FLIGHT DELAY PREDICTION

Sri Gayathri Devi I

College of Engineering, Guindy

Abstract

Analysis of flight delay and its causal factors is crucial in maintaining airspace efficiency and improving the tactical and operational decisions of airports. The primary objective of this project is to build a two-stage predictive machine learning model. Status of flights and their delay duration is predicted by models, their performance was improved with analysis and incorporating techniques such as sampling, feature engineering and hyperparameter tuning.

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. In order to alleviate these negative impacts and satisfy increasing demand, an accurate prediction of flight delay is required. Delays occur due to weather disruptions, seasonal demands, airline policies and technical issues among which, weather conditions are often cited as one of the main reasons. Thus, the weather report of flights is utilised to build a predictive model.

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience. **Supervised machine learning** is a task of learning a function that maps an input to an output based on sample input-output pairs. It approximates a function that maps the features to the ground truth in the training data. For instance, a datapoint containing status of flights with its influencing features can be trained to predict future delays. Here, the feature set of delayed and on time flights are well labelled. The first stage deals with classifying flights to be delayed or not, followed by prediction of delay duration using regression.

2. Dataset and data preprocessing

The dataset is taken from World Weather Online API which provides hourly synoptic weather observations and BTS (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 arrival delay prediction.

- Weather data: Weather data comprises of 25 hourly features and 6 daily features. The features considered are shown in Table 1.

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

Table 1: Weather Attributes

- Flight data: Flight data consists of on-time performance of flights. The features in Table 2 are taken into consideration.

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

Table 2: Flight Attributes

The data was then preprocessed to focus the prediction of arrival delay. The flight details are merged with its corresponding weather information based on flight date, departure time and origin airport so that, a flight flying at a particular time, originating from a particular airport, will have the corresponding hour's weather information.

3. Classification

Classification is the process of categorizing a given set of data into classes. The flights are categorized into delayed and on time flights with ArrDel15 as target variable.

3.1 Feature Selection

The features which contribute the most for prediction of arrival delay is selected using following feature selection techniques.

- Univariate Selection (SelectKBest with ANOVA): The strength of relationship of the features with its target variable is estimated as scores using a statistical test ANOVA. ANalysis Of VAriance is a parametric statistical hypothesis test for determining whether the means from two or more samples of data come from the same distribution or not. Among 30 features, 20 best features are extracted based on their scores, higher the score, greater is its strength of relationship with target variable as shown in Table 3.
- Feature Importance: The *importance score* for each feature is calculated based on its ability to make key decisions with boosted decision trees. The relative rank (i.e. depth) of a feature used as a decision node in a tree is used to assess the relative importance of that feature with respect to the predictability of the target variable. 20 features are extracted based on their scores where higher the score, more important is the feature as shown in Figure 1.

Features	Scores
DepDelayMinutes	7.869E5
DepTime	4.516E4
CRSDepTime	3.046E4
time	2.980E4
DayofMonth	1.602E4
humidity	1.456E4
WinddirDegree	1.349E4
WindGustKmph	1.200E4
pressure	1.016E4
DestAirportID	4.306E3
DewPointF	2.699E3
weatherCode	2.652E3
tempF	2.324E3
WindChillF	1.900E3
Month	1.007E3
Quarter	9.549E2
windspeedKmph	8.859E2
Year	7.021E2
OriginAirportID	6.950E2
visibility	1.953E2

Table 3: Scores of Univariate selection

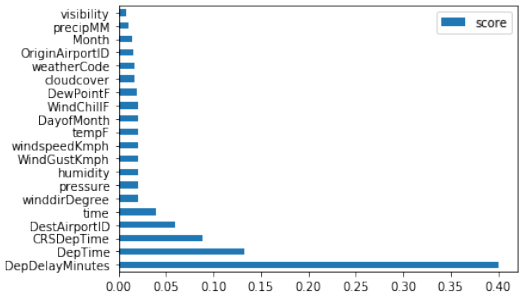

Figure 1: Feature Importance

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

Table 4: Top 17 correlated features

The foremost 17 features from the intersection of above methods are considered as final set of features as shown in table 4.

3.2 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).

3.3 Classification metrics:

- Confusion Matrix: A summary of prediction results which contains the number of predicted samples that belong to,
    1. **True Positive** - Instances are observed and predicted as on time flights.
    2. **False Negative** - Instances are observed as on time but predicted to be delayed.
    3. **True Negative** - Instances are observed and predicted as delayed flights.

4. **False Positive** - Instances are observed as delayed but predicted to be on time.

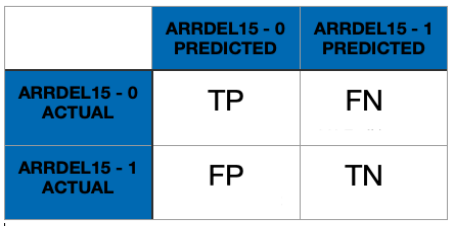

Figure 2: Confusion matrix

- Precision

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{1}$$

- Recall

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2}$$

- F1 score

$$F1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3}$$

The performance of Logistic regression, XGBoost, DecisionTree, GradientBoost and ExtraTrees classifiers are shown in Table 5. Recall scores are considered for model evaluation as avoiding the prediction of on time flights as delayed is essential and of more concern to our case of delay prediction. Extra trees classifer has the highest recall score.

