FLIGHT DELAY PREDICTION

GROUP 15

AY2022/2023 SEM 1

Jovinder Singh

Jung Hyun Park

Kevin Lin

Yuqi Zhang

National University of Singapore

Foundations in Business Analytics (DBA 5106)

Dr. Wong Liang Ze

18 November 2022

1 Introduction

Flight delay was an exacerbating issue in recent years, especially since the spread of the COVID-19 pandemic where lockdown and quarantine policies are widely adopted by many countries. These cross border policies lead to disturbance on flight schedules and unforeseen flight delays or cancellation. According to statistics from the Bureau of Transportation in the U.S., flight delay has increased greatly from 16.78% in 2021 to 21.26% in 2022, and the trend is projected to be higher in the future (United States Department of Transportation, 2022). Besides the impact of the external reasons, such as global pandemic, flight delay can also be caused by internal management policies, explained by failed attempts to optimise the tradeoff between enough buffer time and maximise fleet utilisation to earn higher profit (Wu, 2005). Moving beyond the causes of flight delay, the cost of flight delay has also been studied vigorously by many managers and economists. According to the data from Airlines for America, the average cost of flight delay for the U.S. in 2021 was $80.52 per minute, and the major direct airline costs include labour expenses, fuel costs, as well as maintenance and aircraft ownership (Airlines for America, 2022). Indirect costs are more to do with consumers, such as the loss of trust and compromised labour productivity for business travellers. Hence, the same study claims that the net welfare in the U.S. would increase by $17.6 billion for a 10 percent reduction in flight delay (Peterson et al., 2022).

In this report, we will adopt an analytical perspective to predict flight delay due to external and internal management reasons. Prediction results will also be interpreted with business contextual knowledge in terms of profit-loss matrix to generate profit values that can be better utilised by managers in decision-making.

2 Literature Review

Numerous studies have been dedicated to model flight scheduling and predict potential delay. Most of them choose to study the effect of different flight conditions and how they will contribute to flight delay. These studies can be summarised into three categories, including statistical methods, operational methods and machine learning methods. Due to the nature of this course, this report will focus on operational and machine learning methods.

One of the best studies in operational method investigates schedule optimization using Markov Chain algorithm to simulate the stochastic nature of aircraft operations in a network (Peterson et al., 2022). It proposed an important concept called "propagation of flight delays" from a delayed flight to other flights via affected schedules, passengers, and crew in the same network. It concludes that with more buffer time in between flights, whenever the delay can be predicted, flight and passenger cost can be minimised.

Besides operational methods, there are also many machine learning methods, converting techniques, such as deep learning, SVR algorithm, logistic regression, decision tree regression, LSTM-AM, and clustering. A deep learning paper deals with highly complex and massive flight information data by automatically extracting important features from an imbalance dataset (Yazdi et al., 2020). With under-sampling and de-noising methods, the proposed model has higher accuracy than other models. Another paper chooses a long short-term memory network with attention mechanism (LSTM-AM) model to predict flight delay (Wang et al., 2022). It distinguishes direct factors (i.e. weather and holidays) from indirect factors (i.e. delay from previous schedule) and uses prediction results to formulate actionable suggestions in face of flight delay. A generalised machine learning paper compares almost all popular machine learning methods, such as logistic regression and decision tree, and concludes that random forest regression was the best model in predicting delays (Meel et al., 2020).

The above papers are doing well in models and statistics. However, they all face the same limitation of being unable to translate the prediction results in dollar-and-cents terms. For example, the most popular metrics among all papers is accuracy score. However, accuracy alone is insufficient to translate into business terms. Therefore, our paper aims to explore different machine learning methods with a cost-benefit matrix to quantify the prediction results. The research questions can be summarised below:

Research Question 1: Which machine learning model is the most accurate and efficient in prediction flight delay.

Research Question 2 How to use the prediction results in minimising the delay cost?

3 Methodology

The methodology deployed in this study follows the flow chart (Figure 1). The study first investigated the aviation industry regarding flight delays to achieve business understanding. As discussed earlier, flight delays incur immense economic loss, not only to airlines but also to societal welfare. To materialise the loss, we built the costs and benefits matrix with regard to flight delays.

