Applying Machine Learning Classifiers to the Airline Dataset

2nd December 2022

Table of Contents

1 Exploratory Data Analysis

This first section of the report focuses on Exploratory Data Analysis to understand the information provided in the dataset such as its size and the features included. Additionally, this section aims to identify any inconsistencies such as class imbalance and outliers which may hinder the outcomes of applying machine learning classifiers.

1.1 Dataset Size and Attributes

The dataset contains 129879 instances and has 23 features. Overall, the dataset aims to provide information on how customers' satisfaction is affected by different aspect of the airline service. It used to have a 24th ID feature which was removed as it does not provide any valuable information. Read sections 1.6 and 2.1 for more detail on feature selection.

The first 5 features are details about the customer, which includes their Age and Gender, as well as their flight details, including Class, Type of Travel and whether a customer is Loyal or Disloyal. Additionally, the Flight Distance and the Delay in Departure and Arrival are also recorded. Age, Flight Distance and Arrival and Departure Delay contain continuous data while the other features mentioned contain nominal data.

The rest of the features appear to measure the quality of the different aspects of the airline service based on a survey where they were rated on a scale from 0 to 5. As this data ranks the quality of the services provided this data is ordinal. Those featured include Inflight Entertainment, Leg Room and Food and Drink service to name a few.

Additionally, there is an overall Satisfaction feature containing ordinal data. This feature will be used to classify how the other features determine overall satisfaction through machine learning.

1.2 Class Imbalance

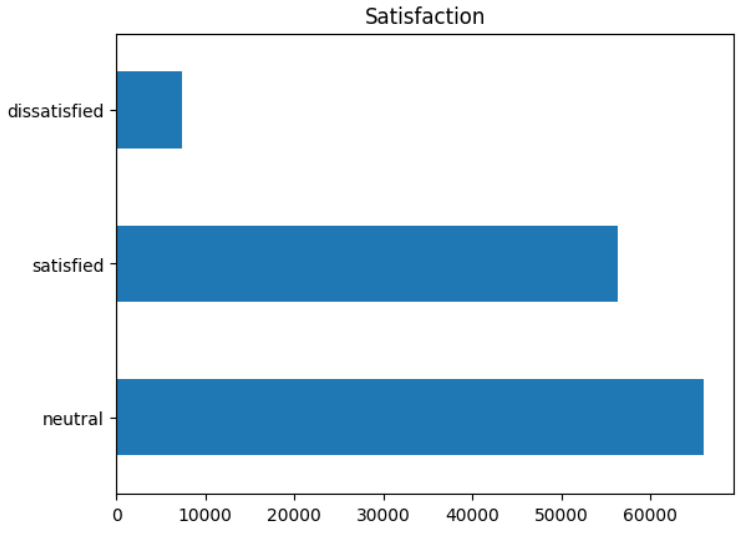

Figure 1.1

The bar graph in Figure 1.1 above shows the class distribution of overall customer satisfaction. There is a clear class imbalance as most of the customers are either satisfied or neutral about their experience with the airline. For information on dealing class imbalance read section 2.4.

