Data Mining

IMAT3613

Author
Himanshu Abburu
(P2503016)

Date

Organics Assignment

(Word count- 2060)

Lecturer: Dr Anthony Williams

Table of contents

Contents

Summary 3
Business Problem 3
Data Mining Representation 3
Methodology - data mining approach 3
Data Exploration 5
Data partition creation of model sets 7
Data Modelling 8
Regression 8
Neural Network 10
Development of models 10
Model performance 10
Neural network architecture of best model 11
Decision Tree 12
Development of models 12
Performance of models 13
Overfitting and limitations 17
Analysis of the best model 18
Conclusion 20
Recommendation 21
References and Bibliography 21
Appendix 21
My Reflections on the Patchwork assignment 23

Summary

This report presents the results of using the _SEMMA_ data mining framework, to build models that will help the management of a supermarket concentrate their resources on targeting customers that are most likely to purchase organic products. Three binary __predictive__ models were generated using __regression__ analysis, decision ___trees___ and __neural network_. __AGE, AFFL__ and _GENDER(f)_ were identified as the most important predictors. All three models were predictive and have a _similar_ level of performance. __Decision Tree__ was chosen as the champion models based on performance and also the techniques ability to provide an non-technical explanation of the model with a lift value of __3__times greater than selecting customers at random. Several recommendations are made for improving data quality for the next cycle of data mining.

Business Problem

The business problem is to identify the customers that are most likely to _purchase_ organic products in the supermarket. A data models will be built data set (organics.xls) collected during the supermarket incentive period. By identifying _customers_ who are _willing_ to purchase organic products the company will be able to target its marketing efforts more effectively which should result in more sales per marketing advertising spend.

Data Mining Representation

The business problem, identification of customers who are __most likely__ to __purchase__ organic products is a type of data mining representation known as a __binary classification__ problem. The most suitable target variable is _ORGYN_ which is identified as a **binary variable**. The remaining variables will be given roles __input__ and are assigned the default measurement levels with the exception of AFFL which has been changed from interval to ordinal.

Methodology - data mining approach

The process to be adopted is the first flow in the virtuous cycle of data mining which has four distinctive steps:

1. Identify the business problem or opportunity.
2. Mining data to transform it into actionable information.
3. Acting on the information this is outside the remit of the brief (marketing initiative driven by the model, eg targeted marketing offer)
4. Measuring the results of the marketing initiative this is outside the remit of the brief (measure profitability of the pilot marketing study using the model).

The data mining framework used in the generation of models for steps 1 and 2 of the virtuous cycle of data mining is based upon __SEMMA___, Sample, Explore, Modify, Model and Assess. In the brief the data set has already been provide, so there is no need to sample or collect the data. However limitations in data collection may become apparent in the data mining process, these are discussed in the recommendations section.

Explore:

Explore the raw data as provided. This will result in a brief overview of the variables so that they can be classified as qualitative data (binary, nominal, ordinal), or quantitative data (discrete, interval) with a brief description. Establish a Target variable (which variable can be used to establish the required results).

Modify:

Assess the data quality (what changes can be made to the data classifications/levels, model roles). If necessary modify and transform the data, impute missing values, transformations to normalise distributions of heavily skewed data.

Models:

Create models of data (regression, neural networks, decision trees) to identify patterns relationships and parameters that can predict the target variable. Each model representation has their own advantages and disadvantages.

Assess:

Asses the model performance in order to identify the champion model and investigate their limitations. If necessary make changes to the data set, model tuning through a second cycle of data mining to improve on the previous model results. The aim of the second cycle is to upon improve model reliability, robustness, performance and avoid over fitting.

In summarising the results from the models generated arrive at a conclusion, and make recommendations for future data gathering by the business. This should ensure that the future data mining results will be of benefit to the business and refine the data mining process.

Data Exploration

A full meta data description of the data is provided in the appendix (table 1 and 2) only important features and insights that have a bearing on data mining process will be described. These insights are generated from an exploratory data analysis of the data. The analysis has been divided into two sections class variables which are qualitative and interval variables which are quantitative.

From an analysis of the business problem the target variable, the variable to be predicted has already been identified as __ORGYN__. It is noted that the ratio of evidence for organic customer purchasers to non-purchasers is __0.76:0.24__, making this a difficult data mining problem.

An indication of data quality by identifying the level of missing data is also presented, any variable with more than 50% of the data missing is normally considered unsuitable for data mining. Variables that have missing data have a bearing on regression analysis and neural network models. __AFFL__, __AGE__ and __GENDER__ have between 5 and 11% missing data, the data quality for remaining variables is less than 5% missing. Interval variables which are heavily skewed > 3 may be transformed to normal distributions to comply with model assumptions.

Variables with outliers, extreme values are also identified. These anomalous observations may be outside the scope of models and may give indications for the predictive parameters. However the nature of outliers is that they represent only a handful of observations and insights may not be applicable to the majority of observations. The impact of outliers on modelling was found to be negligible. Model performances were practically unaffected by the removal of outliers using the filter node. The variable __BILL__ was heavily skewed and has the majority of extreme values.

Variable	Variable Description	Role	Level	Comments on data
AFFL	The status of person in-terms of wealth	Input	Interval	People with more wealth purchased more organic products.
AGE	Age of the customer, in years	Input	Interval	People aged 50 and above did not purchase organic products. And has the second highest percent of missing values (6.58%).
BILL	The total amount spent	Input	Interval	98% of the customers spent between £0 and

	on the purchase, for both organic and non-organic products			£29631. Only a small number of people made purchases more than this.
GENDER	Male or Female	Input	Interval	Has the highest percent of missing values of 11.25%
CLASS	This variable mentions the status of the customer based on their previous purchases	Input	Nominal	The maximum number of non-organic customers are of the class platinum.
CUSTID	A unique identification number given to each customer	Rejected	Interval	This variable is does not hold importance while building a model. But it will be helpful after developing a model to use the ID numbers as target customers.
LCDATE	Loyalty card application date	Rejected	Interval	This variable just mentions the date a customer applied for the loyalty card. It doesn't help in data modelling as the graph has a sudden increase towards the end of the 1900's as there is a lot of missing data between the years 1909 & 1958 and the statistics show a negative skewness.
LTIME	Time as loyalty card member	Time ID	Interval	Customers have stayed as a loyalty card member for an average of 6 years.
NGROUP	Type of residential neighbourhood	Input	Nominal	People living in the neighbourhood C are the highest number organic products purchasers.
OAC	Output Area classification	Input	Nominal	People living in the Sub-urban areas are the highest number of purchasers for organic products.

ORGYN	Organics purchased or not	Target	Binary	Approximately, only 24% of the total customers purchased organic products.
REGION	Graphic region	Input	Nominal	Approximately, 38% of the customers that purchased organic products are from South-East.
S_CONV	Convenience food	Input	Interval	On an average, customers spend £10.7 on convenience food.
S_FVEG	Fruits and Vegetables	Input	Interval	Nearly, 30% purchased fruits and vegetables spending between £27 and £34.
S_MT	Meat	Input	Interval	The average expenditure on meat was calculated to be 25. There is some variability in the data, as may be seen by the standard deviation of 8.227.
S_TOIL	Toiletries	Input	Interval	The average customer spent on toiletries is 19.217. And it shows a positive skewness of 0.247.

Table 1 Summarising observations, measurement levels and data mining roles.