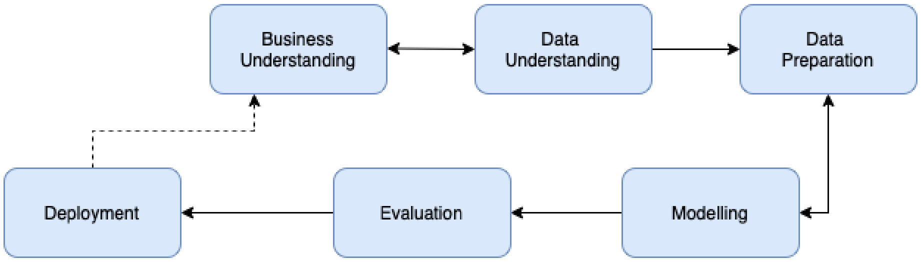


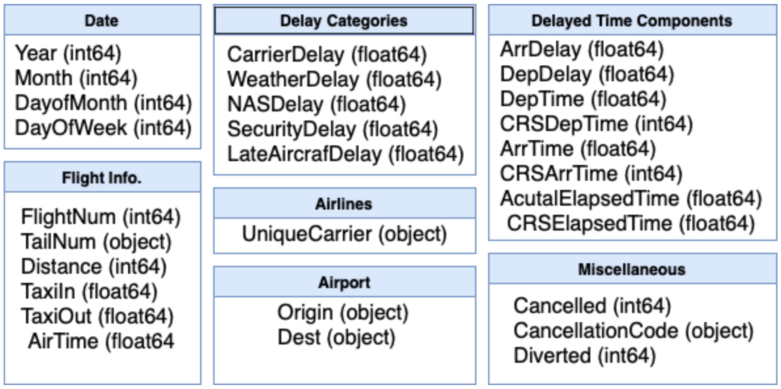
Figure 1. Analysis Flow Chart

The data used in this study was obtained from Kaggle. The dataset was originally created by the Bureau of Transportation Statistics (BTS) under the US Department of Transportation (DOT). BTS built this specific dataset to analyse the on-time performance and causes of flight delays of major airlines in the US in 2008. The target of this study is to predict whether a particular flight will be delayed before its departure. This information is critical for future resource allocation optimization and delay cause analysis to reduce flight delays. The details of the data features and the data preparation process for the study will be further discussed in the Data Preparation section under EDA.

For modelling and evaluation, various models and techniques, such as logistic regression, support vector machine (SVM), decision trees, and random forest classifiers, to name a few, were tried to find out which model performed the best with the dataset. The "Classification Report" in scikit-learn was used to generate different metrics to evaluate each model's performance. Furthermore, the deployment stage has been produced by adopting the train-test split approach. The model that demonstrates the best performance will be further evaluated based on the precision-recall curve, the ROC curve, the AUROC score, and the feature importance test technique.

3.1 Exploratory Data Analysis

The data used in this study were acquired from the Kaggle website (Gonzalez, 2019). The dataset contains twenty-nine different features, which include five different delay categories defined by the BTS along with the twenty major airline companies in the US (Figure 2). The airports analysed in this dataset are three hundred and three at the origin airport and three hundred and four at the destination airport. The features included in the data are as follows:



Date
Year (int64)
Month (int64)
DayofMonth (int64)
DayOfWeek (int64)

Flight Info.
FlightNum (int64)
TailNum (object)
Distance (int64)
TaxiIn (float64)
TaxiOut (float64)
AirTime (float64

Delay Categories
CarrierDelay (float64)
WeatherDelay (float64)
NASDelay (float64)
SecurityDelay (float64)
LateAircrafDelay (float64)

Airlines
UniqueCarrier (object)

Airport
Origin (object)
Dest (object)

Delayed Time Components
ArrDelay (float64)
DepDelay (float64)
DepTime (float64)
CRSDepTime (int64)
ArrTime (float64)
CRSArrTime (int64)
AcutalElapsedTime (float64)
CRSElapsedTime (float64)

Miscellaneous
Cancelled (int64)
CancellationCode (object)
Diverted (int64)

Figure 2. Tables of Features

3.1.1 Data Investigation

The first two rows of the dataset were indices and the year the data were collected, respectively. The dimension of the dataset is 1,936,758 by 29 when opened with DataFrame, excluding the index column. As the target of the study is to predict flight delays before departure, the features delineating the delay need to be described in detail. There are five components, selected by the BTS that determine whether a flight record to be classified as delayed. The components are "CarrierDelay", "WeatherDelay", “NASDelay (National Aviation System Delay)”, “SecurityDelay”, and “LateAircraftDelay” (“Airline On-Time Performance”, 2022). They are defined as following:

- CarrierDelay: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- WeatherDelay: Significant meteorological conditions (actual or forecasted) that, in the judgement of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- NASDelay: Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- SecurityDelay: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of twenty nine minutes at screening areas.
- LateAircraftDelay: A previous flight with the same aircraft arrived late, causing the present flight to depart late.

