Machine Learning for Flight Delay Prediction

D. Prasanna Kumaran

Solarillion Foundation

INTRODUCTION

Thousands of passengers commute everyday by flight. In the modern world air traffic has become very common due to the increase in the number of passengers. In this fast moving world time is everything and people can't risk of losing it. The arrival and departure of a flight is dependent on several factors and one major factor would be the weather condition. This project discusses about the accurate methods involved for classifying whether a flight will be delayed and predicting the arrival delay in minutes using Machine learning.

DATASETS

The data sets considered here are the Flight on time performance data set and Weather Dataset. The details of the data sets are discussed below.

1. FLIGHT DATASET

The Flight dataset contained the details of every flight in USA which was recorded all over the year for 2 years between 2016 and 2017. The data set consisted of over 50 features. Some of the features are DepDel15, DepdelayMinutes and FlightDate.

2. WEATHER DATASET

The Weather data set consisted hourly weather conditions in an airport for the years 2016 and 2017,15 airports were considered for the Dataset. Some of the features are windspeedKmph, winddirDegree and pressure.

PREPROCESSING DATA

The Flight data set is in comma separated values (csv) format and it contains several features that are not related to the weather data and need to be excluded before analysis.

The weather data set obtained is in JavaScript Object Notation (json) format. After careful analysis of the json file the desirable features were extracted from the dataset.

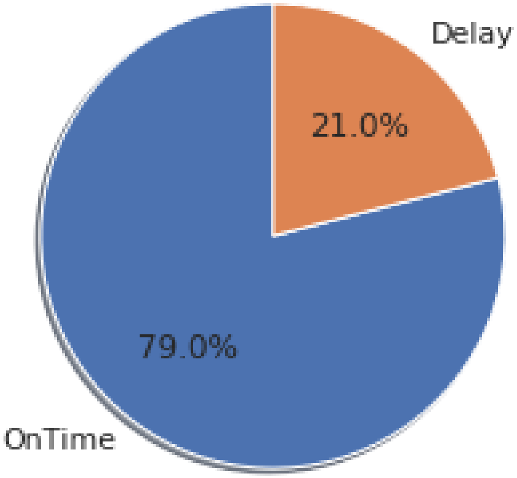
The pre-processed datasets need to be combined based on features that are occuring in both. The data sets were merged considering date, origin, time from flight and weather data sets.

From Fig 1 it is clear that only 21 percent of the flights considered were classified to be delayed. This implies that the data set has flights that arrive on time as majority class.

Heatmap (Refer Fig 2) is a correlation matrix or representation of data in the form of a map or diagram in which data values are represented in different colours based on the correlation between any two features.

Univariate feature selection (Refer Table 1) is used to get a better understanding of the data, its structure and characteristics. It can be used to select features that would improve the model the most. The features are chosen based on the univariate statistical tests such as chi-squared test.

Fig. 1. Ratio of Delayed and On Time Flights



FEATURE SELECTION

The best features from the heatmap and univariate selection method were considered for the machine learning model. Two methods were considered to narrow down the features to be considered and hence improve accuracy of the model.

The Features considered from analysing Heatmap and uni variate selection methods are listed in Table 2 and Table 3.

CLASSIFICATION

A classification is a division or category in a system which divides things into group or types. Column ArrDel15 is the target variable that is considered for this data set.ArrDel15 contains Boolean values indicating whether the arrival was delayed by 15 minutes or more. The data is classified into two classes, 1 if the flight is delayed and 0 if flight is not delayed. The data set was split into training and test data in the ratio 80:20.

For Classification Decision Tree Classifier, Extra Tree Classifier,GradientBoost Classifier were considered.

CLASSIFICATION METRICS

A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. A false positive is an outcome where the model incorrectly predicts the positive class, and a false negative is an outcome where the model incorrectly predicts the negative class.

Precision also called positive predictive value is the fraction of relevant instances among the retrieved instances.

Recall also known as sensitivity is the fraction of the total amount of relevant instances that were actually retrieved.