1.3 Missing/Null values

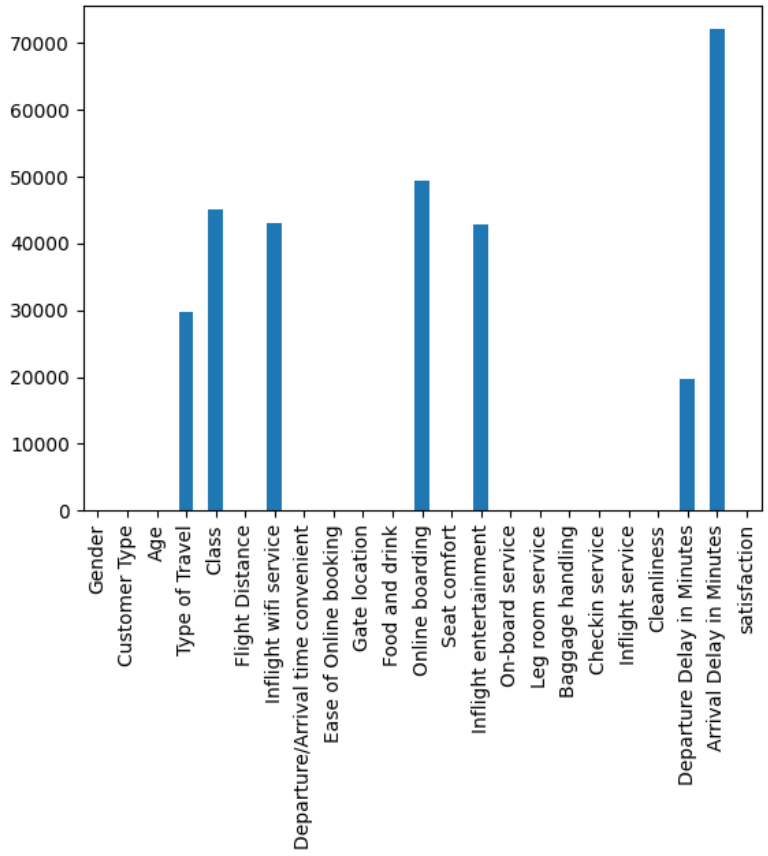

Figure 1.2

The bar graph in Figure 1.2 above shows the number of missing values for each attribute. There is a considerable number of missing values for Type of Travel, Class, Inflight Wi-Fi service, Online Boarding, Inflight Entertainment, Departure Delay and especially Arrival Delay. The number of missing values ranges between just under 20 thousand to over 70 thousand missing values. For information on dealing with missing values read section 2.2.