Furthermore, we discovered that the dataset was imbalanced as we conducted the exploratory data analysis. This imbalance was present due to the definition of flight delay stated by the BTS. According to the organisation, a flight is counted as "delayed" if it operated fifteen minutes later than the scheduled time shown in the carriers' computerised reservations systems (CRS) (“Airline On-Time Performance”, 2022). Therefore, if the delay time exceeds fifteen minutes, which is determined by a combination of the above five delay features, the flight is classified as delayed (equivalent to 1.0) during the initial EDA (figure 3).

To address the imbalanced data issue, we have adopted a suggestion provided by OAG, a UK-based global travel data provider. According to OAG, it is more appropriate and realistic to set the flight delay threshold to thirty minutes instead of fifteen minutes (“Defining Late”, 2019). The reasoning behind this suggestion is that, psychologically, most passengers do not regard a flight as delayed if the delay time is around fifteen minutes. That is because it is common to have a delay within the fifteen-minute range of the scheduled time for numerous flights. Only after the flight exceeds the thirty-minute mark will passengers consider the flight delayed, according to OAG. After incorporating this new delay time threshold, we were able to settle the imbalanced data problem (figure 4). As shown in the histogram below, the delay class is now in the minority, and the classes are not as highly imbalanced as before.

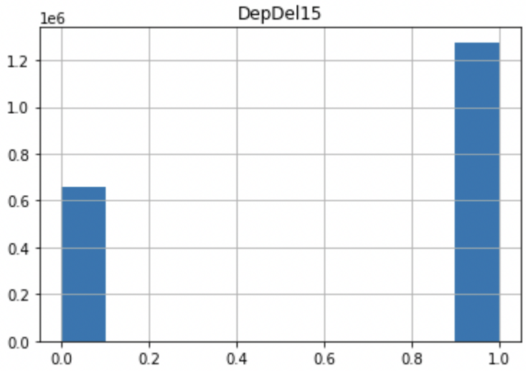


Figure 3. Class Histogram Before Adjustment

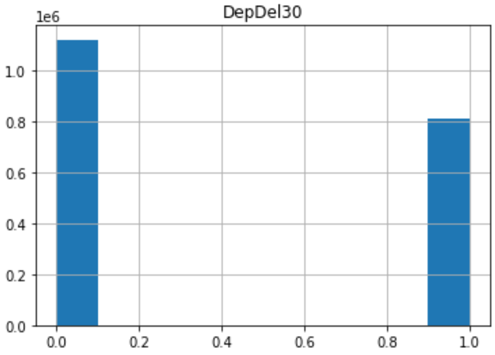


Figure 4. Class Histogram After Adjustment

3.1.2 Trends and Pattern

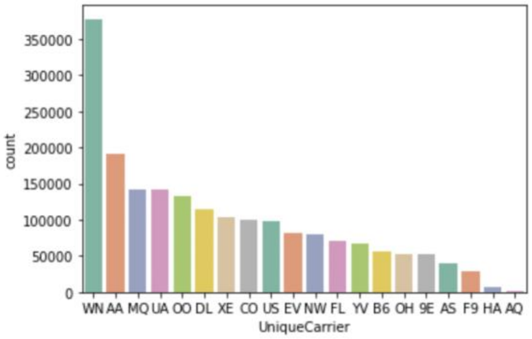


Figure 5. Flight Frequency of the Twenty Unique Carriers

There are twenty unique carriers present in our dataset. Among them, WN, which is the IATA (International Air Transportation Association) code for Southwest Airlines, is the most popular airline. The number of flights operated in 2008 is denoted by a count on the y-axis (figure 5). We can see that the number of flights operated by WN is almost twice as many compared to AA (American Airlines), the next most-operated airline, and more than twice as many relative to the rest of the airline companies in our dataset.

Though this study focuses on the delay before departure, regardless of whether it is an after-arrival or before-departure delay, we were able to identify that the delays were skewed to the right (figure 6). This indicates that most flight delays are short in duration.

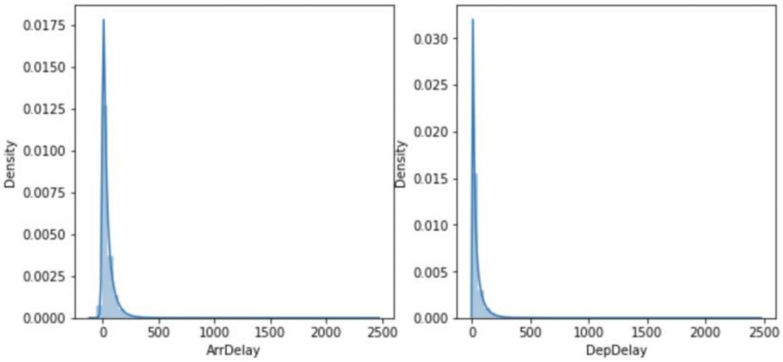


Figure 6. Delay Records Distribution

Furthermore, each aircraft carries a unique identifier, which is often called the tail number of an aircraft. An analysis between the tail number of each aircraft and the delay time reveals that the aircraft that are used more frequently (shown on the left) tend to have a consistent delay time of more than thirty minutes (figure 7). On the other hand, aircraft that are less used display fluctuating delay times, as shown on the right pie chart.

