Data Mining Assignment

Latisha Panwar
D17124379
MSc in Data Analytics

Business understanding

Retail markets are a crucial part of everyone's life. They have made the lives of their customers easier by offering a huge variety of products at the same place. Over the centuries, the retail business has tremendously grown and taken the shapes of large supermarkets and sophisticated malls. Retail markets aim to improve their business to achieve high profits and better business by offering variety of products, new products etcetera to attract more customers and profit.

Business objective

The data set considered for the analysis belongs to a supermarket, which is offering a new line of organic products. They can increase the sales and profits by targeting the correct customers for its product. So the management of the supermarket aims to investigate the customers who are likely to purchase the organic products.

Data mining goals

The data mining goal is to come up with the best classification model which can determine the target segment of audience/ customers who are likely to buy the products. The classification models implemented to determine the target group are decision trees, regression model and neural networks. The aim will be to come up with the best possible model for the supermarket. The best model will be tried to choose based on its accuracy, complexity to implement and performance.

The important or target variable to be considered in the data set is 'ORGYN', which tells if the customer bought the organic product or not. Other variables which can prove to be important to identify the target audience are demographic information(age, neighborhood area, region etc), gender, class, affluence grade.

Data understanding

The supermarket has a customer loyalty program in which they provided coupons for the organic products to all the participant of the program. They have collected the data that includes various details about the customers like demographic information (age, neighborhood area, region etc), gender, class, affluence grade, etc and whether these customer bought any of the organic products or not.

Collect initial data

The data set was imported to the SAS Enterprise Miner. Measurement levels were defined for the variables in the data set as Unary for EDATE, Binary for ORGYN(as it can have only two values, 0 or 1), Interval for continuous variables AGE, ORGANICS, BILL, AFFL and date variables DOB, EDATE, LCDATE, and categorical variables such as CUSTID, GENDER, AGEGRP1, AGEGRP2, TV_REG, NGROUP, NEIHBORHOOD, REGION, CLASS were defined as Nominal variables. These variables are defined under Describe data section.

Describe data

The dataset consists of 22223 observations and 18 attributes. The variables in the data set are defined below.

The target variable for the analysis is ORGYN which gives information if the organic product was purchased or not. It is a binary variable with values as 0 and 1. 0 means the organic product is not purchased and 1 means the product was bought.

Other variable in the data set are CUSTID(unique identification number for each customer), GENDER(Male, Female or Unknown), DOB(date of birth of the customer), AGE(in years of the customer), AGEGRP1(ages with intervals of 20 years such as less than 20, 20 to 40, 40 to 60, 60 to 80), AGEGRP2(ages with intervals of 10 years such as 10-20, 20-30, 30-40 and so on till 70-80), TV_REG(television region), NGROUP(neighborhood group divided between 7 codes A to F and U), NEIGHBORHOOD(type of residential neighborhood), LCDATE(date for loyalty card application), LTIME(number of years since a customer has been a loyalty card member), ORGANICS(number of organic products purchased), BILL(total amount spent), Region(Geographic Region), CLASS(define the loyalty status of the customer, divided between tin, silver, gold, or platinum), AFFL(richness of the customer on the scale of 1 to 30).

Explore data

After rejecting these variables, descriptive analysis was carried out on the data set to observe possible patterns and missing values before building any model. For this purpose StatExplore node(from Explore tab) was used to generate descriptive statistics.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	AGEGRP1	INPUT	5	1508	40-60	43.80	60-80	33.89
TRAIN	AGEGRP2	INPUT	8	1508	50-60	23.46	60-70	20.87
TRAIN	CLASS	INPUT	4	0	Silver	38.57	Tin	29.19
TRAIN	GENDER	INPUT	4	2512	F	54.67	M	26.17
TRAIN	NEIGHBORHOOD	INPUT	56	674	52	5.42	27	4.22
TRAIN	NGROUP	INPUT	8	674	C	20.55	D	19.70
TRAIN	REGION	INPUT	6	465	South East	38.85	Midlands	30.33
TRAIN	TV_REG	INPUT	14	465	London	27.85	Midlands	14.05
TRAIN	ORGYN	TARGET	2	0	0	75.23	1	24.77