Data partition creation of model sets

Partitions help reduce the computation time of the initial model runs instead of running on the whole data set. This comes in handy when the size of the data set is large. The data partition node is used to split the Organics data into 3 sets, namely – Training, Validation and Test sets for the data mining techniques.

- **Training** – Training data set is used as a preliminary model fitting. This will be the main data set that will be used to make model and checked if the model fits the data training set. 40% percent of the total data will only be used, as defaulted by the SAS Enterprise Miner.
- **Validation** – This data set will be used to monitor and tune the model weights while estimating and can also be used for model assessment. 30% percent of the total data will only be used, as defaulted by the SAS Enterprise Miner.
- **Testing** – This can be used as an additional data set for model assessment. 30% percent of the total data will only be used, as defaulted by the SAS Enterprise Miner.

Data Modelling

Modelling is the collection of inputs that connects the outcomes or targets with a set of rules.

- To deal with outliers and skewed data, filter and transform variables node is used.
    - The BILL and LTIME variable have many observations with the value 0. Therefore, spreading the values through the distribution will be helpful.
    - BILL has outliers that could possibly alter the
- To deal with missing values for the regression node, impute node will be used.
    - The impute node allows to replace missing values for interval variables. It ensures that all the data in the training data set is used while building a regression model.
    - By default, interval variables are replaced by mean and class variables are replaced with the most frequent value.

Regression

The workflow diagram shows various regression models constructed.
Initially, I have constructed the basic regression models – Full, Forward, Backward, Stepwise regression.
Full regression: I have not considered this model because it considers all the variables. From this we cannot pick the important variables required for the logistic regression equation, therefore we cannot focus on the business problem by targeting the right customers.
Forward Regression: Starts with a null model and then selects the next significant variable. This continues until the point where no more variables have a p-value less than the specified significance level.
Backward Regression: Starts with the full model and then removes the least significant variables. This continues till the P- value is less than the specified significance level,
Stepwise Regression: Similar to a forward regression, but when a new variable is introduced, the variables are removed that have lower significance than that of the set significance value and continue to run when no variable outside the model has p-value smaller that the set significance level and all the variables have reached the set significance level.

Further, whilst using the impute, filter and transform nodes, I have considered only the stepwise regression. This is considered a better selection from the above definitions, since they can provide the best variables after refining.

Classifier	Missing values removed	Missing values imputed	Outliers removed	Transformation to remove skew	Scope	% True positives	% Precision	AUC ROC	Estimate lift @ 25% depth
Full Regression	-	-	-	-	23.21%	31.64%	71.96%	0.77	1.28
Forward Regression	-	-	-	-	24.25%	32.73%	73.54%	0.77	1.39
Backward Regression	-	-	-	-	24.25%	32.73%	73.54%	0.77	1.39
Stepwise Regression	-	-	-	-	24.25%	32.73%	73.54%	0.77	1.39
Filtered Stepwise Regression	-	-	Yes	-	26.34%	36.06%	72.13%	0.79	1.13
Imputed Stepwise Regression	Yes	Yes	-	-	24.33%	34.52%	73.04%	0.79	1.61
Transformed Stepwise Regression	-	-	-	Yes	24.33%	33.69%	73.21%	0.77	1.36

Classifier	%True - 00	%False - 01	%True +11	%False +10
Full Regression	96.03%	3.96%	31.64%	68.35%
Forward Regression	96.21%	3.78%	32.73%	67.26%
Backward Regression	96.21%	3.78%	32.73%	67.26%
Stepwise Regression	96.21%	3.78%	32.73%	67.26%
Filtered Stepwise Regression	95.01%	4.98%	36.06%	63.93%
Imputed Stepwise Regression	95.9%	4.09%	34.52%	65.47%
Transformed Stepwise Regression	96.03%	3.96%	33.69%	66.3%

From the above tables and the appendix table1 we can conclude that the imputed stepwise regression model is the best compared to the other models.
This is justified by the data seen in the appendix table1. It has a low misclassification rate, stating that the wrong predictions made are less. It has the highest lift value of 3.64 which helps select customers easily than at random. It has the highest area under the curve in the ROC chart and tends to be moving towards the top left corner and has

the highest true positive percentage of 35%, leading, us towards the right customers, and making it the best pick.

Chosen Regression equation
Logit P = -2.2078 + AFFL * 0.2604 - AGE * 0.0500 + GENDER(f) * 0.8599

If there were no variables at the 95% confidence interval then the contribution to the log of the odds ratio is -2.2078 [1]. The negative sign shows that it does not have great importance.
Looking at the next set of parameters GENDER(f) and AFFL make the biggest contribution to an observation likely to not purchase organic products.
AGE, GENDER(f) and AFFL are included in the final model indicating that these have a higher importance in attracting customers to purchase organic products.

Neural Network

Development of models

Three models of Neural network have been developed:
1. Default Neural Network - Connected directly to the partition node.
2. Imputed Neural Network – Connected to the impute node to handle missing values.
3. Filtered Neural Network – Connect to the filter node to handle outliers.
4. 5 Hidden Nodes – Connected to the data partition node and property changed to have 5 hidden nodes. Higher the number of units, multiple patterns can be recognised to train the network well.

Model performance

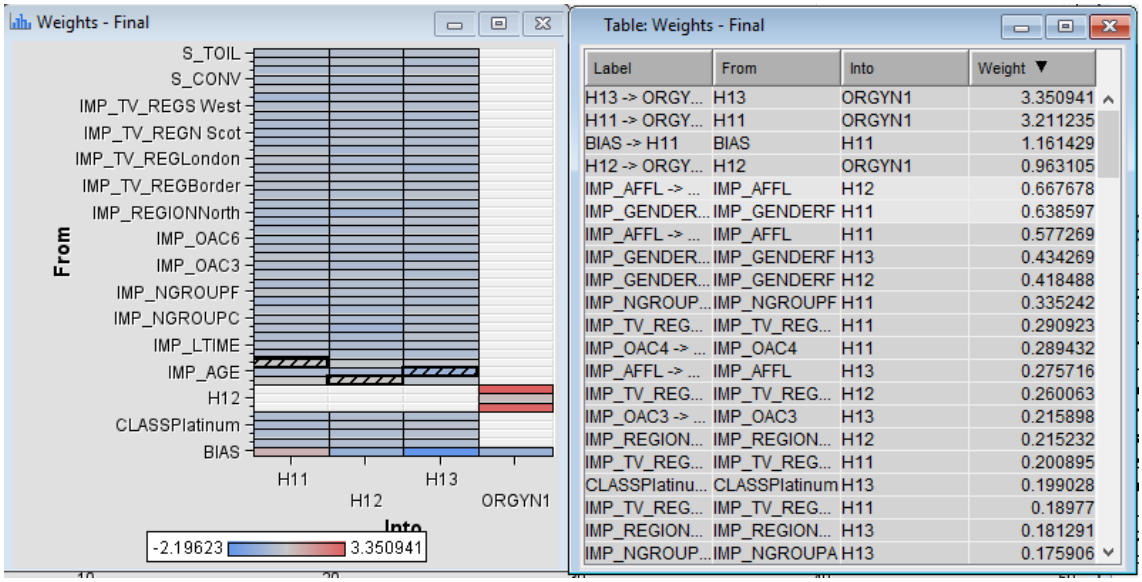
Classifier	Missing values removed	Missing values imputed	Outliers removed	Transformation to remove skew	% True positives	% Precision	AUC ROC	Estimate lift @ 25% depth
Default NN	-	-	-	-	33.8%	72.4%	0.76	1.5
5 Hidden Node NN	-	-	-	-	33.5%	69.6%	0.76	1.1
Imputed NN	-	Yes	-	-	33.8%	72.6%	0.79	1.3
Filtered NN	-	-	Yes	-	36.4%	70.6%	0.78	1.3
Transformed NN	-	-	-	Yes	33.9%	70.8%	0.75	1.1

