Assignment Two

Semester 1 2023

Student Name and ID: Susan Philip – 21139319
Student Name and ID: Samantha Heuss – 21141141

PAPER NAME: Foundations of Data Science

PAPER CODE: COMP615 **–** Semester Two

Due Date: 4th Jun 2022 (midnight NZ time)

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section**.

Contents

List of Figures

Analysing Customer Behaviour and Preferences: Exploring Coupon Acceptance Patterns for Dining Experiences Through In-Vehicle Recommendation Systems

Introduction

> *You are expected to provide information on the dataset you are assigned to use for your assignment. Provide a statement of the problem, outlining the problem that your dataset addresses.*

A research article from TVI MarketPro3 describes how business owners in the automotive industry are always seeking to find new ways to bring in new customers [1] and utilising coupons might be the way to go.

The in-vehicle coupon recommendation dataset provides valuable analytics into customer behaviour and preferences in relation to accepting coupons for nearby coffee houses, restaurants, takeaway and bars. This dataset provides an opportunity to investigate the factors that impact an individual's choice to accept a suggested coupon when it is presented through their car's mobile recommendation device.

The research question that this report will aim to answer is based on the relationship between age and income and how these two factors influence the acceptance of coupons among individuals.

This report will seek to understand whether individuals' age and income levels play a significant role in determining the likelihood of accepting coupons offered through in-vehicle recommendation systems and provide insights into the preferences and behaviours of various age groups and income brackets. By examining this relationship, we can determine whether certain age groups or income levels are more receptive to coupons and how other factors might influence their decision-making process. This analysis can be valuable for business and marketing firms, in designing targeted coupon strategies that align with the preferences and characteristics of specific age and income groups.

Data Exploration

The in-vehicle coupon recommendation dataset was collected from a survey on Amazon Mechanical Turk [2], this dataset was retrieved from the UCI Machine Learning Repository. The dataset comprises of 23 attributes that contain diverse aspects of the data, with a range of data types including strings and integers. Demographic information like age, gender, income, marital status, education and occupation are included. Attributes describing trip details are also stored such as 'direction opp', which store either a 1 or 0 to describe whether the restaurant or bar of the coupon is in the same direction as the driver's current destination, 1 indicating yes and 0 indicating no. 'Direction same' is also similar but stores whether the restaurant or bar of the coupon is in the same direction of the driver's current destination. The toCoupon_GEQ attributes store an integer either 1 or 0 if the driving distance to the restaurant or bar is greater than 5, 15 or 25 minutes. Weather is stored as a string describing the weather conditions either sunny, rainy or snowy. The temperature is stored as an integer describing the temperature as 30 degrees, 55 degrees or 80 degrees. The passengers are included as an attribute which describe whether there are any passengers in the car, either a friend, kids or if they are alone. Destination stores information of where the driver is going, either no urgent place, home or work. The number of times the driver goes to a coffeehouse, restaurant or carry away is stored as categorical datatypes. The expiration of the coupon, type of coupon and whether the coupon is accepted is also stored in the dataset. For the analysis, only certain attributes will be utilised that are compatible with the specific analysis we aim to perform as some attributes have incompatible datatypes.

In terms of the data cleaning and pre-processing stages, cleaning and changes in datatypes were needed to help reduce the impact of noise in the data. In total the dataset comprises of 12,684 instances with a total of 23 attributes. There are six attributes in the dataset that contain missing values which add to about 13,370 missing values. To ensure integrity of the data, the decision has been made to remove these null values from the dataset. By eliminating the missing values, we can ensure that the data is free from any inconsistencies.

As part of the data pre-processing stage. We made certain changes to the datatypes of specific attributes to ensure an efficient analysis. Age was initially stored as strings with values like 21, 46, 26, 31, 36, 41, below 21 and 50 plus. To ensure consistency and

simplicity we converted the instances listed as ‘below 21’ and ‘50plus’ from strings to integers. To achieve this, a for loop was used to iterate through the instances with the value ‘below21’ to 20m assuming it represents the age group below 21. Similarly, the instances denoted as ‘50plus’ was changed to just 50, assuming this represents the age group 50 and above. These modifications were made to streamline the analysis process and ensure the attributes datatype aligns with the subsequent analysis steps.

A similar approach was utilised for the income attribute, where income was changed from range values to single numbers to provide consistency, simplify calculations and allows for easy comparisons. The approximate mean value was taken for each instance which replaced the range value, this in turn creates a representative income value that is fair and reflective of the range. This approach allows the income attribute to be treated as a numerical attribute. Similarly, the expiration attribute’s data was altered to change it to be all numerical. The only values originally in this dataset for this attribute was “1d” or “2h”. This was changed again using a for loop to iterate through all instances, changing “1d” to “24” and “2h” to “2”. The data has been made numerical, with the expiration in hours.

Figure 1 provides summary statistics for the numerical attributes in the dataset. The ‘temperature’ attribute has a count of 12684 instances. The mean temperature is 63.3 and a standard deviation of 19.15. The minimum temperature recorded is 30 and maximum is 80. The ‘has children’ attribute has a mean of approximately 0.41, this indicates that 41% of the instances have children.

The toCoupon_GEQ5min has a mean of 1, this suggests that all the instances in the dataset have a driving distance that is greater than 5 minutes to redeem the coupon.

With the toCoupon_GEQ15min attribute, the mean value is 0.56, this implies that around 56% of the instances have a driving distance greater than or equal to 15 minutes to redeem the coupon at a restaurant or bar, this shows that more than half of the instances have a commute of 15 minutes or more to arrive at the destination of the coupon. Similarly, with toCoupon_GEQ25min the mean is denoted as 0.12 meaning that only 12% of the instances need to travel 25 minutes or more to redeem their coupon at the corresponding restaurant, bar, takeaway or coffee house. The ‘direction same’ attribute has a mean of approximately 0.21 suggesting that around 21% of the instances have been given coupons to a restaurant, bar, takeaway or coffeehouse that is in the same direction as the current destination. The ‘direction opp’ attribute has a mean of 0.79, showing that 79% of the instances in the dataset have been given a coupon that is redeemable at the opposite direction of their destination.

The target variable 'Y' representing the coupon acceptance has a mean of approximately 0.57, indicating 57% of the instances have accepted the coupon.

	temperature	has_children	toCoupon_GEQ5min	toCoupon_GEQ15min	toCoupon_GEQ25min	direction_same	direction_opp	Y
count	12684.000000	12684.000000	12684.0	12684.000000	12684.000000	12684.000000	12684.000000	12684.000000
mean	63.301798	0.414144	1.0	0.561495	0.119126	0.214759	0.785241	0.568433
std	19.154486	0.492593	0.0	0.496224	0.323950	0.410671	0.410671	0.495314
min	30.000000	0.000000	1.0	0.000000	0.000000	0.000000	0.000000	0.000000
25%	55.000000	0.000000	1.0	0.000000	0.000000	0.000000	1.000000	0.000000
50%	80.000000	0.000000	1.0	1.000000	0.000000	0.000000	1.000000	1.000000
75%	80.000000	1.000000	1.0	1.000000	0.000000	0.000000	1.000000	1.000000
max	80.000000	1.000000	1.0	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 1. Table of Summary Statistics

To ensure accuracy and integrity for certain analyses of the data, a filtering process was applied to create a new dataset that includes the appropriate specific attributes. These attributes are age, income, expiration, toCoupon_GEQ5min, toCoupon_GEQ15min, toCoupon_GEQ25min, direction_same, direction_opp and Y. This step is a part of the data pre-processing phase, where the preparation of data for further analysis or modelling tasks occurs. The attributes that have been extracted to a new subset, specifically have no null values which will ensure that the data used for the analysis is complete and reliable. Removing the attributes with missing values also ensures simplicity in the dataset and avoids potential issues during modelling or analysis due to incomplete information.

The visualisation of the in-vehicle coupon recommendation dataset provides valuable insights that can be visualised by bar graphs and box plots. *Figure 2* examines the acceptance of coupons based on the type of coupon given. The coffee house coupon is the most accepted, with approximately 2000 instances. Interestingly, the coupon for restaurants with an average spend of $20 to $50 shows the lowest number of acceptances, if most drivers in this dataset usually tend to visit restaurants in groups, they may not spend an average that exceeds $20 to $50 which could resonate as to why this is the least accepted coupon.

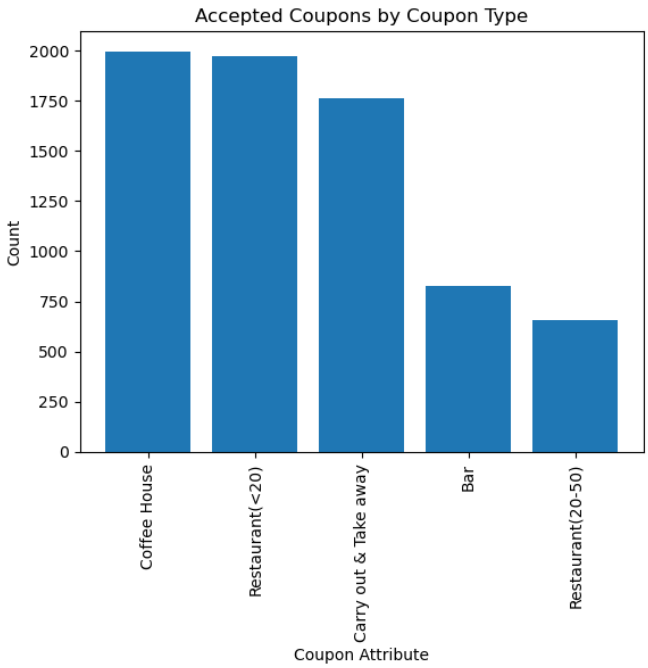

Figure 2. Bar Graph representing the accepted coupons categorised by type.