1.4 Outliers

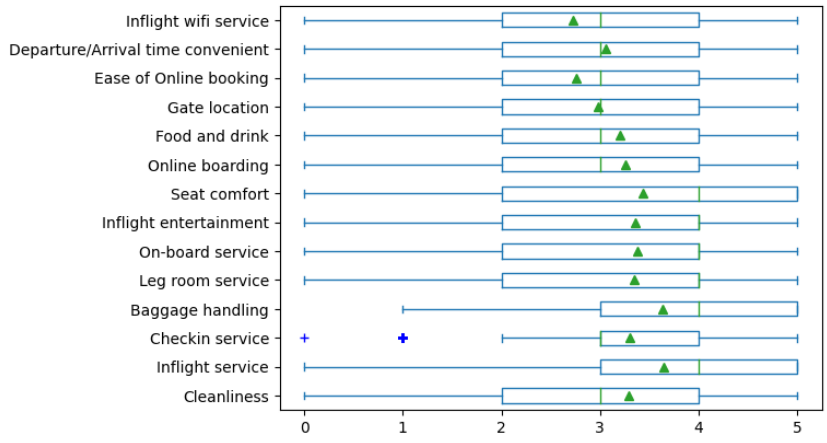

Figure 1.3

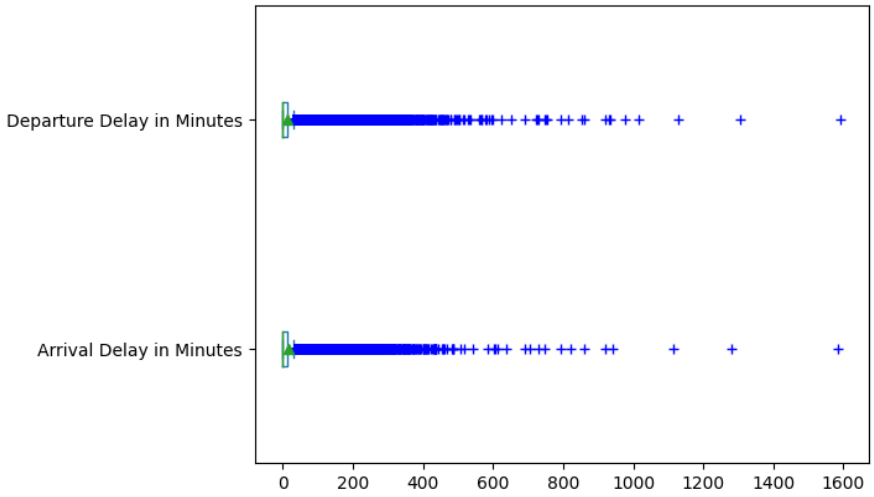

Figure 1.4

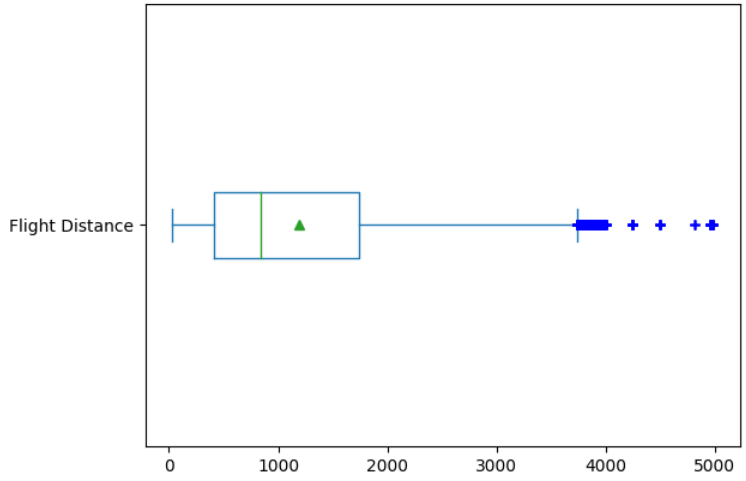

Figure 1.5

The box plots in Figures 1.3, 1.4 and 1.5 above show outliers for attributes that have them marked with blue plus signs. There are some outliers for Check in Service and Flight Distance. There are also many outliers for Arrival Delay and especially Departure delay. Read section 2.3 for more information on outliers.

1.5 Duplicates

The dataset does not contain any data duplicates.

1.6 Correlation

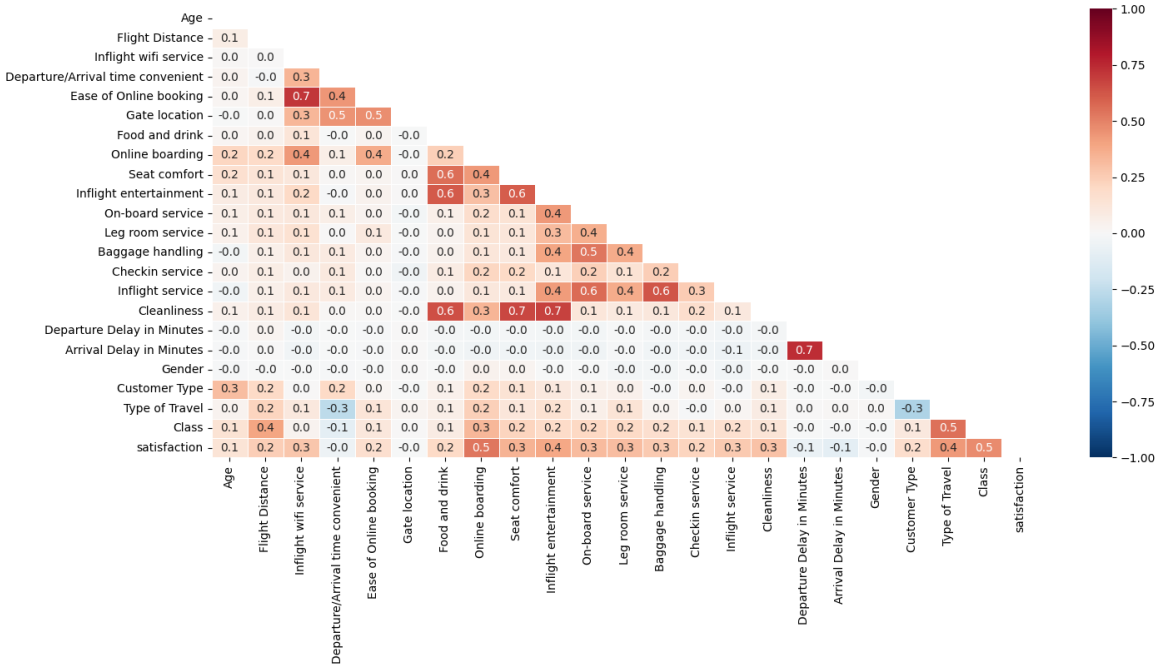

Figure 1.6

The correlation matrix in Figure 1.6 shows a strong positive correlation of >=70% between several features such as Inflight Wi-Fi Service and Ease of Online booking as well as between Cleanliness and a few other features. Not surprisingly, Departure Delay also has a strong positive correlation with Arrival Delay.