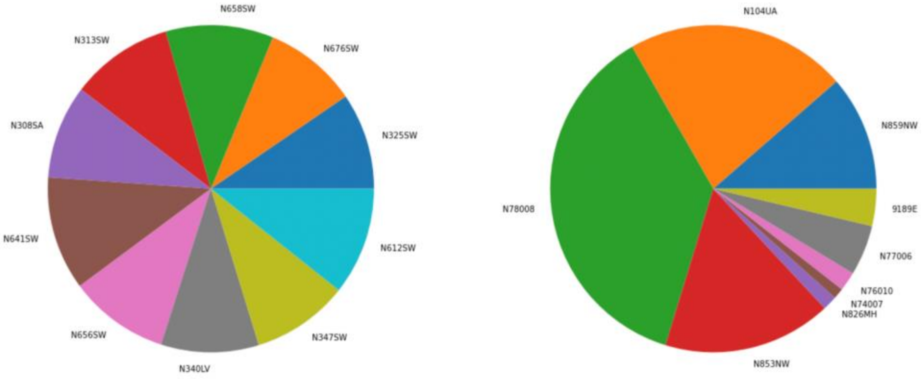


Figure 7. Pie Charts Reflecting Relationship Between Delayed Time and Popularity of Aircraft

We also composed a correlation map regarding the five delay variables to verify a multicollinearity problem among them. The correlation heat map above discloses that there could be some degree of correlation between each delay variable, but the correlation is weak (figure 8).

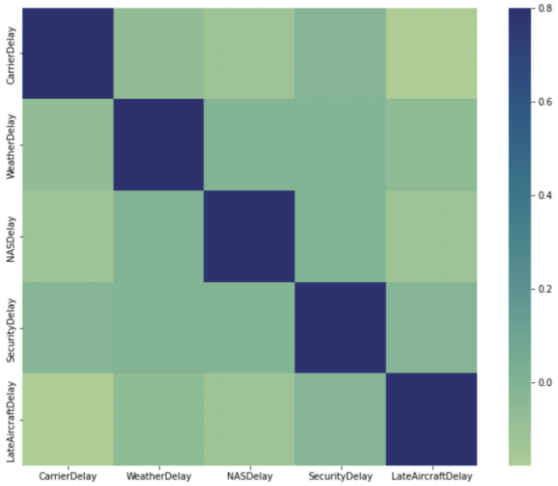


Figure 8. Correlation Heatmap of the Five Delay Variables

3.2 Feature Engineering

The features present in the dataset vary across different types of categories such as time, numerical, and categorical. As the dataset is structured by time from the start to end of 2008, we aim to capture this time relation through a cyclic transformation of time features. These columns are such as scheduled departure/arrival time, day of month, and day of week. To enable this, we split the scheduled departure and arrival times into hours and minutes. Following this, we transform each of them into their sine and cosine components so as to capture the cyclic nature of time.

Assuming that weekends might play an impact on flight delays, we created a new feature 'is_weekend' if the day of the week falls on Saturday or Sunday. As the Distance feature between the origin and destination airports is wide and has a high standard deviation, we have chosen to log it. This enables it to approximately conform to normality.

