Flight Delay Prediction

Rishi Vardhan K

November 13, 2019

Abstract

A Flight delay is when an airline flight takes off and/or lands later than its scheduled time. Several factors affect flight delay with respect to weather anomalies and management errors. The project works with the former on a two stage predictive machine learning engine that predicts the chance of a flight to be delayed and the amount of delay.

1 Introduction

Operating a flight is challenging on so many different levels, largely because of all the different people involved. On the one hand, there are factors that are under the direct control of the carrier, such as aircraft turnarounds between flights, passenger punctuality, technical and crew performance, etc. On the other hand, there are perhaps even more factors that are outside of the airline's control, such as weather, air traffic control, security, airport conditions, etc. The reality is such that so long as airplanes continue flying, flight delays will be a part of the experience. According to the Bureau of Statistics, about 20% of all flights are delayed by 15 minutes or more.

The Project is divided into three modules each with different sets of operations. Module 1 focuses on Data Pre-processing, Module 2 on Classification and Module 3 on Regression. Each Module works incrementally on previous module's output.

2 Data Pre-Processing - Module 1

The layout of Module 1 is designed as to process the data input of flight data and weather data and create an output data file needed for running Module 2's Classification process. The flight data is obtained from the Bureau of Transportation Statistics between years 2016-2017. The corresponding weather data for the years are also collected. Each data-set is filtered to contain features with maximum similarity to the given problem.

The following are the list of airport codes considered for the model. The data is processed so as to work with only the arrival and departure of flights under these airports.

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

The following are the list of weather features considered.

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

The following are the list of Flight features considered.

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

The given flight data and weather data are compared using date, time and airport of weather data and similar arrival features of flight data as merging parameters to merge the two data-sets. The output set is further made void of duplicate features so as to avoid redundancy in data. The categorical values are label encoded to fit the necessary machine learning algorithms. Null values found are removed since they don't cause major data alteration. The final data-set comprises of 23 columns in total.

3 Classification - Module 2

Module 2 works on predicting the chance of flight delay, i.e Classifying flights to be either delayed or not. The standing criteria for a flight to be delayed is set under a threshold of 15 minutes. We consider only arrival delay to be existent for our classification process. The feature **ArrDel15** contains values of Class 1 or Class 0 depending on whether a flight is delayed based on our criteria. Class 1 marking flight delay and Class 0 the negation.
The train set and test set are split in a ratio of 80:20.

The following are the scores of various Classifiers considered.

Classifiers	Class	Precision	Recall	f1	Accuracy
Decision Tree	1	0.68	0.71	0.92	0.86
	0	0.92	0.91	0.92	
Weighted Average		0.87	0.87	0.87	
Extra Trees	1	0.83	0.71	0.76	0.90
	0	0.93	0.96	0.94	
Weighted Average		0.91	0.91	0.91	
Gradient Boosting	1	0.89	0.68	0.77	0.91
	0	0.92	0.98	0.95	
Weighted Average		0.92	0.92	0.91	

There stays a significant difference in values between Class 1 and Class 0 scores. The difference is attributed to the fact in reality that flights are not delayed often and therefore the number of delayed flights is less. In fact the number of flights delayed is just 20% of the whole data. This imbalance causes the classifier to not work properly due to less/more number of records supporting each case. This problem is solved by the inclusion of Sampling methods. Sampling is done to solve data imbalance by creating or deleting records to compensate for the improper class distribution.

Only the training sets are sampled so as not to falsify the scores.

The various over-sampling and under-sampling techniques are applied to each classifier to find the optimal technique. The results are tabulated as follows.

Decision Tree

Samplers	Class	Precision	Recall	f1	Accuracy
SMOTE	1	0.67	0.70	0.68	0.86
	0	0.92	0.91	0.91	
Weighted Average		0.87	0.86	0.87	
ADASYN	1	0.36	0.40	0.38	0.72
	0	0.84	0.82	0.83	
Weighted Average		0.74	0.73	0.73	
Random Over Sampler	1	0.69	0.70	0.69	0.87
	0	0.92	0.92	0.92	
Weighted Average		0.87	0.87	0.87	
Near Miss	1	0.23	0.67	0.35	0.47
	0	0.83	0.42	0.56	
Weighted Average		0.70	0.47	0.51	
Random Under Sampler	1	0.50	0.80	0.62	0.79
	0	0.94	0.79	0.86	
Weighted Average		0.85	0.79	0.81	

On observation we find that Random Over-Sampler and SMOTE has optimal values for scores. The rest techniques fail to fit the optimal criteria. Though the best fitting techniques has optimal values for scores, the values don't seem to vary significantly with the original test result (without sampling). Hence this problem maybe asserted to anomaly in data-set.