Precision = True positive/ True Positive + False Positive

Recall = True positive/ True Positive + False Negatives

F1 score = 2 . Precision · Recall/Precision + Recall

Fig. 2. Heatmap

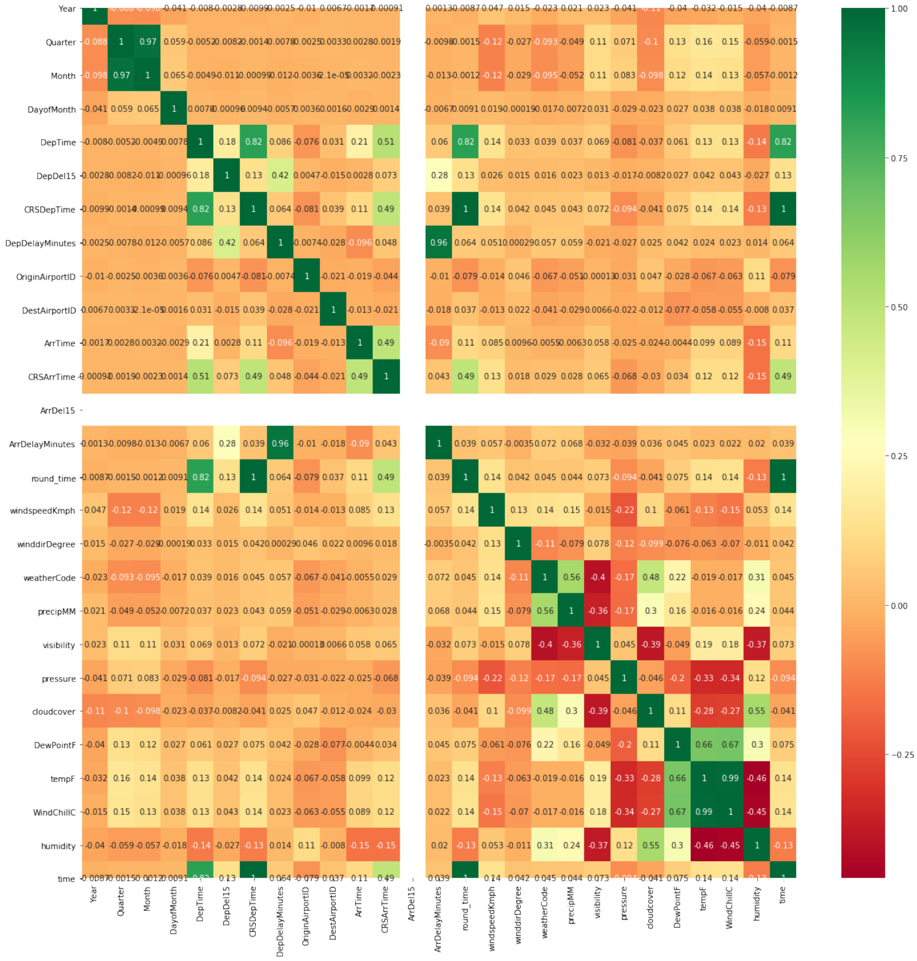


Table 1. Univariate Selection Table

Specs	Score
CRSDepTime	5.901832e+07
roundtime	5.889679e+07
DepTime	4.845862e+07
ArrTime	3.544509e+07
CRSArrTime	3.331966e+07
OriginAirportID	9.559860e+05
DestAirportID	6.031153e+05
DepDelayMinutes	2.060157e+05
ArrDelayMinutes	9.582767e+04
weatherCode	8.864185e+04
windspeedKmph	8.371076e+04
winddirDegree	5.950781e+04
Origin	1.765807e+04
Dest	1.053829e+04
precipMM	4.557246e+03
DepDel15	2.467687e+03
DayofMonth	3.967547e+02
Month	3.847671e+02
Quarter	1.060474e+02
Year	5.652170e-02
ArrDel15	4.664912e-28

Table 2. Flight Dataset

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

Table 3. Weather Dataset

windspeedKmph	winddirDegree	weatherCode	precipMM	visibility
pressure	cloudcover	DewPointF	tempF	WindChillC
humidity	time	date	airport	