Classifiers	Precision		Recall		f1-score	
	On-time	Delayed	On-time	Delayed	On-time	Delayed
XGBoost	0.92	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
Logistic Regression	0.92	0.89	0.98	0.69	0.95	0.77
ExtraTrees	0.92	0.87	0.97	0.71	0.95	0.77

Table 5: Performance of classifiers on unsampled data

A binary classification was performed on label, ArrDel15. Figure 3 shows that the dataset is highly imbalanced where number of instances of class - 0 comprises 79% of data. In such cases, it is assumed that the minority class is class - 1 and majority class is class - 0 and on modelling classifiers, they are likely to predict most of the points as majority class. Sampling techniques are used to overcome the complication of skewed distribution.

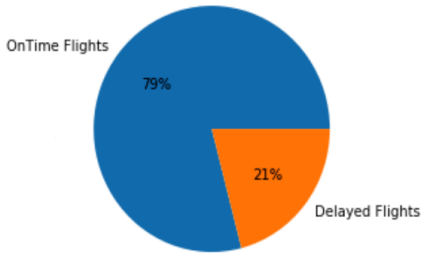

Figure 3: Distribution of delayed and on-time flights

3.4 Data Sampling

Sampling is a technique that is designed to change the distribution of population of classes in a data. Following methods were used to sample the data.

- Random Under-Sampling: Balances the class distribution through random elimination of majority class examples.
- ADASYN:Adaptive Synthetic method produce an appropriate number of synthetic alternatives for each minority class observation. It uses a density distribution as a criterion to automatically decide the number of synthetic samples to be generated. The weights of different minority samples are adaptively changed to compensate the skewed distributions.
- SMOTETomek: In **SMOTE**(oversampling technique), k-nearest neighbours are identified in feature space (typically k=5) and a line is drawn between the examples where new samples are generated at a point along that line. **Tomek links** refers to a method for identifying pairs of nearest neighbors in a dataset that have different classes. Removing all cross-class nearest neighbors in majority class that are closest to the minority class has the effect of making the decision boundary in the training dataset less noisy or ambiguous. In **SMOTETomek**, data is initially oversampled with SMOTE and then Tomek links are applied as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.

Classifiers (RandomUnderSampling)	Precision		Recall		f1-score	
	On-time	Delayed	On-time	Delayed	On-time	Delayed
XGBoost	0.95	0.73	0.92	0.80	0.93	0.76
DecisionTree	0.94	0.50	0.79	0.80	0.86	0.62
GradientBoost	0.94	0.74	0.92	0.79	0.93	0.76
Logistic Regression	0.94	0.74	0.93	0.78	0.93	0.76
ExtraTrees	0.95	0.67	0.89	0.82	0.92	0.74

Table 6: Classification results (Random undersampling)

Classifiers (ADASYN)	Precision		Recall		f1-score	
	On-time	Delayed	On-time	Delayed	On-time	Delayed
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
Logistic Regression	0.95	0.66	0.89	0.79	0.92	0.72
ExtraTrees	0.93	0.81	0.96	0.73	0.94	0.77

Table 7: Classification results (ADASYN oversampling)

Classifiers (SMOTETomek)	Precision		Recall		f1-score	
	On-time	Delayed	On-time	Delayed	On-time	Delayed
XGBoost	0.92	0.89	0.98	0.69	0.95	0.78
DecisionTree	0.92	0.69	0.91	0.71	0.92	0.69
GradientBoost	0.93	0.82	0.96	0.74	0.94	0.78
Logistic Regression	0.94	0.74	0.93	0.78	0.93	0.76
ExtraTrees	0.93	0.83	0.96	0.73	0.94	0.77

Table 8: Classification results (SMOTE-TOMEK combined sampling)

From tables 6, 7 and 8, we infer that Extra trees classifier with Random under sampling gives preferable results based on its recall score. It has a better recall for delayed class compared to other classifiers by a large margin. Correct prediction of delayed flights is essential in our case of delay prediction.

4. Regression

Regression is a technique that predicts values of a desired continuous target quantity by establishing a relationship between dependent and independent variables. Regressors are fit on training set and their performance was evaluated based on following metrics. The behavior of various regressors are shown in Table 9.

4.1 Regression metrics

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Y_i - \hat{Y_i} | \tag{4}$$

- Root Mean Squared Error(RMSE)

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2} \tag{5}$$

- R2 score

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2} \tag{6}$$

where Y refers original values, $\hat{Y}$refers predicted values and n refers to the total samples in a dataset.