Extra Trees

Samplers	Class	Precision	Recall	f1	Accuracy
SMOTE	1	0.80	0.73	0.76	0.90
	0	0.93	0.95	0.94	
Weighted Average		0.90	0.91	0.90	
ADASYN	1	0.79	0.73	0.76	0.90
	0	0.93	0.95	0.94	
Weighted Average		0.90	0.90	0.90	
Random Over Sampler	1	0.83	0.70	0.76	0.90
	0	0.92	0.96	0.94	
Weighted Average		0.90	0.91	0.90	
Near Miss	1	0.41	0.84	0.55	0.71
	0	0.94	0.68	0.79	
Weighted Average		0.83	0.72	0.74	
Random Under Sampler	1	0.67	0.81	0.73	0.87
	0	0.95	0.89	0.92	
Weighted Average		0.89	0.88	0.88	

A similar trend is observed here too as values don't vary significantly as a cause of sampling, hence asserting it to data-set anomaly.

Gradient Boosting

Samplers	Class	Precision	Recall	f1	Accuracy
SMOTE	1	0.83	0.73	0.78	0.91
	0	0.93	0.96	0.95	
Weighted Average		0.91	0.91	0.9	
ADASYN	1	0.80	0.75	0.77	0.90
	0	0.93	0.95	0.94	
Weighted Average		0.91	0.91	0.91	
Random Over Sampler	1	0.73	0.79	0.76	0.89
	0	0.94	0.92	0.93	
Weighted Average		0.90	0.90	0.90	
Near Miss	1	0.45	0.83	0.58	0.74
	0	0.94	0.73	0.82	
Weighted Average		0.84	0.75	0.77	
Random Under Sampler	1	0.73	0.79	0.76	0.89
	0	0.94	0.92	0.93	
Weighted Average		0.90	0.89	0.90	

The same trend observed in previous cases arises here.

Since f1-scores are a measure of both precision and recall they are suitable for deciding on best model. Here since sampled and non sampled data have the same best f1-scores, sampling tends to be non-effective. The conclusion from applying different sampling techniques would be that the different techniques tend to be useless for the current data-set as they result in improper output.

4 Regression - Module 3

The last module of the project deals with predicting the amount of delay occurred, in case of flight delay. The model implements a regression model to predict the necessary output. Since only delayed flights are to be considered, the rest flight details are removed from the set. The set is partitioned into train and test sets in a ratio of 80:20. The results of different models are tabulated as follows.

Model	**MAE**	**RMSE**	**R2**
Linear Regression	12.34	17.77	0.94
XGBoost	11.63	16.88	0.94
Extra Trees	11.85	16.92	0.94
Decision Trees	16.59	24.13	0.88

4.1 Feature Importance

The Feature Importance concept is able to perform the ranking of features based on importance of features to the machine learning model. The first three important features of each model ranked, is as follows.

Decision Tree Regressor - CRSDepTime, DEST, Depdel15

XGBoost Regressor - CRSDepTime, Depdel15, weatherCode.

Extra Trees - CRSDepTime, Dep$_{D}elay, Depdel15$

5 Conclusion

Hence the two-stage machine learning model for flight delay prediction is achieved. The first stage deals with classifying flights to be delayed or not. Due to data-imbalance, sampling techniques are employed to compensate for the improper data distribution. Sampling does not however solve the situation and leads to improper output with limited changes. Hence by consideration of better scores among all other classifiers, the **Gradient Boosting Classifier** sampled under SMOTE proved better with Class 1 - 0.78 % and Class 0 - 0.95 % . The follow up module worked on predicting the amount of delay for delayed flights by running different regression algorithms. Among the various regressors considered, the **XGBoost Regressor** resulted in the most efficient output with minimum MAE (11.63) and RMSE (16.88) score. Thus the project is implemented successfully and the results are verified.