FLIGHT DELAY PREDICTION

Sri Gayathri Devi I

Solarillion Foundation

Abstract

The primary objective is to build a two stage predictive machine learning model that forecasts the on-time performance of flights. Flight delays have a negative impact on passengers, airlines, airport authorities impacting overall economy. Prediction of delays in prior could save many disruptions and avoid any losses. Learning models are analysed and evaluated for greater performance and could act as a prototype for delay prediction.

1. Introduction

Over the last twenty years, air travel has been increasingly preferred among travelers because of its speed and comfort. This has led to phenomenal growth in aviation industry in turn increasing air-traffic congestion causing flight delays. 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. These delays are responsible for large economic and environmental losses. In order to alleviate these negative impacts and satisfy increasing demand, an accurate prediction of flight delays is needed. Delays occur due to weather conditions, seasonal demands, airline policies, technical issue such as problems in airport facilities, luggage handling, mechanical apparatus, and accumulation of delays from preceding flights. Adverse weather conditions are often cited as one of the main reasons. This leads to the integration of weather report with flight information to build a predictive modelling. Hence, the prediction analysis retrieved from this model can optimize airline and airport operations.

1.1 Supervised Learning

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience. In supervised machine learning is a task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data. For instance, a past data containing status of flights along with the features influencing its delay can be trained to predict future delays. The feature set of delayed and on time flights are well labelled.

2. Dataset

The dataset is taken from National Oceanic and Atmospheric Administration which provides hourly synoptic weather observations and UBTS (US Bureau of Transportation Services) which tracks the flight status and provide air traffic delay statistics. The weather and flight details of 15 airports for the years 2016 and 2017 are considered for delay prediction.

- Weather data: Weather data comprises of 25 hourly features and 6 daily features. The weather report accounting for each month contains hourly information for each day. Some of the features are shown in Table1.

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

Table 1: Weather Data Features

- Flight data: Flight data consists of on-time performance of flights with 110 features. Some of the features are displayed in Table2.

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

Table 2: Flight Data Features

3. Data Preprocessing

The weather features that are more related to flight delay is filtered out and basic information of flight timings, date, stations are considered. Subsequently, the flight details are merged with its corresponding weather information so that, a flight flying at a particular time starting from a particular airport will have the corresponding hour's weather report. The primary features for merge are date, time and origin airport. Further, the missing values are handled by removing the instances containing missing labels.

4. Feature Selection

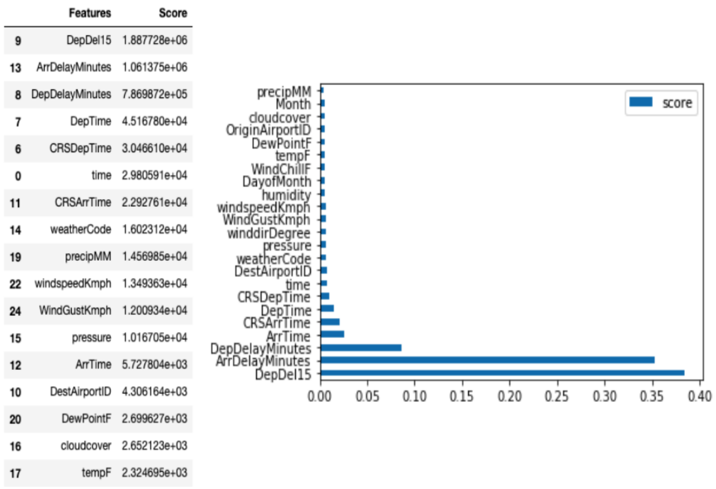
Feature Selection is the process of selecting features which has a strong relationship with the prediction variable or contributes the most for prediction. The target variable being 'ArrDel15' for delay prediction, below techniques are used for feature selection.

- Univariate Selection(SelectKBest with ANOVA): It determines the strength of the relationship of a feature with the response variable using statistical tests such as chi2, ANOVA and Mutual Information Coefficient.
- Feature Importance: The *importance score* for each feature is calculated based on its ability to make key decisions with boosted decision trees. Features are then selected based on their ranks.
- Correlation Matrix with Heatmap: Correlation states how the features are related to the target variable('ArrDel15'). Values range from -1 to +1. In Figure 2, values greater than 1 shows positive correlation, less than 1 shows negative correlation, zeros refers non-linearity in two features.