Classifier	%True -	%False -	%True +	%False +
Default NN	95.8%	4.1%	33.8%	66.1%
5 Hidden Node NN	95.2%	4.7%	33.5%	66.4%
Imputed NN	95.9%	4.0%	33.8%	66.1%
Filtered NN	94.5%	5.4%	36.4%	63.5%
Transformed NN	95.5%	4.4%	33.9%	66.0%

Table 5 Neural network performance

Neural network architecture of best model

From the Appendix Table1, we can see that the Imputed Neural Network and Default Neural Network have a lot of similarities. But the Imputed Neural Network has a higher Area Under the Curve for the ROC chart. The higher the AUC, the curve tends to move closer to the sensitivity value 1, showing that it has higher true positive values. The higher the concavity, the better the model performs. A lower misclassification rate and a better precision percentage, mentioning that the predictions made are correct because of the high true positives, are added points which will help the business choose a better model to target organic product purchasing customers.



From the table weights, we gather that the AFFL (0.67) and GENDER(f) (0.64) variables have a high importance, similar to as seen in the decision tree model. The business can concentrate on the customers whose affluence grade is greater that 4 and females in their customers list to achieve higher profits

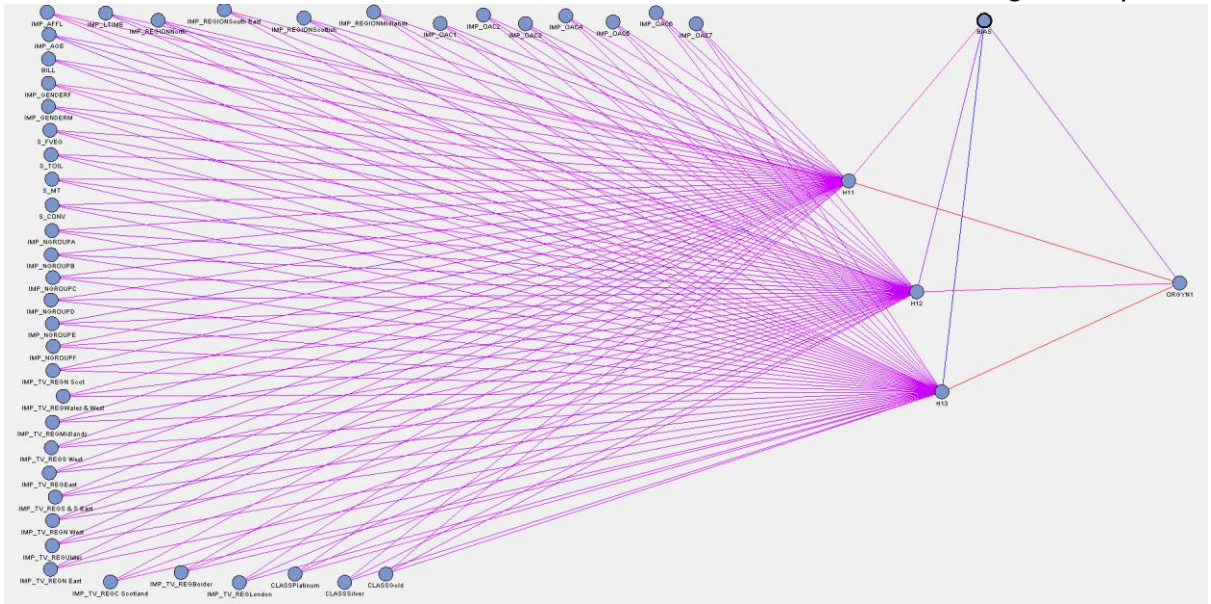


Figure 1 example neural network architecture

Decision Tree

Development of models

Four different decision tree models were developed and compared to choose the best amongst them.

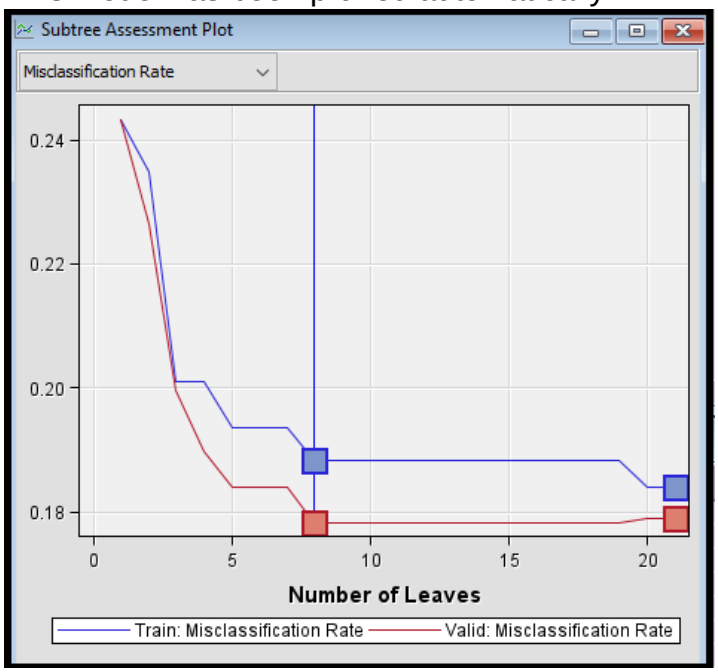
1. A decision tree with default properties is built.
2. A 3 branched split
3. A pruned tree
4. Manual decision tree.

1 . A Default Decision tree

This tree has default property values as follows:
Maximum Branches (allows only restricted subset splitting) = 2
Maximum Depth (maximum number of generations) = 6
Leaf Size (least number of training observations) = 5
Split Size (minimum number of training observations that a node must have to be able to split) = .

2 . Pruned Decision Tree

Pruning: Smaller leaves are merged to remove instability, so that the tree can make predictions.
This model is pre-pruned. Altering the default values of the leaf size and split size will stop the growth of the tree.
This model has been pruned automatically.



The above classification chart shows a tree was constructed with 21 leaves (terminal nodes) but then pruned back to have only 8 leaves. This has been done automatically because the performance remains the same till the 19th leaf. And the validation set seems to increase the misclassification rate, which is not a good sign.

3 . 3-way split Decision Tree

This tree is built to have 3 branches instead of the default 2 branches. This will have a smaller depth and the tree will be wider.
In the properties section, the Maximum Branch is set to 3 under the splitting rule.

4 . Manual Decision Tree

To construct a manual tree, after connecting the decision tree node to the Data Partition node, in the properties section click on the ellipsis against the interactive node. The root node is split based on the logworth of the variables. Logworth gives the purity of the node. Higher the purity, better the split.