Fig. 3. Classification Metrics

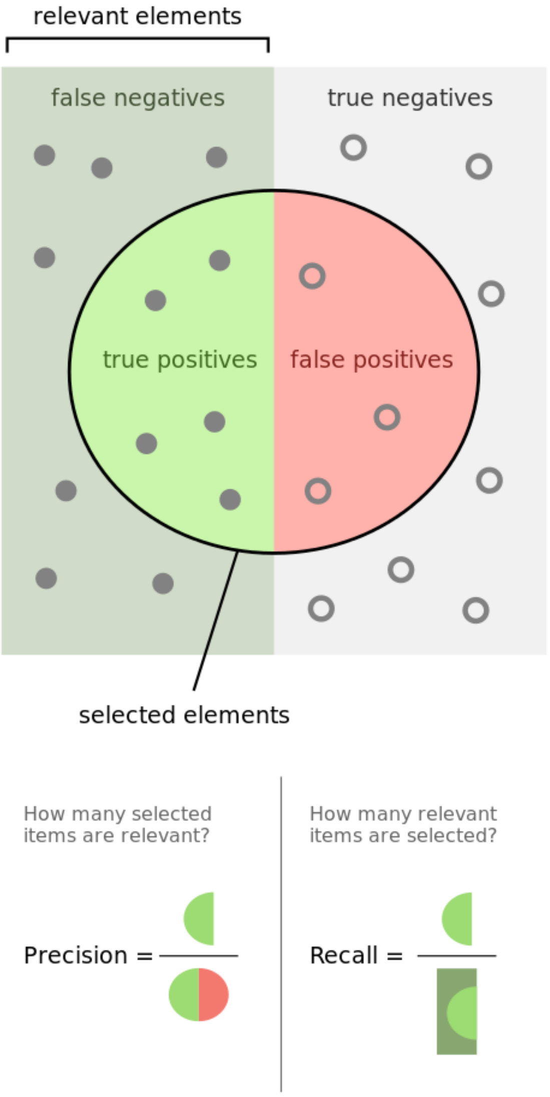


Table 4. PRELIMINARY CLASSIFICATION RESULTS

Classifier	Test	Precision	Recall	f1
DecisionTreeClassifier	0.8692	0.6802	0.7092	0.6944
ExtraTreeClassifier	0.8929	0.8404	0.6044	0.7032
GradientBoostClassifier	0.9166	0.8960	0.6830	0.7751

SAMPLING

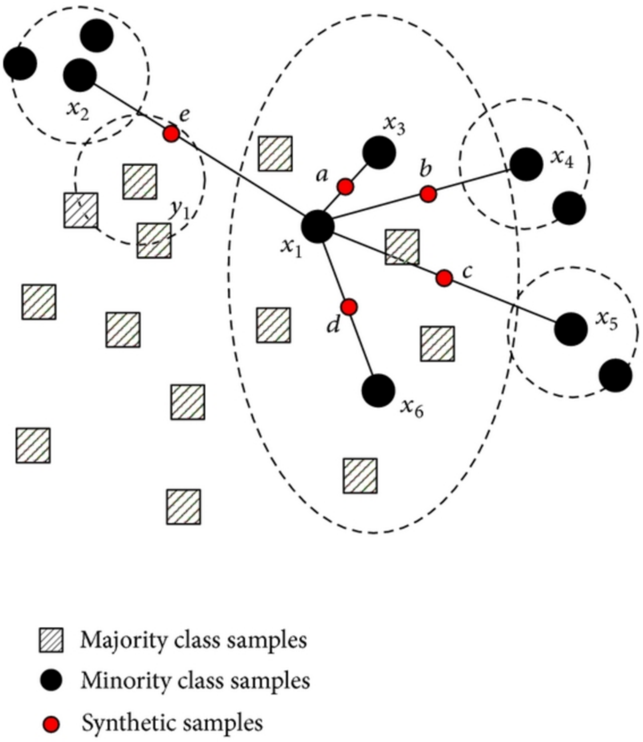
From the obtained pie chart it is clear that there is imbalance in data set and we introduce sampling methods to balance the data set. Sampling is an active process of gathering observations with the intent of estimating a population variable.We can either over sample or under sample data.

ADASYN and SMOTE were used for oversampling and NearMiss for under sampling. The test scores were evaluated and the best model suitable for the given data set was determined from both sampled and unsampled data.

In Over Sampling the minority sample count is increased to match the sample count of the majority sample and in under sampling the majority sample count is brought down to match the minority sample count.

In SMOTE the k nearest neighbours of the minority class of each sample is identified and new samples are generated along the lines joining the neighbouring point and minority sample. The pictorial representation of the working of SMOTE is shown in Fig 4, where synthetic samples are generated along the line joining the K nearest neighbours of a sample.

Fig. 4. SMOTE



ADASYN is similar to SMOTE but the samples have more variance to the parent i.e they are a bit more scattered.