There is a more detailed discussion on feature selection and why a strong correlation between features is detrimental to machine learning in Section 2.1.

2 Data pre-processing

This section of the report focuses on pre-processing of the data to eliminate any erroneous data such as missing values irrelevant features to maximise the accuracy of machine learning models with the goal of achieving an accuracy of 90%. Seeing as a lot of features in this dataset contain ordinal data, algorithms that produce tree diagrams will likely be used, hence the secondary objective will be to demonstrate that the Random Algorithm is a superior algorithm to the Decision Tree Algorithm

2.1 Feature selection

Feature selection is the process of only using the most relevant features of the dataset to train the machine learning algorithm. Filtering methods were used to identify the least relevant features and remove them.

Some features may not be as useful if they correlate highly with other features other than the target feature, meaning that it can be predicted by other features, making its existence redundant and add to the complexity of the dataset. This is backed up by Kotsiantis (2011) where they state that “perfectly correlated features are truly redundant in the sense that no additional information is gained by adding them.”

Another indicator of a feature that isn’t useful is a feature with a low variance. If a feature has the same value throughout the column (0 variance) it is a constant and does not need to be considered, therefore, features with variance below a certain threshold may be worth eliminating. The VarianceThreshold library in ski-kit learn can be used to carry out thresholding.

Feature selection was applied and the following features were removed: Gender due to low variance; Cleanliness due to high correlation of >70% with multiple other features; Departure/Arrival time convenient due to high correlation of >70% with Inflight Wi-Fi Service; and finally, Arrival Delay due to high correlation of >70 % with Departure Delay as well as due to the feature having over 70,000 missing values, as it would have been easier to remove rather than impute missing values which could have skewed the machine learning process, given that the same information could be extracted from Arrival Delay.

2.2 Dealing with missing values

Missing values need to be dealt with as they produce biased and inaccurate models and because many machine learning algorithms are unable to process them.

Missing values could be eliminated by imputing them with the mean value for a feature or by simply removing all the records with missing values. It is also an option to remove a feature with too many missing values, which is one of the reasons why Arrival Delay was removed in feature selection.

An article by T. Emmanuel et al (2021) states that “caution should be made for continuous-based techniques when imputing categorical data as this may lead to biased results”. Seeing as most of the

data in this dataset is ordinal and not continuous, it may be worth experimenting with both imputing missing values and removing records containing them, although imputing seems more appropriate as deleting data means that machine learning classifiers have less information to work with. Section 3.2 discusses which method was more effective in practice.

2.3 Dealing with outliers

Outliers are data points that differ significantly from other data points and are considered anomalies that sabotage the machine learning process which produces inaccurate models. However, since instances which may be considered outliers may be representative of real-life situations such as flight delays one could argue that these outliers are not actually anomalies and are worth keeping as they produce a model which is more realistic and representative of the situations that occur in real life.

This is backed up by V. Ilango at al (2012) in their article which states that "there are 'good' outliers that provide useful information that can lead to the discovery of new knowledge and 'bad' outliers that include noisy data points."

Section 1.4 contains box plot which show that there are many outliers for Departure and Arrival Delay as well as Flight Distance and some for Check in Service. Those instances can be removed but given the information above about how outliers may be helpful, it may be best to leave them in. A version of the dataset was made where the outliers were deleted and both versions were tested, and Section 3.2 discusses the results.

2.4 Dealing with class imbalance

It is important to ensure that the dataset does have class imbalance as this causes machine learning classifier to produce inaccurate models that are biased towards the majority classes. R Mohammed et al (2020) states that because of class imbalance "the models lean more to the majority class and eliminate the minority class."

As seen in Section 1.2, there is significantly less records of dissatisfied customers, which is good for the airline but bad for producing a model that allows exploration of what features are likely to determine future customers' satisfaction as machine learning classifiers will favour classifying data for sets of features as satisfied or neutral, with dissatisfied being underrepresented.

This can be corrected through over sampling the minority class or under sampling the majority class until both classes are equal. Over sampling produces exact duplicates of data in minority class which can result in overfitting in a machine learning model, whereas under sampling can remove a large portion of the dataset meaning that the machine learning classifiers have less data to work with.

In the case of this dataset, there is a difference of around 57,000 instances between the majority and the minority class in this dataset. Therefore, over sampling will likely be more appropriate. There is also an option to apply a combination of over sampling and under sampling which will be tested.