Interval Variable Summary Statistics
(maximum 500 observations printed)

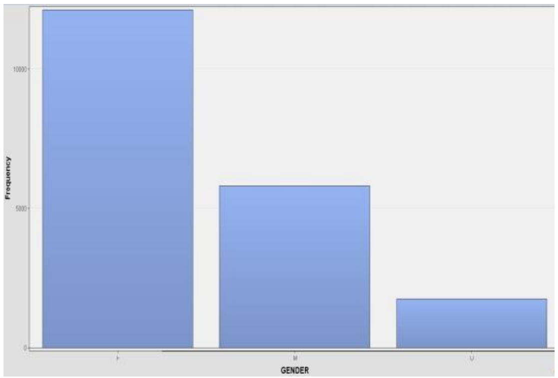
Data Role=TRAIN

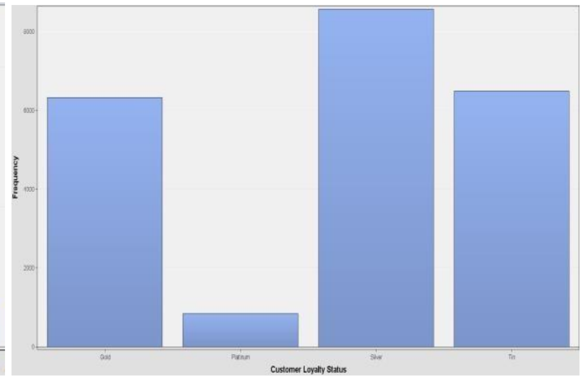
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
AFFL	INPUT	8.711893	3.421125	21138	1085	0	8	34	0.891684	2.09686
AGE	INPUT	53.79715	13.20605	20715	1508	18	54	79	-0.07983	-0.84389
BILL	INPUT	4420.59	7559.048	22223	0	0.01	2000	296313.9	8.037186	184.8715
DOB	INPUT	-5877.32	4825.523	22223	0	-15266	-5842	7190	0.077679	-0.84924
LTIME	INPUT	6.56467	4.657113	21942	281	0	5	39	2.28279	8.077622
ORGANICS	INPUT	0.29474	0.562831	22223	0	0	0	3	2.021011	4.245531

Now from the results above we can clearly see that our data set contains missing values for most of the variables. So the next step will be to replace these values.

Key points from the descriptive statistics shown above:

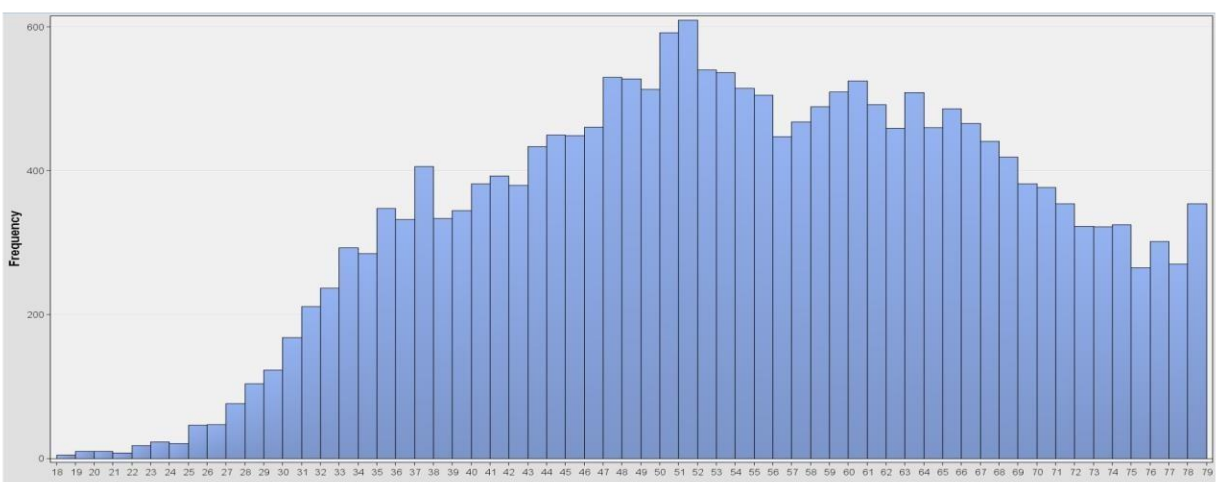
- For the target variable ORGYN, it can be said(based on Mode Percentage) that 75% of the customers did not buy the organic products.
- Most customers who bought these products are female (almost 55%), lies between the age of 50 to 60 and belongs to Silver class.