DepDelay Minutes	DepTime	CRSDep Time	time	Dest	Origin
weather Code	pressure	winddir Degree	WindGust Kmph	Dayof Month	Quarter
windspeed Kmph	WindChillF	DewPointF	humidity	tempF	Month

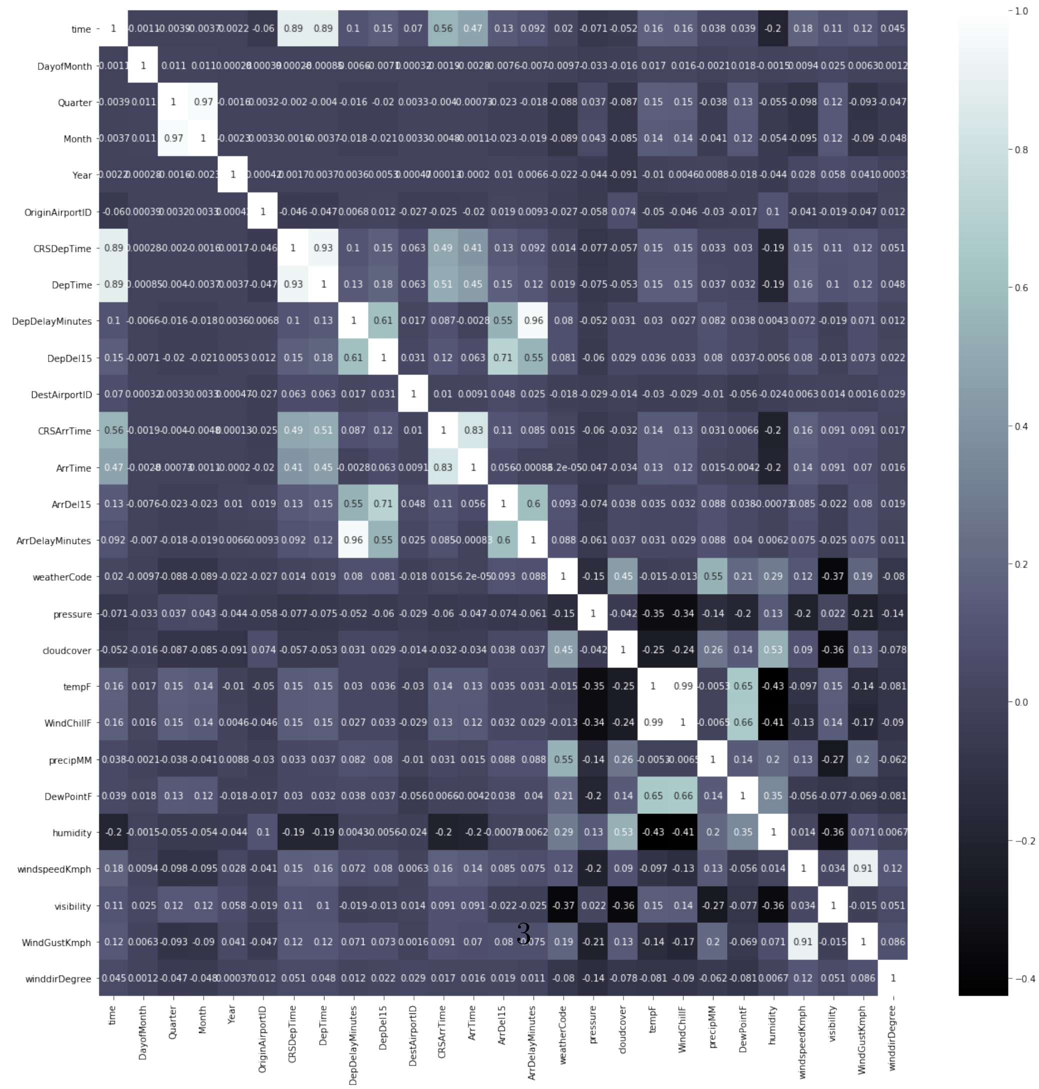
Table 3: Filtered Features

Among the above methods, best features from univariate selection and feature importance were filtered out and shown in Table3.