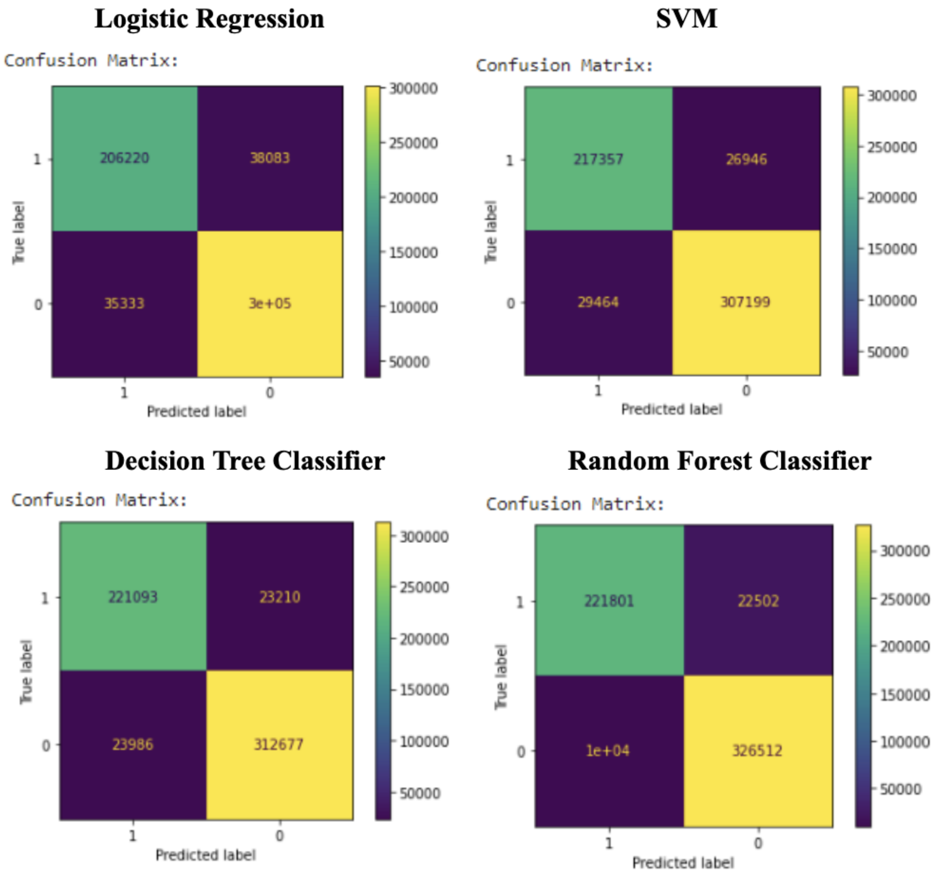
The Carrier, Origin and Dest columns are of a categorical nature with the last two having immensely high cardinality. For the Carrier feature, after a train-test split, we explored the value counts of the airlines and selected the 25th and 75th quantile to bin the airlines based on its popularity: popular, average, unpopular. This was hardcoded and this transformation was done in the pipeline. The Origin and Dest features contained over 200+ airports with IATA codes. To categorise them, we cross referenced the IATA code with airport types such as small, medium, and large and binned them respectively. In addition, we calculated the grouped mean of the Delay by the Origin and Dest airports in the training dataset and added it as a numerical feature. The TailNum categorical feature with over 5000+ unique values, was converted into a numerical feature using the mean of the departure delay grouped by each individual tail number in the training dataset. As the categorical columns did not contain an ordinal ranking, they were one-hot encoded. We used features that are available during the time of departure for our model training thus to not induce 'leakage' into our model. The last step in feature engineering was to Min-Max scale the numerical features.

4 Model Selection and Model Evaluation

We experimented with four different models to capture flight delay, and they are: logistic regression, SVM, decision tree classifier, and random forest classifier. We used ***DepDel30*** as the target, and a flight delay is defined as when the actual departure time of a flight is greater than 30 minutes from its scheduled departure time. In this circumstance of flight delay, ***DepDel30*** will take the value of 1 and vice versa. As class = 1 is a minority class and is the class we are interested to predict, we observe the evaluation scores for this class in particular (Table 1) and respective confusion matrices (Figure 9).

	Cross validated mean F1 Score	Balanced Accuracy	F1 Score
Logistic Regression	0.853	0.87	0.849
SVM	0.885	0.901	0.885
Decision Tree Classifier	0.902	0.917	0.904
Random Forest Classifier	0.93	0.939	0.931