The accepted coupons categorised by income is shown in *Figure 3* as a bar graph. Drivers with an income bracket of $75,000 to $87,499 accept the least number of coupons, while those with an income bracket of $25,000 to $37,499 have the highest acceptance rate. An interesting observation in this bar graph is that the drivers with the highest income in the dataset exhibit the lowest acceptance rate of coupons. A possible reason as to why could be that individuals with higher income levels may be less inclined to seek out discounts from coupons.

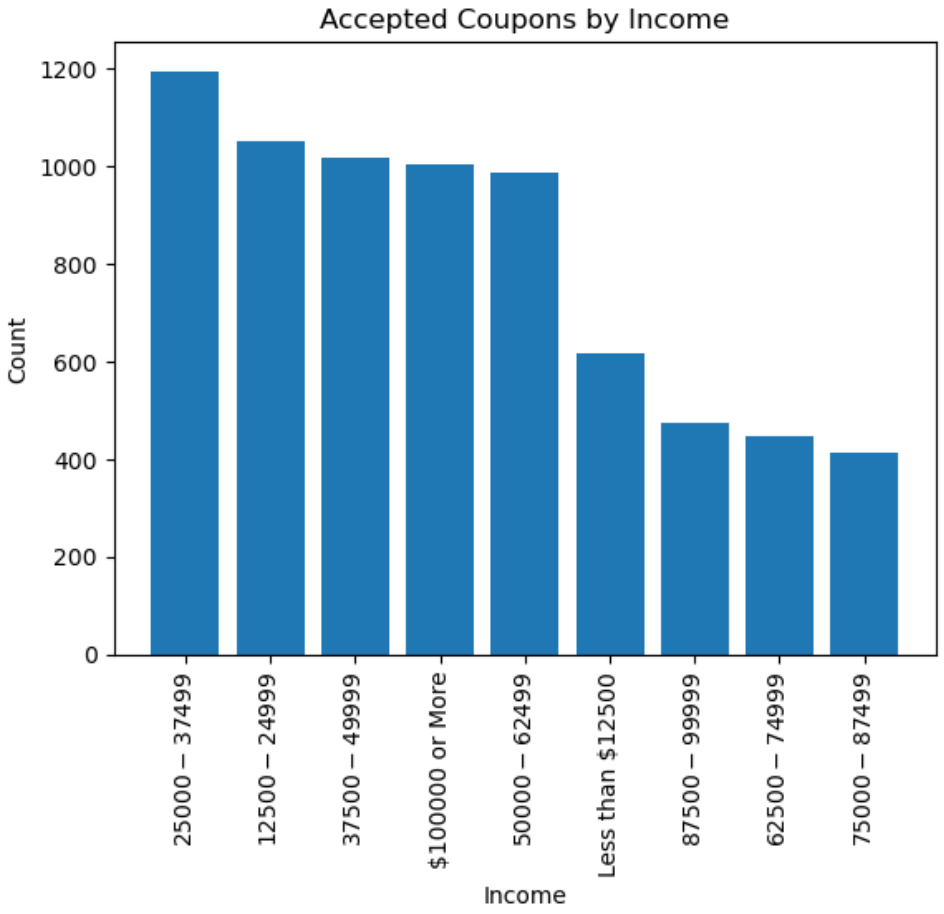

Figure 3. Bar graph representing the accepted coupons by Income.

The accepted coupons by age are also displayed in a bar plot as *Figure 4*, assuming that under 21 are denoted as 20 but represent individuals who are under 21 and 50 plus individuals are set to 50 but represent 50 plus, it is intriguing to see that those who are aged 21 accept the most coupons, following closely are individuals who are 26. Those who are aged less than 21 accept the least coupons out of all ages assessed. This could be because there are less under 21's on the road than those who are aged 21 to 26.

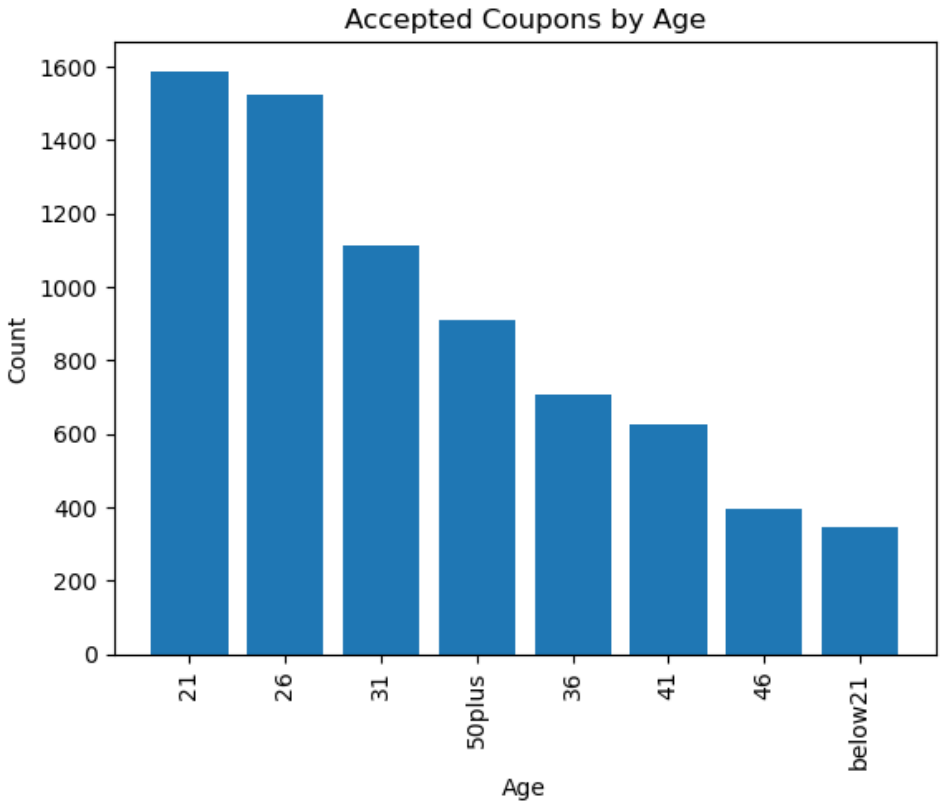

Figure 4. Bar graph representing accepted coupons by age.

The acceptance of coupons based on the time until expiration is depicted in *Figure 5*. It is clear to see that coupons with a longer expiration period of one day are more likely to be accepted compared to coupons that are offered to individuals with an expiry of 2 hours. Over 4000 coupons with a 1-day expiry date have been accepted, while the number of coupons with a expiry date of 2 hours decreases to less than 3000 which is still significantly high for such a limited amount of time. A possible reason for this trend is that individuals who are en route to their destination may find a takeaway or coffee coupon appealing at the time. However, it is logical for most individuals to accept a coupon with a longer validity as this allows individuals to redeem the coupon flexibly according to their schedule.

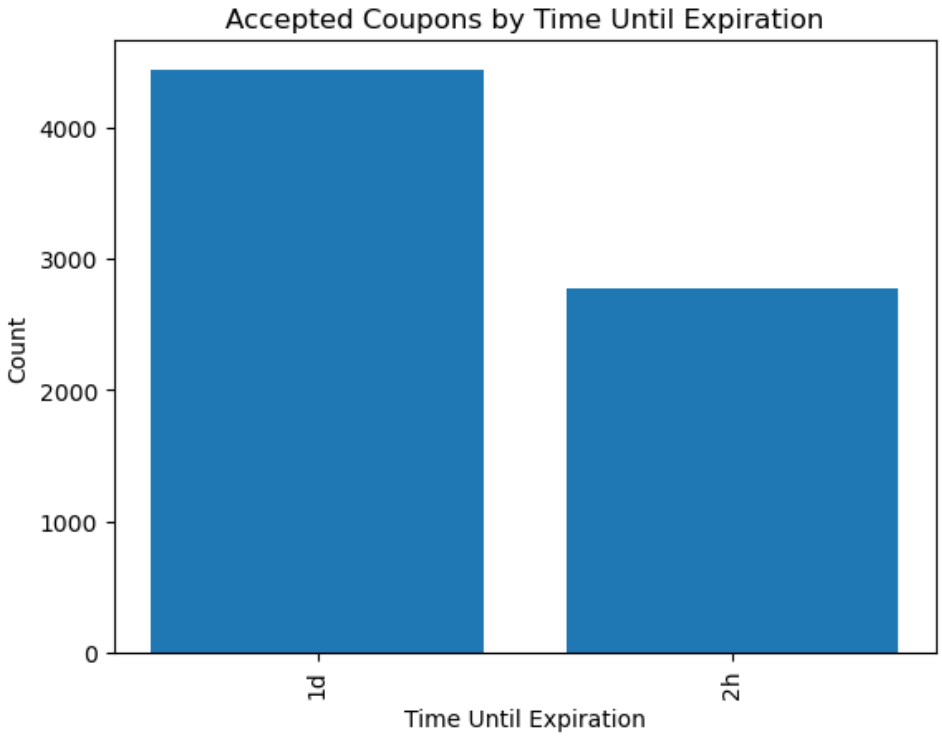

Figure 5. Bar graph representing the accepted coupons by time until expiration.

