Project 1: Flight Arrival Delay Prediction

ENSF 544: Data Science for Software Engineers

By: Maram Elsayed (30071200) and Madhu Selvaraj (30061979)

Date: 24th of October 2022

Introduction

In the world of Commercial Aviation, flight delays have become a problematic phenomenon that puts both passengers and airlines in risky situations. Consequences for airlines and the flying public include increased operating costs for airlines, decreased passenger welfare, greater fuel consumption, and increased emissions [3]. This subject has become widely investigated, which intrigued many data scientists to pursue it as a topic of research. This report will be based on the study of two papers: “A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines” and “Predictive Modelling of Aircraft Flight Delay”, both versed in the prediction and modelling of flight delays [1][2]. We explore the methods used in both papers, perform modifications, and then compare the results of various metrics such as efficiency, accuracy, etc. The remainder of this paper is divided into the following sections: Project Objectives, Used Datasets, Summary of the Methods Proposed, Proposed Modification/Method, Results, and Conclusion.

Project Objectives

The project objectives include extracting methods from the previously mentioned papers, modifying the processes, and comparing the results achieved to one another, as well as comparing them to the paper’s results. The project follows a certain procedure in order to accomplish these objectives. First, to organise the project’s outcomes, we thoroughly study the previously mentioned papers to develop an understanding of the sequence of processes used to achieve their results. There are two main sections, data preprocessing and method application. Each of the chosen papers have their own techniques of analysing the same dataset, which will be further explored in the upcoming sections.

Used Datasets

For both papers, we are using the “Airline On-Time Performance Data” dataset provided by the Bureau of Transportation Statistics [4]. To follow the approaches in the two research papers, only the years 2015 and 2016 datasets were extracted. The utilised features are displayed in **Table 1** with their descriptions.

Feature	**Description**
YEAR	Year
QUARTER	Quarter (1-4)
MONTH	Month
DAY_OF_MONTH	Day of Month
DAY_OF_WEEK	Day of Week
OP_UNIQUE_CARRIER	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
OP_CARRIER_AIRLINE_ID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
TAIL_NUM	Tail Number

OP_CARRIER_FL_NUM	Flight Number
ORIGIN	Origin Airport
ORIGIN_WAC	Origin Airport, World Area Code
DEST	Destination Airport
DEST_WAC	Destination Airport, World Area Code
CRS_DEP_TIME	CRS Departure Time (local time: hhmm)
DEP_TIME	Actual Departure Time (local time: hhmm)
DEP_DELAY	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DEP_DELAY_NEW	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DEP_DEL15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
TAXI_OUT	Taxi Out Time, in Minutes
TAXI_IN	Taxi In Time, in Minutes
CRS_ARR_TIME	CRS Arrival Time (local time: hhmm)
ARR_TIME	Actual Arrival Time (local time: hhmm)
ARR_DELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ARR_DELAY_NEW	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ARR_DEL15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
CANCELLED	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	Specifies The Reason For Cancellation
DIVERTED	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
ACTUAL_ELAPSED_TIME	Elapsed Time of Flight, in Minutes
AIR_TIME	Flight Time, in Minutes
DISTANCE	Distance between airports (miles)
CARRIER_DELAY	Carrier Delay, in Minutes
WEATHER_DELAY	Weather Delay, in Minutes
NAS_DELAY	National Air System Delay, in Minutes
SECURITY_DELAY	Security Delay, in Minutes
LATE_AIRCRAFT_DELAY	Late Aircraft Delay, in Minutes

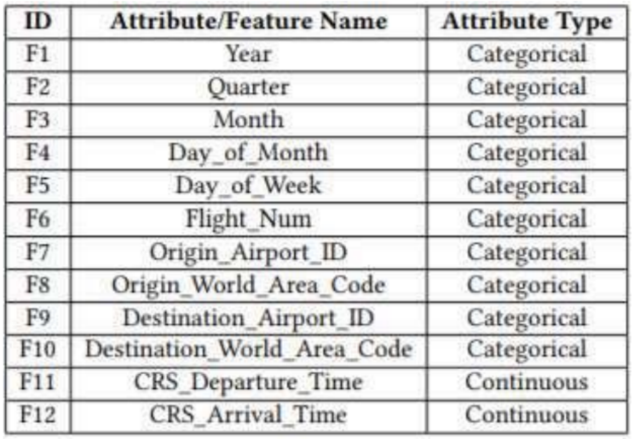
Table 1 : Features of the dataset and their descriptions [4].

Summary of the Methods Proposed

Paper 1: A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines

In this study, the primary aim was to determine the accuracy of using machine learning to predict whether a flight would be delayed on arrival. To perform the prediction, the author utilised the Gradient Boosting Classifier model, which is an Ensemble Learning method that aggregates the results of multiple estimators while also learning from their mistakes and improving on errors made by predecessors [5].

This paper limits its analysis to 2015 and 2016 American Airlines flights originating and arriving in the five following busiest US airports: Hartsfield-Jackson Atlanta International Airport, Los Angeles International Airport, O'Hare International Airport, Dallas/Fort Worth International Airport and John F. Kennedy International Airport. From the Bureau of Transportation Statistics dataset, the attribute that describes whether a flight has been declared as a late arrival is "Arr_Del _15", which is 1 if a flight's arrival time is 15 minutes or greater than scheduled, and 0 otherwise (i.e. no arrival delay). Therefore, this was used as the label in the classification model. Ten additional attributes were used as features in the model. The features are summarised in **Figure 2**, however following the paper, F1 and F2 were dropped due to low variability.