Target Variable: ORGYN

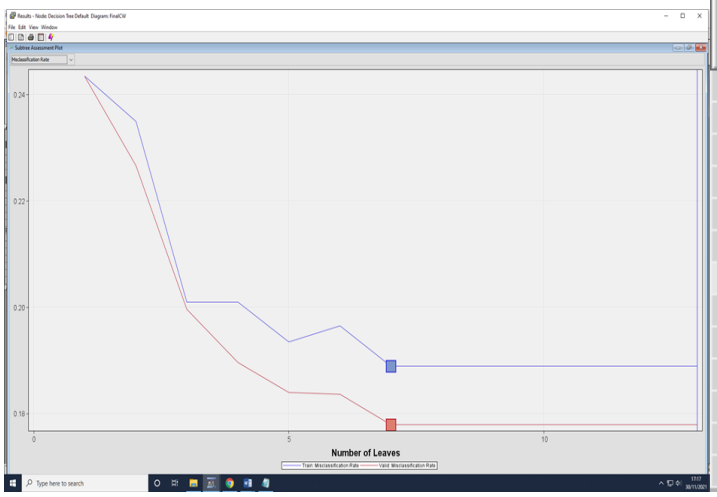
Variable	Variable Description	-Log(p)	Branches
AFFL	AFFL	0.0	2
AGE	AGE	0.0	2
BILL	BILL	0.0	2
LTIME	LTIME	0.0	2
S_CONV	S_CONV	0.0	2
S_FVEG	S_FVEG	0.0	2
S_MT	S_MT	0.0	2
S_TOIL	S_TOIL	0.0	2
CLASS	CLASS	0.0	2
NGROUP	NGROUP	0.0	2
OAC	OAC	0.0	2
REGION	REGION	0.0	2
TV_REG	TV_REG	0.0	2

Continuing the split on the lighter/impure nodes, I have stopped at 7 leaves. From the above the table, we can see that the logworth for the remaining variables is 0. Log (0) = undefined. There is no reason for further splitting of the nodes.

Performance of models

Classifier	% True positives	% Precision	AUC ROC	Estimate lift @ 25% depth
Default Decision Tree	91.1%	66.7%	0.81	1.69
Decision Tree Pruned	62.4%	66.0%	0.76	1.54
Decision Tree 3way split	92.8%	68.7%	0.79	1.81
Manual Decision tree	94.9%	72.2%	0.71	0.86

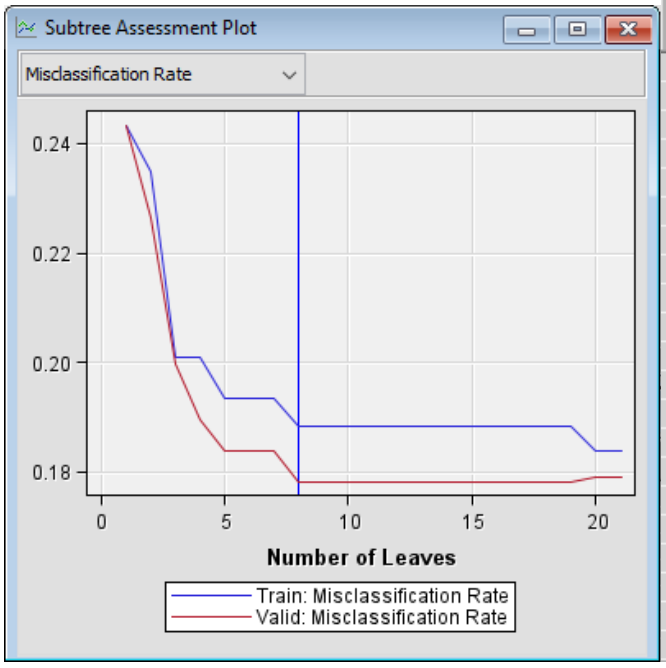
Default Decision Tree Misclassification Rate



Number of Leaves	Train: Misclassification Rate	Valid: Misclassification Rate
1	0.2435	0.2433
2	0.2350	0.2267
3	0.2010	0.1997
4	0.2010	0.1897
5	0.1935	0.1840
6	0.1965	0.1837
7	0.1890	0.1780
8	0.1890	0.1780
9	0.1890	0.1780
10	0.1890	0.1780
11	0.1890	0.1780
12	0.1890	0.1780
13	0.1890	0.1780

The above chart shows that the node splits performed well till the 3rd node. From the 7th node no big difference is seen in either the training set or the validation set.
The table shows that a 7-leaf model has a misclassification rate of 17.80% on the validation data set. This is a good model because the misclassification rate for the validation set doesn't seem to increase.

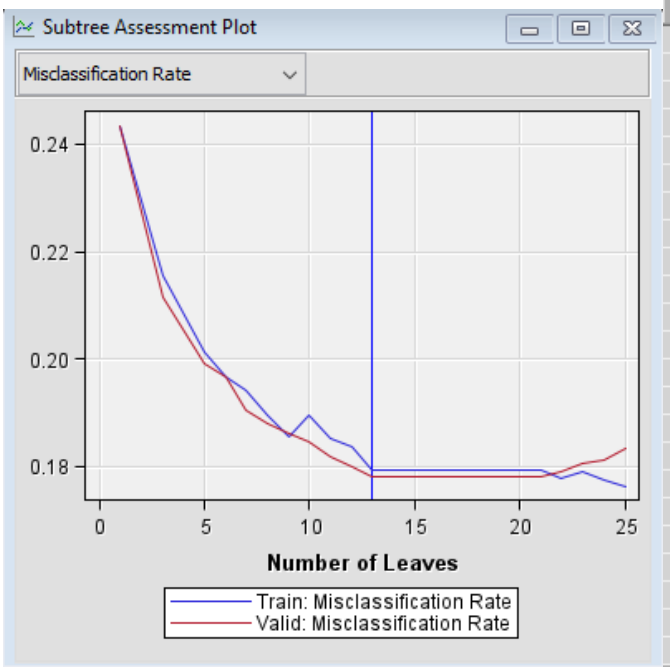
Pruned Decision Tree



Number of Leaves	Train: Misclassification Rate	Valid: Misclassification Rate
1	0.2435	0.2433
2	0.2350	0.2267
3	0.2010	0.1997
4	0.2010	0.1897
5	0.1935	0.1840
6	0.1935	0.1840
7	0.1935	0.1840
8	0.1883	0.1783
9	0.1883	0.1783
10	0.1883	0.1783
11	0.1883	0.1783
12	0.1883	0.1783
13	0.1883	0.1783
14	0.1883	0.1783
15	0.1883	0.1783
16	0.1883	0.1783
17	0.1883	0.1783
18	0.1883	0.1783
19	0.1883	0.1783
20	0.1840	0.1790
21	0.1840	0.1790

The above chart shows that the node splits performed well till the 3rd node. From the 8th node no big difference is seen in either the training set or the validation set till the 19th leaf. Then there is an improvement in the 20th node in the training set but a deterioration in the performance of the validation set.

