FLIGHT DELAY PREDICTION

Sri Gayathri Devi I

Solarillion Foundation

ABSTRACT

The objective of this briefing is to present an overview of the machine learning techniques.A two stage predictive machine learning engine that forecasts the on-time performance of flights is discussed in this project. Classifier and regressor models are built and analysed to predict flight delays.

Introduction

In the fast paced world, time prevails as the most powerful and invaluable thing. To our concern here, Flight delays are considered to affect passengers, airlines, goods' transport and indirect circular impact on rest of the economy.Furthermore, in the domain of sustainability, it causes environmental harm by rise in fuel consumption and gas emissions. Hence, these factors indicate how crucial it is to understand the delays in prior. As a result, air passengers' travel decisions can be influenced by delay information and could save many disruptions.

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the flight status and provide a summary of the arrival delays, departure delays, on-time arrival, etc in their monthly report. This study aims at analyzing flight information of US domestic flights covering top 15 busiest airports of US.

A flight delay is said to occur when an airline lands or takes off 15 minutes later than its scheduled arrival or departure time respectively. Notable reasons are adverse weather conditions, air traffic congestion, maintenance and security issues.

Supervised Learning:

Machine learning is a method of analytical model building that automatically learn and improve from experience.*In Supervised machine learning, we train the machine using well "labeled" data*. It is categorized into,

- Regression: Regression technique predicts a single continuous output value using training data typically used in predicting, forecasting, and finding relationships between quantitative data. In our case,by how many minutes a particular flight is delayed is predicted.
- Classification: Classification intends to group the output inside a class.Modelling a classifier that predicts whether a flight is delayed or not is projected here.

Dataset:

The dataset comprises of weather and Flight information individually. We are interested in the weather and flight details of 15 airports for the years 2016 and 2017.

- Weather data: It includes weather features such as WindSpeedKmph, Pressure, Cloudcover, tempF. There were 25 hourly features and 6 daily features. Some of the significant features were,

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

Table 1: Weather Features

- Flight data: Each file holds the On-Time Performance of flights. There are around 110 features. Some of the prominent features were,

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

Table 2: Flight Features

Data Preprocessing

The objective is to merge the flight details with its corresponding weather information. As a result, a flight flying at a particular time starting from a particular airport will have the corresponding hour's weather report.

- Weather- The .json files are in the form of nested dictionaries. Each file includes the weather report for a month containing hourly information for each day. All these files are processed into a single .csv file.

- Flight - The files are found in csv format and combined accordingly into a single dataframe.

- Merge - The primary features for merge would be date, time and airport.

- Handling Missing Values: Instances containing missing labels are removed from the dataset.

A well-defined dataset is procured from the raw data.

Feature Selection

Here we understand which features are not helping ML pipeline and are removed. Techniques used to select the best set of features are,

- 1.Principal Component Analysis: A dimensionality reduction technique where the data is initially standardised and converted into principal components using covariance matirx. From this, a feature vector is formed with fewer significant features.
- 2.Univariate Selection(SelectKBest with ANOVA): It determines the strength of the relationship of a feature with the response variable using correlation tests such as chi2, ANOVA, Mutual Information Coefficient.
- 3.Feature Importance: The *importance score* for each feature is calculated based on its ability to make key decisions with boosted decision trees. They are ranked and compared with each other.
- 4.Correlation Matrix with Heatmap: A matix plot depicting the correltion between features.

	Features	Score
9	DepDel15	1.887728e+06
13	ArrDelayMinutes	1.061375e+06
8	DepDelayMinutes	7.869872e+05
7	DepTime	4.516780e+04
6	CRSDepTime	3.046610e+04
0	time	2.980591e+04
11	CRSArrTime	2.292761e+04
14	weatherCode	1.602312e+04
19	precipMM	1.456985e+04
22	windspeedKmph	1.349363e+04
24	WindGustKmph	1.200934e+04
15	pressure	1.016705e+04
12	ArrTime	5.727804e+03
10	DestAirportID	4.306164e+03
20	DewPointF	2.699627e+03
16	cloudcover	2.652123e+03
17	tempF	2.324695e+03

