Flight Delay Prediction

S Saikrishnan

July 10, 2020

Abstract

The Federal Aviation Administration (FAA) considers a flight to be delayed when it takes off and/or lands 15 minutes later than its scheduled time. This project aims to predict if a flight is delayed and also find how long the flight was delayed using a two stage predictive model. This model is designed based on the data containing flights in USA and the weather data pertaining to 15 airports, both from 2016 and 2017. Based on this data, classifier predicts if the flight is delayed or not. If delayed, regressor predicts the delay. Out of various algorithms, random forest classifier (F1 score (0.78)) proved to be the best for classification and random forest regressor (R2 score (0.944)) proved to be the best for regression.

1 Introduction

Delays in flights can be caused due to various factors such as traffic volume, aircraft type, aircraft maintenance, airline operations and weather conditions. Studies show that weather conditions have contributed to 69% of delays in flights. This project leverages the influence of the weather conditions in affecting the flight schedule, to predict if the flight was delayed and also find how long the flight was delayed. The project is divided into 3 modules,

- Data Preprocessing
- Classification
- Regression

2 Data Preprocessing

Airports for which this data has been collected are represented by their corresponding Airport Codes as shown in Table 1.

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

Table 1: List of Airports

The flight data is collected for 24 months (2016-2017), from CSV files corresponding to each month and collectively made into a single dataset. However, this dataset would contain many attributes out of which only those shown in Table 2 are filtered to make the final flight data.

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

Table 2: List of flight features

Weather data is given in the form of a JSON file from 2013-2017, month wise. This also made into a dataset containing hourly data from years (2016-2017) with features represented in Table 3.

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

Table 3: List of weather features

Since, the weather data is hourly, the departure time of the flights is approximated to the nearest hours. Flight data and weather data are merged based on Departure Airport, Departure Time and Date. This dataset is now considered as the final processed data that will be used for further modeling and predictions. It is ensured that there are no duplicates in this dataset to avoid redundancy, and the final data has 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 classification, where the flights are classified as delayed or not delayed.

3.1 Metrics Used

TP (True Positive) denotes the delayed flights predicted correctly,
TN (True Negative) denotes not delayed flights predicted correctly,
FP (False Positive) denotes not delayed flights predicted incorrectly,
FN (False Negative) denotes delayed flights predicted incorrectly.

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

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

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

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

3.2 Classifier Results

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

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 flights that are not delayed are more than the number of flights that are delayed. Distribution of the dataset is depicted in Fig 1.

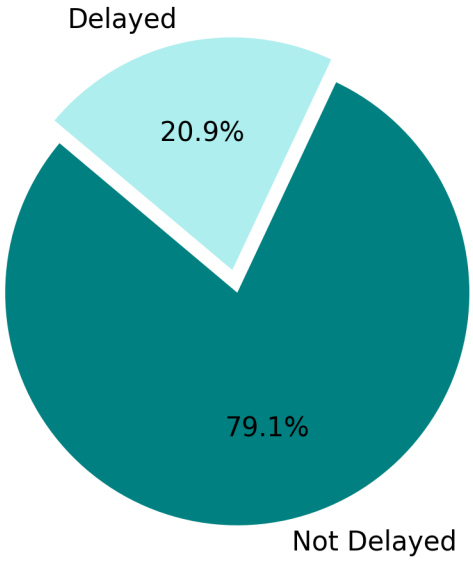

Figure 1: Pie Chart of ArrDel15 Distribution

This data imbalance calls for the need of sampling. There are 2 types of sampling namely, Oversampling and Undersampling. These are techniques used to adjust the class distribution of a dataset. Oversampling methods derive new examples from the existing datapoints in the minority class, whereas undersampling balances the dataset

by eliminating datapoints from majority class. SMOTE was used for oversampling and Random Under Sampler was used for undersampling.

Synthetic Minority Oversampling Technique (SMOTE) works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Table 5: Scores after Oversampling

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

In **Random Under-Sampling**, the majority class instances are discarded at random until a more balanced distribution is reached. It is repeated until the desired class distribution is achieved in the training dataset, such as an equal split across the classes.

Table 6: Scores after undersampling

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

3.3 Observations

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 scores for class labelled 1, in Logistic Regression.
Table 6 shows results after undersampling and it is observed that the scores for class labelled 1 have improved after sampling. It is also inferred that the oversampled 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 on arrival (in minutes), for those flights which are classified under the "Delayed" class by the Classifier. Since only the delayed flights are considered, the remaining flights are removed to make a subset of the original dataset, which is again split in the ratio 80:20 for train and test sets respectively of the regressor model.

4.1 Metrics Used

y_1, y_2, y_3, ..., y_n are values predicted by the regressor.
x_1, x_2, x_3, ..., x_n are actual values.
n is the number of observations.

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

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

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

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

4.2 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.

4.3 Regression Analysis

The dataset was split into ranges of **ArrDelayMinutes** and performance of the model is tested in the ranges given in Table 8.

Ranges	**RMSE**	**MAE**
0-100	13.69	10.25
100-200	18.35	14.43
200-500	27.06	19.34
500-1000	20.98	16.01
1000-2000	62.61	25.69

Table 8: Regression Testing

From Table 8, it is clear that the errors increase with increase in arrival delay minutes. This is because, the number of datapoints in the range (0-100) and (100-200) is much more compared to the higher ranges, shown in frequency distribution plot (Figure 2). But the relative errors decrease with increasing ranges. 25.69 MAE for range(1000 - 2000) is better than 10.25 MAE for (0-100).