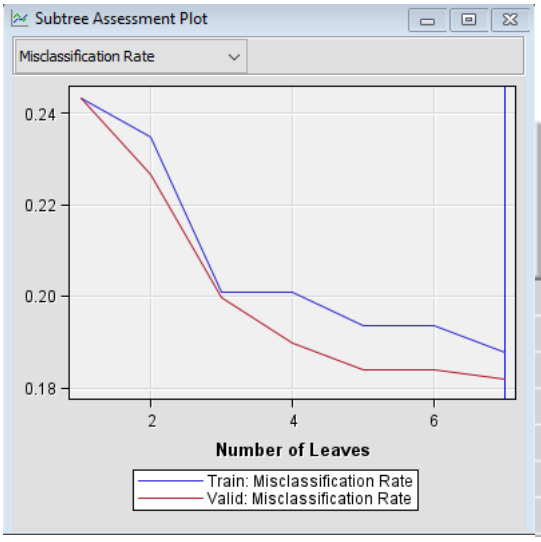
3-way Split decision tree



Number of Leaves	Train: Misclassification Rate	Valid: Misclassification Rate
1	0.2435	0.2433
3	0.2155	0.2117
5	0.2013	0.1993
6	0.1968	0.1967
7	0.1943	0.1907
8	0.1898	0.1880
9	0.1855	0.1863
10	0.1898	0.1847
11	0.1853	0.1820
12	0.1838	0.1800
13	0.1795	0.1783
14	0.1795	0.1783
15	0.1795	0.1783
16	0.1795	0.1783
17	0.1795	0.1783
18	0.1795	0.1783
19	0.1795	0.1783
20	0.1795	0.1783
21	0.1795	0.1783
22	0.1780	0.1790
23	0.1790	0.1807
24	0.1775	0.1813
25	0.1765	0.1833

The above chart shows that the node splits performed well till the 13th node validation set. From the 13th node no big difference is seen in either the training set or the validation set till the 21st leaf. Then there is an improvement in the 22nd node in the training set but a deterioration in the performance of the validation set. Not very different to the pruned decision tree.

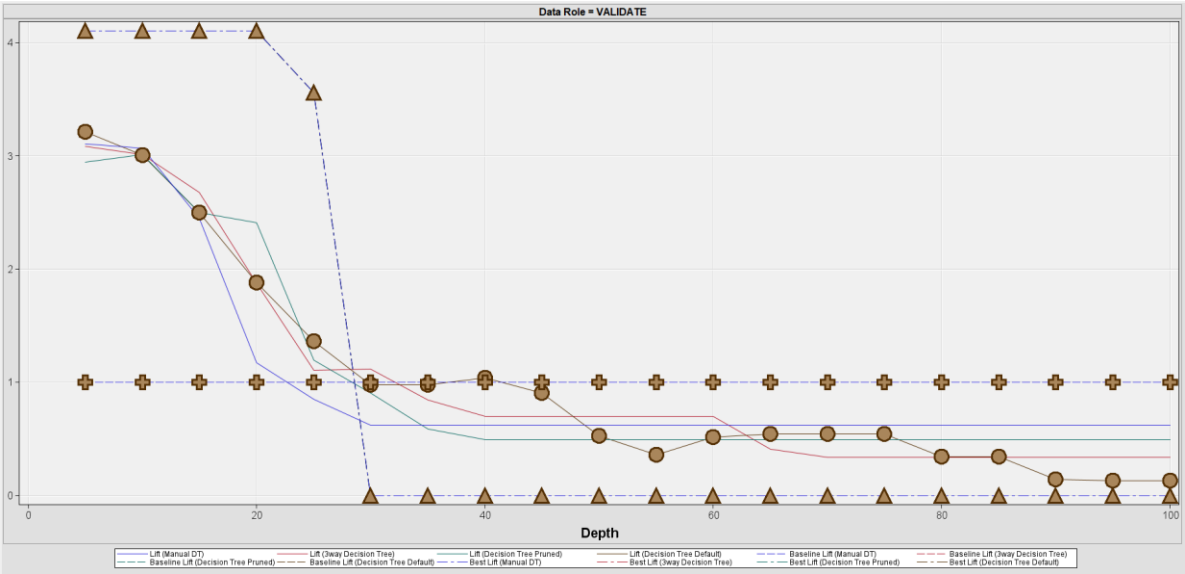
Manual Decision Tree



Number of Leaves	Train: Misclassification Rate	Valid: Misclassification Rate
1	0.2435	0.2433
2	0.2350	0.2267
3	0.2010	0.1997
4	0.2010	0.1897
5	0.1935	0.1840
6	0.1935	0.1840
7	0.1878	0.1820

The above chart shows that the node splits performed well till the 3rd node in both sets. The overall classification rate seems to be reducing for both the training set and the validation set. Near the 7th leaf, the also seem to be converging. This seems to be like a good model.

Model Comparison of all the decision trees constructed



The above is the Lift chart of the decision trees considered. The Default decision tree is the best. This model has the least misclassification rate, high lift at 5% and the highest accuracy among the other models (Appendix Table1).

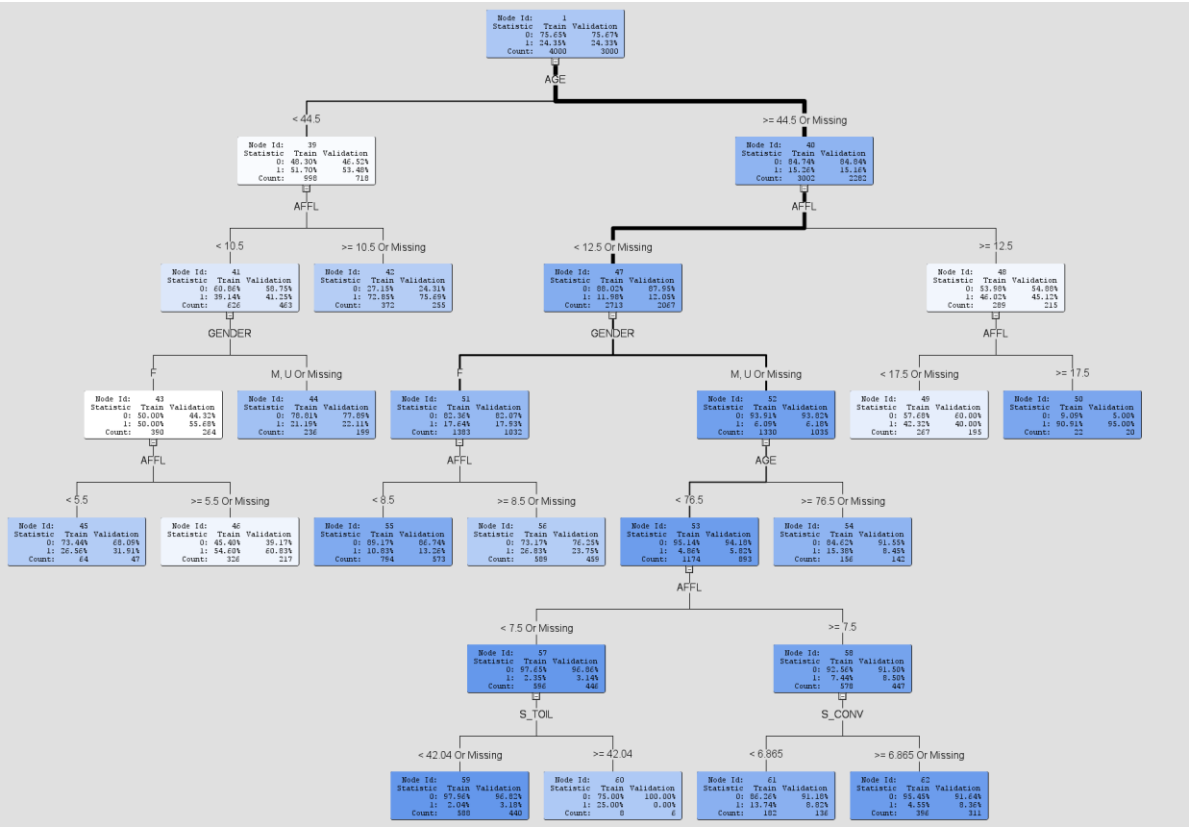
Though the Area Under the curve for the Manual Decision tree (0.86) is the highest in the ROC chart, the misclassification rate is high and the has the least True Positive rate amongst all the decision trees (from below table). Therefore, Manual Decision tree is not the best model.