ID	Attribute/Feature Name	Attribute Type
F1	Year	Categorical
F2	Quarter	Categorical
F3	Month	Categorical
F4	Day_of_Month	Categorical
F5	Day_of_Week	Categorical
F6	Flight_Num	Categorical
F7	Origin_Airport_ID	Categorical
F8	Origin_World_Area_Code	Categorical
F9	Destination_Airport_ID	Categorical
F10	Destination_World_Area_Code	Categorical
F11	CRS_Departure_Time	Continuous
F12	CRS_Arrival_Time	Continuous

Figure 2. Summary of features [1]

All preprocessing and cleaning steps detailed in the paper were followed. First, all instances that had a missing value for Arr_Del_15 were removed (a total of 1614 rows). Next, the categorical features (F3-F10 in **Figure 2**) were label encoded, which resulted in each unique value in a feature being assigned a number alphabetically. The author then used One-Hot encoding, which was done on the origin and destination data (F8-F10 in **Figure 2**). This resulted in these features being split into multiple columns for each unique category in that feature. The final preprocessing step was to balance the data. As the author discovered, there were far greater instances where Arr_Del_15 = 0 (76,673, majority label) compared to where Arr_Del_15 = 1 (19,871, minority label). This high data imbalance is shown **Figure 3**. During the reimplementation, we find a slight variance in the number of instances for each label, but we still observe the imbalance. Thus, we follow the paper’s methods and use the 200% R-SMOTE balancing technique. The aim of this technique is to create new instances of the minority label using linear interpolation of neighbour data points, which ultimately produces a more balanced dataset. The result is shown in **Figure**

4, where the new number of Arr_Del_15 = 1 instances is 59,459 (around a 200% increase), and the data imbalance is minimised. The final dimensions of the input dataset are 136,132 instances and 26 features.

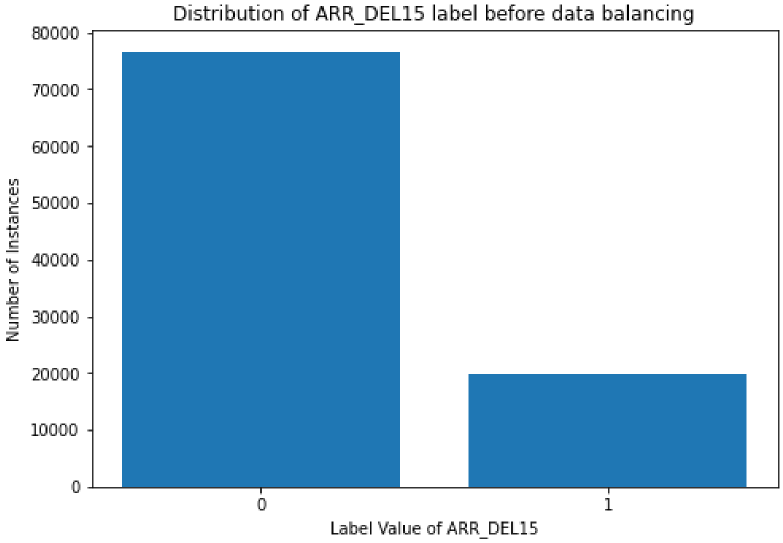


Figure 3. Distribution of the Arr_Del_15 label before data balancing (Taken from reimplementation)

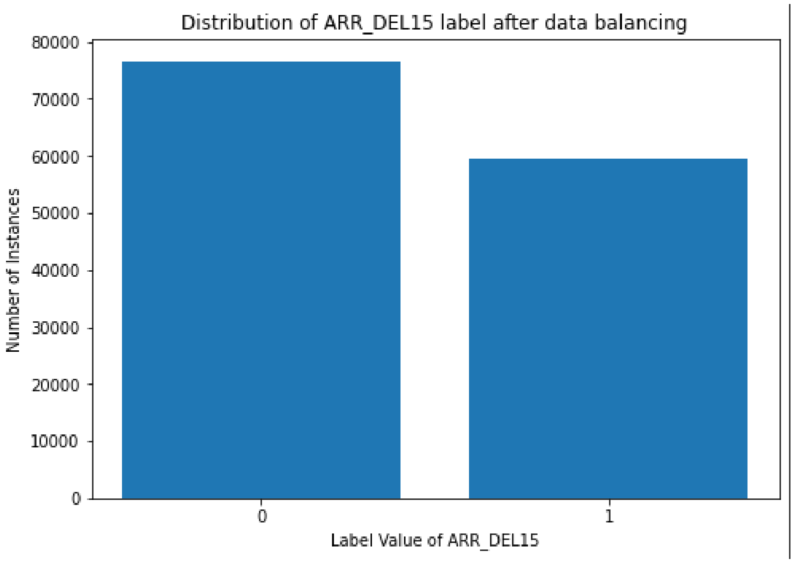


Figure 4. Distribution of the Arr_Del_15 label value after data balancing applied (Taken from reimplementation)

The paper tests two strategies for the classification: 1) Using the data unbalanced, 2) Using the data after applying 200% R-SMOTE balancing. As the author concluded that they obtained more accurate results using strategy 2, we only follow that approach for the reimplementation. Therefore, we then move on to applying the Gradient Boosting Classifier to the dataset. To determine the optimal hyperparameters (number of estimators and max depth), Grid Search was used by the author. This technique is an exhaustive search that tries every hyperparameter value combination. The range of values used by the author in Grid Search were 100-500 (step of 50) for the number of estimators, and 3-5 for the max depth. The result of this search is shown in **Figure 5**, where they find that 400 estimators with a max depth of 5 is the optimal set. We therefore use these values when reimplementing the Gradient Boosting Classifier model.