3 Empirical Evaluation

This section of the report will focus on the empirical evaluation of two machine learning classifiers which have been chosen for this dataset. It will discuss applying different data processing methods and parameters as well as evaluate the effectiveness and efficiency of those algorithms.

3.3 Machine Learning classifiers

The two classifiers that have been chosen are the CART Decision Tree algorithm and the Random Forest algorithm, both of which are examples of supervised algorithms from ski-kit learn. The Random Forest is an example of a black box algorithm and the Decision Tree is a white box algorithm.

The reason for choosing classifiers that produce tree diagrams is because the dataset contains a lot of features with ordinal data (i.e. rating from 0 to 5), as well as categorical data with 2 to 3 different categories, which would likely not map on a graph as well as on a tree diagram. Furthermore, tree diagrams are easy to interpret and force consideration of all possible outcomes before reaching a conclusion.

The Random Forest algorithm is different from the Decision Tree algorithm as it produces multiples different trees, making it less practical as a white box model as executives in the airline company will likely be more interested in results and wouldn't have to time to go through every tree diagram.

3.4 Data processing methods

Section 2 goes into detail about different anomalies and dataset characteristics that can have a detrimental impact on the machine learning process and produce models that are biased or inaccurate – missing values, redundant features, outliers, and class imbalance.

Several different ways of dealing with those variables were discussed and tested with machine learning classifiers chosen for this dataset. This section lists the methods which yielded the best accuracy and performance.

To further experiment with feature selection, more features were removed, this time based on a correlation coefficient with other features other than the target variable of more than >50%. This experiment followed a though process of less features meaning a model that is not overfitted and more adaptable to test data. This led to a decrease in accuracy of at least 3% due to a loss of multiple key features and lots of data, so those features were kept.

Missing values were imputed with the mean using ski-kit learn as removing records with missing values significantly reduced the amount of data in the dataset, meaning machine learning classifiers had a lot less data to train on which yielded poorer accuracy.

Removing outliers did not have a significant impact on the accuracy of the models, so it was decided to leave them in the dataset to preserve the authenticity of the models to real-life situations such as long flights and flight delays which sometimes occur in airlines.

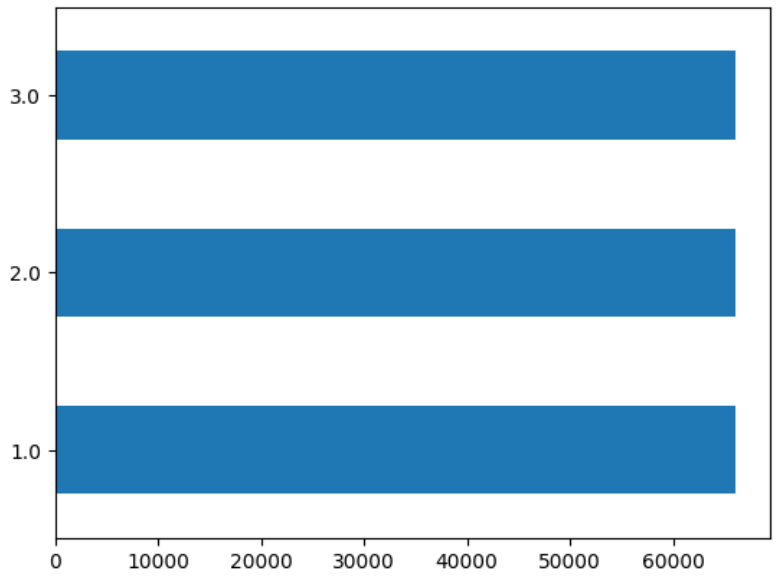

Figure 3.1

Figure 3.1 shows the class distribution of customer satisfaction after a combination of both over sampling and under sampling was used to correct class imbalance. This was done using RandomUnderSampler, SMOTE and Pipeline libraries in ski-kit learn. Although this did not show a significant increase in accuracy over using just over sampling, using both resampling methods seemed more appropriate as this method would balance creating data duplicates and loosing data.

3.3 Hyperparameter tuning

To find the best combination of parameters to apply to the data, Grid Search and Randomized Search algorithms can be used to test different combinations of parameters and find the most effective one. Randomized Search is faster as it only tests randomly chosen sets of parameters, whereas Grid Search exhaustively tests every possible combination of parameters. This makes Randomized Search a better choice for larger datasets, and therefore it was chosen for this dataset.

