Abstract

A **flight delay** is when an airline flight takes off and/or lands later than its scheduled time. Flight delays affect airports, airline and passengers. Multiple factors are responsible for flight delay. This project mainly makes use the weather and flight data to predict the delay of a flight. This project is a two stage model which will be using different algorithms of classification to classify the delayed flights and that data will be used in the regressor to predict the delay of the flight in minutes .

1. Introduction

Modern societies are characterized by a high degree of mobility of goods, services and people. Supply chains are interlinked across countries and people increasingly travel for business and private reasons. As population grows and income increases, the demand for mobility is growing as well (Schafer and Victor, 1997). Overall, the increasing demand for mobility leads to a high dependence on transport systems and their services.

When transport systems are interrupted, delays emerge, introducing uncertainty regarding travellers arrival time. On the one hand, delays directly lead to extra travel time. On the other hand, in reaction to uncertain travel times, travellers may adjust their travelling schedule to ex ante account for potential delays. This represents a disutility and hence additional costs. In general, empirical research shows that individuals have a high preference to avoid delays. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time.

This projects aims to use the real time weather data of collected form 15 airport in the year 2016 and 2017. The aim is to develop a model which uses the departure delay and weather data to predict the arrival delay in minutes. The project is developed incrementally in three different stages. Data pre-processing is done in the first stage of the project. The pre-processed data is classified in the second stage and in the third stage the delayed flight details are used to predict the delay.

2. Dataset description

The two dataset that will be used for this project is flight and weather dataset. Flight dataset contains actual flight time, flight schedule and flight date from United States of America for the years 2016 and 2017. Weather dataset which was recorded periodically for every one hour for the years 2016 and 2017.

3. Data pre-processing

TABLE: 1(Features used from weather data)

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

TABLE: 2(Features used from flight data)

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

The flight dataset is pre-processed to drop the less significant features. The features consider are mentioned in table1 and table 2. The time attribute is rounded off to the nearest hour and all the flight details of the airports which are mentioned in (Table 3) are alone considered. The flight dataset and weather dataset are merged based on date, time and airport.

TABLE: 3

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

4. Regression

This module is about predicting the delay of flight in minutes. The same training data used in the classifier is used for the regressor. The regressor's explored in this module are Linear regressor, Extra tree regressor and XGB regressor. The performance of the models is given below (Table 4)

4.1 Metrics used

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Table 4

	RMSE	MAE	R-SQUARED
LINEAR REGRESSOR	32.9702	10.7310	0.3643
EXTRA TREE REGRESSOR	32.9045	10.8291	0.3675
XGB REGRESSOR	32.9045	10.8290	0.3675

The average RMSE value is 32 after regression. Errors must be penalized and reduced as much as possible. In hopes of reducing the error, classification will be performed and that data will be used in the regressor to find the delay.

5. Classification

This module aims is to predict the possibility of a flight delay. The processed data is fed into a classifier to predict whether the flight is delayed or not. The feature Arrdel15 in the dataset determines whether a flight is delayed by at least 15 minutes. Arrdel15 will be used as the y-axis feature. The dataset is split in the ratio of 80:20 to train and test the models. The algorithms used in this module are XGBoost, Extra tree classifier , Decision tree classifier and logistic regression. Flights that are classified as delayed by the model with best results are pipe-lined to the next model.

5.1 Metrics used

$$\text{Precision} = TP/TP+FP$$

$$\text{Recall} = TP/TP+FN$$

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

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

True Positive (TP): Reality: Flight Delayed.	False Positive (FP): Reality: Flight NOT- Delayed.

Prediction: “Delayed." Outcome: Correct Prediction.	Prediction: “Delayed." Outcome: In-Correct Prediction.
False Negative (FN): Reality: Flight Delayed. Prediction: “Not Delayed." Outcome: In-Correct Prediction.	True Negative (TN): Reality: Flight NOT- Delayed. Prediction: “NOT-Delayed." Outcome: Correct Prediction.

True vs. False and Positive vs. Negative

5.2 Performance analysis before sampling

		PRECISION	RECALL	F1 SCORE
XGBOOST	Flight Not-Delayed	0.92	0.98	0.95
	Flight Delayed	0.90	0.68	0.77
EXTRA TREES CLASSIFIER	Flight Not-Delayed	0.93	0.95	0.94
	Flight Delayed	0.80	0.74	0.77
DECISION TREE	Flight Not-Delayed	0.93	0.95	0.94
	Flight Delayed	0.80	0.74	0.77
LOGISTIC REGRESSION	Flight Not-Delayed	0.92	0.98	0.95
	Flight Delayed	0.88	0.68	0.77

The take away from the above result is that the label Flight Not-delayed is performing better than the label Flight Delayed. On further inspection, an imbalance in the data is found (Figure1). The fudge in the dataset is clear now. Balancing the dataset might give better result. Since the minority class has enough data to over sample. Over sampling will be performed. Synthetic Minority Over-sampling Technique (SMOTE) and random oversampling will be the techniques used in over-sampling.