The 3-way branched Decision tree and the pruned tree are not considered because they have a smaller lift value, area under the curve and a lower precision percentage. A lower precision percentage can affect the decision making for the businesses, because they have low true positive values affecting the model making process.

.

Classifier	%True –	%False –	%True +	%False +
Default Decision Tree	91.1%	8.8 %	55.2 %	44.7 %
Decision Tree Pruned	90.9 %	9.0 %	54.9 %	45.0 %
Decision Tree 3way split	92.8 %	7.1 %	49.0 %	50.9 %
Manual Decision Tree	94.9%	5.0 %	40.9%	59.0 %

Critical path



The bold/thick line is the critical path for the Default decision tree.

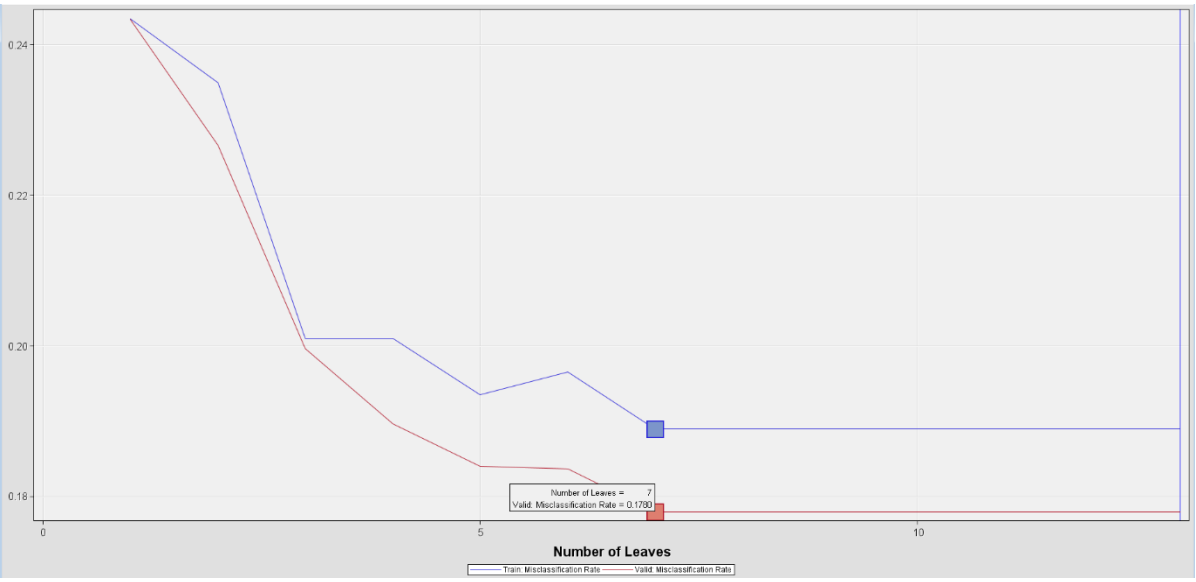
Only 23% of women aged above 44.5 years and affluence between 8.5 and 12.5 have purchased organic products.

Target path of interest

95% of the customers aged above 44.5 years and have an affluence grade more than 12.5 purchase organic products.

Overfitting and limitations

Limitaiton:



From the above misclassification rate graph, we see that the validation set performs better than the training set. From the 7th leaf we see no major improvement, therefore we can consider pruning the decision tree model to the 7th leaf.

Analysis of the best model

Classifier	Missing values removed	Missing values imputed	Outliers removed	Transformation to remove skew	% True positives	% precision	AUC ROC	Estimate lift @ 25% depth
Imputed Stepwise Regression	Yes	Yes	-	-	34.52%	73.04%	0.79	1.61
Imputed Artificial Neural Network	Yes	Yes	-	-	36.9%	72.645%	0.79	1.4
Default Decision Tree	No	No	-	-	55.2%	69.92%	0.81	1.3

Classifier	%True -	%False -	%True +	%False +
Imputed Stepwise Logistic Regression	95.9%	4.0%	34.5%	65.4%
Imputed Artificial Neural Network	95.3%	4.6%	36.9%	63.0%
Default Decision Tree	91.1%	8.8%	55.2%	44.7%

Models	Average Squared Error	Cumulative Lift @5%	AUC_ROC	Accuracy	Precision %	FP	FN	TN	TP
Default Decision Tree	17.80%	3.21	0.81	82.2	69.92	148	386	2122	344
Imputed Stepwise Regression	19.03%	3.64	0.79	80.97	73.04	93	478	2177	252
Imputed Neural Network	19.20%	3.67	0.79	80.8	72.65	93	483	2177	247

