‘Part 3: Report - Decision Trees and Random Forests’

**Clodagh King x15474222**

*Higher Diploma in Science in Data Analytics,*

*National College of Ireland*

*International Financial Services Centre,*

*Dublin.*

**Word Limit for Assignment:** 1,000 **Actual Word Count: 1,000 (Not including Figures and Index)**

**Data and Web Mining**

**19th March 2021**

**Lecturer: Neesha Rathe**

***Introduction***

This report will examine the use of a Decision tree classification model on my chosen dataset, I will then apply the Random Forest technique to increase the proficiency of the model and assess the extent to which this improves the model. In brief the dataset I have chosen records the satisfaction of airline passengers based on a number of variables such as online support, and inflight entertainment, I will discuss in detail all of the variables provided in the dataset and a rationale behind choosing this dataset. The aim of this report is to evaluate how accurate the decision tree and random forest models are at predicting the satisfaction of airline passengers.

A decision tree is a process of dividing data into smaller portions to identify patterns that can be used for prediction. (Lants, 2013)

***Chosen Dataset***

The dataset I have chosen for this assignment records the demographics of airline passengers and shows whether the customer is satisfied or dissatisfied with the airline and their journey. This data set can be found through the following [link](https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction). The dataset contains 129,487 rows and 23 columns/variables.

A factor which influenced me to choose this particular dataset was its scale, the amount of data included should help increase the accuracy of the model produced.

The main motivation behind choosing this dataset on airline customer satisfaction is there are a lot of variables which may be used to try and predict whether or not a customer will be satisfied or not. It is my goal that this model could provide an indication as to which factors (for example, inflight wifi or food and drink) have the most influence over customer satisfaction. In an official scenario the airline could use this model to improve these ‘important’ areas of their business to help increase the level of satisfaction. Included in the Index section at the end of this report is a description of the process to prepare the dataset.

Below in Figure 1 there is a table including the name of each variable and a description of the data it describes including it’s assigned datatype(factor, integer, character).

Table

Description automatically generated

***Decision Tree***

The rpart function was used to form the classification decision tree, and ‘satisfaction’ was chosen as the dependent variable.

Model 1 – Figure 3

![Timeline

Description automatically generated]()

Figure 3 shows a plot of the classification decision tree model formed from the training data. In brief this Model shows us that 40% of customers who were satisfied, rated the inflight entertainment and online booking at over 3.5. On the other hand, customers who rated the inflight entertainment and seat comfort at less than 3.5 were generally dissatisfied. Depending on the accuracy this model produces it suggests that the satisfaction of the customer relies on the inflight entertainment and the comfort of their seat.

Next, I used the test data and the predict function to evaluate this Model and produce a confusion matrix.

Figure 4 – Confusion matrix

|  |  |
| --- | --- |
| True negative | False negative |
| False positive | True positive |

![Table

Description automatically generated with medium confidence]()

The confusion matrix at the top shows the number of values which the model predicted correctly and incorrectly, the values predicted correctly have been highlighted in green and the values in red show the incorrect predictions. The confusion matrix is flipped from what a conventional confusion matrix would show, I have drawn out the true labelled structure beside it.

As gathered by the sensitivity and specificity values the model was slightly better at predicting Satisfaction.

The overall accuracy of the model has been calculated at 86.4%, this would be considered a high level of accuracy, especially in the context of calculating whether a customer will be satisfied or not.

Following this, I attempted to improve the decision tree model by pruning the tree, to do this I found the relevant CP value to prune the tree, this was a value of 0.01. I then applied this and found that the decision tree remained exactly the same.

To test why the pruned model didn’t change from the original model, I changed the CP value up slightly to 0.015, to produce another pruned model which I could then predict a confusion matrix from, so that I could compare to see which model was more accurate.

Figure 5 – Pruned tree model.

![Diagram

Description automatically generated]()

Text

Description automatically generated with medium confidence

Figure 6

Figure 4 above shows the confusion matrix produced by the ‘pruned’ decision tree, despite this model producing an even more accurate result for Satisfaction, overall, it was less accurate than the original decision tree model. This led me to conclude that the reason the original decision tree model did not change when pruned at the optimal CP value of 0.01 is because the original model did not need pruning. And in pruning it at 0.15 leads to an underfitting model.

![Chart, line chart

Description automatically generated]() ![Chart

Description automatically generated]()

Figure 7 – ROC Curve Figure 8 – AUC curve

***Random Forest***

A random forest can be described as a type of classification algorithm that operates by forming a multitude of decision trees that together form an ensemble. Each individual tree in the random forest produces a class prediction and the class with the greatest number of votes becomes the model’s prediction. (Pal, 2015) It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. The individual trees that make up the random forest come together to form a collection that is greater than the sum of its parts. (Biau, 2016)

The process for forming the model used the randomForest function and like with the decision tree model the training set was used to form the original random forest model which was then used against the test data to predict and form a confusion matrix. The results to the confusion matrix were as follows.

Table

Description automatically generated with medium confidence

Figure 7 shows the confusion matrix produced by the random forest model.

The most important observation is that the overall accuracy of this model is 95% which is a considerable difference from the original model which had an accuracy of 86.4%. This is an example of how a random forest model can be stronger than a single decision tree alone. A 95% accuracy in the context of customer satisfaction is considered to be very high.

The confusion matrix shows that the random forest model predicted 17014 dissatisfied customers correctly and 20202 satisfied customers correctly.

The Sensetivity and specificity also show how the random forest model is slightly better at predicting dissatisfied customers.

***Conclusion***

In conclusion decision trees and random forests are both very useful algorithms to use for predicting in this case the satisfaction of airline customer. The decision tree model could be useful to see what specific variables are significant to having satisfied customers, like in this case the in-flight entertainment. The random forest model produced a very high overall accuracy and therefore could be used in a real-life context to help the airline improve their overall satisfaction rating.

***Index***

Process

After importing the dataset, the first step was to check and remove any missing or NA data, luckily the chosen dataset did not contain any missing data. I then looked at a summary of the data, this was useful for showing a general overview of the data such as that there was a pretty even split of male and female customers surveyed and that most of the customers surveyed were travelling for business purposes.

The summary of the data can be viewed in the table below, Figure 2.

Figure 2

A picture containing text, receipt

Description automatically generated

Following this I checked the attributes of the data to show the data type and convert any variables that needed to be. For this dataset the following columns had to be converted into factor variables, ‘Class’, ‘type of travel’, ‘customer type’, ‘gender’, ‘Satisfaction’.

this was necessary as the rpart function for calculating decision trees needs the dependent variable to be a factor.

It was decided that all the variables included in the dataset would be relevant and useful to forming the model and therefore it was not necessary to exclude any columns.

The data is then split into training and testing set – the reasoning for this is so that the training set is used to form the model and then the test set is used to evaluate how well the training model can predict the test data. As this is a large dataset it was decided to go with a split of 70% training and 30% testing.

***Bibliography***

Biau, G., Scornet, E. A random forest guided tour. *TEST* **25,**197–227 (2016).

Brett Lantz, Machine learning with r, Packt Publishing, 2013.

M. Pal (2015) Random forest classifier for remote sensing classification, International Journal of Remote Sensing,