Below *Figure 6* showcases the accepted coupons according to gender categorised as Females and Males. Among both genders, there is a nearly equal number of accepted coupons. More than 3500 have been accepted by Male drivers and about 3500 have been accepted by Females. This observation suggests that coupon acceptance is not significantly influenced by gender.

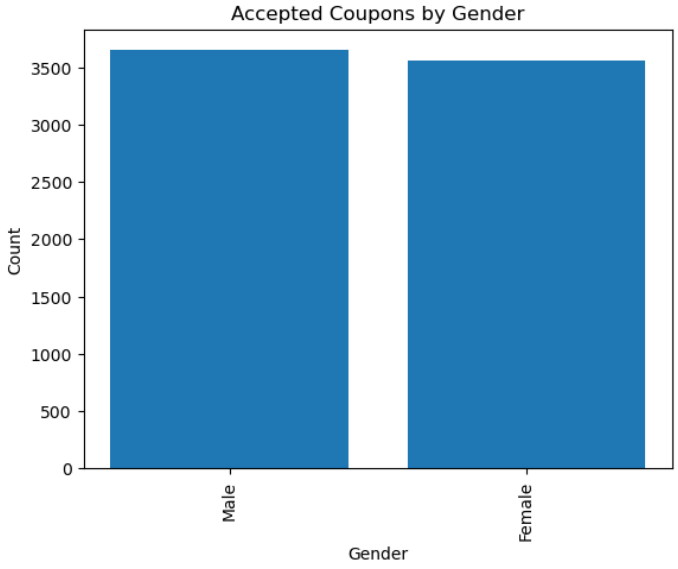

Figure 6. Bar graph showing the accepted coupons categorised by Gender, i.e Female or Male.

The decision to further visualise the age attribute by a box plot was made as it provides additional insights and a more detailed understanding of the distribution and variability within the different age groups in the dataset. *Figure 7* captures a clearer overview of the variability of each age group's acceptance rate. The median, quartiles and potential outliers can be seen through this box plot. The upper quartile for both accepting and not accepting is similar at about the age of 42. The middle quartile is also similar, just above the age of 30. However, the lower quartile for accepting and rejecting the coupons vary, meaning that for the lower quartile of accepting coupons is around 21 which suggests that individuals in this age range are more likely to accept coupons compared to those above 26. This also shows that younger demographics are more receptive to accepting coupons. The lower quartile for the rejected coupons segment is seen as about 26, this implies that individuals who are older than 26 are more likely to reject coupons. This age group may not be as motivated as those who are under 26 to accept a coupon, this could be because they have a busier schedule, so they are less likely to redeem the coupon resulting in declining the coupon.

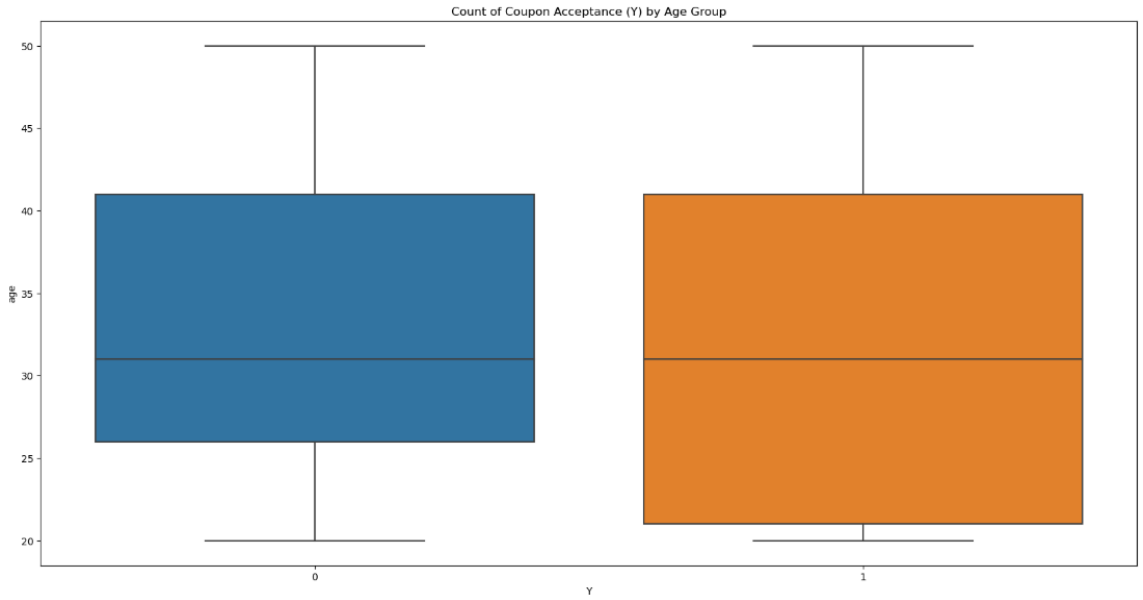

Figure 7. Box plot representing the count of accepted coupons by age group.

Below is a linear regression graph illustrating the relationship between income and the count of accepted coupons. *Figure 8* displays the comparison of the actual values with the predicted values. The majority of actual income categories aligns with most of the predicted income categories however with the income bracket of $100,000 or more the actual value is higher than the predicted value, this is a similar case with the individuals with an income bracket between $50,000 and $62,499, where the predicted number of accepted coupons falls around 800 but the actual number of accepted coupons in this income category is above 1000.

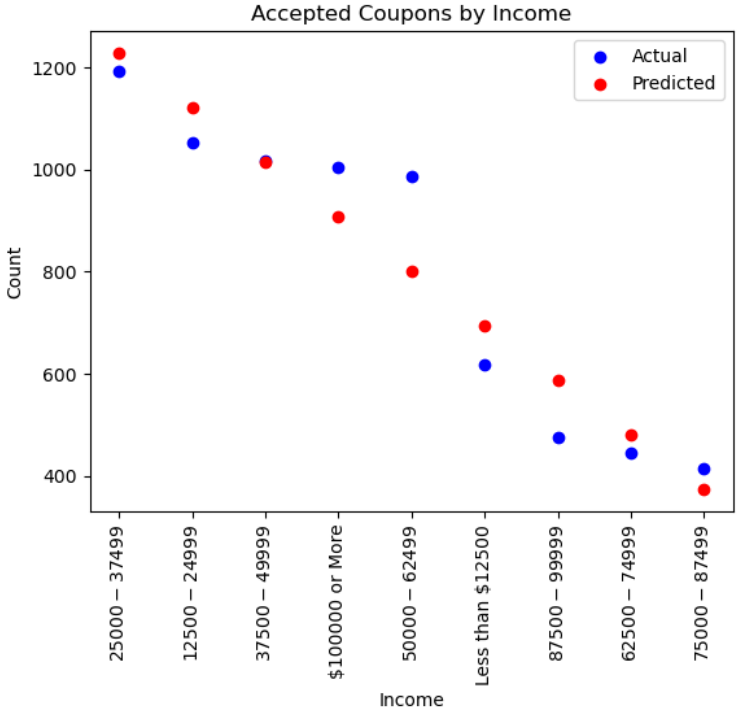

Figure 8: Linear regression analysis graph showing the predicted number of accepted coupons by income and the actual number of accepted coupons by income.

Decision Tree Classifier

> *You are required to build a model using the **Decision Tree Classifier** and answer the following questions based on the model built. In building the model, use the 10-fold cross validation option for testing. Your answers need to be supported by suitable evidence, wherever appropriate. Some examples of suitable evidence are Confusion Matrices, Model Visualizations, and Model Summary Report.*

To further analyse this data set, a series of decision tree models have been produced. Decision trees are a useful tool to break down complex data sets such as this one into more manageable components for analysis. To achieve this, the decision tree classifier that will be used, amongst other tools, requires the data being used to be all numerical. Therefore, the filtered dataset mentioned above in the data exploration section of this report, will be used. To begin, the target (output) and the predictors (inputs) need to be established. The

target value has been identified as the 'Y' attribute. The predictors/inputs are therefore the remaining attributes in the filtered data set that will affect the outcome of the target: age, income, expiration, toCoupon_GEQ5min, toCoupon_GEQ15min, toCoupon_GEQ25min, direction_same, and direction_opp. Next, the data needs to be split into testing and training sets. The training set will be used to train the model(s), and the testing set will be used to elaborate on the performance of the model(s). The data has been split to train 70%, and test 30%.

