Extra Tress Regressor

5.2 Performance Metrics

The following performance metrics were used to evaluate the regressor $M.S.E = (\frac{1}{n})\sum_{i=1}^{n}(y_i - y_i')^2$

$$R.M.S.E = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - y_i')^2}$$

$$M.A.E = (\frac{1}{n}) \sum_{i=1}^{n} |(y_i - y_i')|$$

$$R^2 Score = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{\sum_{i=1}^{n} (y_i - U)^2}$$

5.3 Model Performance

Model	M.S.E	R.M.S.E	M.A.E	R Squared
Linear Regression	311.87	17.66	12.22	0.93
XGB Regressor	286.499	16.92	11.67	0.94
Random Forest Regressor	294.160	17.15	12.07	0.94
Artificial Neural Network	289.008	17.00	11.77	0.94
Extra Trees Regressor	587.89	24.24	16.57	0.88

Table 6: Regression results

5.4 Regression Analysis

The arrival delay ranged from 0 to 2142 minutes. The dataset was split into ranges of arrival delay minutes and performance of Linear Regression model is studied in these ranges.

Range	Count	R.M.S.E	M.A.E
0-100	274294	14.59	10.81
100-200	47481	27.58	17.61
200-500	14085	42.58	35.23
500-1000	1169	42.89	35.10
1000-2000	168	69.51	65.58

Table 7: Regression analysis

Most data points have ArrDelayMinutes ranging between 1 - 100 minutes and hence, MAE (10.81) and R.M.S.E (14.59) are least in this range. As the range increases, the number of data points decreases and as result, the values of RMSE and MAE increase. This reduction in data points takes place drastically as the range increases so we can safely summarise that the R.M.S.E and M.A.E values are good in ranges where the mojority of datapoints are present. The model chosen to be the Regressor is the XGB Regressor because of having the lowest R.M.S.E and M.A.E overall.

6 Pipeline

We have chosen the XGB classifier and XGB regressor in step one and two respectively and trained them on the data accordingly. Now we send the data through the pipeline.

6.1 Working

Initially all the data points are sent to the classifier which classifies them as either "not delayed" or "delayed". All the data points which have been classified as "delayed" are sent to the regressor which predicts the delay in minutes. The chosen regressor has as r.m.s.e score of 16.92. This along with the error in the classifier will give us a slightly higher r.m.s.e score for the pipelined model.

6.2 Results

Metric	Value
R.M.S.E	18.055
M.A.E	12.76
R^2	0.94

Table 8: Pipeline results

7 Conclusion

The weather and flight data were combined into a single data set for training the models.Due to data-imbalance, Sampling techniques are employed to compensate for the skewed data distribution. Sampling does not however solve the situation and is ineffective in helping to predict the flight delay for the given data-set with a higher accuracy.The classifier chosen was XGB classifier having the highest F1 Score (0.78) accuracy (0.90) and recall(0.80).The Regression Model chosen was XGB regressor having the highest R^2 Score (0.94) and the least R.M.S.E (16.92). The pipelined model was designed using the aforementioned classifier and regression Models and performed with a reasonable accuracy.