Enhancing Flight Delay Predictions through Combined Classification and Regression Methods

P.Prithvikiran

ABSTRACT

Flight delays have significant consequences for passengers, airlines and the aviation industry. Predicting these delays accurately is crucial for optimizing airline operations and improving the overall passenger experience. This involves a two-stage machine learning engine that forecasts the on-time performance of flights. Through the integration of classification for categorical delay outcomes and regression for continuous delay duration, our methodology offers a comprehensive and precise prediction framework. This Project contributes to the improvement of aviation operations and passenger experience by enhancing the accuracy of flight delay forecasts.

1 Introduction

The worldwide importance of air travel underscores the critical requirement for precise predictions of flight delays. Flight delays, which can result from a multitude of factors, presenting significant challenges for both airlines and passengers. Traditional prediction models often categorize delays or estimate duration independently, neglecting the complexity of delay occurrences. Our project addresses this gap, combining classification and regression techniques to create a comprehensive flight delay prediction model. The aim is to develop a unified framework that not only classifies flights into delay categories but also estimates delay durations Leveraging historical flight data and other relevant features, this approach promises to enhance decision-making for airlines and improve passenger experiences. This project's significance lies in its potential to optimise airline operations and passenger satisfaction. The following sections include the methodology, data processing, model development and implications, setting stage for a holistic approach to flight delay prediction.

2 DATASET

The dataset contains extensive information on flight data, including arrival and departure details of flights. Additionally, it includes weather data providing information on various parameters for specific airports during particular time periods. This dataset is sourced from the Transtats data library, which encompasses data from 15 airports within the USA during the years 2016 and 2017.

Data preprocessing was instrumental in optimising our dataset for machine learning analysis. This extensive dataset underwent rigorous preprocessing procedures, Feature extraction, label encoding and normalisation, streamlined the data, making it more manageable and model-ready. Removal of null values enhanced data completeness, reducing the risk of erroneous predictions. Merging datasets using the inner method ensured accurate alignment of flight and weather information. Collectively, these tasks improved data quality and paved the way for a more accurate and effective flight delay prediction model. Tables 1, 2, and 3 below display the airport codes along with their corresponding flight and weather columns for these 15 different airports.

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

Table 1: Airport Codes

WindSpeedKmph	WindDirDegree	WeatherCode	precipMM
Visiblity	Presssure	Cloudcover	DewoPoint
WindGustKmph	tempF	WindChillF	Humidity
date	time	airport	

Table 2: Recommended Weather Columns:

FlightDate	Quarter	Year	Month
DayofMonth	DepTime	CloudCover	DewPointF
DepDelMinutes	OriginalAirportID	DestAirportID	arrTime
CRSArrTime	ArrDel15	ArrDelayMinutes	

Table 3: Recommended Flight Columns:

3 CLASSIFICATION

In machine learning, classification is a supervised learning task where algorithms learn to categorise data points into predefined classes based on input features, often employing mathematical decision boundaries. In a classification task, the target variable, also referred to as the dependent variable or label, represents the outcome or category of interest that the machine learning model aims to predict. Binary classification involves two possible classes (e.g., “Delay” or

”On Time”) . The target variable serves as the basis for the model’s learning process, offering ground truth labels for training data, enabling pattern and relationship recognition between input features and the desired category. The ultimate objective of classification is to construct a model capable of accurately assigning class labels to new or unseen data instances. Here *ArrDel15* is the target variable.

Classification metrics are performance measures used in machine learning and statistics to evaluate the performance of a classification model. Here are some of the metrics used .

• Precision : Precision is the ratio between the true positives and all the positives. In context to the problem statement, that would be the flights that were correctly identified as delayed of all the flights which were actually delayed.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

• Recall : The recall is the measure of the model correctly identifying true positives. Thus, for all the flights that were delayed, recall shows how many flights were correctly identified as delayed.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

• F1-score : The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean.

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

• Accuracy : Accuracy is the ratio of the total number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

3.1 DATA IMBALANCE

A classification dataset with imbalanced class proportions is referred to as imbalanced dataset. The classes that constitute a significant proportion of the dataset are known as majority classes, while those with a smaller representation are considered minority classes. In our observation, imbalanced dataset for the target variable *ArrDel15* where 0 represents On Time and 1 indicates Delayed flights is identified.

From Figure 1, significant disproportion in the data points is observed , with 0 representing approximately 79.90% and 1 accounting for about 20.1%. This observation confirms that the data is highly imbalanced.Such imbalanced data can negatively impact the training model, resulting in highly biased predictions favoring the majority class. To address this issue, data imbalance sampling techniques is carried out, which balances the imbalanced data by ensuring a

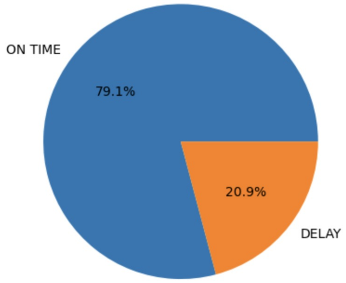

Figure 1: Data before SMOTE

50:50 ratio for both data points within the variable. The different types of sampling techniques are :

• Under-sampling : In this sampling technique, the number of samples in the majority class is decreased while keeping the minority class unchanged.