The first decision tree model has been produced (see Figure 9). To create this model, the criterion, used to specify the type of splitting used, has been set to Gini. The Gini Index is useful to determine how to split the features of this dataset during the construction of the decision tree (Karabiber, n.d.). Additionally, the Gini Index *"measures the probability for a random instance being misclassified when chosen randomly."* (Dash, 2022). For a lower chance of misclassification to occur, the Gini Index should also be a low value. The max_depth, which sets the depth/number of levels of the tree, has been set to the arbitrary value '4'. This value gives a good starting value to assess the accuracy of the model, and therefore what changes should be made to it to improve this. This tree has four levels plus the root. There are 31 nodes (including the root). The Gini Index ranges between 0.26 and 0.5. The model accuracy score with criterion gini index is 0.593.

```
Model accuracy score with criterion gini index: 0.593
```

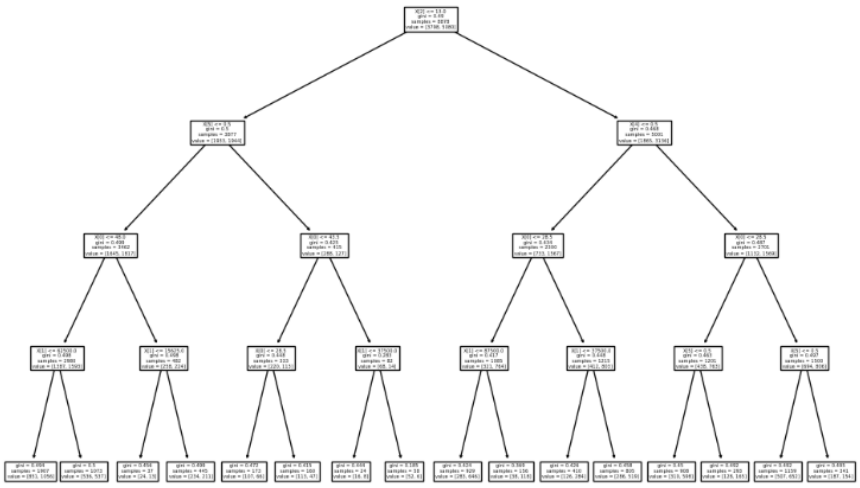

Figure 9: Decision tree model of initial specification where max depth is the arbitrary value four.

A confusion matrix has been produced (see Figure 10). This confusion matrix indicates that for class '0', 336 instances were correctly classified as '0', whilst 1340 instances were incorrectly predicted as '1'. It also indicates that for class '1', 1920 instance were correctly classified as '1', whilst 210 instances were incorrectly predicted as '0'. To further support these claims, the model performance summary has been calculated in the python file used for all calculations and models produced in this analysis report. The matrix suggests that this model has a relatively low accuracy in identifying those instances that belong to class

'0'. A higher precision score implies that fewer negatives will be misclassified as a positive. As the precision score here is 0.62 and 0.59 for classes '0' and '1' respectively, these positive values are likely to misclassified, as observed in the classification of instances for class '0'. Similarly, a higher recall value means that fewer positives will be misclassified as negatives. This is clearly observed in the confusion matrix. The recall score for class '0' is only 0.20, whilst for class '1' it is far greater at 0.90. Out of 1676 total instances of class '0' (correctly and incorrectly classified), only 336 have been correctly classified. This is opposed to 1920 instances out of a total of 2130 total instances correctly classified in the class '1'. A clear difference is observed here.

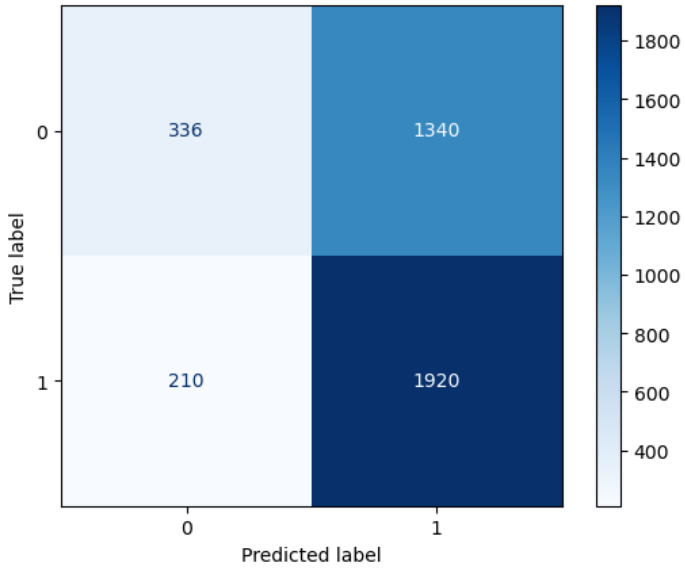

Figure 10: Confusion matrix plot generated from the specifications assigned to produce the initial decision tree model.

To improve the accuracy of the model, the parameters of the decision tree can be altered. Changing the depth of the decision tree, will affect how the complexity and accuracy of the tree; a deeper tree will increase the complexity but if it is too deep it can overfit the data and create noise. To determine the best depth to use, a 10-fold cross-validation has been used (see Figure 11 and Figure 12). For this, max depths from a depth of 1 to 9 have been tested, with the Average 10-Fold CV Score for each of these calculated. This score shows the accuracy of each of these depths tested. It also displays a count of the total nodes in each of these different max_depth models. This plot, along with the printed values, show that the highest score obtained is 0.594 (3 sig. figs). This value occurs when the max_depth value is set to '3'.

```
max_depth=1 Average 10-Fold CV Score:0.5719809681484962 Node count:3
max_depth=2 Average 10-Fold CV Score:0.5879069686506428 Node count:7
max_depth=3 Average 10-Fold CV Score:0.5939764165131639 Node count:15
max_depth=4 Average 10-Fold CV Score:0.5849888632843865 Node count:31
max_depth=5 Average 10-Fold CV Score:0.5907449667265763 Node count:63
max_depth=6 Average 10-Fold CV Score:0.5854613036420541 Node count:125
max_depth=7 Average 10-Fold CV Score:0.5804188946312578 Node count:233
max_depth=8 Average 10-Fold CV Score:0.5689840605757781 Node count:407
max_depth=9 Average 10-Fold CV Score:0.5632308158887124 Node count:627
```

Figure 11: Output of average 10-fold cv scores from 10-fold cross validation testing max depths from 1 to 9.

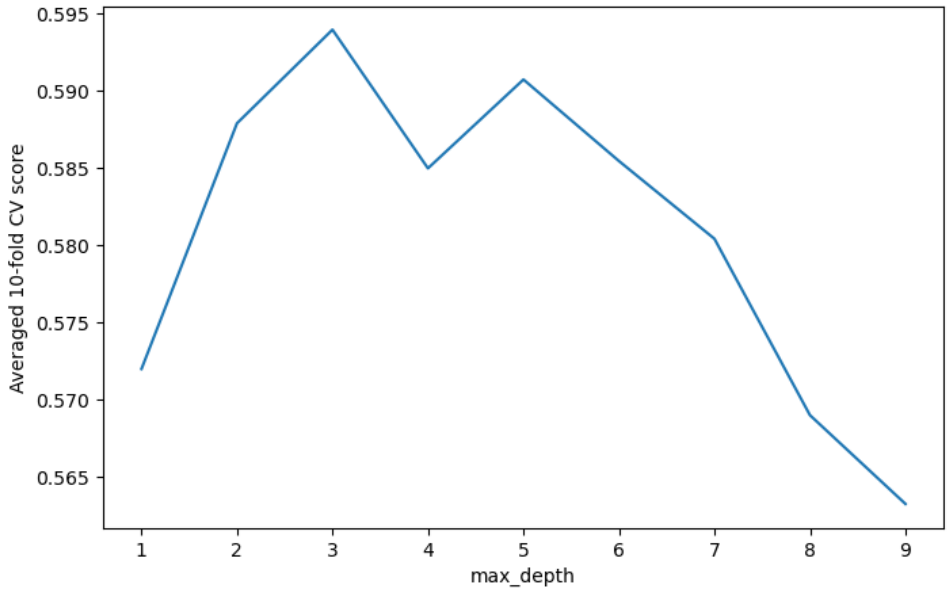

Figure 12: Line graph of average 10-fold cv scores from 10-fold cross validation testing max depths from 1 to 9.