Table 1. Evaluation Scores of Different Models



Logistic Regression

SVM

Decision Tree Classifier

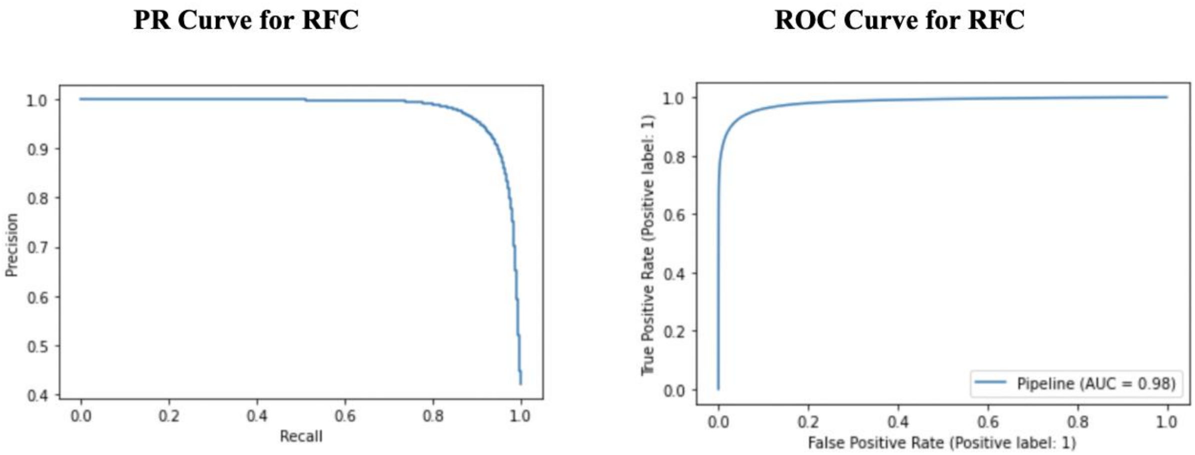
Random Forest Classifier

Figure 9. Confusion Matrix of Different Models

Exploring the cross-validation scores among the trained classifiers, the Random Forest Classifier (RFC) scored the highest at 0.93. Based on the test dataset, the RFC also has the best performance with the highest test F1 score for the positive class. It also has the highest test Balanced Accuracy.

In the context of this project, False Negatives (FN) mean that there is an actual delay that was not predicted, this will have the highest cost as shown by the cost benefit matrix below. Thus as FNs are highly costly, we aim to have a high recall score for the positive class.

Moreover, we will further explore the precision-recall (PR) curve and receiver operating characteristic (ROC) curve of RFC (Figure 10). The PR and ROC curves above are in relation to the positive class (flight delays) of the random forest classifier. In these graphs, the Recall and TPR have the same meaning. Thus it can be viewed from the PR and ROC curves that the model performs well. The area under the ROC curve (AUROC) is 0.985, which is very high.


PR Curve for RFC

ROC Curve for RFC

Figure 10. PR and ROC Curves of RFC

4.1 Feature Importance

We extracted the feature importances for the RFC and noticed that the time based transformed features are important (Figure 11). The remaining transformed features also contribute to the model's predictions albeit at a lower rate. The delay propagation due to late aircraft delay is the most important feature because one late arriving carrier is the direct cause of the following departure delay.

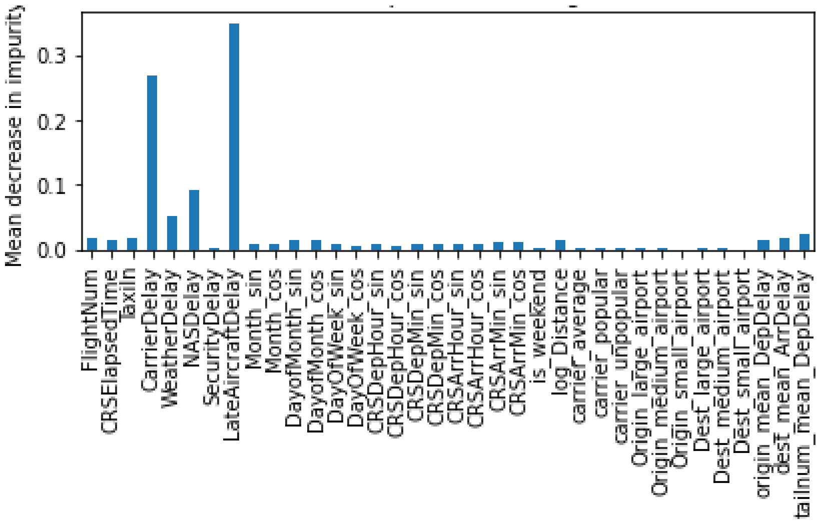


Figure 11. Feature Importances Using MDI

In addition, CarrierDelay being within the control of the air carrier due to awaiting passengers, crew, slow boarding, and aircraft inspections is the second most important feature. This rationale makes for a practical understanding that a flight can be potentially delayed.