	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



	time	DayofMonth	Quarter	Month	Year	OriginAirportID	CRSDepTime	DepTime	DepDelayMinutes	DepDel15	DestAirportID	CRSArrTime	ArrTime	ArrDel15	ArrDelayMinutes	weatherCode	pressure	cloudcover	tempF	WindChillF	precipMM	DewPointF	humidity	windspeedKmph	visibility	WindGustKmph	winddirDegree
time	1	-0.0011	-0.0039	-0.0037	0.0022	-0.06	0.89	0.89	0.1	0.15	0.07	0.56	0.47	0.13	0.092	0.02	-0.071	-0.052	0.16	0.16	0.038	0.039	-0.2	0.18	0.11	0.12	0.045
DayofMonth	-0.0011	1	0.011	0.011	0.00028	0.00039	0.00028	-0.00085	-0.0066	-0.0071	0.00032	-0.0019	-0.0028	-0.0076	-0.007	-0.0097	-0.033	-0.016	0.017	0.016	-0.0021	0.018	-0.0015	0.0094	0.025	0.0063	0.0012
Quarter	-0.0039	0.011	1	0.97	-0.0016	0.0032	-0.002	-0.004	-0.016	-0.02	0.0033	-0.004	-0.00073	-0.023	-0.018	-0.088	0.037	-0.087	0.15	0.15	-0.038	0.13	-0.055	-0.098	0.12	-0.093	-0.047
Month	-0.0037	0.011	0.97	1	-0.0023	0.0033	-0.0016	-0.0037	-0.018	-0.021	0.0033	-0.0048	-0.0011	-0.023	-0.019	-0.089	0.043	-0.085	0.14	0.14	-0.041	0.12	-0.054	-0.095	0.12	-0.09	-0.048
Year	0.0022	0.00028	-0.0016	-0.0023	1	0.00042	0.0017	0.0037	0.0036	0.0053	0.00047	0.00013	-0.0002	0.01	0.0066	-0.022	-0.044	-0.091	-0.01	0.0046	0.0088	-0.018	-0.044	0.028	0.058	0.041	0.00037
OriginAirportID	-0.06	0.00039	0.0032	0.0033	0.00042	1	-0.046	-0.047	0.0068	0.012	-0.027	-0.025	-0.02	0.019	0.0093	-0.027	-0.058	0.074	-0.05	-0.046	-0.03	-0.017	0.1	-0.041	-0.019	-0.047	0.012
CRSDepTime	0.89	0.00028	-0.002	-0.0016	0.0017	-0.046	1	0.93	0.1	0.15	0.063	0.49	0.41	0.13	0.092	0.014	-0.077	-0.057	0.15	0.15	0.033	0.03	-0.19	0.15	0.11	0.12	0.051
DepTime	0.89	-0.00085	-0.004	-0.0037	0.0037	-0.047	0.93	1	0.13	0.18	0.063	0.51	0.45	0.15	0.12	0.019	-0.075	-0.053	0.15	0.15	0.037	0.032	-0.19	0.16	0.1	0.12	0.048
DepDelayMinutes	0.1	-0.0066	-0.016	-0.018	0.0036	0.0068	0.1	0.13	1	0.61	0.017	0.087	-0.0028	0.55	0.96	0.08	-0.052	0.031	0.03	0.027	0.082	0.038	0.0043	0.072	-0.019	0.071	0.012
DepDel15	0.15	-0.0071	-0.02	-0.021	0.0053	0.012	0.15	0.18	0.61	1	0.031	0.12	0.063	0.71	0.55	0.081	-0.06	0.029	0.036	0.033	0.08	0.037	-0.0056	0.08	-0.013	0.073	0.022
DestAirportID	0.07	0.00032	0.0033	0.0033	0.00047	-0.027	0.063	0.063	0.017	0.031	1	0.01	0.0091	0.048	0.025	-0.018	-0.029	-0.014	-0.03	-0.029	-0.01	-0.056	-0.024	0.0063	0.014	0.0016	0.029
CRSArrTime	0.56	-0.0019	-0.004	-0.0048	0.00013	-0.025	0.49	0.51	0.087	0.12	0.01	1	0.83	0.11	0.085	0.015	-0.06	-0.032	0.14	0.13	0.031	0.0066	-0.2	0.16	0.091	0.091	0.017