Now with the best value for the max_depth calculated as '3', a second decision tree model can be plotted (see Figure 13). This tree has three levels plus the root. It contains 15 nodes including the root. The Gini Index ranges between 0.357 and 0.497. The model accuracy score with criterion Gini index is also 0.594. Compared to the first model, the model accuracy doesn't differ much. From the 10-fold cross-validation scores, a max_depth of three, concluded to be the best fit, only has a difference of 0.0089875532287774 between itself and a max_depth of four. Furthermore, these scores only range between 0.56 and 0.59 (approximately) for all 10 max_depth's tested. The confusion matrix produced from this second decision tree model (see Figure 14), and the classification report calculated, also show little changes to the first model. Interestingly, the precision, recall, accuracy, and f1-scores only ever so slightly change between the models. The arbitrary value '4' initially chosen for the max_depth by coincidence has an accuracy score almost identical to the best max_depth value identified as '3'.

```
Model accuracy score with criterion gini index: 0.594
```

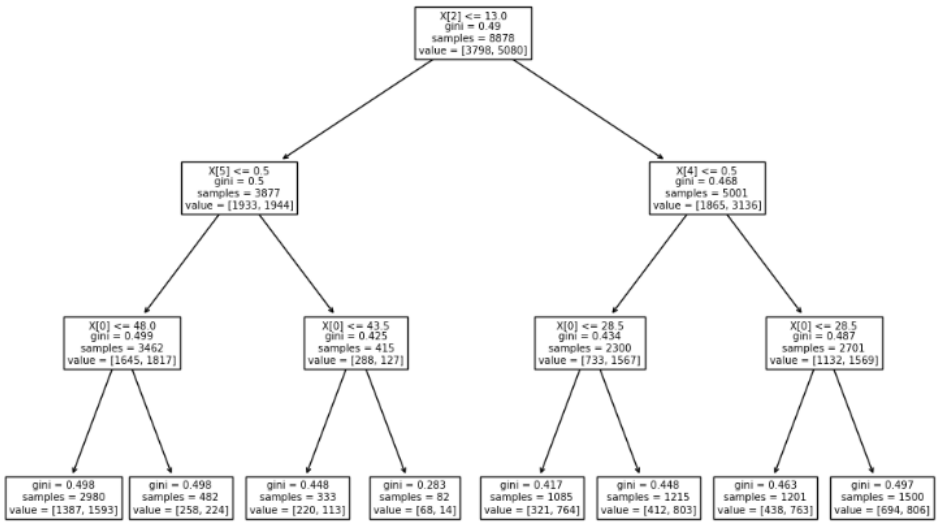

Figure 13: Decision tree model with max depth as best value calculated from 10-fold cross validation testing, three.

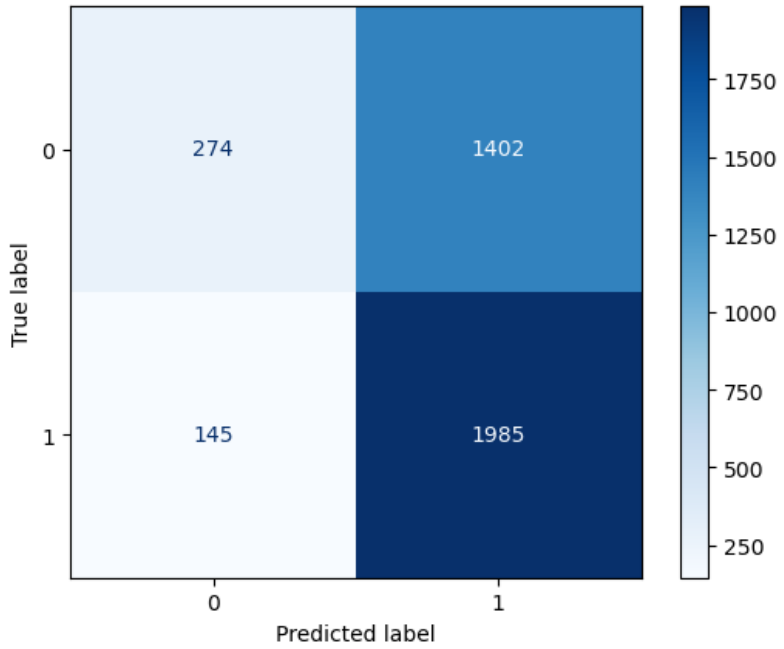

Figure 14: Confusion matrix produced from specifications for decision tree model with max depth as best value calculated from 10-fold cross validation testing, three.

Changing further parameters can also increase the accuracy of the model. The second parameter chosen to alter is the max_leaf_nodes. In a similar process to determining the best max_depth, a decision tree has once again been created. In this model, the max_depth has been set to the concluded best value '3'. The criterion has remained as 'gini', and the random state as '0'. The max_leaf_nodes an addition to the criteria of the tree. In the second decision tree produced (see Figure 13), the number of leaf nodes were eight. This value has therefore been used as the starting point for the max_leaf_nodes. The decision tree produced

(see Figure 15) has three levels plus the root. It contains 15 nodes including the root, and eight leaf nodes. The model accuracy score with criterion gini index is 0.594. This is the same as the previous decision tree model. Furthermore, as no changes have been made to the structure of the tree between decision tree model's two and three, the confusion matrix and classification report scores are also identical.

```
Model accuracy score with criterion gini index: 0.594
```

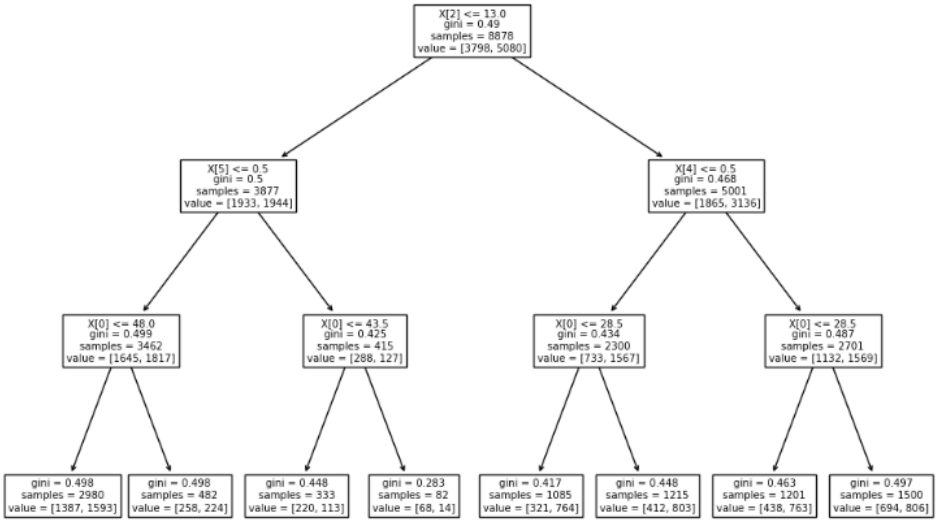

Figure 15: Decision tree model with max depth as best value calculated from 10-fold cross validation testing, three, and the max leaf nodes as value eight.

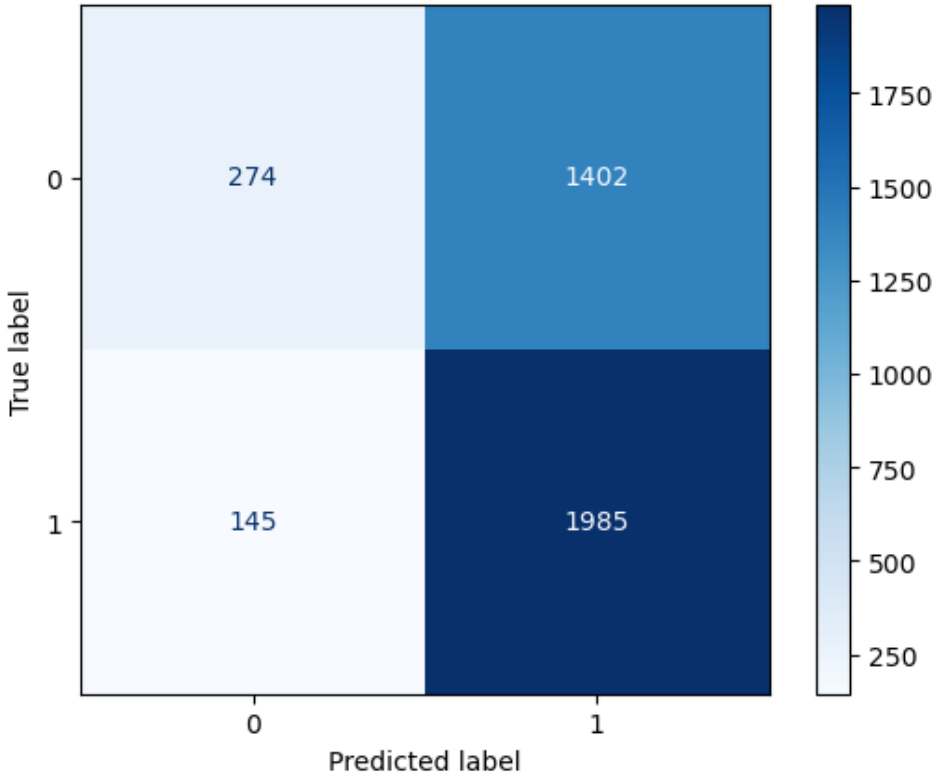

Figure 16: Confusion matrix produced from decision tree model with max depth as best value calculated from 10-fold cross validation testing, three, and the max leaf nodes as value eight.