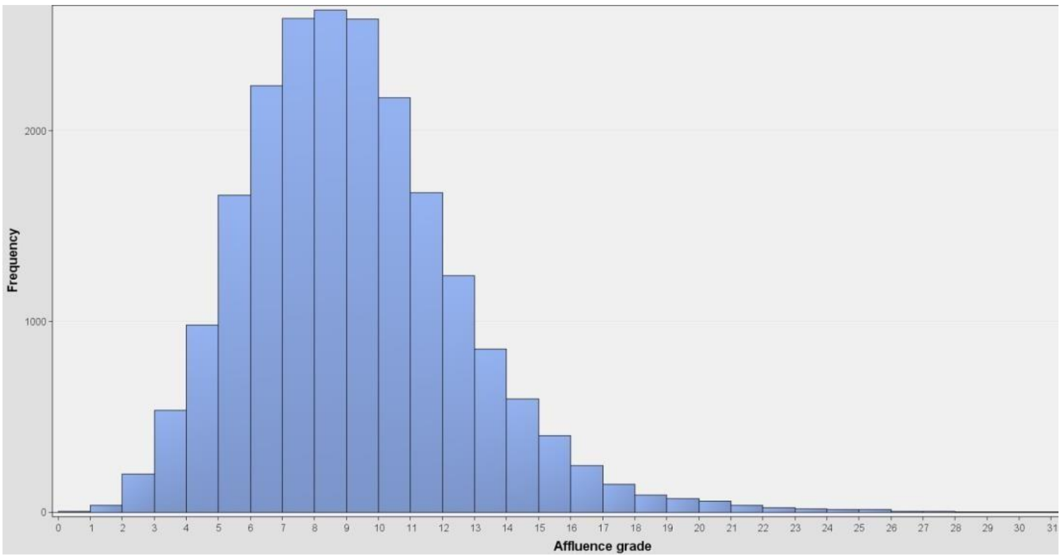


- For interval variable 'Bill', it can be said that it is very positively skewed and cannot be considered normal. Also, the kurtosis value is very large which means not only the data is skewed but the peak formed at the right is very tall. This means that a certain amount has been paid by many customers. This can be interesting to know why the values are like that. It is observed from the graph that a large number of people have spent as less as 10 cents and customers have spent up to forty thousand on the products(hence the large standard deviation).
  It was observed from the data set that that all the customers with loyalty status as 'tin' are the ones who spent only 10 cents. Possible reasons can be that these customers belong to very low affluence grade, or they used the coupon over a very small bill amount and hence obtained a huge discount, etcetera. These reasons should be investigated by the retailer to get any interesting insight and work on that.

- Age is also an important factor with large standard deviation as there are customers from as less as 18 years old to 79 years old. Skewness for age is slightly negatively skewed which means that the data is left skewed, so it can be said that older customers are spending more on the products as can be seen below:



- Another important factor to consider is 'AFFL' which is the affluence grade of customers on the scale of 1 to 30. As can be seen from the graph below, it is positively skewed with more number of people lying in the initial range of scale. This information can help determine which category of people should be targeted more.



Data preparation

Select data

After going through the data set it was realized that some attributes are more important for the analysis than other. Hence, attributes which do not hold much importance in the analysis of determining the target customers were rejected in the beginning. These variables are:

CUSTID - This attribute is a nominal variable which is a customer loyalty identification number and is provided by the retailer. So, it should not be considered a factor to determine the target group of audience.

EDATE - This is a unary variable which provides the date extracted from the daily sales data base. As this value is same for all the observations and has no variation so it cannot be considered important for the analyses.