The parameter max_depth is the number of splits a tree algorithm performs before it stops. Value of 4 and 12 were tested for both Decision Tree and Random Forest, which is against the recommended settings produced by the Random Search Algorithm of 30 and 200 respectively. Larger tree diagrams are harder to interpret, which defeats the point of using a white box algorithm. However, generating larger tree diagrams produce better results in terms of accuracy and performance, so it may be worth sacrificing readability for accurate classification as this information is more important for the airline business.

The parameter n_estimators is the number of trees the Random Forest algorithm generates. A value of 25 was chosen manually by running the algorithm multiple times, incrementing the value by 5 each time until no further improvements in accuracy could be observed.

Any other parameters were set to default as changing those values either lead to either negligible or detrimental results to accuracy.

3.4 Results and Validation

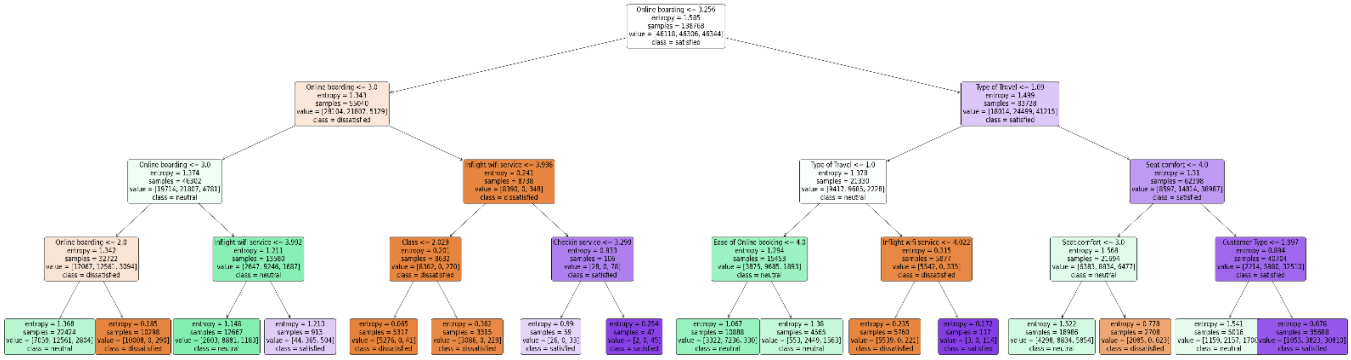

Figure 3.2

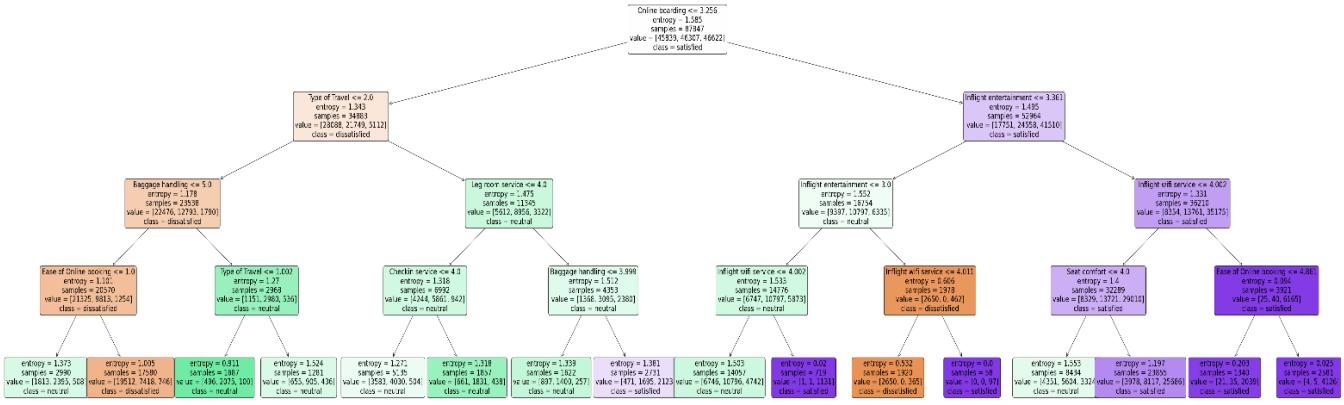

Figure 3.3

Figure 3.2 above is a tree diagram produced by the Decision Tree algorithm and Figure 3.3 is one of the tree diagrams produced by the Random Forest Algorithm. Those diagrams were generated with a maximum depth of 4 splits for demonstration purposes, as larger decision trees become unreadable. In practice using a depth of 12 splits is better as this makes the models more accurate.