ArrTime	0.47	-0.0028	-0.00073	-0.0011	-0.0002	-0.02	0.41	0.45	-0.0028	0.063	0.0091	0.83	1	0.056	-0.00083	-6.2e-05	-0.047	-0.034	0.13	0.12	0.015	-0.0042	-0.2	0.14	0.091	0.07	0.016
ArrDel15	0.13	-0.0076	-0.023	-0.023	0.01	0.019	0.13	0.15	0.55	0.71	0.048	0.11	0.056	1	0.6	0.093	-0.074	0.038	0.035	0.032	0.088	0.038	-0.00073	0.085	-0.022	0.08	0.019
ArrDelayMinutes	0.092	-0.007	-0.018	-0.019	0.0066	0.0093	0.092	0.12	0.96	0.55	0.025	0.085	-0.00083	0.6	1	0.088	-0.061	0.037	0.031	0.029	0.088	0.04	0.0062	0.075	-0.025	0.075	0.011
weatherCode	0.02	-0.0097	-0.088	-0.089	-0.022	-0.027	0.014	0.019	0.08	0.081	-0.018	0.015	-6.2e-05	0.093	0.088	1	-0.15	0.45	-0.015	-0.013	0.55	0.21	0.29	0.12	-0.37	0.19	-0.08
pressure	-0.071	-0.033	0.037	0.043	-0.044	-0.058	-0.077	-0.075	-0.052	-0.06	-0.029	-0.06	-0.047	-0.074	-0.061	-0.15	1	-0.042	-0.35	-0.34	-0.14	-0.2	0.13	-0.2	0.022	-0.21	-0.14
cloudcover	-0.052	-0.016	-0.087	-0.085	-0.091	0.074	-0.057	-0.053	0.031	0.029	-0.014	-0.032	-0.034	0.038	0.037	0.45	-0.042	1	-0.25	-0.24	0.26	0.14	0.53	0.09	-0.36	0.13	-0.078
tempF	0.16	0.017	0.15	0.14	-0.01	-0.05	0.15	0.15	0.03	0.036	-0.03	0.14	0.13	0.035	0.031	-0.015	-0.35	-0.25	1	0.99	-0.0053	0.65	-0.43	-0.097	0.15	-0.14	-0.081
WindChillF	0.16	0.016	0.15	0.14	0.0046	-0.046	0.15	0.15	0.027	0.033	-0.029	0.13	0.12	0.032	0.029	-0.013	-0.34	-0.24	0.99	1	-0.0065	0.66	-0.41	-0.13	0.14	-0.17	-0.09
precipMM	0.038	-0.0021	-0.038	-0.041	0.0088	-0.03	0.033	0.037	0.082	0.08	-0.01	0.031	0.015	0.088	0.088	0.55	-0.14	0.26	-0.0053	-0.0065	1	0.14	0.2	0.13	-0.27	0.2	-0.062
DewPointF	0.039	0.018	0.13	0.12	-0.018	-0.017	0.03	0.032	0.038	0.037	-0.056	0.0066	-0.0042	0.038	0.04	0.21	-0.2	0.14	0.65	0.66	0.14	1	0.35	-0.056	-0.077	-0.069	-0.081
humidity	-0.2	-0.0015	-0.055	-0.054	-0.044	0.1	-0.19	-0.19	0.0043	-0.0056	-0.024	-0.2	-0.2	-0.00073	0.0062	0.29	0.13	0.53	-0.43	-0.41	0.2	0.35	1	0.014	-0.36	0.071	0.0067
windspeedKmph	0.18	0.0094	-0.098	-0.095	0.028	-0.041	0.15	0.16	0.072	0.08	0.0063	0.16	0.14	0.085	0.075	0.12	-0.2	0.09	-0.097	-0.13	0.13	-0.056	0.014	1	0.034	0.91	0.12
visibility	0.11	0.025	0.12	0.12	0.058	-0.019	0.11	0.1	-0.019	-0.013	0.014	0.091	0.091	-0.022	-0.025	-0.37	0.022	-0.36	0.15	0.14	-0.27	-0.077	-0.36	0.034	1	-0.015	0.051
WindGustKmph	0.12	0.0063	-0.093	-0.09	0.041	-0.047	0.12	0.12	0.071	0.073	0.0016	0.091	0.07	0.08	0.075	0.19	-0.21	0.13	-0.14	-0.17	0.2	-0.069	0.071	0.91	-0.015	1	0.086
winddirDegree	0.045	0.0012	-0.047	-0.048	0.00037	0.012	0.051	0.048	0.012	0.022	0.029	0.017	0.016	0.019	0.011	-0.08	-0.14	-0.078	-0.081	-0.09	-0.062	-0.081	0.0067	0.12	0.051	0.086	1

