FLIGHT DELAY PREDICTION

S Saikrishnan

June 10 2020

Abstract

A **flight delay** is when an airline flight takes off and/or lands later than its scheduled time. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time. Flight Delay prediction is a project developed in python, to predict delay in flights (if any) and hence the amount of delay using a two stage predictive model, based on the dataset of flights in USA and the weather data pertaining to 15 airports, both from 2016 through 2017.

1 Introduction

Delays in flights can be caused due to factors such as traffic volume, aircraft type, aircraft maintenance, airline operations, weather conditions, customer service issues, and late aircraft or crew arrival, to name a few. The results show that weather has contributed to **69 %** of the delays. This project focusses on various aspects of the weather conditions in particular to make predictions. The implementation of this model is broken down into 3 stages namely-

- Data Preprocessing
- Classification
- Regression

2 Data-Preprocessing

Initially there are a set of airports whose airport codes are given in Table 1. Each of these airports have data pertaining to them in files each of which is extracted to make a data frame. However, this data frame would contain many attributes all of which would not be required for the prediction. Hence only the features shown in Table 2 would be extracted, to make the final flight data.

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

Table 1: **List of airport codes**

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

Table 2: **List of flight features**

The weather data is also collected in a similar manner and made into a data frame which contains the aspects as listed in Table 3.

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

Table 3: **List of weather features**

Flights for which weather data is available is found and appended to the data frame corresponding to the flight data, based on **Departure Airport, Departure Time and Date**. This data frame is now considered as the final processed data that will be used for further modelling and predictions. It is ensured that there are no duplicates in this data frame to avoid redundancy, and is found to have a total of 28 features and 18,34,170 data points. 80% of these are used for training the model and the remaining 20% is used as the test data to test the accuracy of the model.

3 Classification

The first step of the two-stage predictive model is the Classifier which classifies the flights into 2 classes namely- Delayed and Not Delayed. These are represented as 1 and 0 respectively in the data frame in the column titled *'ArrDel15'*, the target variable.

3.1 Metrics Used

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

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

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

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

3.1.1 Terms used

- **TP** = True Positive,delayed flight predicted correctly
- **TN** = True Negative,not delayed flight predicted correctly
- **FP** = False Positive,not delayed flight predicted incorrectly
- **FN** = False Negative,delayed flight predicted incorrectly

3.2 Classifiers used

- Logistic Regression
- ExtraTrees Classifier
- DecisionTree Classifier
- Random Forest Classifier

3.3 Results

Classifier	Class	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0	0.92	0.98	0.95	0.91
	1	0.89	0.67	0.76	
Decision Tree	0	0.92	0.91	0.92	0.87
	1	0.68	0.71	0.69	
ExtraTree	0	0.93	0.96	0.94	0.91
	1	0.83	0.73	0.78	
Random Forest	0	0.92	0.98	0.95	0.92
	1	0.89	0.70	0.78	

Table 4: **Classifier scores**

The observed scores for class 1 are much less compared to those of class 0.This is because the number of 0's (i.e the flights that are not delayed) is much more than the number of 1's (i.e the flights that are delayed).

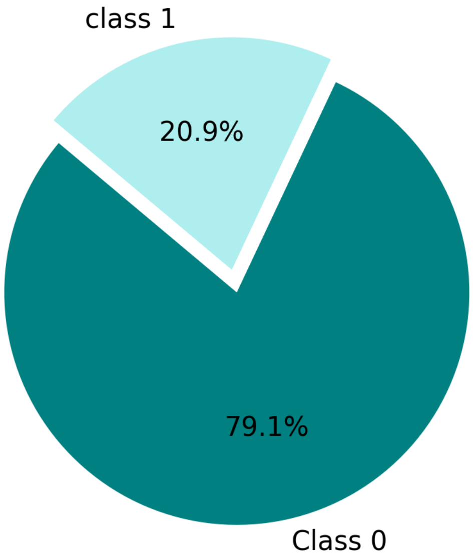

Figure 1: **Class Distribution**

This data imbalance calls for the need of **sampling**. There are 2 types of sampling namely-

Oversampling.
undersampling.
both of which are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes), in data analysis.
Oversampling methods duplicate new synthetic examples in the minority class, whereas undersampling methods merge examples in the majority class.

3.4 Oversampling

Synthetic Minority Oversampling Technique, or **SMOTE** for short, has been used to oversample the dataset.