• Over-sampling : In oversampling, all of the data in the majority class is kept while increasing the size of the minority class.

In our case, an Over sampling technique called SMOTE (synthetic minority oversampling technique) an Under sampling technique called Random Under sampler are implemented.

- SMOTE is one of the most commonly used oversampling methods to address the imbalance problem. Its goal is to balance the class distribution by increasing the representation of the minority class. This is achieved by generating synthetic instances within the minority class, essentially creating new data points between existing minority instances.

- It achieves this by synthesizing new minority instances through linear interpolation. It randomly selects one or more of the k-nearest neighbors for each example in the minority class to generate these synthetic training records.

-RANDOM UNDER SAMPLER is an under sampling technique used in the context of addressing class imbalance in binary classification problems. It's a method designed to mitigate the impact of class imbalance by randomly reducing the number of instances in the majority class to match the number of instances in the minority class. This balancing of class distribution helps prevent the classifier from being biased toward the majority class.

-It randomly selects a subset of instances from the majority class to make it equal in size to the minority class. This random selection can involve simply discarding some instances from the majority class until both classes have roughly the same number of samples

- Following the sampling process, the data is reconstructed, and various classification models can be applied to the processed data.

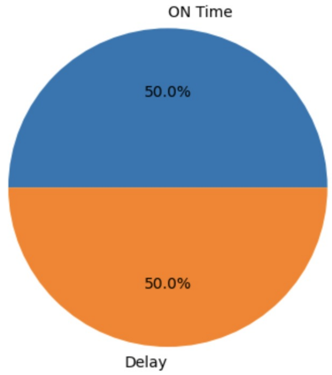

Figure 2: Data after SMOTE

After resampling, the data is evenly distributed with a 50:50 ratio, as depicted in Figure 2.

For classification, different models, including Logistic Regressor, Decision Trees Classifier, Random Forest Classifier, Extra Trees Classifier, and Gradient Boosting Classifier are implemented. The results of these classification models both oversampling and under sampling applying SMOTE and RANDOM UNDER SAMPLER are provided below. Notably, better results are obtained using the oversampling technique SMOTE.

In Table 5, after the SMOTE process, the Logistic Regression model achieves approximately 60% accuracy, the Decision Tree Classification model achieves about 79.1% accuracy, the Random Forest Classification model achieves around 93.2% accuracy, the Gradient Boosting Classification model attains about 69.2% accuracy, and the Extra Trees Classifier model reaches about 85.1% accuracy. Based on these observations, it is concluded that Random Forest Classifier is the most suitable classification model for this data set.

Classifier	Precision		Recall		f1-score		Accuracy
	0	1	0	1	0	1	
Logistic Regressor	0.61	0.61	0.61	0.61	0.61	0.61	0.61
Decision Trees	0.79	0.79	0.78	0.79	0.79	0.79	0.79
Extra Trees	0.84	0.87	0.87	0.84	0.86	0.85	0.85
Random Forest	0.90	0.95	0.96	0.90	0.93	0.92	0.93
Gradient Boost	0.69	0.70	0.69	0.70	0.69	0.69	0.69

Table 4: Result Analysis of Classification after SMOTE

Classifier	Precision		Recall		f1-score		Accuracy
	0	1	0	1	0	1	
Logistic Regressor	0.82	0.97	1.00	0.15	0.90	0.26	0.82
Decision Trees	0.92	0.67	0.91	0.70	0.91	0.69	0.87
Extra Trees	0.87	0.77	0.96	0.46	0.92	0.58	0.86
Random Forest	0.90	0.91	0.98	0.57	0.94	0.70	0.90
Gradient Boost	0.85	0.98	1.00	0.36	0.92	0.52	0.86

Table 5: Result Analysis of Classification after RANDOM UNDER SAMPLER

4 REGRESSION

Regression is a method employed to analyze the relationship between independent variables or features and a dependent variable or outcome. This analysis allows for the prediction of outcomes once the connection between independent and dependent variables has been established. Regression serves as an approach to predict continuous outcomes in predictive modeling, making it valuable for forecasting and predicting outcomes based on data. In this project, regression is used to predict the arrival delay of flights through the training of regression models.

The regression predictions are analysed by various metrics such as:

• MAE: MAE takes the average of the absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - \hat{y}_i|\right) \tag{5}$$

• RMSE : RMSE measures the root mean squared value of the errors and is concerned with the deviation of predictions from the actual value.

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

• R2 score : R-squared score is a metric that is used to measure the performance of the model. It indicates how much of the variation of a target variable is explained by the independent variables in a regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}i\right)^2}{\sum i = 1^n \left(y_i - \bar{y}\right)^2} \tag{7}$$

The dataset is divided into training and testing sets, test-train split as 4:1. By fitting the training dataset to the regression models, predictions are made using the testing dataset. Subsequently, the accuracy is calculated by comparing the predicted values to the actual values from the testing dataset. The regressors used includes Linear Regressor, Random Forest Regressor, Extra Trees Regressor, and Gradient Boosting Regressor. The results of the regression models are presented in Table 6. The predictions are evaluated using the r-squared metric since it is context-independent and easily interpretable.