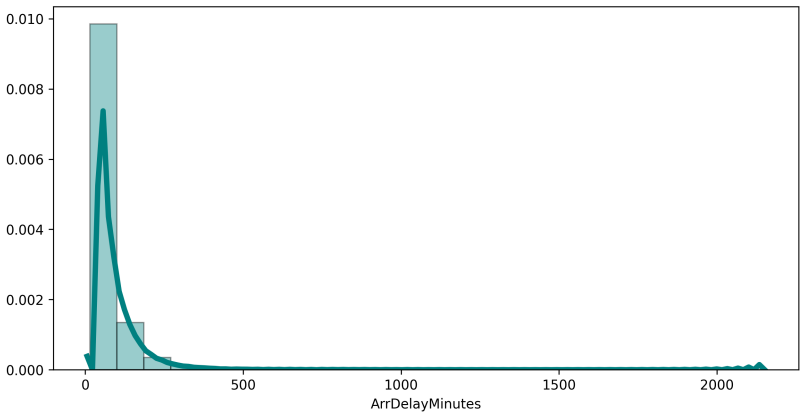

Figure 2: Frequency Distribution of ‘ArrDelayMinutes‘

5 Pipeline

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(shown in the flowchart given below) and hence completes the two-staged model.

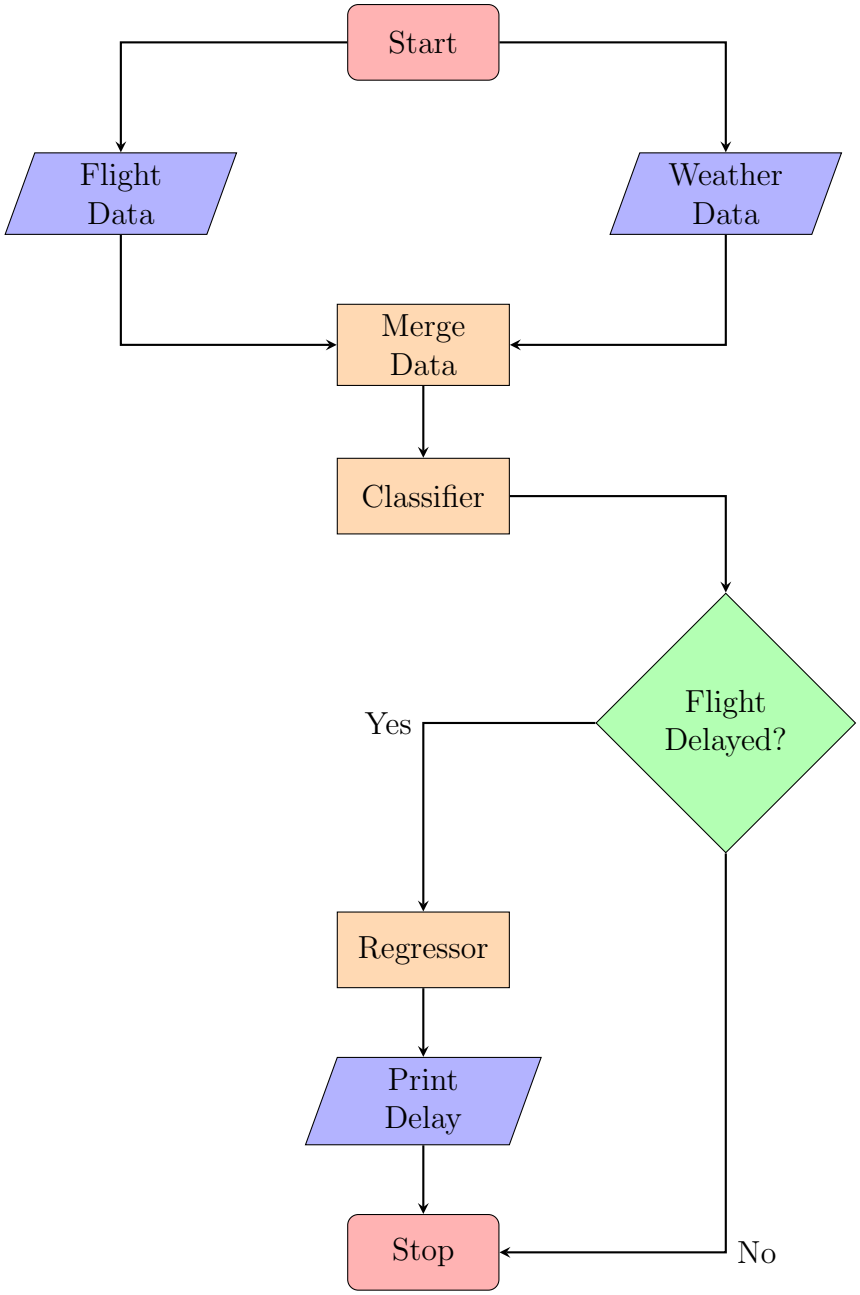

Table 9: Scores after Pipeline

Regressor	**RMSE**	**MAE**	**R2 Score**
Random Forest	13.24	8.79	0.97

6 Conclusion

The flight data and the weather data were processed and merged based on airport, date and departure time. The classifier was trained using this data and it was found that classifier performed poorly for the minority class in the dataset due to data imbalance. Sampling the data points using SMOTE and random undersampling improved the scores for the minority class. The best classifier, Random Forest Classifier having F1 score (0.78) and accuracy (0.92), was chosen and pipelined with the best regressor, Random Forest Regressor having R2 score (0.944) and RMSE (16.64) to form the pipelined architecture. The performance of the two staged pipelined model was analyzed. The pipeline model had higher scores, RMSE (13.24) and R2 score (0.97).