As done so to determine the best max_depth, the 10-fold cross-validation can also be carried out to determine the best value for the max_leaf_nodes. Doing so, as observed in Figure 17 and the plot (see Figure 18), the best value for the max_leaf_nodes is '6'. However, observing Figure 17 and the plot in Figure 18 where the line levels off, the best value may also be '7', '8', '9', or '10'. As any of these values can be used according to these calculations, the first value '6' has been used to produce the final decision tree. This model has three levels plus the root, and eleven nodes including the root node. As the max_leaf_nodes value has been set to '6', the number of leaf nodes in this model is six. The model accuracy score with criterion gini index is also 0.594. This is the same value that has been produced by decision tree models two and three. The confusion matrix and classification report produced from this model is also the same as those produced from decision tree models two and three. After identifying and applying the value for the best max_depth, the best value for max_leaf_nodes has no effect on the model accuracy score when the max_leaf_nodes value is greater than or equal to '6'.

```
max_leaf_nodes=2 Average 10-Fold CV Score:0.5719809681484962 Node count:3
max_leaf_nodes=3 Average 10-Fold CV Score:0.5864885289343307 Node count:5
max_leaf_nodes=4 Average 10-Fold CV Score:0.5879069686506428 Node count:7
max_leaf_nodes=5 Average 10-Fold CV Score:0.5879069686506428 Node count:9
max_leaf_nodes=6 Average 10-Fold CV Score:0.5939764165131639 Node count:11
max_leaf_nodes=7 Average 10-Fold CV Score:0.5939764165131639 Node count:13
max_leaf_nodes=8 Average 10-Fold CV Score:0.5939764165131639 Node count:15
max_leaf_nodes=9 Average 10-Fold CV Score:0.5939764165131639 Node count:15
max_leaf_nodes=10 Average 10-Fold CV Score:0.5939764165131639 Node count:15
```

Figure 17: Output of average 10-fold cv scores from 10-fold cross validation testing max leaf nodes from 2 to 10.

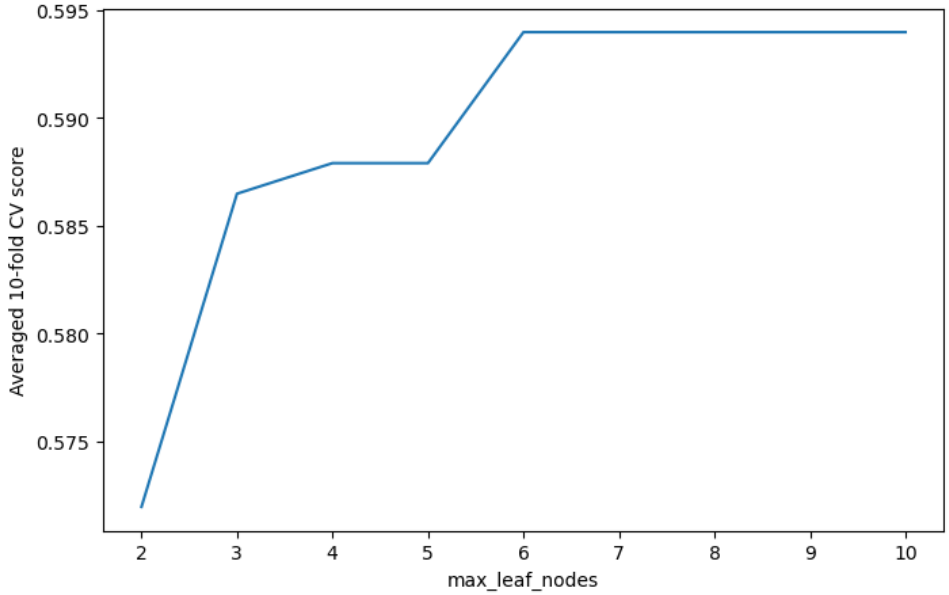

Figure 18: Line graph of average 10-fold cv scores from 10-fold cross validation testing max leaf nodes from 2 to 10.

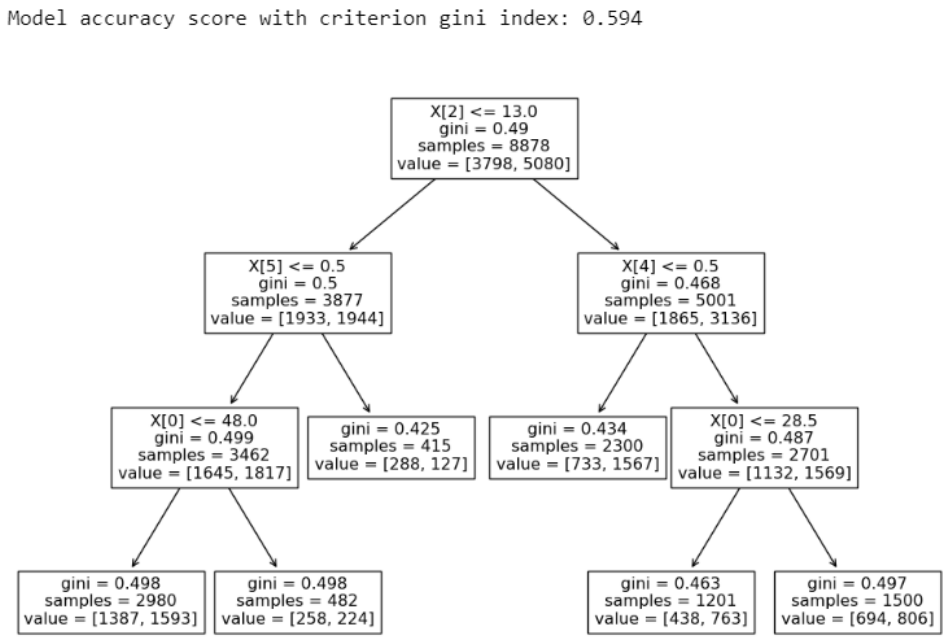

Figure 19: Final decision tree model with max depth as three, and max leaf nodes as six.

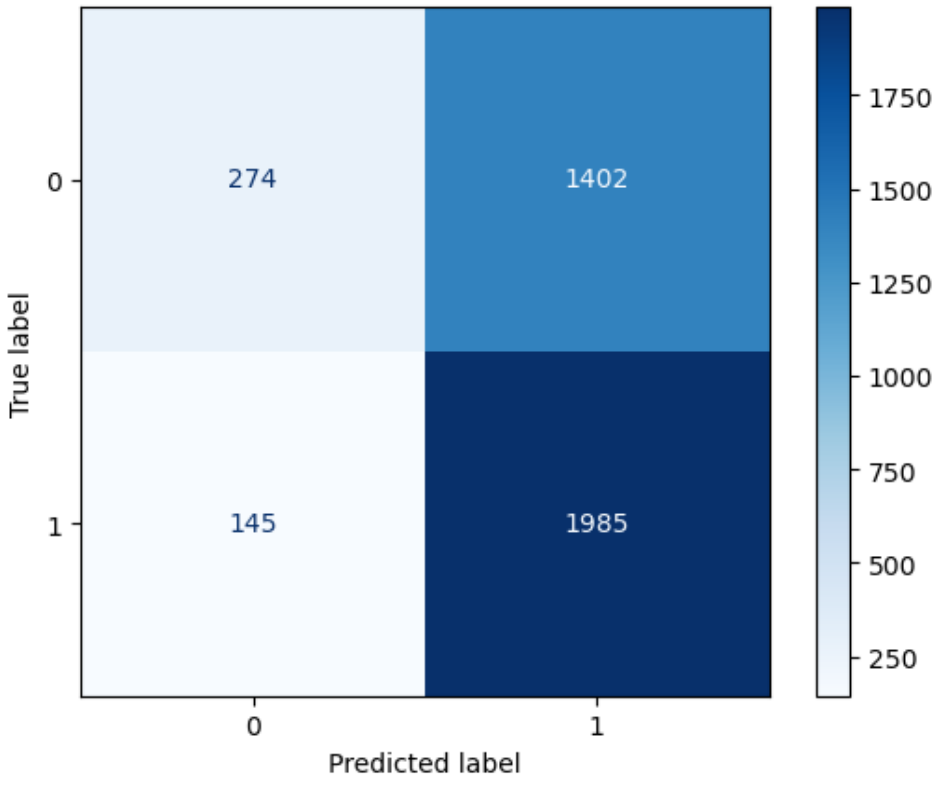

Figure 20: Confusion matrix of final decision tree model with max depth as three, and max leaf nodes as six.

	precision	recall	f1-score	support
0	0.65	0.16	0.26	1676
1	0.59	0.93	0.72	2130
accuracy			0.59	3806
macro avg	0.62	0.55	0.49	3806
weighted avg	0.62	0.59	0.52	3806

Figure 21: Classification report of final decision tree model with max depth as three, and max leaf nodes as six.

From this, it can be concluded that the best max_depth value is '3', and the best max_leaf_nodes value is any value from '6' to '10'. These chosen parameters have had little effect on improving the accuracy of the model. The first parameter, max_depth, was altered to determine which value produces the highest accuracy, based on the maximum depth of the decision tree. As mentioned before, knowing the best fit for this parameter allows the control of the accuracy of the model, and preventing the decision tree from learning all the training data which would lead to over-fitting (Gulati, 2022). The second parameter altered was the maximum number of leaf nodes in the decision tree. Controlling this too gives the ability of the researcher to control the complexity of the model, thereby controlling the accuracy and limiting the model's ability to overfit the data. The values concluded to be the best fit for this dataset have been determined through 10-fold cross-validation testing, and supported by evidence of confusion matrixes and classification reports. These values may be replicated in other datasets through coincidence, but it is unlikely. It is because the dimensions of the dataset; the number of attributes, columns, rows, etc. in the dataset that have led to these values being most favourable. For other datasets with the exact same number of attributes and number of '1's' from the target class for each attribute, then it may be possible to use these same values for these parameters. It is always best to calculate and support these calculations with evidence, which is why it is better to carry out a process like the one done for these decision tree models, to determine the best values for the parameters, even if the datasets are similar.