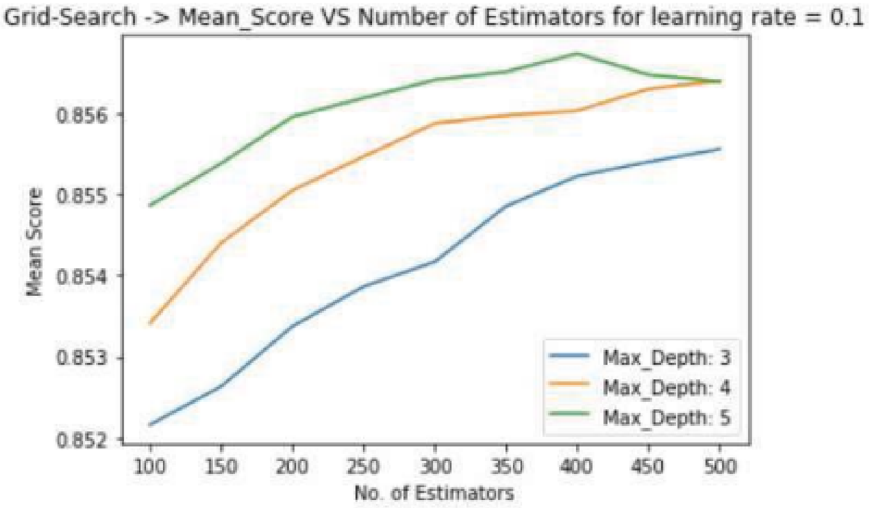
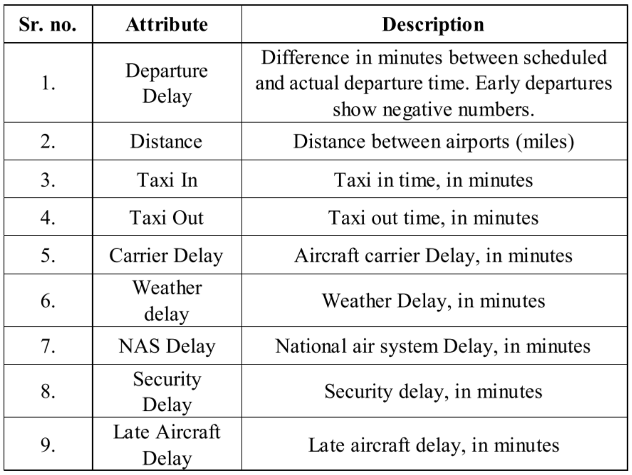


Figure 5. Grid Search results (Taken from [1])

Additionally, as specified in the paper, we set 20% of the data to be used for testing. The model evaluation and accuracy of this classification is described in the Results section.

Paper 2: Predictive Modelling of Aircraft Flight Delay

In this paper, the main objective is to analyse the on-time performance of domestic flights from January 2016 to December 2016 to develop a better predictive model to predict flight delays. There are three mainly proposed methods: Multiple Regression, Decision Tree, and Random Forest. The paper did not propose any procedure of preprocessing the dataset. The variables utilised in the research paper are outlined in **Figure 6** with their descriptions**.**



Sr. no.	Attribute	Description
1.	Departure Delay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
2.	Distance	Distance between airports (miles)
3.	Taxi In	Taxi in time, in minutes
4.	Taxi Out	Taxi out time, in minutes
5.	Carrier Delay	Aircraft carrier Delay, in minutes
6.	Weather delay	Weather Delay, in minutes
7.	NAS Delay	National air system Delay, in minutes
8.	Security Delay	Security delay, in minutes
9.	Late Aircraft Delay	Late aircraft delay, in minutes

Figure 6. Variable Description [2].

Before moving on to modelling, some preliminary analysis was carried out. The predictor plot shown in **Figure 7** gives the Pearson's constant (*r*) for all the features shown in **Figure 6.** The Pearson's correlation coefficient shows the linear relation between two variables, where if $r = 1$, then there is a perfectly positive correlation, and if $r = -1$, then there is a perfectly negative correlation. This method is used to detect the existence of multicollinearity within the variables. If any value of *r* exceeds 0.5, then it means that there is serious multicollinearity between these variables. Another method used for preliminary analysis of all the variables is the scatterplot shown in **Figure 8**. A scatterplot is a data visualisation tool used to visualise and identify the data trends between variables. This technique is carried out by plotting two variables against each other in a single plot.

	Departure Delay	Distance	Taxi In	Taxi Out	Carrier Delay	Weather Delay	NAS Delay	Security Delay	Late Aircraft Delay
Departure Delay	1	-0.00844	0.074082	0.0777637	0.052680	0.2485282	0.300320	-0.016650	0.467740449
Distance	-0.00844	1	0.071508	0.0711831	0.014189	-0.027744	-0.00145	0.0008507	0.023294442
Taxi In	0.074082	0.071508	1	0.066415	0.008970	0.0398598	0.235657	-0.001850	0.000618343
Taxi Out	0.077763	0.071183	0.066415	1	0.010281	0.0926416	0.447768	-0.009421	-0.02913716
Carrier Delay	0.526808	0.014189	0.008970	0.0102819	1	-0.050103	-0.07636	-0.025475	-0.13387436
Weather Delay	0.248528	-0.02774	0.039859	0.0926416	-0.05010	1	0.018261	-0.006999	-0.05024643
NAS Delay	0.300320	-0.00145	0.235657	0.4477682	-0.07636	0.0182611	1	-0.017283	-0.08231478
Security Delay	-0.01665	0.000854	-0.001850	-0.00994	-0.02547	-0.006999	-0.01728	1	-0.04244976
Late Aircraft Delay	0.46774	0.023294	0.0006183	-0.029137	-0.13387	-0.050246	-0.08231	-0.042449	1

Figure 7. Predictor Plot Correlation [2].