4.2 Hyperparameter Tuning

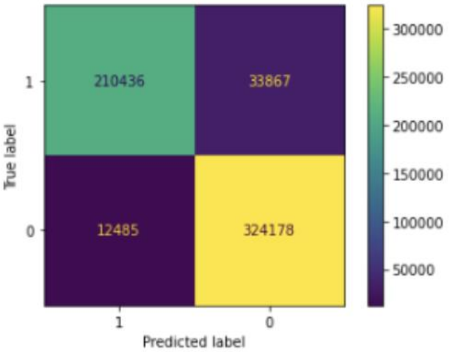
In order to further improve our Random Forest Classifier, we implemented hyperparameter tuning. We utilise the HalvingGridSearchCV to speed up the hyperparameter tuning selection as our dataset contained a large number of rows. This method uses Successive Halving where it uses a subset of the data early in the process to find some of the best performing parameter combinations and gradually increases the amount of data used as it narrows in on the best combinations. We used the following hyperparameters for tuning:

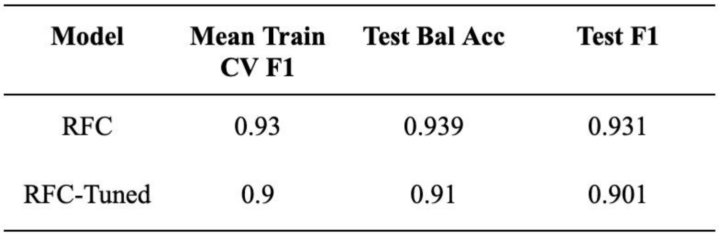
- Number of Estimators: [50, 100, 200]
- Max Features: ['sqrt', 'log2']
- Max Depth: [3, 5, 8]
- Min Samples Split: [2, 5, 10]

The results are as follows: Best Hyperparameters = {

'max_depth': 8, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 100

}. The result confusion matrix of the tuned RFC and evaluation scores can be seen below (Figure 12).





Model	Mean Train CV F1	Test Bal Acc	Test F1
RFC	0.93	0.939	0.931
RFC-Tuned	0.9	0.91	0.901

Figure 12. RFC-tuned Confusion Matrix and Evaluation Scores

The use of HalvingGridSearchCV for hyperparameter tuning led to a decrease in the mean CV score on the training data, test F1 score and test Balanced Accuracy. In addition, the Precision and Recall scores for both classes decreased marginally. The numbers of TP, TN, and FP decreased while the FN increased marginally. The high number of max_depth and reduced evaluation scores on the test dataset suggests that the hyperparameter tuned model might be overfitting on the training dataset. Thus it is not looking good for the tuned model.

5. Business Application

Moving beyond the machine learning model, we will proceed to the business application of our model. We will combine the confusion matrix with a well-researched cost-benefit matrix to see how capable our model performs in cost-cutting.

5.1 Cost Matrix

Research suggests that flight delay cost has two components, and they are airline costs and customer cost. A detailed cost illustration can be seen from the introduction part of this paper. According to the most recent data from Airlines for America in 2022, the average cost of aircraft delay per minute for the airline company was $80.52. Similarly, the average cost in terms of passenger's time and trust will be $0.78 per minute (Airlines for America, 2022). Studies show that when a flight delay is announced in advance, passenger cost will be reduced by 35.56%, because they can be better prepared to reschedule their agendas. At the same time, the delay cost for airline companies will be reduced by up to 80% with proper traffic management in place to mitigate the impact of a forecasted delay (Wang et al., 2022). Hence, our cost matrix is constructed as follows (Figure 13): if a delay is not predicted, it will bear the full cost of $81.3 per minute (see the blue region). If a delay is predicted, both airline and customer costs will be discounted at 20% and 64.44% respectively, and the final cost will only be $16.6 per minute (see the green region). However, if a delay is predicted but the flight arrives on time, only a small discounted customer cost, $0.5 per minute, will be incurred (see the red region), and there is no extra fuel or crew cost. Lastly, when there is no delay and no announcement, the cost will be $0 (see the grey region). As we are determining the price of the cost incurred with this matrix, we aim to minimise the expected value by combining the cost matrix and our prediction results.

		PREDICTED	
		PP (predict delay)	PN (predict no delay)
ACTUAL	P (delay)	\$80.52×20% + \$0.78×64.44% = **\$16.6**	\$80.52 + \$0.78 = **\$81.3**
	N (no delay)	\$0.78×64.44% = **\$0.5**	**\$0**

Figure 13. Cost Matrix Calculation and Values

5.2 Expected Values

This project aims to minimise the expected cost value, so a smaller expected value is more preferred. By calculating the respective expected values of different models (Table 2), we find out that the Random Forest Classifier (RFC) has an expected value of 9.5, so it has the best practical business application in terms of cost-cutting. However, with hyperparameter tuning, the expected value increases to 10.76, as a sign of over-fitting. Although the tuned model has high evaluation scores and a high expected value, the number of False Negatives also increased through the hyperparameter tuning and their associated cost undermines an already well-performing model. In view of this, the default RFC model offers sufficient business value, and hyperparameter tuning will be unnecessary.