SMOTE stands for *Synthetic Minority Oversampling Technique*. This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input.

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population

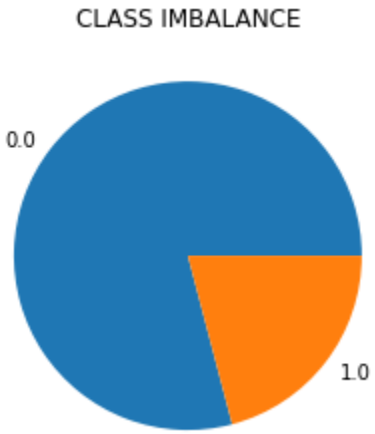

Figure 1

0.0 represents the not delayed flights.(79%)
1.0 represents the delayed flights.(21%)

5.3 Performance analysis after oversampling(SMOTE)

		PRECISION	RECALL	F1 SCORE
GNB	Flight Not-Delayed	0.94	0.93	0.93
	Flight Delayed	0.74	0.76	0.75
EXTRA TREES CLASSIFIER	Flight Not-Delayed	0.94	0.94	0.94
	Flight Delayed	0.77	0.76	0.77
DECISION TREE	Flight Not-Delayed	0.92	0.91	0.92
	Flight Delayed	0.68	0.70	0.69
LOGISTIC REGRESSION	Flight Not-Delayed	0.94	0.93	0.93
	Flight Delayed	0.74	0.78	0.76

5.4 Performance analysis after oversampling(Random Sampling)

EXTRA TREES CLASSIFIER	Flight Not-Delayed	0.93	0.95	0.94
	Flight Delayed	0.80	0.74	0.77
DECISION TREE	Flight Not-Delayed	0.92	0.92	0.92
	Flight Delayed	0.70	0.70	0.70

Recall score will be used as the deciding performance metric, since predicting a Not-Delayed flight as delayed flight is better the predicting a delayed flight as Not-Delayed flight. As going by the recall score, Logistic regression sampled using SMOTE has the best results. Hence the results of Logistic regression will be pipelined.

6. Pipelining

This module uses the pipelined data from the best classifier in the pre-trained regressor to find the duration by which the flight is delayed. The performance of this model is given below(Table 5). Figure 2 is the flowchart of this process.

Figure 2

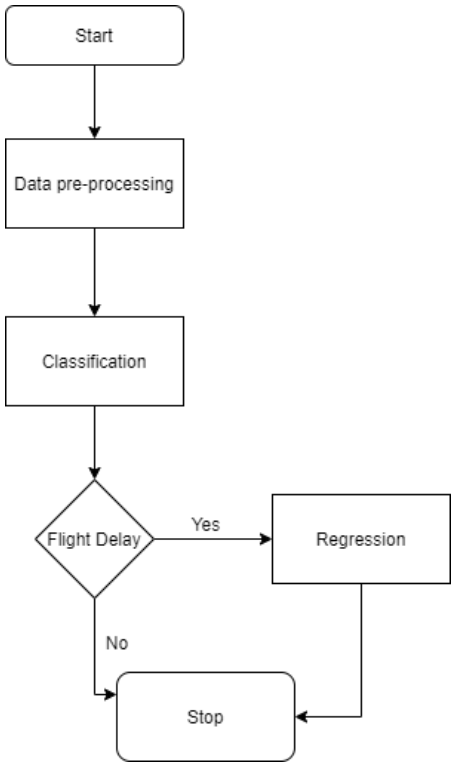

Table 5

	RMSE	MAE	R-SQUARED
LINEAR REGRESSOR	18.8519	14.0365	0.9397
EXTRA TREE REGRESSOR	19.3984	14.6161	0.9350
XGB REGRESSOR	19.2290	14.4818	0.9373

Linear regressor has given a MAE of 14.0365 and RMSE of 18.8519 which is considered low when compared to the actual arrival delay.

7. Conclusion

The flight and weather data was processed and merged into a single dataframe. The data was classified without performing sampling and from the results of the classification it was found that the data was imbalanced. The data was re-sampled and then classified. The Extra tree classifier gave the best results when compared to the other classifiers with the precicion score of 0.93 to the flight's which are not delayed and 0.80 to the flights which are delayed. Among the explored regressor's, Linear regressor was able to predict the arrival delay of flights with a mean absolute error of 14 minutes and root mean squared error of 18 minutes. The predicted time is good when compared to the data as the error goes as large as 1200 minutes.