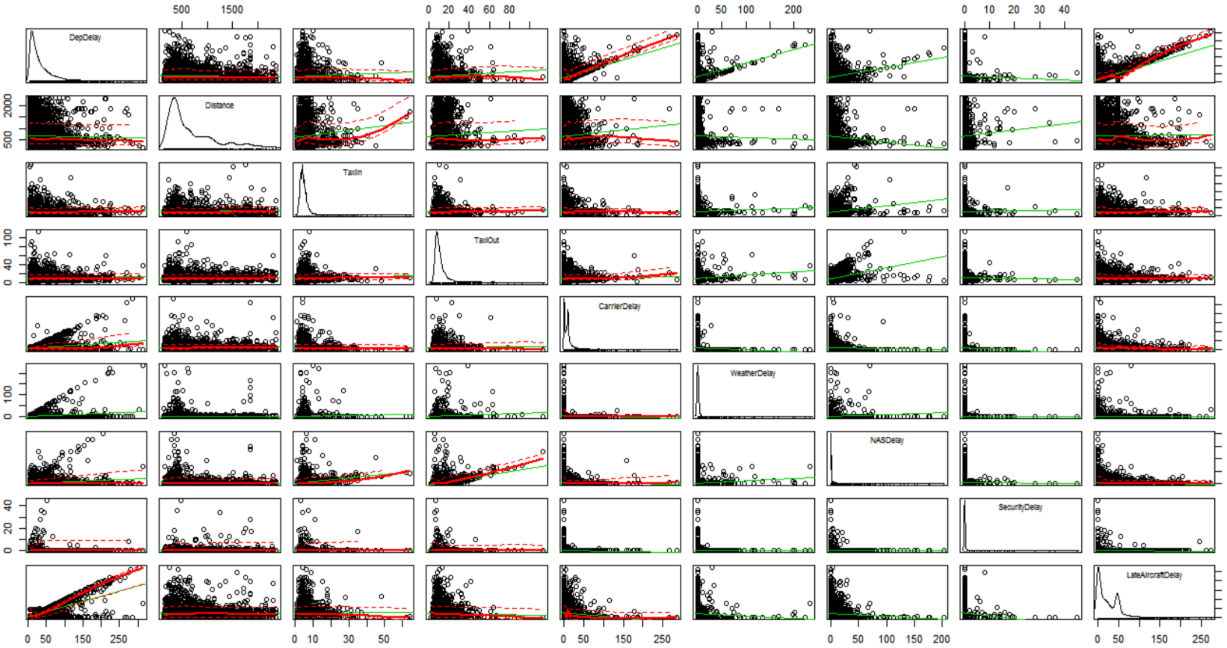


Figure 8. Scatter plot matrix of all the variables [2].

Multiple (Linear) Regression (MLR) is an uncomplicated approach that predicts a quantitative response (Y) from multiple predictor variables (X), this model assumes that there is a direct linear correlation between X and Y [2]. Decision Trees (DT) are appropriate to use in this case since a linear relationship only exists between some variables and not variables in the scatterplot in **Figure 8** [2]. Decision Trees involve the division of the predictor space into different regions and to make predictions for certain observations [2]. In **Figure 9**, we can see a decision tree where Late aircraft delay is the topmost significant variable/node and the values below (in the square boxes) are the prediction of departure delay estimates[2]. Lastly, Random Forest (RF) is a method originated and extended from Decision Trees [2]. RF consists of a collection of decision trees that grow in parallel to each other and help reduce the variance in the model [2]. In the paper, a total of 500 decision trees were constructed by the RF algorithm [2].

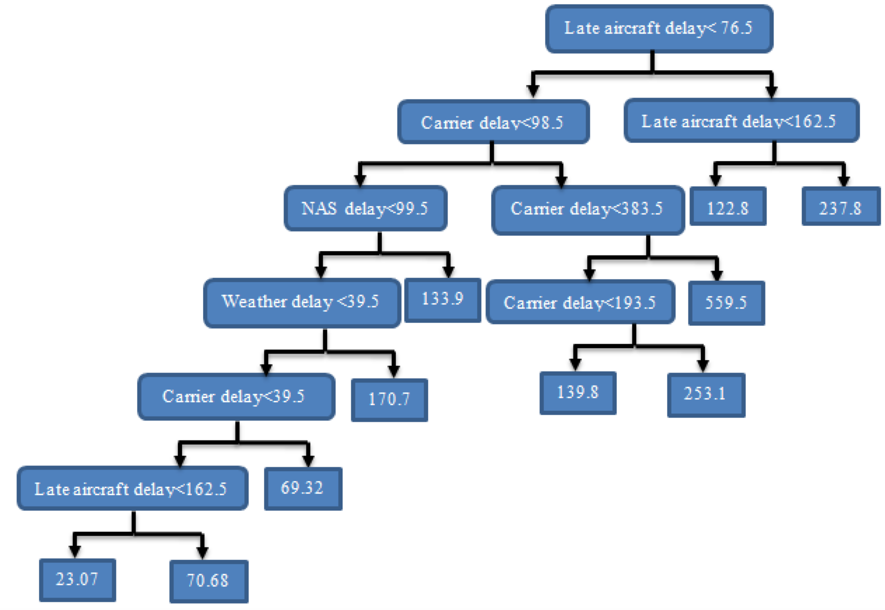


Figure 9. Decision Tree [2].

Proposed Modification/Method

Paper 1: A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines

We decided to modify the hyperparameter tuning technique for the Gradient Boosting Classifier. The paper's approach of using Grid Search is highly exhaustive as it tries every possible hyperparameter value combination. This in turn can make it computationally expensive, as well as unrealistic to run on a large dataset such as the one we are modelling [6]. We therefore propose to use Random Search to determine the optimal hyperparameters. In contrast to Grid Search, only a random selection of hyperparameter combinations are tried, and the number of iterations can be controlled depending on the available resources or dataset size. Although the randomised nature may lead to certain combinations not being tested, it is less costly in terms of time and can allow us to test a larger range of hyperparameters, which in turn increases the possibility of discovering values that improve model performance. Therefore, we also modify the range of the number of estimators to be 300-650 (step of 50), and the max depth to be 4-7.

The resulting optimal hyperparameters discovered by Random Search, and the comparison of running the Gradient Boosting Classifier using these values rather than the values determined by Grid Search is described in the Results section.