Figure 22 shows the feature importance in a tabular format for the final decision tree model. This layout has been chosen to clearly show each feature in the dataset used in the decision tree models, and its corresponding value of importance. This importance value indicates the level of input of each feature on the decision tree. It is observed in this that the toCoupon_GEQ5min attribute has no input at all on the decision tree, as its importance value is '0'. This compares to income having the greatest level on input with an importance value of '0.327948'. The attributes are ranked in the table by level of input, the attribute with the greatest level of input first.

	Feature	Importance
1	income	0.327948
0	age	0.300774
2	expiration	0.123231
4	toCoupon_GEQ15min	0.0904199
5	toCoupon_GEQ25min	0.0822118
7	direction_opp	0.0410102
6	direction_same	0.0344055
3	toCoupon_GEQ5min	0

Figure 22: Table of feature importance for the final decision tree model.

Artificial Neural Network (ANN)

In this part, you are required to explore various architectures for building an Artificial Neural Network (ANN). In building the model, use the 10-fold cross-validation option for testing.

a) *Use an appropriate feature selection method to identify the top **five most significant features**. State the method used and list the features produced. Compare the list produced in the previous section by the Decision Tree model. Identify similarities and differences. Discuss any differences.*

We selected the ANOVA method to identify the top 5 most significant features. Figure 23 represents the significance levels of the 8 attributes in the dataset, from this figure it is possible to see the top 5 significant features clearly where 2, 5, 4, 0, 1 have the highest significance respectively.

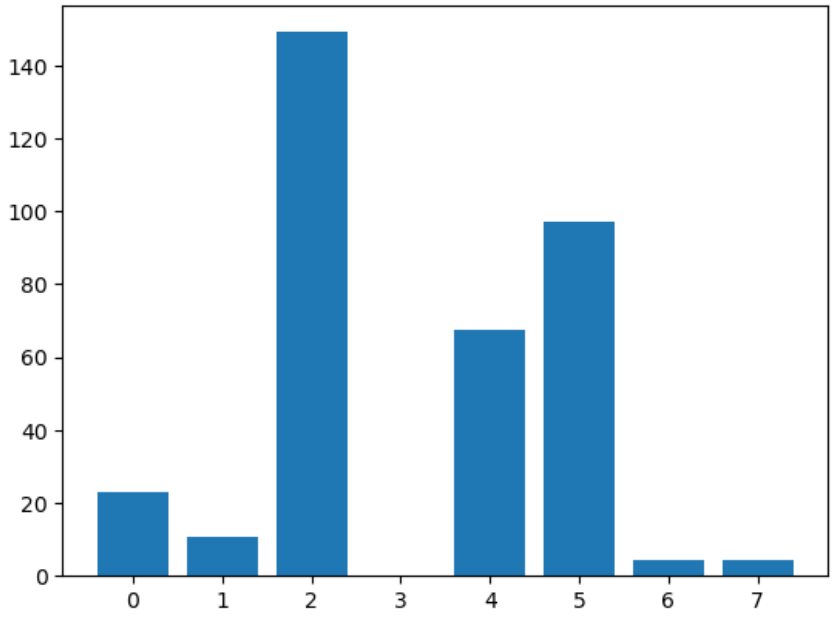

Figure 23: Bar graph which represents all 8 features significance of the in-vehicle coupon dataset.

The top 5 features with its corresponding significance score are stored in a table shown as Figure 24. The feature with the highest score is expiration with a score of 154.789, meaning that there is a high correlation with the target variable y, which is the acceptance of coupons. Hence, the expiration feature is expected to have a high impact in predicting the acceptance of coupons. For the toCoupon_GEQ25min feature, there is also a high score of 90.277. This represents a high correlation with the target variable y, where the driving distance to the restaurant or bar for using the coupon is greater than 25 minutes. The high score indicates there is a significant impact of toCoupon_GEQ25min on the target variable y: the acceptance of coupons. The toCoupon_GEQ15min feature produced a score of approximately 56.246; this similarly shows a high score suggesting a meaningful role in predicting coupon acceptance. Age has a score of 36.390 which holds a moderate correlation with the target variable y, meaning that there is some influence of age, but it is not as strong as the other features. Income also has a somewhat low score of 11.335, meaning there is a relatively weaker impact compared to the other 4 features.

Top 5 Features:

Feature Number	Feature Name	Score
2	expiration	154.789632
5	toCoupon_GEQ25min	90.277073
4	toCoupon_GEQ15min	56.246900
0	age	36.390162
1	income	11.335267

Figure 24: shows a table of the top 5 features including the feature number, name and significance score which is ordered from highest significance to lowest significance.

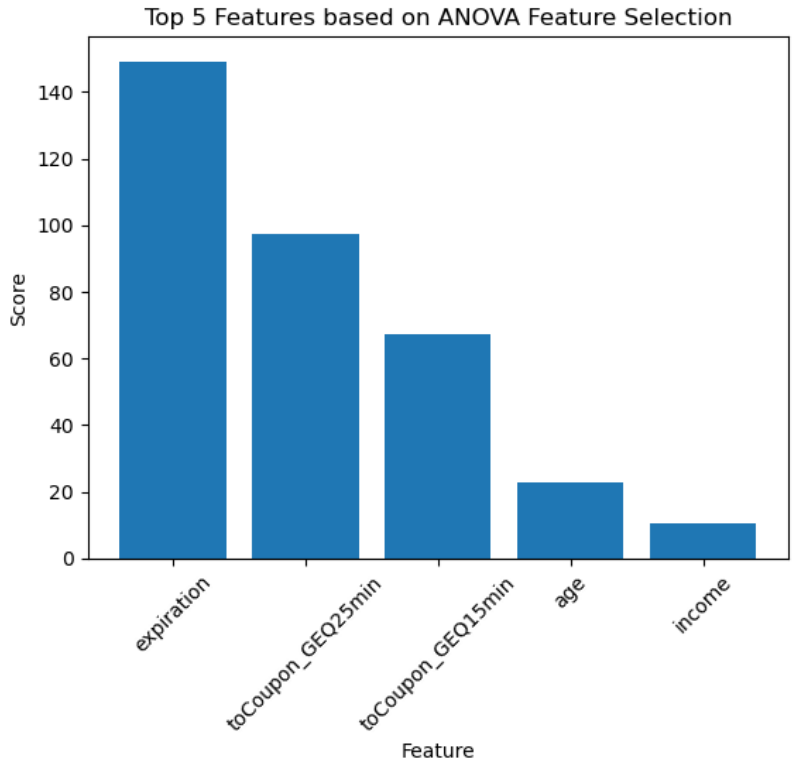

Figure 25: displays another representation of the top 5 attributes along with the significance score in a bar graph.

If we compare Figure 22 and Figure 24, it is clear to say that both have outputted different scores for the corresponding features. The ANOVA model displays five significant features, expiration, toCoupon_GEQ25min, toCoupon_GEQ15min, age and income. The decision tree model showing the features indicated that income and age were the most important features, followed by expiration then toCoupon_GEQ15min and toCoupon_GEQ25min. There is quite a difference in rankings between both ANOVA model and Decision Tree model. This can be accounted for, as ANOVA is known to analyse the variance in the target variable with respect to each feature while the Decision Trees consider the feature's ability to split the data effectively.

b) *Use the sklearn.MLP Classifier with default values for parameters and **a single hidden layer** with k neurons (k <=25). Use default values for all parameters other than the number of iterations. Determine and report the best number for iteration that gives the highest accuracy. Use this classification accuracy as a baseline for comparison later parts of this question.*

The MLP learning algorithm is utilised to classify data, and tries different numbers of iterations, measuring the accuracy of the model using cross-validation. The MLP classifier is configured with a hidden layer of 25 neurons. To determine the optimal number of iterations, cross-validation is used. This then iterates over a range of values: 100, 200, 300, 400, 500. The model is trained using the specific number of iterations, and the accuracy of the model is assessed through the cross validation with 10 folds. Taking these steps, the best number of iterations is retrieved along with the highest accuracy.

The best number of iterations is displayed as 100, meaning that out of the five iterations, the iteration value of 100 resulted in the best number. The highest accuracy achieved with the 100 iterations is about 53.76%.

```
The best number of iterations: 100
Highest accuracy achieved: 0.5376126126126126
```

c) *Enable the loss value to be shown on the training segment and track the loss as a function of the iteration count. You will observe that even when the loss value decreases the error value increases between consecutive iterations. Conversely, the error value may decrease when the loss increases between consecutive iterations. How do you explain this?*

If the model is overfitted, meaning when the model starts to memorise noise or certain patterns in the training process, the loss value decreases while the error value increases. This can also occur when the model chooses to ensure correct classification rather than minimising the overall loss or error. It is also useful to note that error and loss would have an inverse relationship since one increases while the other decreases.

Figure 26 below displays an inverse exponential graph which represents the loss count against the iteration count. The inverse exponential loss curve implies that the loss value decreases rapidly at the beginning of the training process and then slowly levels off.