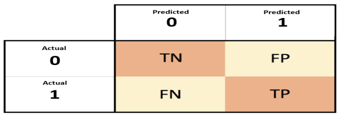
Figure 2: Heat map

5. Classification

It is the process of categorizing a given set of data into classes. Classification model maps a function from input variables to discrete output variables. The target variable is "ArrDel15" denoting 1,0 for delayed and on time flights respectively.

5.1 Training and testing sets

The dataset is split with 80% of data as training set (which is trained by the model) and 20% of data as testing set (which is used to evaluate the trained model).



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Figure 3: Confusion matrix

5.2 Classification metrics:

- Confusion Matrix: It is a summary of prediction results which contains the number of predicted samples that belong to TruePositive (observation is positive and is predicted to be positive),FalseNegative (observation is positive, but is predicted negative), TrueNegative(observation is negative, and is predicted to be negative), FalsePositive(observation is negative, but is predicted positive).

- Accuracy score

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

- Precision

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

- Recall

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

- F1 score

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

The performance of XGBoost, DecisionTree, GradientBoost, ExtraTrees, and Stochastic Gradient Descent classifiers are shown in Table4.

Classifiers	Precision		Recall		f1-score	
	0	1	0	1	0	1
XGBoost	0.91	0.90	0.98	0.69	0.95	0.78
DecisionTree	0.92	0.68	0.91	0.71	0.92	0.69
GradientBoost	0.92	0.90	0.98	0.68	0.95	0.78
Stochastic Gradient Descent	0.95	0.67	0.89	0.81	0.92	0.73
ExtraTrees	0.92	0.87	0.97	0.70	0.95	0.77

Table 4: Classification results of unsampled data

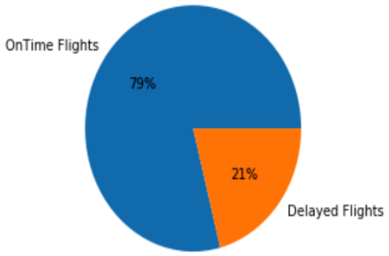


Figure 4: Distribution of delayed and ontime flights

A binary classification was performed on label, ArrDel15. The Figure4 shows that the dataset is highly imbalanced where number of instances of class-0 comprises 80% of data. In such cases, it is assumed that the minority class is positive class and majority class is negative class and on modelling classifiers, they are likely to predict everything as negative(the majority class). Sampling techniques are used to overcome this problem.

6. Class imbalance and Sampling

Sampling is a technique that is designed to change the distribution of population of classes in a data. The data can either be under-sampled or over-sampled.

- Random Under-Sampling: Balances the class distribution through random elimination of majority class examples.

- 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.

- SMOTETomek: Tomek links are formed between two instances of different class where the instances are each others' nearest neighbour. Tomek links are then used to locate all cross-class nearest neighbors in majority class that are closest to the minority class and removed. In SMOTETomek, Tomek links are applied to SMOTE training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.

- 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 different minority samples to compensate for the skewed distributions.

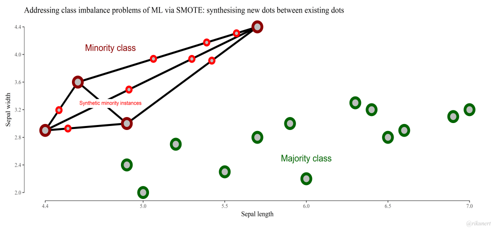
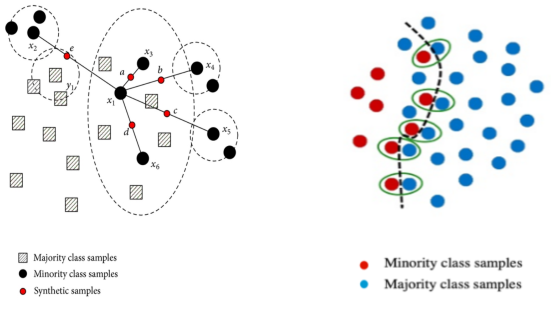


Figure 5: SMOTE



(a) ADASYN

(b) Tomek Links

Classifiers	Precision		Recall		f1-score	
(RandomUnderSampling)	0	1	0	1	0	1
XGBoost	0.85	0.92	0.93	0.83	0.89	0.88
DecisionTree	0.94	0.60	0.79	0.80	0.86	0.62
GradientBoost	0.94	0.74	0.92	0.79	0.93	0.76
Stochastic Gradient Descent	0.94	0.70	0.91	0.80	0.93	0.74
ExtraTrees	0.95	0.67	0.89	0.82	0.92	0.74