Paper 2: Predictive Modelling of Aircraft Flight Delay

Due to the lack of implementation in paper 2, we have taken the initiative of making a deal of modifications [2]. Firstly, the preprocessing procedure followed the steps below:

1. Finding the percentage of null values in each column.
2. Filling the null cells of types of delays (Carrier, Weather, National Air System, Security, and Late Aircraft Delays) with zero. Because when it's null, it means it has no impact on the delay.
3. Dropping the rows with missing values.
4. Dropping unneeded data.
5. Label encoding features to change categorical variables into numerical ones.

There were a few modifications applied to the variables in **Figure 6**. Instead of using the nine variables mentioned, we used 18 (9 on top of the ones mentioned in the paper), they are shown in **Table 3** (their descriptions can be found in **Table 1**).

Features (variables)	
DAY_OF_WEEK	SCHEDULED_ELAPSED_TIME
AIRLINE	ELAPSED_TIME
ORIGIN	AIR_TIME
DEST	DISTANCE
DEP_DELAY	CARRIER_DELAY
TAXI_OUT	WEATHER_DELAY
TAXI_IN	NAS_DELAY
ARR_DELAY	SECURITY_DELAY
DIVERTED	LATE_AIRCRAFT_DELAY

Table 3 : Proposed Modification of Features.

Lastly, instead of using 500 trees for Random Forest as proposed in the paper, only 10 were used as the process was really vast and took too long to execute. The execution time with 100 trees exceeded 30 minutes, and exceeded the limited RAM available for Google's Colab tool.

Results

Paper 1: A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines

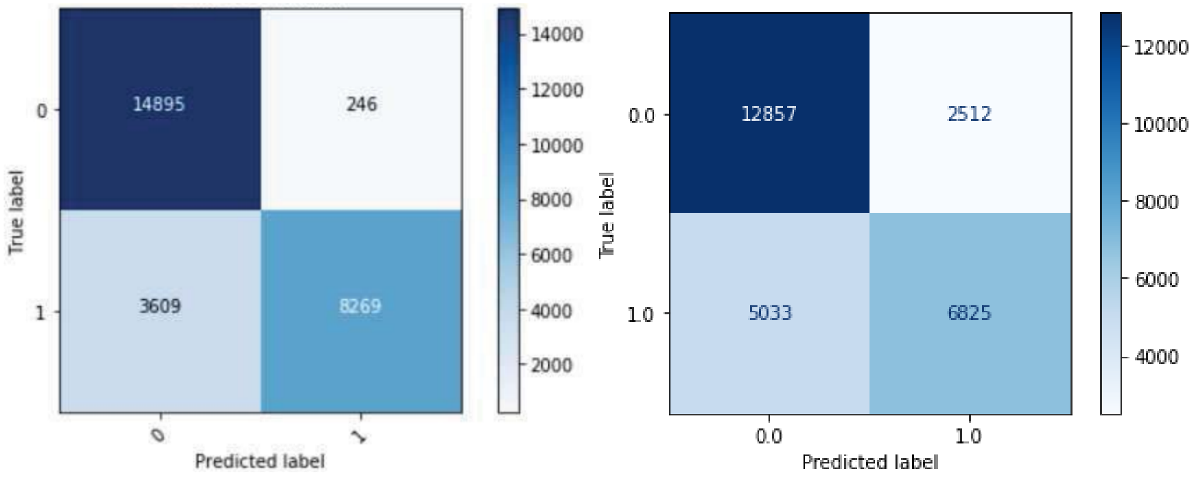
The metrics used in the paper to evaluate the Gradient Boosting Classifier are training accuracy, testing accuracy, recall, precision, F1-Score, and the area under the ROC curve. The scores for all three approaches (original paper, reimplementation, and modification) are shown in **Table 3.**

Approach	**Training Accuracy**	**Validation Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUROC**
Original Paper [1] (Strategy 2)	86.68%	85.73%	0.86	0.88	0.85	0.90
Reimplementation	74.49%	72.23%	0.58	0.73	0.64	0.79
Modification	86.06%	78.90%	0.69	0.79	0.74	0.86

Table 3. Model Evaluation Summary

To compare the original results from the paper with our reimplementation results, we first look at differences in their respective evaluations scores in **Table 3**. Across all scores, our reimplementation does not measure as well as the paper, particularly in the recall scores (0.58 vs 0.86). We also compare confusion matrices shown in **Figure 10**, and area under the ROC plots shown in **Figure 11**, where we observe differing values for AUROC (0.79 for reimplementation vs 0.90 for the original). One possible reason why the reimplementation did not produce the same results as the paper may be due to the data balancing implementation. In paper, the specific R-SMOTE model used was not specified, so we chose to use SMOTE-NC, as the data had both categorical and continuous data.

We then observe the effects of modifying the paper's methods. The results of running Random Search to obtain optimal hyperparameters for the Gradient Boosting Classifier was 650 estimators and a max depth of 7. Therefore, when we compare the results obtained after replacing Grid Search with Random Search, we see in **Table 3** that the model performance improved greatly compared to our direct reimplementation. It however still does not measure as well as the original paper. Looking at **Figure 11**, the area under ROC score increased to 0.86, which is much closer to the paper's result of 0.90 when compared to the reimplemention's score of 0.79. Training accuracy also improved by a large amount to almost the same level as the original paper, but less so for testing accuracy.