Table 8 summary results of the best performing models

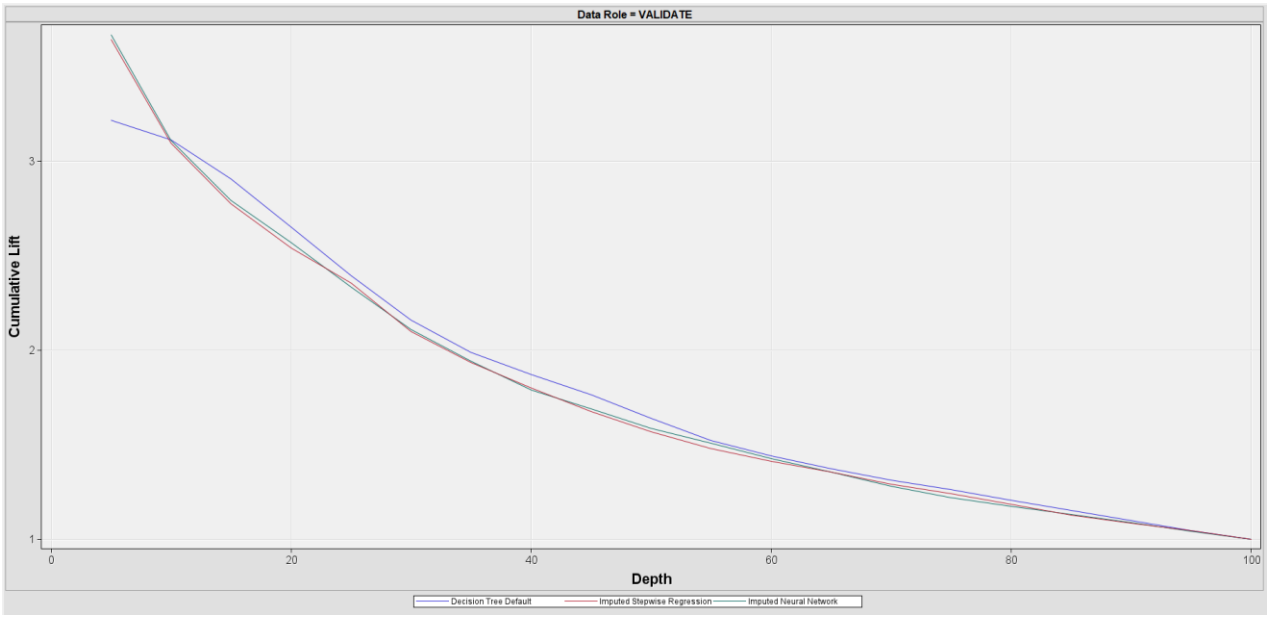


Figure 2 Cumulative lift chart of the best performing models

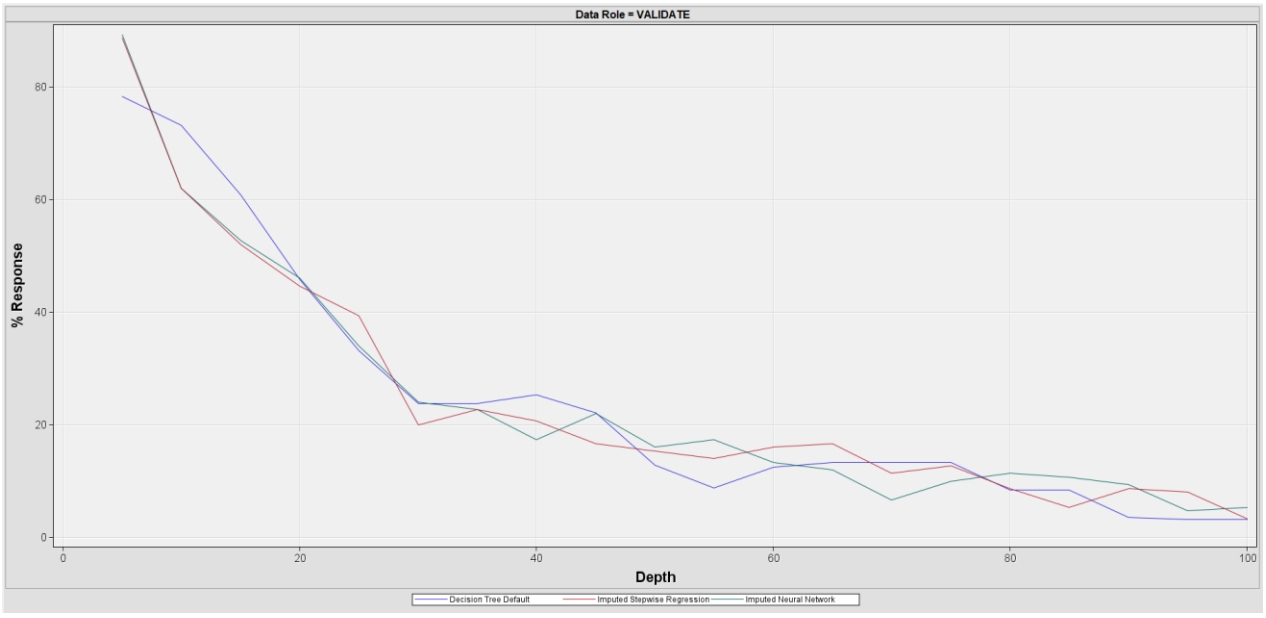


Figure 3 Non-cumulative lift chart of the best performing models

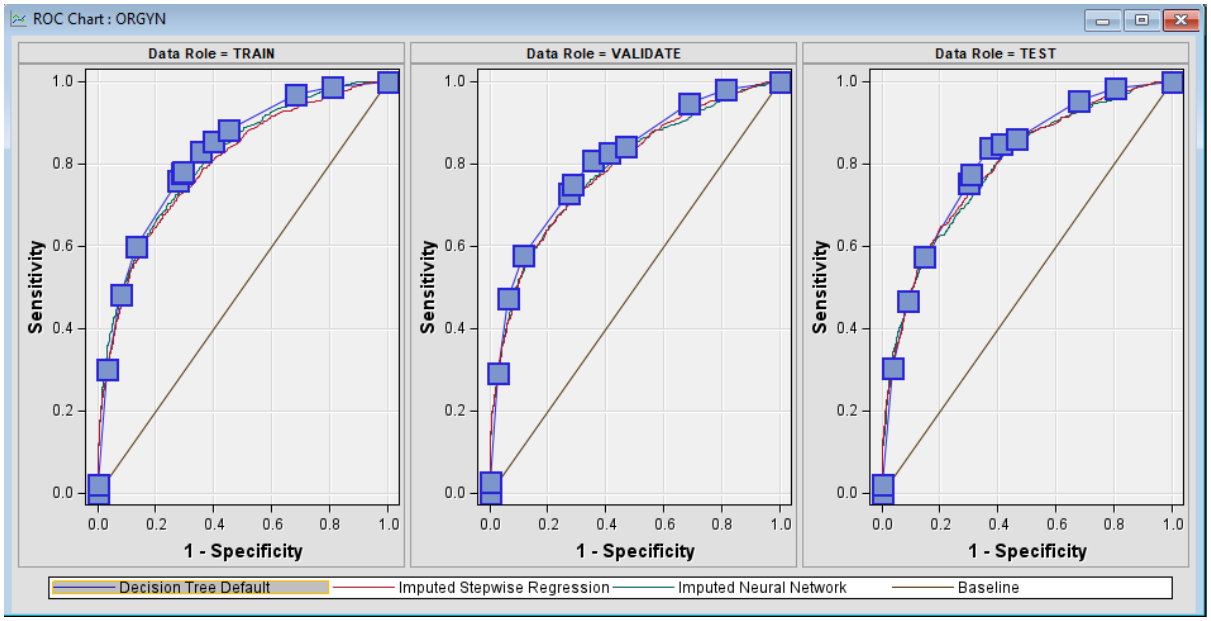


Figure 4: ROC chart for best performing model

Frequency| Percent | Row Pct | Col Pct	0	1	Total
0	2069	201	2270
	68.97	6.70	75.67
	91.15	8.85	
	84.14	37.15	
1	390	340	730
	13.00	11.33	24.33
	53.42	46.58	
	15.86	62.85	
Total	2459	541	3000
	81.97	18.03	100.00

Table 9 summary of confusion matrix for test set

Conclusion

From the above Table 8, we see that the decision tree is the best performing model. It has the least misclassification rate, has more area under the ROC curve and makes a curve towards the upper left corner and a better accuracy. predicting more accurately compared to regression and the neural network. Women aged more than 44.5 years and AFFLUENCE less than 12.5 purchase organic products. The company can target the customers fitting this description.

Recommendation

There are a few variables that I believe do not hold importance for this data mining problem.

The LCDATE and the LTIME are co-related. LCDATE can be removed because the data has a lot of data towards the end of the 1900's. That means loyalty card holders have only started applying since recently, for which the time is small when considered to the current date.

The NGROUP variable can have better naming system instead of A, B, C… The better naming can help us to link the neighbourhood region to the Region and the TV region to target the customers better. If they are similar to OAC, we can consider discarding the variable.

ETHNICITY, of each customer, is a variable that can be introduced to help the business to target the customers. There is a possibility where younger adults are more inclined to purchasing organic products. If this is true, we can show the same results to the customers and attract the younger generation to purchase more organic goods.

References and Bibliography

1. Williams, A., n.d. *Regression II*.
2. Williams, A., 2018. *(Binary) Logistic Regression Part 2*.
3. SAS Institute Inc. 2013. Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach, Third Edition. Cary, NC: SAS Institute Inc
4. Wiliams, A., 2021. *Assessment*.
5. Wielenga, D., 2017. [online] SAS Communities. Available at: <https://communities.sas.com/t5/SAS-Data-Mining-and-Machine/Gains-vs-response-charts/td-p/387615>.