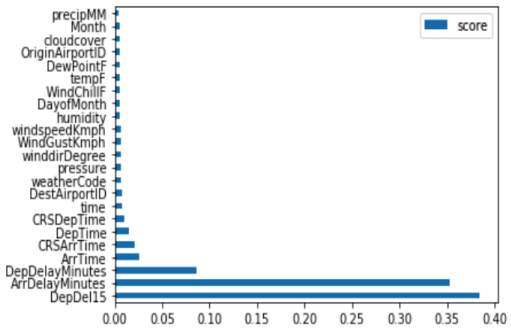


Figure 1: Univariate Selection and Feature Importance

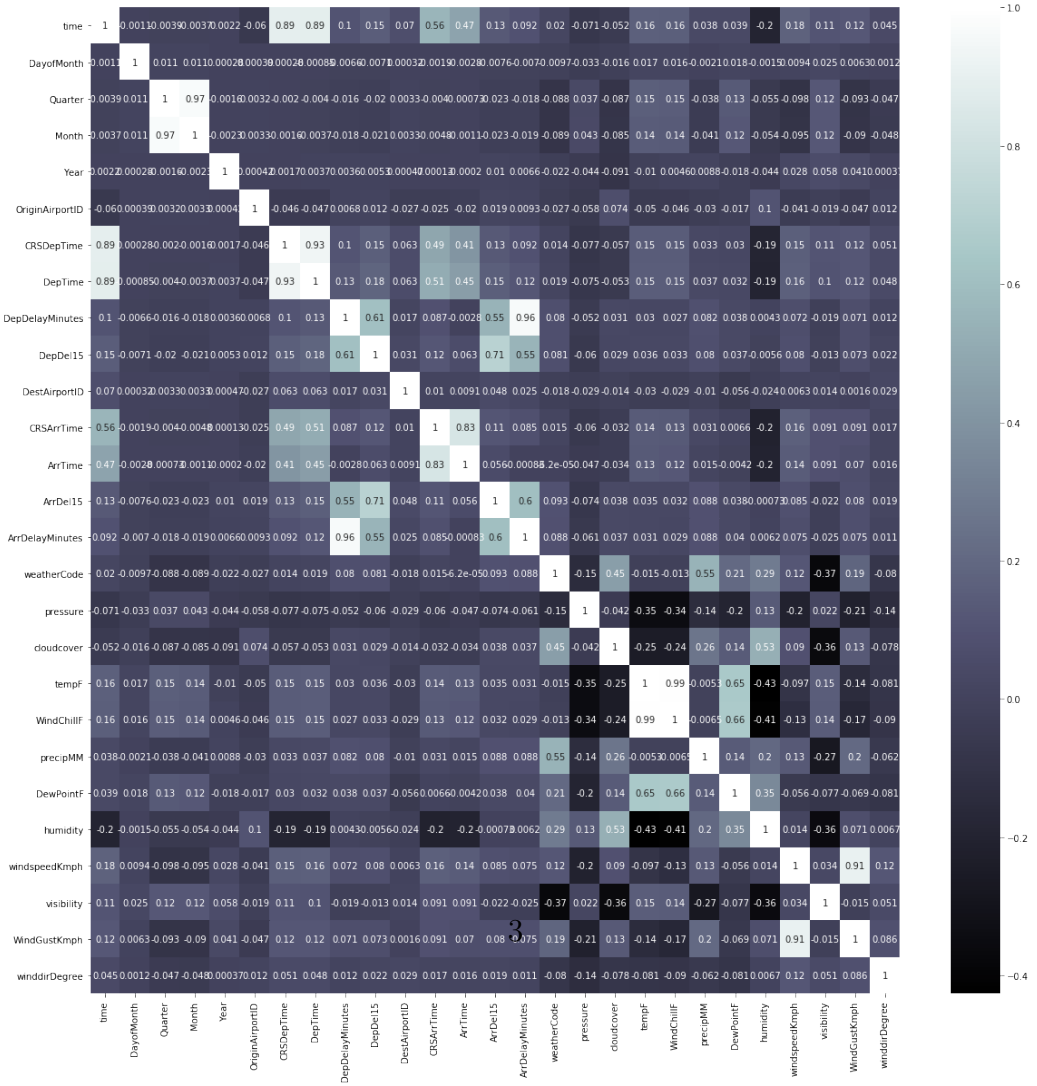


Figure 2: Heat map

Among the above methods, best features from Univariate and Feature Importance were filtered out.