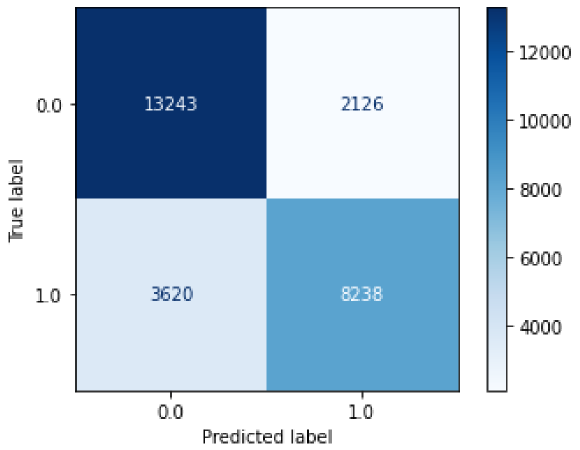


Figure 10: Confusion Matrices for original paper [1] (top left), reimplementation (top right), and modification (bottom centre)

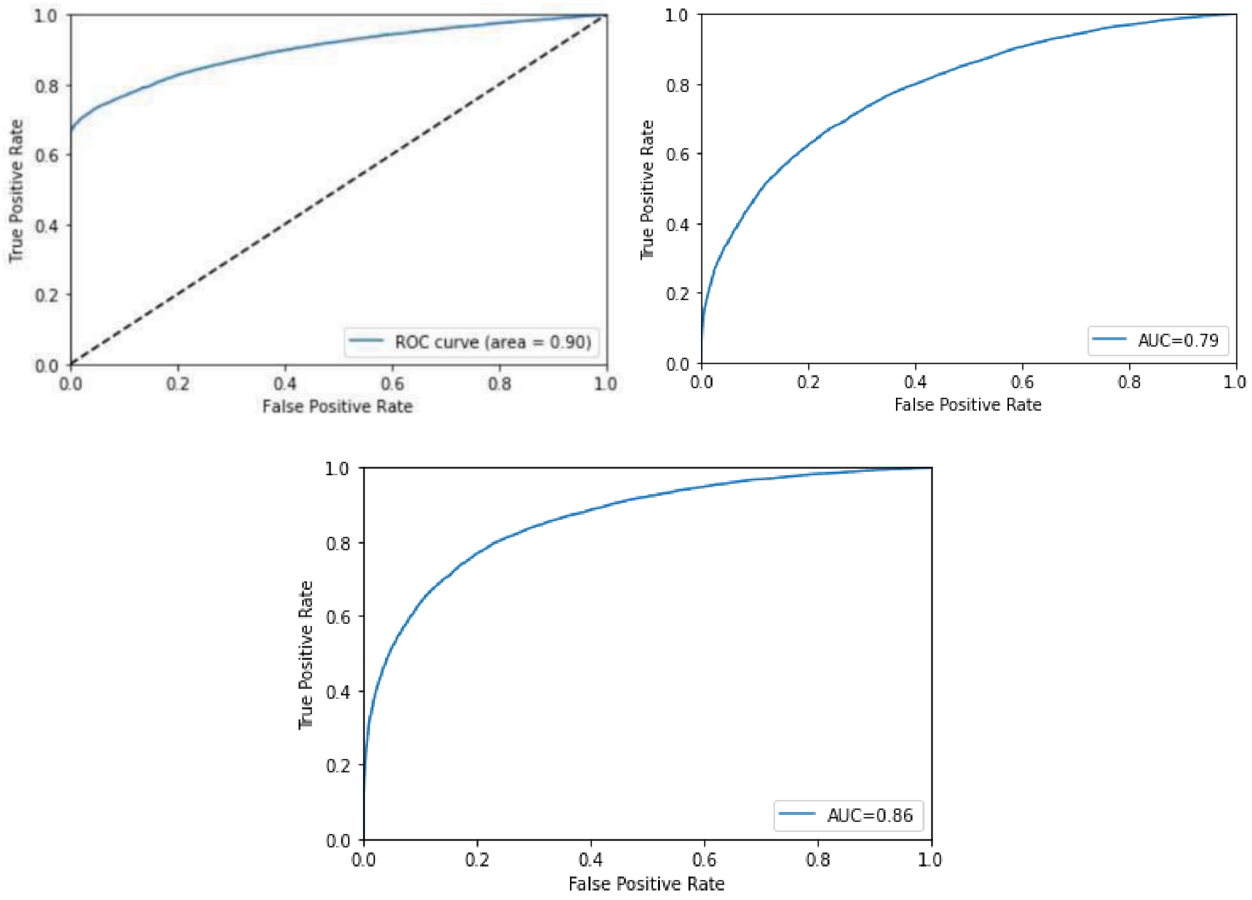


Figure 11. ROC Curve for original paper [1] (top left), reimplementation (top right), and modification (bottom centre)

Overall, we can see that the new hyperparameters obtained by Random Search did make a significant difference in the model performance, and that tuning these values correctly is important to obtain accurate prediction results.

Paper 2: Predictive Modelling of Aircraft Flight Delay

Approach	**R-squared Score**	**Mean Absolute Error (or RMSE in paper) (minutes)**
MLR (Paper)	0.84	21.2
DT (Paper)	~0.80 (for 12 splits)	26.5
RF(Paper)	0.94	12.5
MLR (Modification)	0.9999999995788214	2.2018607992789854e-06
DT (Modification)	0.9979946574458742	0.48023635338022747
RF (Modification)	0.9989853044854837	0.3347003366468358

Table 4 : Comparison of results.