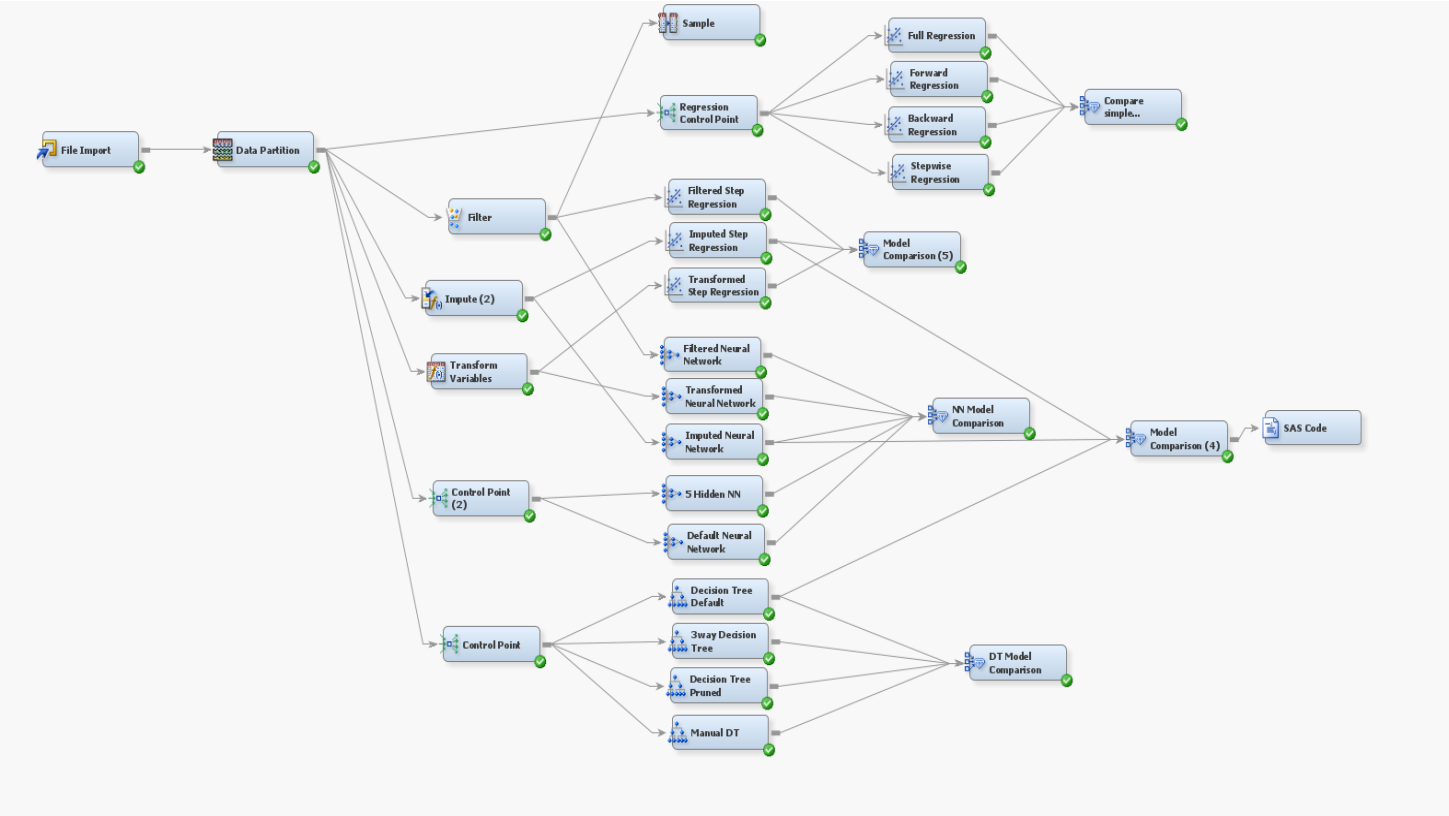
Appendix

General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	3016
Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	09/11/21 13:43
Run ID	1280d684-f1b1-4ab7-a374-1
Last Error	
Last Status	Complete
Last Run Time	29/11/21 13:15
Run Duration	0 Hr. 0 Min. 8.77 Sec.
Grid Host	
User-Added Node	No

Seed Generator

Name	Role
AFFL	Input
AGE	Input
BILL	Input
CLASS	Input
CUSTID	Rejected
GENDER	Input
LCDATE	Rejected
LTIME	Input
NGROUP	Input
OAC	Input
ORGYN	Target
REGION	Input
S_CONV	Input
S_FVEG	Input
S_MT	Input
S_TOIL	Input
TV_REG	Input

Data Mining Roles



Workflow diagram

Models	Misclassification Rate	Lift @ 5%	AUC_ROC	Accuracy	Precision %	FP	FN	TN	TP
Default DT	17.80%	3.21	0.81	82.2	69.92	148	386	2122	344
Pruned DT	17.83%	2.95	0.76	82.17	66.06	206	329	2064	401
3way DT	17.83%	3.08	0.79	82.17	68.71	163	372	2107	358
Manual DT	18.20%	3.21	0.86	81.8	72.22	115	431	2155	299
Full Reg	19.63%	3.56	0.77	80.37	71.96	90	499	2180	231
Forward Reg	19.23%	3.53	0.77	80.77	73.54	86	491	2184	239
Backward Reg	19.23%	3.53	0.77	80.77	73.54	86	491	2184	239
Stepwise	19.23%	3.53	0.77	80.77	73.54	86	491	2184	239
Filtered Step Reg	20.51%	3.34	0.79	79.49	72.13	80	367	1525	207
Imp Step Reg	19.03%	3.64	0.79	80.97	73.04	93	478	2177	252
Transformed Step Reg	19.13%	3.56	0.77	80.87	73.21	90	484	2180	246
Imp NN	19.20%	3.67	0.79	80.8	72.65	93	483	2177	247
Default NN	19.23%	3.69	0.76	80.77	72.43	94	483	2176	247
5 Hidden Node NN	19.73%	3.62	0.76	80.27	69.6	107	485	2163	245
Filtered NN	20.74%	3.34	0.78	79.26	70.61	87	365	1518	209
Transformed NN	19.47%	3.53	0.75	80.53	70.86	102	482	2168	248

Table1: Models Performance data

My Reflections on the Patchwork assignment

I had difficulties installing and running the SAS Enterprise Miner Workstation on my personal computer. Then, after a few trials I managed to run it. But I was getting this error to update my JAVA version every time I ran it. So, I decided to work from the university for the coursework. Initially, the lectures and the labs seemed to be very interesting and easy. They were straightforward. As we got deep into the module, I found the topics to be a little difficult, so I had to spend more time in the labs to catchup with the others. Alex, the lab tutor, was of great help when I was unable to progress with the lab sheets. The discussion board was a great tool to discuss questions. Many queries I had were faced by my peers as well. The quick responses from the tutor helped us progress well.

The weekly journal was helpful in tracking my progress. Since I was not able to answer many questions in them, I decided to restart all the activities in the course work during the enhancement week, though I was up to date with the labs. I had to revisit the lectures and the labs to gain more confidence. I was able to appreciate the learning materials much better than before. The recordings were also very useful in completing the course work.