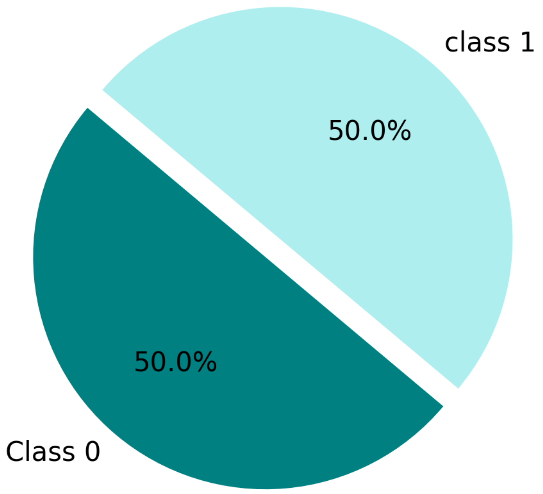

Figure 2: **Class Distribution after oversampling**

Table 5: **Scores after Oversampling**

Classifier	Class	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0	0.94	0.93	0.93	0.89
	1	0.74	0.78	0.76	
Decision Tree	0	0.92	0.91	0.91	0.87
	1	0.67	0.70	0.69	
Extra Tree	0	0.94	0.95	0.94	0.91
	1	0.80	0.75	0.77	
Random Forest	0	0.93	0.97	0.95	0.92
	1	0.85	0.72	0.78	

Table 5 shows the results for the different models used for the Classifier after Oversampling and it is observed that there is an increase in Recall for class labelled 1, in Logistic Regression. However, such a change is not observed in the other 3 as they are designed to work on imbalanced classes.

3.5 Undersampling

RandomUnderSampler has been used to undersample the dataset

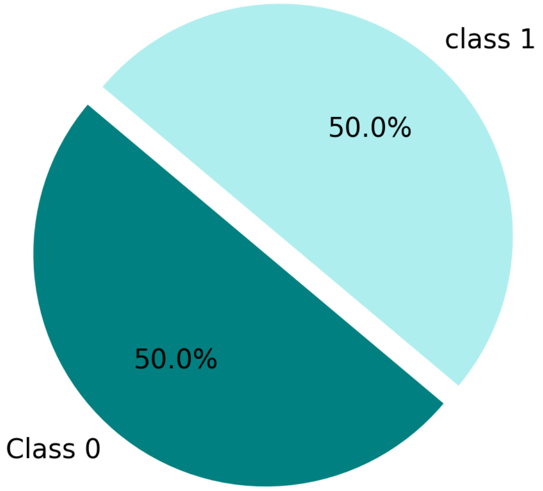

Figure 3: **Class Distribution after undersampling**

Table 6: **Scores after undersampling**

Classifier	**Class**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Logistic Regression	0	0.94	0.93	0.93	0.90
	1	0.74	0.78	0.76	
Decision Tree	0	0.94	0.79	0.86	0.79
	1	0.51	0.80	0.62	
Extra Tree	0	0.95	0.89	0.92	0.88
	1	0.67	0.82	0.74	
Random Forest	0	0.95	0.91	0.93	0.89
	1	0.71	0.81	0.76	

Table 6 shows results after Undersampling and it is observed that it doesn't yield satisfactory results.It is also inferred that the Random Forest model is the best classifier as it has the highest F1 score corresponding to class 1

4 Regression

The second of the two-stage predictive model is the **Regressor** which predicts the delay in arrival time (in minutes), for those flights which are classified under the 'Delayed' class by the Classifier. Since only the delayed flights are conidered, the remaning flights are removed to make a subset of the original dataset, which is again split in the ratio 80:20 fo train and test sets respectively of the Regressor model.

4.1 Metrics Used

Mean$(\overline{x}) = (\frac{1}{n})\sum_{i=1}^{n}(x_i)$

Root Mean Square Error(RMSE) $= \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}$

Mean absolute Error(mae) $= (\frac{1}{n})\sum_{i=1}^{n}|y_i - x_i|$

R Squared(R2) $= 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \overline{x})^2}$

4.2 Regressors Used

- Linear Regression
- ExtraTrees Regressor
- Random Forest Regressor
- Gradient Boost Regressor

4.3 Regressor Results

Regressor	RMSE	MAE	R2 Score
Linear Regression	17.50	12.19	0.938
ExtraTree	16.83	11.87	0.943
Random Forest	16.64	11.72	0.944
Gradient Boost	16.80	11.66	0.943

Table 7: **Regressor Scores**

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable in a regression model. Table 7 gives the performance of the different models used for the Regressor and it is observed that the Random Forest has the best performance with R2 value 0.944 and RMSE 16.64

5 Pipelining

The best classifier is chosen and fed to the best regressor, (here both being Random Forest), forms the basis for the functioning of the pipeline and hence completes the two-staged model.

Table 8: Scores after Pipelining

Regressor	**RMSE**	**MAE**	**R2 Score**
Random Forest	23.38	21.48	0.615

6 Conclusion

To summarise, we begin with converting the flight data and the weather data into data frames, followed by merging them into one which is used for the two-stage model predictions. The classsifier is fit into the data and it is found that it classifies the data set in a way that is biased towards the Not Delayed class. Oversampling the data points in the Delayed class using SMOTE helps overcome this biasness. The best classifier Random Forest Classifier is chosen and pipelined with the Random Forest Regressor to get accurate results in the prediction of delay in time for the flights