The tree diagrams above suggest that the machine learning algorithms may be performing redundant operations as it continues splitting even if the accuracy of the prediction at a node is accurate enough. For example, at the rightmost nodes of Figure 3.3, the decision tree going further even when entropy is 0.094. Entropy in machine learning is a measure of disorder, the closer it is to 0 the more accurate a prediction is. The airline may not need predictions more accurate than entropy of 0.094.

This exposes a flaw in the algorithm which could be solved through applying some pruning methods to stop the algorithm from doing redundant operations after a certain degree of accuracy is achieved.

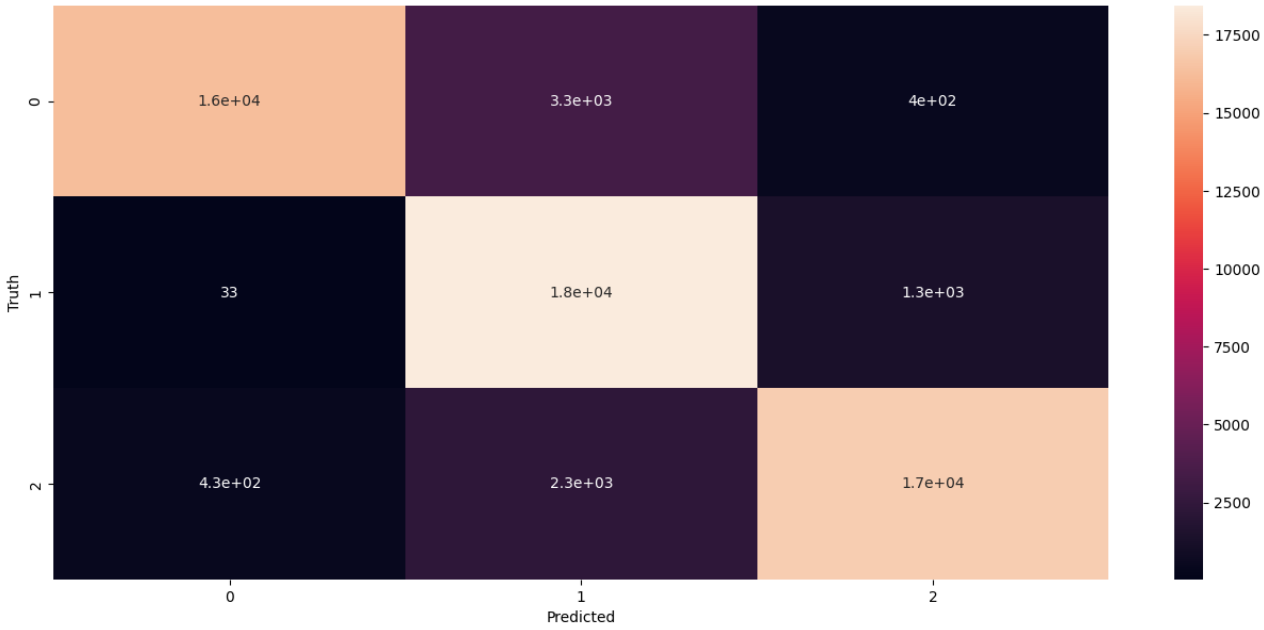

Figure 3.4

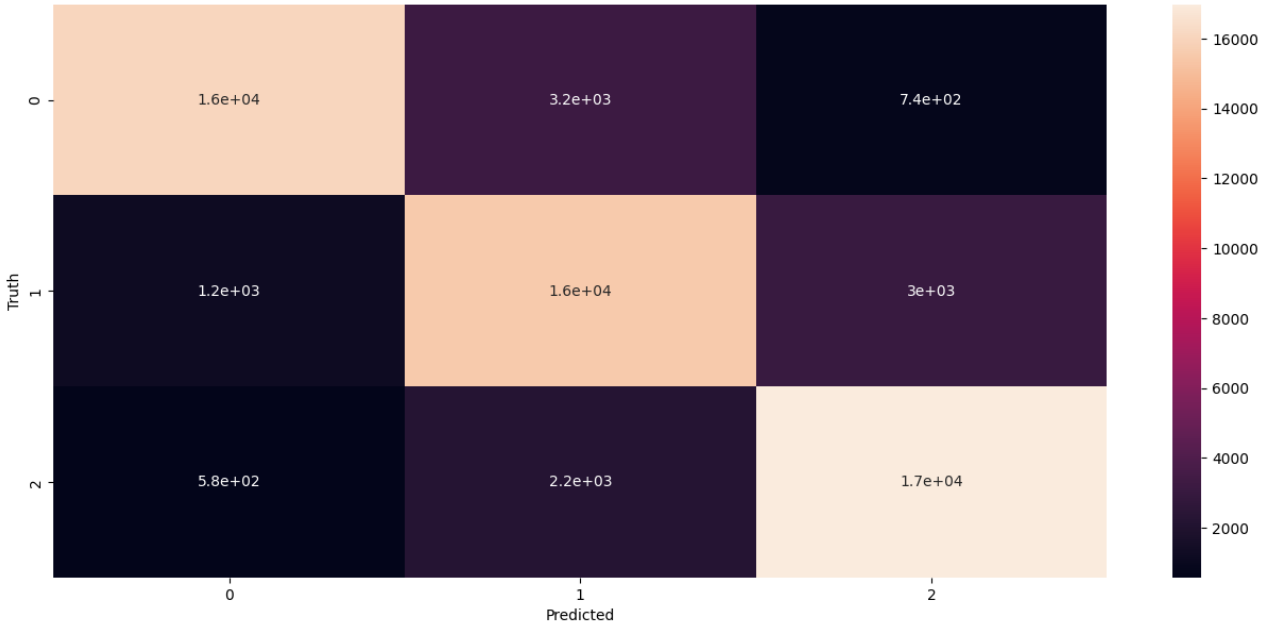

Figure 3.5

Hold out validation was applied to the machine learning algorithms with 5 folds, meaning the dataset was split into 5 smaller datasets, 4 of which were for training and 1 for testing. A maximum depth of 12 was used for both algorithms as no further significant improvements were observed by increasing the maximum depth beyond this point.

The Decision Tree classifier had a mean accuracy of 86% while the Random Forest classifier had an accuracy of 91%. This means that the objective of obtaining an accuracy of 90% was only achieved

for the Random Forest Algorithm. This also completes the secondary objective of proving that the Random Forest algorithm is a superior classifier.

The algorithms were tested for precision, which is a measure of quality of positive predictions made by the algorithms. The Decision Tree classifier had a precision of 97% for dissatisfied, 82% for neutral and 92% for satisfied (91% precision overall). The Decision Tree classifier had a precision of 97% for dissatisfied, 84% for neutral and 94% for satisfied (92% precision overall).

This shows that the Random Forest classifier is more precise and that both algorithms sometimes wrongly classified instances of dissatisfied and satisfied customers as neutral. This is reflected in confusion matrix diagrams in Figure 3.4 for Decision Tree and Figure 3.5 for Random Forest.

4 Conclusion

Overall, the objective of applying machine learning algorithms and achieving an accuracy of 90% was partially achieved as only the Random Forest algorithm had an accuracy of >90%. There is also sufficient evidence of the fact that Random Forest is a superior classifier to the Decision Tree, which completes the secondary objective.

Given more time on this project, different pruning methods would have been experimented with to reduce redundant operations carried out by the classifiers, reduce overfitting and produce diagrams which are more readable. In addition, a third machine learning classifier would have been tested as well as the performance in terms of efficiency for all machine learning models.

References

S. B. Kotsiantis, 2011. Feature selection for machine learning classification problems: a recent overview, p10.

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago & Oteng Tabona, 2021. A survey on missing data in machine learning. Journal of Big Data.

V. Ilango, R. Subramanian, V. Vasudevan, 2012. A Five Step Procedure for Outlier Analysis in Data Mining, p1.

Roweida Mohammed, Jumanah Rawashdeh, Malak Abdullah, 2020. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results.