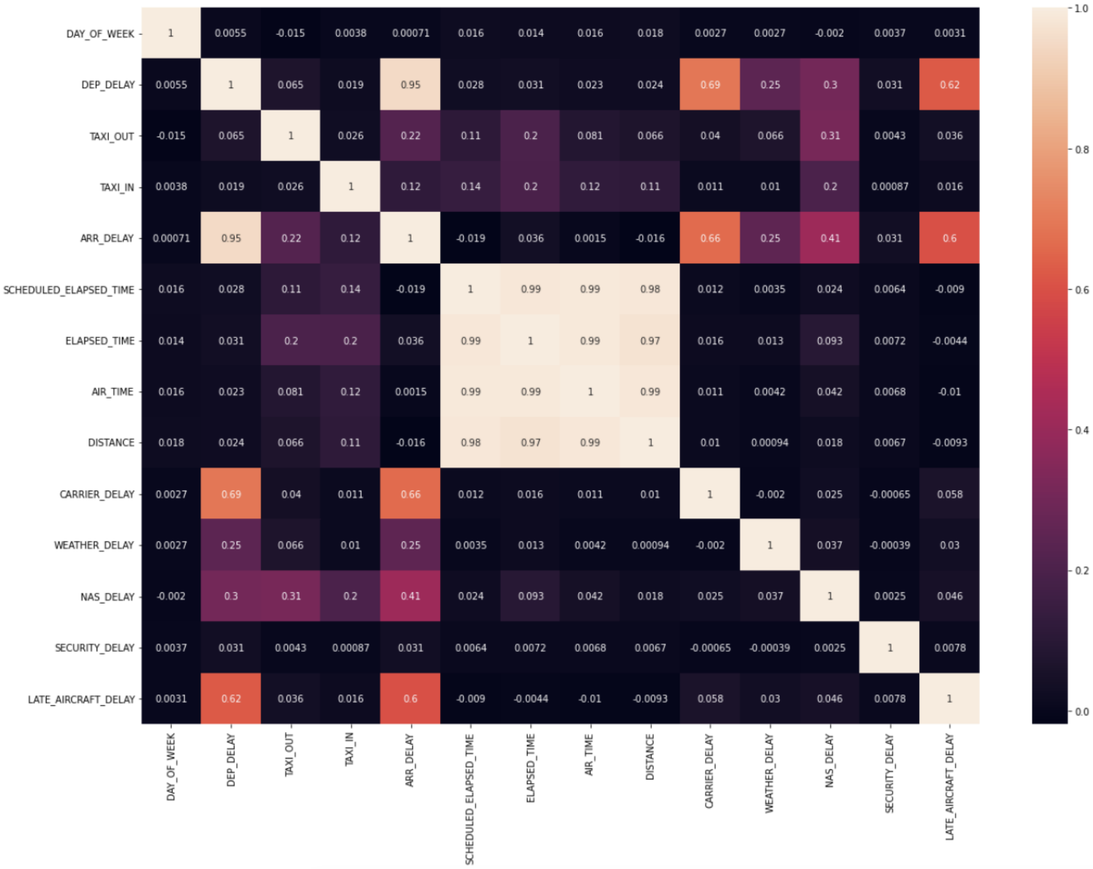


Figure 12. Predictor Plot Correlation.

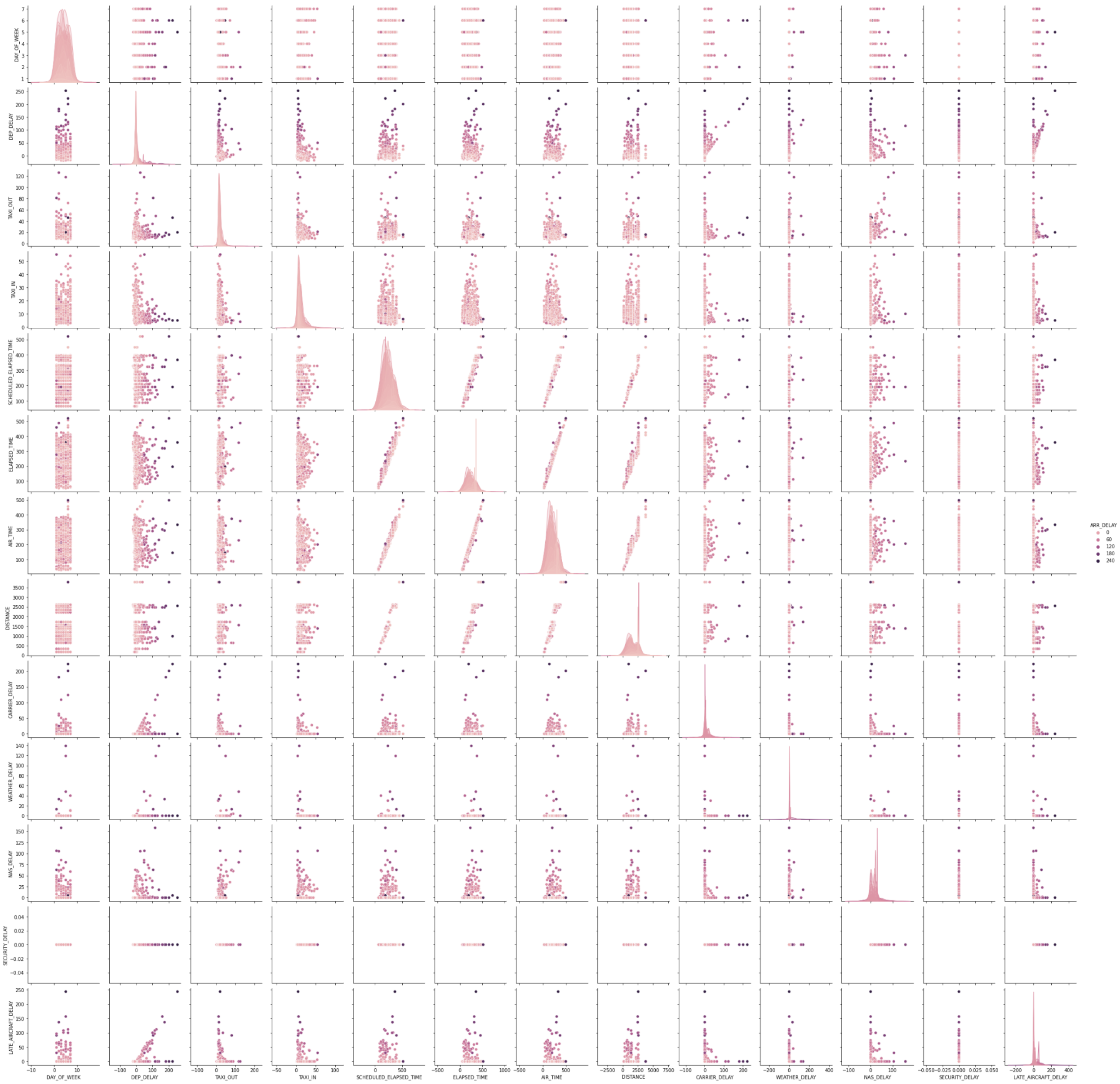


Figure 13. Scatter plot matrix of all the variables.