The NearMiss under sampling method calculates the distances between all instances of the majority class and the instances of the minority class. K instances of the majority class that have the smallest distances to those in the minority class are selected. If there are n instances in the minority class, the nearest method will result in k*n instances of the majority class.

Despite using various sampling methods there is no improvement in the test results. Overall GradientBoostClassifier on unsampled data gave the best result and the f1 score was found to be 0.7751.

Table 5. CLASSIFICATION RESULTS SAMPLED DATA

Classifier	Test	Precision	Recall	f1
DecisionTreeClassifier(SMOTE)	0.8672	0.6754	0.7060	0.6903
DecisionTreeClassifier(ADASYN)	0.8671	0.6759	0.7040	0.6897
DecisionTreeClassifier(NearMiss)	0.5953	0.3227	0.8463	0.4672
GradientBoostClassifier(SMOTE)	0.9042	0.7787	0.7592	0.7688
GradientBoostClassifier(NearMiss)	0.8028	0.2133	0.1330	0.1638
GradientBoostClassifier(ADASYN)	0.9038	0.7760	0.7607	0.7683
ExtraTreeClassifier(SMOTE)	0.8920	0.7950	0.6538	0.7175
ExtraTreeClassifier(NearMiss)	0.5711	0.3055	0.8214	0.4454
ExtraTreeClassifier(ADASYN)	0.8915	0.7931	0.6529	0.7162

REGRESSION

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables. In this dataset the arrival delay in minutes was estimated using various regressors. For regression Linear,Random Forest and Gradient Boost regressor were used. The R2,MSE,RMSE and MAE errors were calculated and the results are shown in Table 6.

METRICS

The various metrics used for regressors are R2,MAE,MSE,RMSE.

RMSE(Root Mean Squared Error): is the square root of the sum over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} x(i) - y(i)^2}$$

MAE(Mean absolute Error): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y(i) - y(p)|$$

Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. Both the MAE and RMSE can range from 0 to INF. They are negatively-oriented scores: Lower values are better.

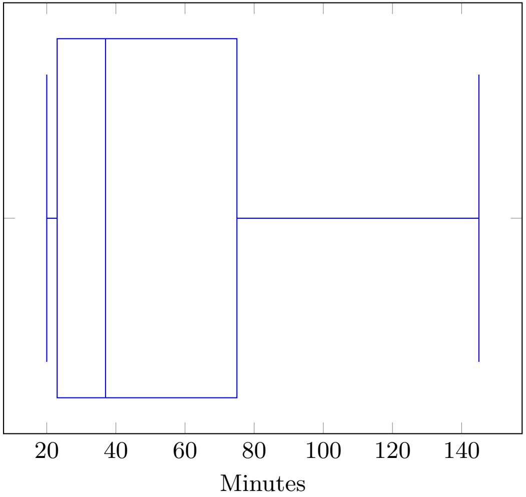
Table 6. REGRESSION RESULTS

Regressor	R2	MSE	RMSE	MAE
Linear Regression	0.94	308.98	17.58	12.26
RandomForest	0.94	311.26	17.64	12.50
GradientBoost	0.95	276.06	16.62	11.50
Voting	0.95	260.40	16.14	11.18

REGRESSION ANALYSIS

GradientBoostingRegressor was considered to be the best model as it gave the best results. It was chosen based on the RMSE and MAE values shown in Table 6. The RMSE and MAE scores of GradientBoostingRegressor was 16.62 and 11.50 respectively.

Fig. 5. Box Plot for actual delayed minutes



From Fig.4 obtained it is clear that the Inter Quartile range lies between 22 and 78 . This range tells that for most delayed samples the delay in minutes lies in the Inter Quatile range. The Inter Quartile range obtained from GradientBoostRegressor lies in similar range and the range is between 28 and 75.In practical approach the model is trained to predict values primarily lying between Inter Quartile Range and the MAE values obtained for the GradientBoostRegressor is 11.50 and therefore comparing the Inter Quartile Range and MAE score the model is accurate and reliable.

CONCLUSION

From Figures 4,5 and 6 the model that fits the given data set are GradientBoostClassifier for classifying whether a delay exists and GradientBoostRegressor for predicting arrival delay for flights that are classified to be delayed. From this project we have trained a machine learning model for data prediction and classification using the features which correlate with each other the most and thus improving overall accuracy. Accuracy can be further improved when features and data are further added to the data set.