Table 5: Classification results of UnderSampled data

Classifiers	Precision		Recall		f1-score	
(ADASYN)	0	1	0	1	0	1
XGBoost	0.92	0.89	0.98	0.69	0.95	0.78
DecisionTree	0.92	0.67	0.91	0.70	0.91	0.69
GradientBoost	0.94	0.76	0.94	0.77	0.94	0.76
Stochastic Gradient Descent	0.92	0.67	0.91	0.71	0.91	0.69
ExtraTrees	0.93	0.81	0.96	0.73	0.94	0.91

Table 6: Classification results of OverSampled data

Classifiers	Precision		Recall		f1-score	
(SMOTETomek)	0	1	0	1	0	1
XGBoost	0.92	0.89	0.98	0.69	0.95	0.78
DecisionTree	0.92	0.91	0.91	0.92	0.92	0.92
GradientBoost	0.93	0.82	0.96	0.74	0.94	0.78
Stochastic Gradient Descent	0.90	0.97	0.98	0.57	0.94	0.71
ExtraTrees	0.93	0.83	0.96	0.73	0.94	0.77

Table 7: Classification results of CombineSampled data

XGBoost classifier with SMOTETomek sampling is found to give better results.

7. Regression

A predictive modelling technique that predicts values of a desired continuous target quantity establishing a relationship between dependent and independent variables. The target variable for prediction of arrival delay period is “ArrDelayMinutes”.

7.1 Regression metrics

- Mean Absolute Error (MAE)

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

- Root Mean Squared Error(RMSE)

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

- R2 score

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

where Y refers original values, Y(cap) refers predicted values and n refers to the total samples in a dataset.

- Cross Validation (K-Fold echnique): This approach involves randomly dividing the set of observations into k-folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds.

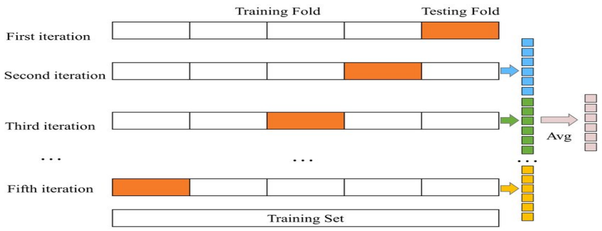


Figure 6: K-fold cross validation

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 8: Regression results

7.2 Regression Analysis

Gradient Boosting Regressor is found to perform well. The cross-validated r2 scores of the regressor is more than 0.94 for 75% of data and RMSE is 17.09. Figure7 shows that the 50% of actual arrival delay minutes data range from 23-75 minutes. Similar values are found with gradient boosting regressor. Thus, the model is accurate.
Further, the delay minutes at different intervals of time are predicted and regressor is found to perform well when the delay ranges from 15-300 minutes . The performance is poor when delay is greater than 1400 minutes as shown in Figure7. The former constitutes large insatnces of data whereas the latter only a few. Thus, the regressor performs well for good amount of data.

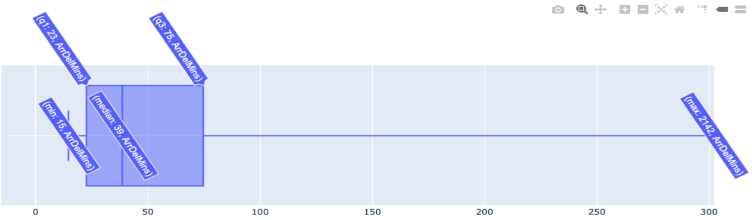


Figure 7: Analysis of actual delay minutes

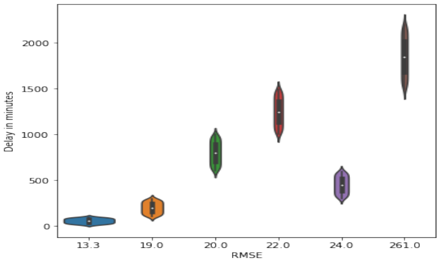


Figure 8: Regressor performance at different intervals

8. Conclusion

This project presented a methodology for predicting flight delays by exploring supervised learning methods. The designed prediction involve classification and regression tasks in which XGBoost classifier and Gradient boosting regressor is found to perform well with 92% and 94% accuracy. The dataset was explored which showed how various features influenced the target variable. The analysis shows how each clasifers perform. Accuracy can further be improved with data imputation, feature engineering and hyperparameter tuning.