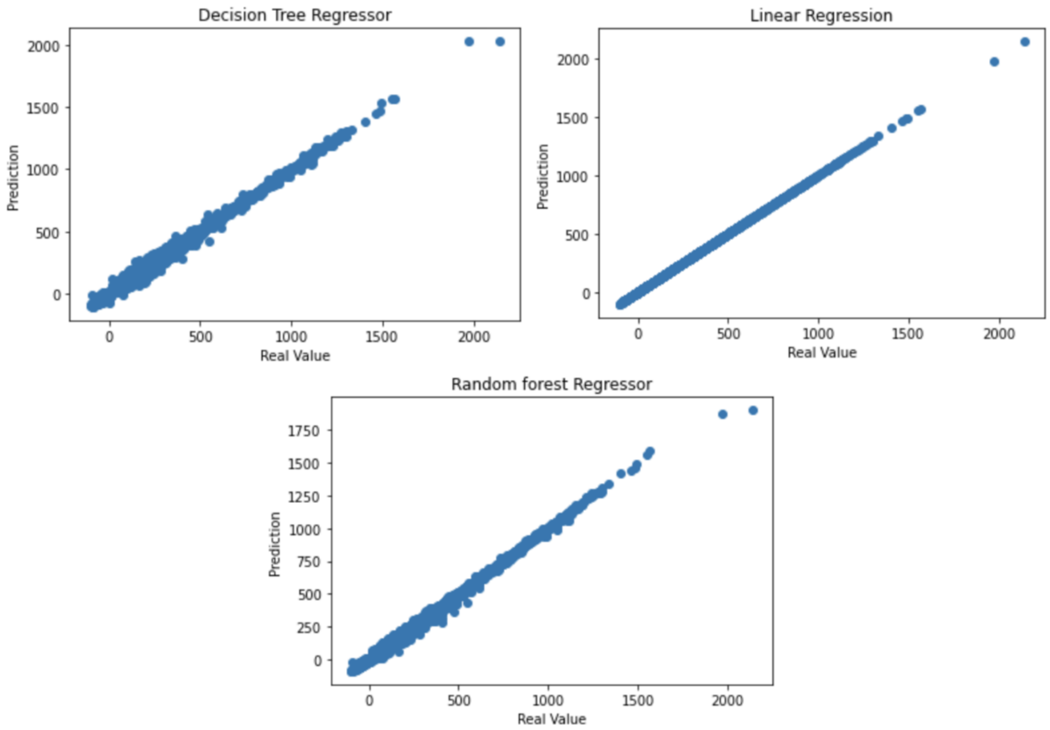


Figure 14. Correlation between Predicted Value and Real Value for each Model.

Unlike the paper, our best model was the Multiple (Linear) Regression model; instead of the proposed Random Forest Regression model. This was deduced by observing the R-squared Score, as well as the Mean Absolute Error of all the methods used. For both evaluating criterias in **Table 4**, MLR has exceeded the other proposed models by a great deal. Furthermore, when observing **Figure 14**, it seems that MLR has the most accurate Prediction-to-Real Value correlation out of the three models assessed.

We can also conclude that the Arrival Delay (ARR_DELAY) has a very strong correlation to the Departure delay (DEP_DELAY) by looking at the Predictor Plot Correlation in **Figure 12.** The figure outlines the high Pearson's constant coefficient between the two variables of $r = 0.95$.

Conclusion

The main motivation for the two papers we studied are the negative consequences flight delay has on both airlines and passengers. The Airline On-Time Performance dataset from the Bureau of Transportation Statistics has shown to be useful for developing delay prediction models, as it contains a large variety of flight information [4]. The study "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines" [1] attempted to classify whether a certain subset of flights would be delayed upon arrival using the Gradient Boosting Classifier model, which provided them with relatively high accuracy scores (85.73%). After reimplementing their methods, we obtained less accurate results (72.23%). However, after modifying the implementation by replacing Grid Search with Random Search for the hyperparameter tuning, we saw accuracy increase significantly (78.90%). For the study "Predictive Modelling of Aircraft Flight Delay", the author implemented and developed three predictive models for flight delay prediction, and found that Random Forest performed the best with an r-squared score of 0.94.

After making modifications to the preprocessing procedure, using additional features from the dataset in the analysis, and changing model parameters, we found that instead Multiple Linear Regression produced the best results (r-squared value of ~0.99). Overall, we see that although the feature space used in both studies was different, the modifications made to the methods in paper 2 provided the most accurate results. Obtaining an accurate prediction model is a promising sign for future analysis and prevention of commercial flight delays.

References

[1] Chakrabarty, N. (2019, March). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines. *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*. https://doi.org/10.1109/iemeconx.2019.8876970

[2] Kalliguddi, A. M., & Leboulluec, A. K. (2017, October). Predictive Modeling of Aircraft Flight Delay. *Universal Journal of Management*, *5*(10), 485–491. https://doi.org/10.13189/ujm.2017.051003

[3] Shadare, W. (2022, February 15). *Flight delays cost more than just time, airlines' reputation at stake*. Aviation Metric. Retrieved October 23, 2022, from https://aviationmetric.com/flight-delays-cost-more-than-just-time-airlines-reputation-at-stake/

[4] *Bureau of Transportation Statistics*. (n.d.). TranStats. Retrieved September 29, 2022, from https://www.transtats.bts.gov/Homepage.asp

[5] Aliyev, V. (2020, October 7). Gradient boosting classification explained through python. *Medium*, https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d

[6] Worcester, P. (2019, June 6). A comparison of grid search and randomized search using Scikit learn. *Medium*, https://medium.com/@peterworcester_29377/a-comparison-of-grid-search-and-randomized-search-using-scikit-learn-29823179bc85