Model	Expected Value	Tuned Expected Value
Logistic Regression	11.25	N.A.
SVM	10.01	N.A.
DTC	9.59	N.A.
RFC	9.5	10.76

Table 2. Expected Cost Value of Models

5.3 Cost Curve

Since our cost benefit matrix outlines cost and we aim to minimise the expected value, we have renamed the Profit Curve to Cost Curve (Figure 14). A maximum expected cost value of 35 is achieved with a class probability of between 1.0 and 2.0 for the random forest classifier. It can be observed that despite class probabilities being at 0, there is a minimum cost incurred. The expected values of 9.5 and 10.76 are relatively close as compared to the maximum. Thus the threshold can be adjusted to a lower value to reduce the expected cost.

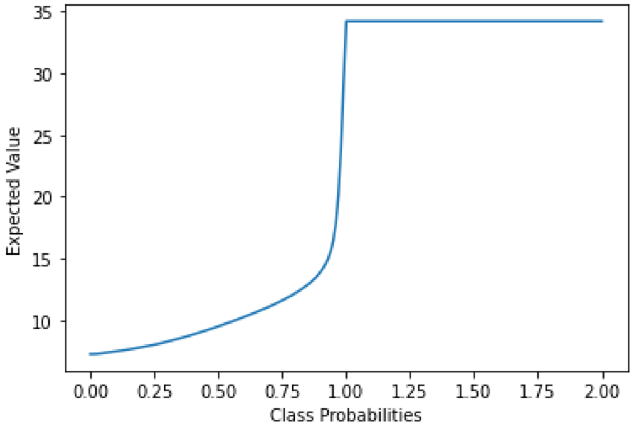


Figure 14. Cost Curve

6 Conclusion

In conclusion, we explored four different machine learning models in prediction flight delay and found out that Random Forest Classifier is the best model with the highest F1 and accuracy score. To further improve on the model, we also tried hyperparameter tuning to search for the ideal model architecture. However, the result is less than satisfying as all the evaluation scores dropped due to over-fitting. Thus, we decided to stick to Random Forest Classifier as the ideal mode. We also converted the prediction results into monetized values to achieve a more comprehensive view in the business field. This is done via a well-researched value matrix that breaks down the delay cost into airline and passenger cost. Combining the evaluation results with cost matrix, we re-affirmed that Random Forest Classifier is the best cost-cutting model.

7 References

Airline On-Time Performance and Causes of Flight Delays. (2022). Bureau of Transportation Statistics.

Retrieved November 13, 2022 from https://www.bts.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays

Defining Late. (2019). OAG Aviation Worldwide Limited. Retrieved November 10, 2022, from https://www.oag.com/airline-on-time-performance-defining-late

Dou, X. (2020). Flight Arrival Delay Prediction And Analysis Using Ensemble Learning. *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 836–840. https://doi.org/10.1109/ITNEC48623.2020.9084929

Gonzalez, G. (2019). *Airlines Delay*. Kaggle. Retrieved October 15, 2022, from https://www.kaggle.com/datasets/giovamata/airlinedelaycauses?datasetId=355

Lastname, C. (2008). Title of the source without caps except Proper Nouns or: First word after colon. *The Journal or Publication Italicized and Capped*, Vol#(Issue#), Page numbers.

Lastname, O. (2010). Online journal using DOI (digital object identifier). *Main Online Journal Name*, Vol#(Issue#), 159-192. https://doi.org/10.1000/182

Lastname, W. (2009). *Title of webpage*. Site Name. Retrieved July 3, 2019, from http://www.example.com

Meel, P., Singhal, M., Tanwar, M., & Saini, N. (2020). Predicting Flight Delays with Error Calculation using Machine Learned Classifiers. *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, 71–76. https://doi.org/10.1109/SPIN48934.2020.9071159

Peterson, E. B., Neels, K., & Barczi, N. (2022). *The Economic Cost of Airline Flight Delay*. *47*, 16.

U.S. Passenger Carrier Delay Costs. (n.d.). Airlines For America. Retrieved 18 November 2022, from https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/

United States Department of Transportation. (2022). *Bureau of Transportation Statistics* Retrieved 18 November 2022, from https://www.transtats.bts.gov/homedrillchart.asp

Wang, F., Bi, J., Xie, D., & Zhao, X. (2022). Flight delay forecasting and analysis of direct and indirect factors. *IET Intelligent Transport Systems*, *16*(7), 890–907. https://doi.org/10.1049/itr2.12183

Wu, C.-L. (2005). Inherent delays and operational reliability of airline schedules. *Journal of Air Transport Management*, *11*(4), 273–282. https://doi.org/10.1016/j.jairtraman.2005.01.005

Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, *7*(1), 106. https://doi.org/10.1186/s40537-020-00380-z