Regressor	MAE	RMSE	R2 Score
LinearRegression	14.64	20.01	0.9213
ExtraTreesRegressor	12.03	17.17	0.9422
GradientBoostingRegressor	11.76	16.97	0.9436
RandomForestRegressor	11.90	17.07	0.9429

Table 9: Regression results

The residual metrics MAE and RMSE are used to compare the performance of regressors and r2 score signifies how well the model fits to dependent variables. From Table 9, is can be inferred that Gradient Boosting Regressor has the highest r2 score (0.943) and minimum MAE (11.74) and RMSE (16.97).

4.2 Cross validation and hyperparmater optimization

In order to validate the stability of the model, K-fold cross validation was performed on each regressor. The entire dataset is split into k-folds(k=10). The first fold is treated as a validation set, and the model is fit on the remaining k-1 folds and repeated until every k-fold serve as the validation set as depicted in Figure 4. Figure 5 shows the cross validated r2 scores of Linear, Extra trees, Gradient Boosting and Random forest regressors.

Gradient boosting regressor performs well with inter quartile range lying between 0.942 to 0.947. Thus, Gradient boosting regressor has made reliable predictions.

Figure 4: K-fold cross validation

Hyperparamter optimization is the process of choosing a set of optimal hyperparameters for a learning algorithm. These hyperparamters are tuned to optimize the performance of the model. GridSearchCV was performed on gradient boosting regressor where a model is built for each possible combination of all defined hyperparameter values, evaluated and the parameter set which produces the best results was selected. The tuned algorithm from GridSearchCV gives an r2 score of 0.96.

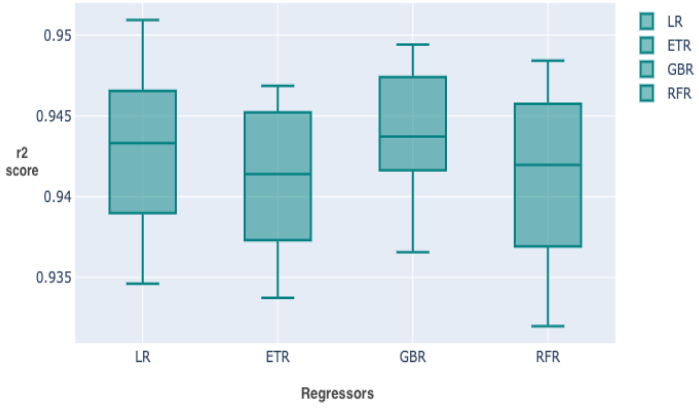

Figure 5: Performance of Regressors through pipeline

4.3 Regression Analysis

The distribution of actual delay minutes is shown in Figure 6. Though the delay value ranges from 0-2142, the frequency of each range is different and unevenly distributed. In order to evaluate the performance of model in each interval, gradient boosting regressor was tested on various range of inputs and their results are shown in Table 10.

Arrival Delay Minutes	15-100	100-300	300-600	600-1000	1000-2200
RMSE	13.34	18.01	21.22	19.29	114.87
MAE	9.94	13.53	15.24	14.87	107.67

Table 10: Gradient Boosting Regressor at different delay intervals

It can be seen from Figure 6 that most of the datapoints are concentrated in the range 15-300 minutes. The error values in this range are very low (closer to 0), inexact 13 RMSE for 100 minutes delay and 18 RMSE for 300 minutes delay. 18 minutes deviation for 300 minutes is a small value which implies that the predicted delays are closer to observed delays. Such low values mean that the regressor makes good predictions in this range. Further, the residuals are minimum for delay greater than 600 minutes which comprises 0.2% of data (900 instances). This implies that the model performs well not only on abundant data but also on sparse region of data. Therefore, the model is robust in such unevenly distributed data.

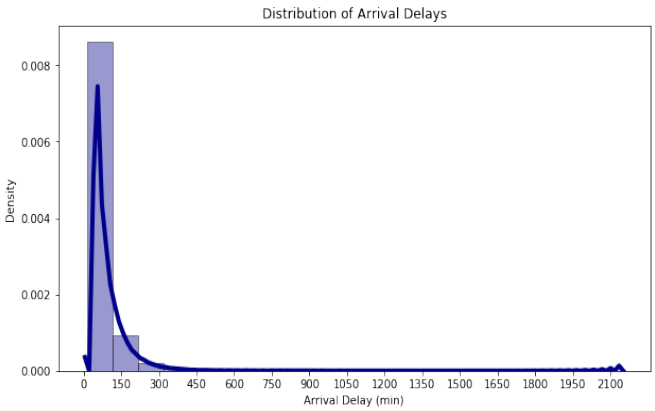

Figure 6: Distribution of actual delay minutes

5. Conclusion

The paper presented a methodology for flight delay prediction by exploring supervised learning methods. Extra trees classifier performs well with 90% accuracy and recall of 0.82 on de1ayed class. Gradient boosting regressor is reliable with minimum MAE (11.74) and RMSE (16.97). Its performance was further optimized with hyperparameter tuning where the r2 score improved from 0.94 to 0.96. Hence, the predictive analysis inferred from this model can help in optimizing airline and airport operations.