Classification

Classification predictive modeling is the task of approximating the mapping function from input variables to discrete *output variables*. The main goal is to identify which class/category the new data will fall into. Thus, the target variable is "ArrDel15" denoting 1,0 for delayed and on time flights respectively.

Training and testing sets:The dataset is split(4:1) into training set(which is trained by the model) and testing set(which is used to evaluate the trained model).

Figure 3: Train and Test sets

Classification metrics: Performance metrics are used to evaluate the model.

- Confusion Matrix: It is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions.It contains the number of predicted samples that belong to TN, FP, FN and TP. Ideal case is when False Negatives and False Positives are zeros.
- Accuracy score: number of correct prediction over all predictions.

$$accuracy = \frac{TruePositives + TrueNegatives}{TotalSamples} \tag{1}$$

- Precision: the probability that the decision is correct.

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

- Recall: the proportion of actual positives that was identified correctly.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3}$$

- AUC curve(Area Under Curve): It tells how much a model is capable of distinguishing between classes.

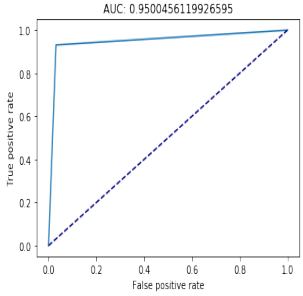


Figure 4: AUC

- F1 score: F1 Score is the weighted average of Precision and Recall.

$$f1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4}$$

Modelling a classifer: The dataset is now fit into classifiers such as XGBoost, DecisionTree, GradientBoost, ExtraTrees, and Stochastic Gradient Descent.Below is the performance of each classifier.

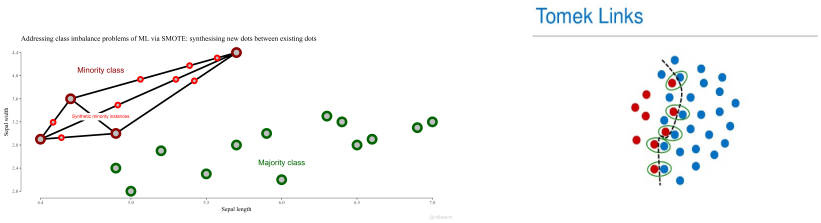
Class imbalance and Sampling

A binary classification was performed on label,ArrDel15. But, the ratio of number of instances with label 0 to label 1 is found to be 4:1. Hence, the dataset is highly imbalanced. This might lead to incorrect accuracy. Some of the techniques used to overcome class imbalance are,

- Random Under-Sampling: Balances the class distribution by randomly eliminating majority class.
- SMOTE(Synthetic Minority Oversampling Technique): K-nearest neighbours are identified in feature space and a line is drawn between the examples in the feature space where new samples are generated at a point along that line.

Classifiers	Accuracy	Precision	Recall	f1-score
XGBoost	0.919	0.899	0.692	0.785
DecisionTree	0.868	0.678	0.705	0.691
GradientBoost	0.917	0.895	0.681	0.76
Stochastic Gradient Descent	0.876	0.669	0.808	0.73
ExtraTrees	0.868	0.678	0.705	0.743

Table 3: Unsampled Classification



(a) SMOTE (b) Tomek Links

- SMOTETomek: Performs over-sampling using SMOTE and cleaning using Tomek links.
- ADASYN (ADAptive SYNthetic):The algorithm uses Euclidean distance for KNN Algorithm. The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions.

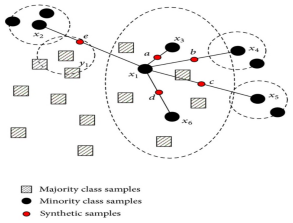


Figure 5: ADASYN

XGBoost with SMOTETomek sampling is found to perform well.