From the table, the Linear Regressor achieved an r-squared score of 93.3%, the Gradient Boost Regressor scored 93.4%, the Extra Trees Regressor achieved a score of 93.5%, and the Random Forest Regressor obtained a score of 94.2%. Based on this data, it is concluded that the Random Forest Regressor is the most suitable regression model for this dataset.

Regressor	RMSE	R-square	MAE
Linear Regressor	10.71	0.93	5.69
Extra trees Regressor	10.34	0.94	5.53
Random Forest regressor	10.30	0.94	5.61
GradientBoosting	10.54	0.93	5.69

Table 6: Result Analysis of Regression

5 PIPELINE ARCHITECTURE

In machine learning, pipelining refers to the process of creating a sequence or pipeline of data processing and transformation steps that are applied in a specific order to a dataset. These steps typically include data preprocessing, feature extraction, model training, and model evaluation.

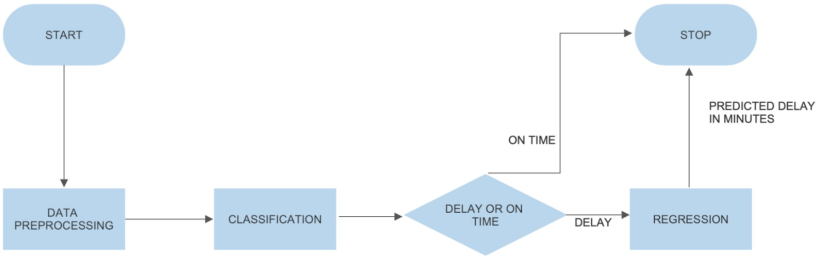

Figure 3: Pipeline Architecture

In the pipelining process, the Random Forest Classifier, which demonstrated the best performance among the classification models in the main dataset, is trained using the entire preprocessed dataset. The predicted output serves as a replacement for the *ArrDel15* column in the main data frame. The delayed flights, predicted by the classifiers are integrated with the main data frame.Subsequently, regression analysis is conducted, and the results are assessed using the r-squared (r2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) metrics.

Classifier	Precision		Recall		f1-score		Accuracy
	0	1	0	1	0	1	
Random Forest	0.99	0.91	0.98	0.95	0.98	0.93	0.97

Table 7: Pipeline Model Classification Result Analysis

Regressor	RMSE	R-square	MAE	Support
Random Forest Regressor	10.98	0.97	6.90	374918

Table 8: Result Analysis of Regression

6 REGRESSION ANALYSIS

The Random Forest Regressor, which proved to be the best-performing regression model in the pipelining process, is utilized to predict the delay minutes for delayed flights. The predicted results are categorized into intervals ranging from 15 to 100, 100 to 200, 200 to 500, 500 to 1000, 1000 to 2000, and above 2000. The best regressor is trained within each range, and the model provides predictions for each interval. These predictions are then analyzed using the r-squared (r2) score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) metrics, with the results presented in Table 8.

Range	RMSE	R-square	MAE	Support
15-100	9.85	0.80	6.59	290596
100-200	15.48	0.67	9.00	48814
200-500	17.19	0.93	9.59	14348
500-1000	13.41	0.97	8.19	1118
1000-2000	13.90	0.99	9.64	157

Table 9: Pipeline Model Regression Result Analysis

In the interval 15 to 100, R-squared (R2) score of 80.0% is achieved, which is comparatively better than the interval 100 to 200, where the R2 score is 67.0%. This difference is primarily due to the larger sample size in the 15 to 100 range, with 290,596 samples, compared to the smaller 48814 samples in the 100 to 200 range. Conversely, for intervals from 200 to 2000, the regressor excels in predicting delay minutes in the higher range. It yields an R2 score of 93.0% for the interval from 200 to 500, with a sample size of 14348, and an even higher R2 score of 97.0% for the interval from 500 to 1000, with 1,118 samples. However, the interval from 1000 to 2000 has 157 samples and the highest R2 score of 99.0%.

7 CONCLUSION

The estimation of flight delays involved a two-stage pipelining process. In the first stage, classification was employed to predict DELAYED or ON TIME flights. This was achieved by addressing data imbalance using SMOTE and Random Undersampler. The most suitable classification model, the Random Forest Classifier, achieved an accuracy of 97.0%.In the second stage, regression was utilized to predict the delay minutes for delayed flights. Among the regression models, the Random Forest Regressor emerged as the best fit with an impressive r2 score of 94.2%. Subsequently, the regression results were analyzed across various prediction intervals. While satisfactory results were obtained in all intervals, the range of 1000 to 2000 displayed exceptional performance with an r2 score of 99.0%.In conclusion, this flight delay project report offers a comprehensive analysis of the factors contributing to flight delays, their impact on passengers and airlines, and potential strategies for mitigating delays.