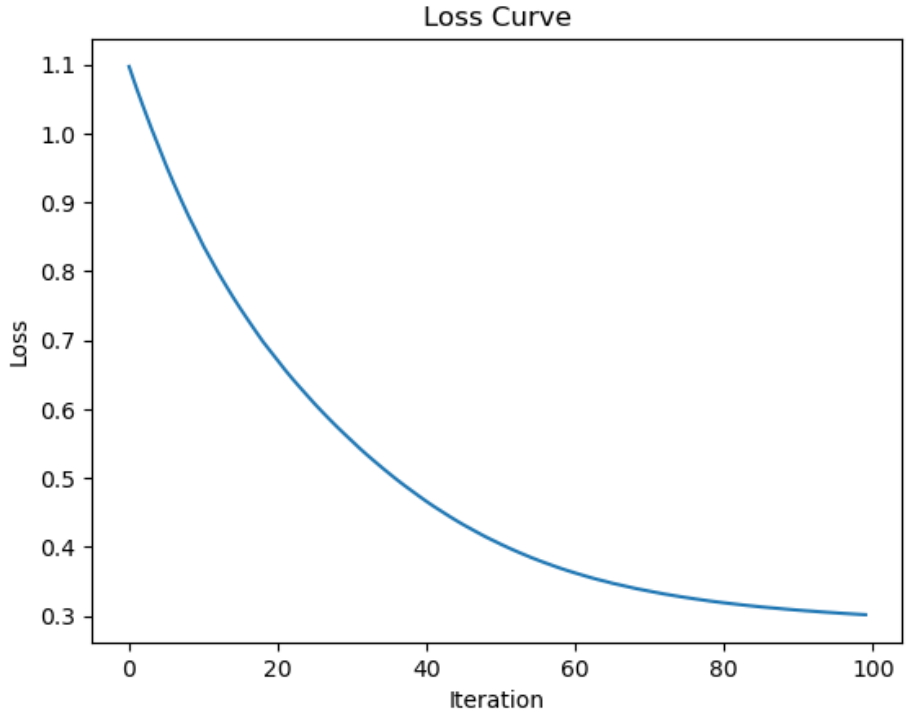

Figure 26: Representing the Loss Curve against the iterations during training.

From the classification table, shown below as Figure 27, it is observed that the MLP accuracy is 83%. The precision score for '0' representing those who haven't accepted the in-vehicle coupon is 0.78. This means that out of all instances the model predicted for 0,

78% of them as true positives. For '1' which represents those who did accept the coupons, the precision score is 0.87 which means that out of all instances the model predicted 87% of them as actually true positives. The recall score measures the proportion of the actual positives which were correctly detected. For '0' the recall score is 0.85 this means out of all instances 85% were correctly identified as positive. For '1' the recall score is 0.81, meaning 81% was also correctly identified as positive. The F-1 score represents the combination of precision and recall; The F-1 value grants more clarity between both scores. '0' has an F-1 score of 0.82 and '1' has an F-1 score of 0.84, this means that the model has achieved a high level of accuracy as well as a good balance between precision and recall. This implies that the model's predictions are reliable and accurate.

```
MLP Accuracy: 83.00%
MLP Classification report:

              precision    recall  f1-score   support

           0       0.78      0.85      0.82        89
           1       0.87      0.81      0.84       111

    accuracy                           0.83       200
   macro avg       0.83      0.83      0.83       200
weighted avg       0.83      0.83      0.83       200

MLP Training set score: 87.88%
MLP Testing set score: 83.00%
```

Figure 27: shows a table of the MLP classification report.

d) *Experiment with **two hidden layers** and experimentally determine the split of the number of neurons across each of the two layers that gives the highest classification accuracy.*

The table, Figure 28 seen below, shows the neuron combinations and their corresponding classification accuracies obtained by experimenting with the two hidden layers. The table reflects (19,7) to have the highest accuracy at 87.9% and with (9,17) the accuracy is the lowest at 87.1%

Neuron Combination	Accuary
(25, 1)	85.10%
(24, 2)	86.40%
(23, 3)	85.90%
(22, 4)	87.00%
(21, 5)	85.10%
(20, 6)	86.80%
(19, 7)	87.90%
(18, 8)	86.10%
(17, 9)	86.30%
(16, 10)	85.00%
(15, 11)	85.20%
(14, 12)	85.20%
(13, 13)	85.90%
(12, 14)	86.00%
(11, 15)	85.60%
(10, 16)	85.60%
(9, 17)	84.60%
(8, 18)	86.80%
(7, 19)	87.10%
(6, 20)	86.50%
(5, 21)	85.80%
(4, 22)	86.20%
(3, 23)	86.70%
(2, 24)	85.80%
(1, 25)	86.40%

Figure 28: Table of two Hidden Layers and Accuracies

e) *From the table created in part d, you will observe the accuracy variation with the split of neurons across the two layers. Give explanations for some possible reasons for this variation.*

The table above seen in Figure 28, shows the two hidden layers and their calculated accuracies. The left-hand side shows how the neurons (25) are split between the two levels; (layer 1, layer 2). For example (23, 2) means there are 23 neurons in layer 1 and 2 neurons in layer 2. As observed, the highest accuracy occurs when the 25 neurons are split (19, 7). There is much variation between the accuracies of the different neuron splits, and there are various reasons for this.

Firstly, the representation of the information has an impact on the variation of accuracy. How the neurons are spread across the two layers affects how the network can obtain and signify the variation in the dataset. The way in which the neurons are splits gives the ability to control how the accuracy is altered. Secondly, overfitting may occur. This is would occur when the neural network becomes too specialised, thereby failing to generalise any unseen data. Altering the neurons in the levels would affect this, which is another reason variation in accuracy may be occurring.

Performance Comparison

Compare the performance of the Decision Tree and MLP Classifiers on your dataset. Choose the best-performing model for your dataset and explain why you have chosen it. Discuss the findings from your experiments and provide your opinion about these two classifiers.

In this analysis report, two methods have been used analyse the dataset; a Decision Tree and MLP Classifiers. The decision tree was used fist. In this, multiple models were constructed to achieve the best accuracy this model could obtain. The parameters that were altered to do this were max depth and max leaf nodes. A series of 10-fold cross validation tests were carried out to determine the best values to use for these parameters, one at a time. The final decision tree produced (see Figure 19) has a max depth of three, and a max leaf nodes value of six. During the testing to determine the best value to use for the max leaf nodes, it was found any value from six to ten could be used as they all produced the same accuracy score. The highest accuracy the decision tree could produce, after all the tuning and testing, was 0.59 or 59%. MLP Classifiers were used next. In this process, ANOVA selection was used to determine the five most significant features from the dataset in use. Interestingly, the top five features with the most importance from the decision tree analysis method were the same as those from the MLP classifiers. The order of importance/significance does differ between these methods, with the income being the most important and the toCoupon_GEQ25min being the lest important for the decision tree, whereas the ANOVA feature selection found the expiration was the most significant and the income was the least, and by quite a fair bit.

Looking at only the feature selection of these two methods doesn't tell much more than which features have the greatest influence on the models. For this information to be useful, it requires more context and further analysis to determine which model is better for this dataset. To do this, the accuracy scores of these models are essential. As stated above, the accuracy score of the final decision tree is 0.59%. This is significantly lower than the 0.83 or 83% obtained from the MLP classifiers. Furthermore, when the 25 neurons are split (19, 7) during testing of two hidden layers, the accuracy reaches as high as 87.9%. Another way to compare these models is to look at the classification reports, and review the scores other than accuracy. For class '0' in the decision tree, the precision score is 0.65, the recall score is 0.16, and the f1-score is 0.26. These are considerably lower than those obtained in the MLP classifier, where the precision score for class '0' is 0.78, the recall score is 0.85, and the f1-score is 0.82. Similarly, the scores are also higher for class '1' in the MLP classifier than it is in the decision tree. The only score that doesn't follow this trend is the recall for class '1'. However, as the rest of the score all improve, and this score is still relatively high, it doesn't need to be amended for the purposes of this report.

From the significantly higher scores, especially accuracy, obtained from the MLP classifier, it is clear to see that this is the best-performing model for the dataset. This is only for this dataset though. There are many advantages and disadvantages to both methods of analysis. This report is not a one-size fits all analysis, as previously explained in the decision tree section about the use of the chosen parameters to tune the models. The MLP classifier has been proven to be the better model for this dataset, but for other datasets it may or may not be the decision tree model, or even an entirely different model. The most important thing, which has been relentlessly carried out in this analysis, is to constantly tune and test the models to obtain the best accuracy possible for the dataset.

References

1. Dash, S. (2022, November 2). *Decision trees explained-entropy, information gain, Gini Index, CCP pruning.*. Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c
2. Fleming, T. (2021). Coupons Differentiate Dealers and Incentivize Consumers. *TVI MarketPro3* https://www.tvi-mp3.com/blog/coupons-provide-differentiation-for-dealers-and-incentivize-consumers/
3. Gulati, H. (2022, February 11). *Hyperparameter tuning in decision trees and random forests*. Hyperparameter Tuning in Decision Trees and Random Forests. https://www.section.io/engineering-education/hyperparmeter-tuning/
4. Karabiber, F. (n.d.). *Gini impurity*. Gini Impurity . https://www.learndatasci.com/glossary/gini-impurity/#:~:text=Gini%20Impurity%20is%20a%20measurement,nodes%20to%20form%20the%20tree
5. *UCI Machine Learning Repository: in-vehicle coupon recommendation Data Set (n.d)* *https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation#*