Classifiers	Accuracy	Precision	Recall	f1-score
XGBoost(RandomUnderSampling)	0.881	0.922	0.834	0.812
XGBoost(ADASYN)	0.918	0.891	0.693	0.783
XGBoost(SMOTETomek)	0.919	0.893	0.694	0.781
DecisionTree(RandomUnderSampling)	0.791	0.501	0.805	0.621
DecisionTree(ADASYN)	0.866	0.672	0.704	0.692
DecisionTree(SMOTETomek)	0.916	0.913	0.919	0.910
GradientBoost(RandomUnderSampling)	0.895	0.735	0.786	0.765
GradientBoost(ADASYN)	0.900	0.760	0.767	0.766
GradientBoost(SMOTETomek)	0.910	0.818	0.737	0.788
ExtraTrees(RandomUnderSampling)	0.791	0.501	0.802	0.743
ExtraTrees(ADASYN)	0.866	0.672	0.704	0.775
ExtraTrees(SMOTETomek)	0.916	0.913	0.919	0.778

Table 4: Sampled Classification

Regressor	MAE	RMSE	R2 Score
LinearRegression	14.64	20.01	0.9213
ExtraTreesRegressor	12.03	17.17	0.9422
GradientBoostingRegressor	11.74	17.09	0.9426
RandomForestRegressor	11.93	16.97	0.9433

Table 5: Regressor Performance

Regression

Regression model predicts values of a desired target quantity.The target variable is "ArrDelayMinutes" .

Regression mterics:

- Mean Absolute Error (MAE): It is the average difference between the Original Values and the Predicted Values

$$MAE = \frac{1}{n}\sum \mid Y - \hat{Y} \mid \tag{5}$$

- Root Mean Squared Error(RMSE):

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{d_i - f_i}{\sigma_i}\Big)^2} \tag{6}$$

- R2 score: provides an indication of the goodness or fit of a set of predicted output values to the actual output values.

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{7}$$

- Cross Validation Technique and K-Fold Technique: It shows how a model would generalize to an independent data set.

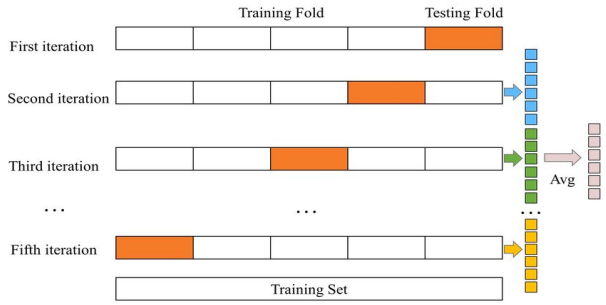
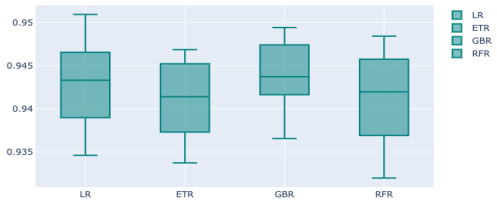
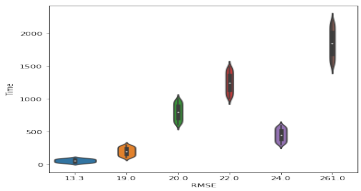


Figure 6: K-fold cross validation



(a) Regression Analysis



(b) RMSE for each interval

Regression Analysis:

Gradient Boosting Regressor is found to perform well. The cross-validated r2 scores of GBR is more than 0.94 for 75% of data. RMSE score of GBR is 17.09 and thus, model is efficient. Further,the delay minutes at different intervals of time are predicted in order to assess the performance of regressor at different intervals of time.

The accuracy can still be improved with hyper parameter tuning using GridSearchCV or RandomSearchCV.

Conclusion

The dataset was explored and intended features were extracted based on its correlation and machine learning models were implemented appropriately. Prediction of arrival delay and its range is successfully established with good amount of accuracy with XGBoost Classifer and Gradient Boosting Regressor.