LCDATE - This interval variable tells about loyalty card application date. This information is useful however, this data can be obtained from LTIME variable, which clearly gives time as loyalty card member in years and does not contain of missing values as well.

Cleaning data

The data set contains NAs, missing values and it is observed that missing information for one variable affects another. For example, 'REGION' variable has the same fields as missing as that of 'TV_REGION', similarly 'AGEGRP1' and 'AGEGRP2' has missing values for each field where 'AGE' variable has NAs. Hence, after data exploration, the next step is to transform the variables to avoid large standard deviation or outliers, skewness and missing values to prepare our data for modeling.

Replacement nodes(from Modify tab) were implemented and connected to the data source to replace NA and missing values. For the replacement nodes 'Default Limits Method' under properties of the nodes for interval variables was selected to none from default 'Standard deviation from the mean', as one of the ideas for replacement is to reduce large deviations. Missing values were first replaced by an unformatted character '?' in first node and then replaced with the value '_MISSING_'. Following table shows the number values that were replaced.

Total Replacement Counts

Variable	Role	Label	Train
REP AGEGRP1	INPUT	Replacement: Age Group 1	1508
REP AGEGRP2	INPUT	Replacement: Age Group 2	1508
REP GENDER	INPUT	Replacement: GENDER	2512
REP NEIGHBORHOOD	INPUT	Replacement: Type of Residential Neighborhood	674
REP NGROUP	INPUT	Replacement: Neighborhood Group	674
REP REGION	INPUT	Replacement: Geographic Region	465
REP TV REG	INPUT	Replacement: TV Region	465

After replacing the missing values, data was then partitioned between trained data(for preliminary model fitting) and validation data(to empirically test the model) in 50-50 ratio.
However, the data is still incomplete and might not yield good results for every model. A solution to this is imputing the missing values with appropriate values and then transforming them to stabilize skewness and improve model response.

Feature Selection Task:

What happens if ORGANICS feature is used as an input feature when building models?
When ORGANICS feature is used as an input feature in the model, a perfect model is generated. For example, let us consider the following SAS output:

Fit Statistics

Target=ORGYN Target Label=Organics Purchased?

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	11112	11111
MISC	Misclassification Rate	0	0
MAX	Maximum Absolute Error	0	0
SSE	Sum of Squared Errors	0	0
ASE	Average Squared Error	0	0
RASE	Root Average Squared Error	0	0
DIV	Divisor for ASE	22224	22222
DFT	Total Degrees of Freedom	11112	.

It can be observed that misclassification rate is zero, which is possible when the model is 100% accurate. In a real situation achieving a model with 100% accuracy is not possible. Now, the reason for this can be that ORGANICS feature is not giving any other useful information than what ORGYN variable is giving. Most of the values contained in the columns are duplicate.

What happens when this feature is not used?
When ORGANICS feature is dropped from the evaluation, following changes were observed in the fit statistics:

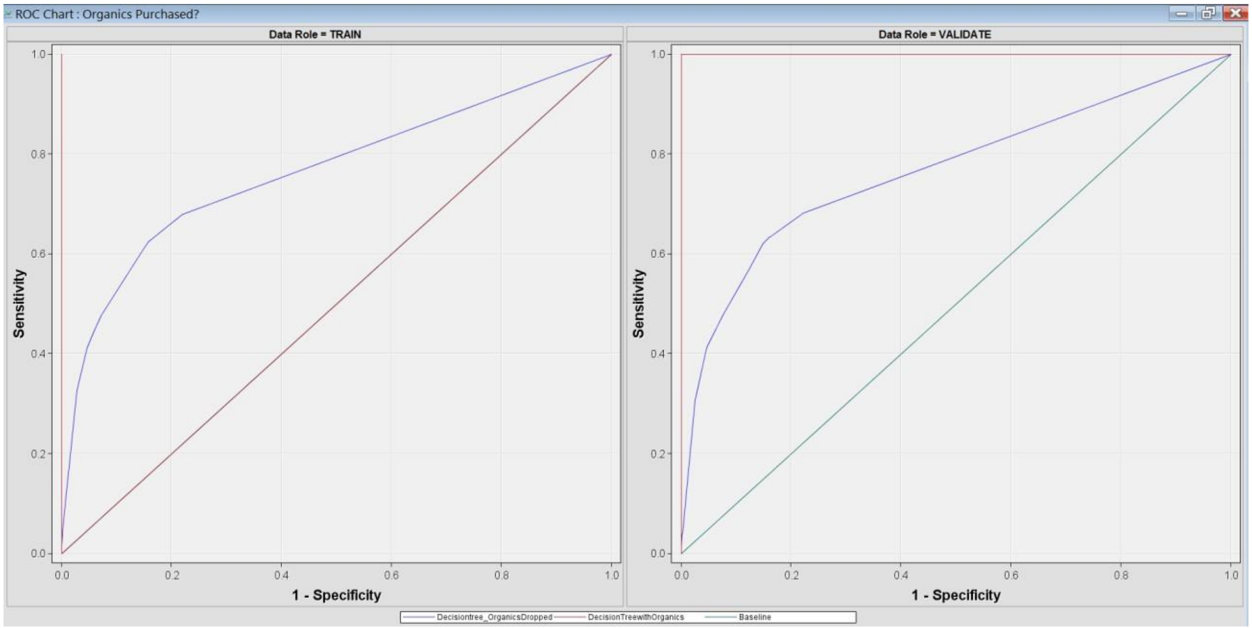
Fit Statistics

Target=ORGYN Target Label=Organics Purchased?

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	11112.00	11111.00
MISC	Misclassification Rate	0.18	0.18
MAX	Maximum Absolute Error	0.92	0.92
SSE	Sum of Squared Errors	3061.43	3065.69
ASE	Average Squared Error	0.14	0.14
RASE	Root Average Squared Error	0.37	0.37
DIV	Divisor for ASE	22224.00	22222.00
DFT	Total Degrees of Freedom	11112.00	.

All the fit statistics have changed and have become more accurate and believable. It shows that 18% of the cases in the dataset came wrong.

Comparison between the two outcomes:



When the two models(one with ORGANICS and another when ORGANICS is dropped) are compared, it confirms again that the that the model with ORGANICS proves to be 100% accurate. This can be said as the line for 'DecisionTreewithOrganics' is as far as it can be from the baseline. However, the line for 'DecisionTree_OrganicsDropped' is reasonably far away from the model. More details can be observed below from the fit statistics:

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Tree	Tree	Decision...	ORGYN	Organics...	0	11112	0	0	0	0	0
Tree2	Tree2	Decisiont...	ORGYN	Organics...	0.181262	11112	0.180526	0.921569	3061.432	0.137753	0.371151

Predecessor Node	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic
Tree	22224	11112	11111	0	0	0	0	0	22222	1	1	1
Tree2	22224	11112	11111	0.181262	0.921569	3065.686	0.137957	0.371426	22222	0.765	0.531	0.465

Should this feature be used or left out?
This feature should be left out as it is hampering the results of the model and affecting its accuracy to predict the results. As can be seen from the fit statistics above, the model with ORGANICS has values as 0 and 1 for all the statistics.

Modeling

A number of classification models and techniques, such as decision trees, regression model, neural networks are used to determine which customers are likely to purchase the new line of products.

Decision Trees

Decision Tree is one of the classification models used for the analysis. As missing values can be handled by decision trees, so missing values are not imputed for this model. This is because surrogate splitting rule can be used to select other variable values when splitting variable values are missing.

ORGANICS was dropped as a feature after proving that it's not worth to be considered in the model. Now, various decision trees were implemented with different settings to obtain the best model.

Tree1 - is the model with the default settings which are provided by the machine. Other trees with modified settings can be compared to this tree and a comparison between machine generated model and the one with manually changed settings can be made.

Tree2 - 'Missing values' under 'Splitting Rule' in properties panel was changed to 'Largest Branch' which means that during split search observations with missing values will be replaced by largest number of training observation.

Tree3 - 'Interval Target Criteria' was changed to 'Variance' and other setting were left the same.

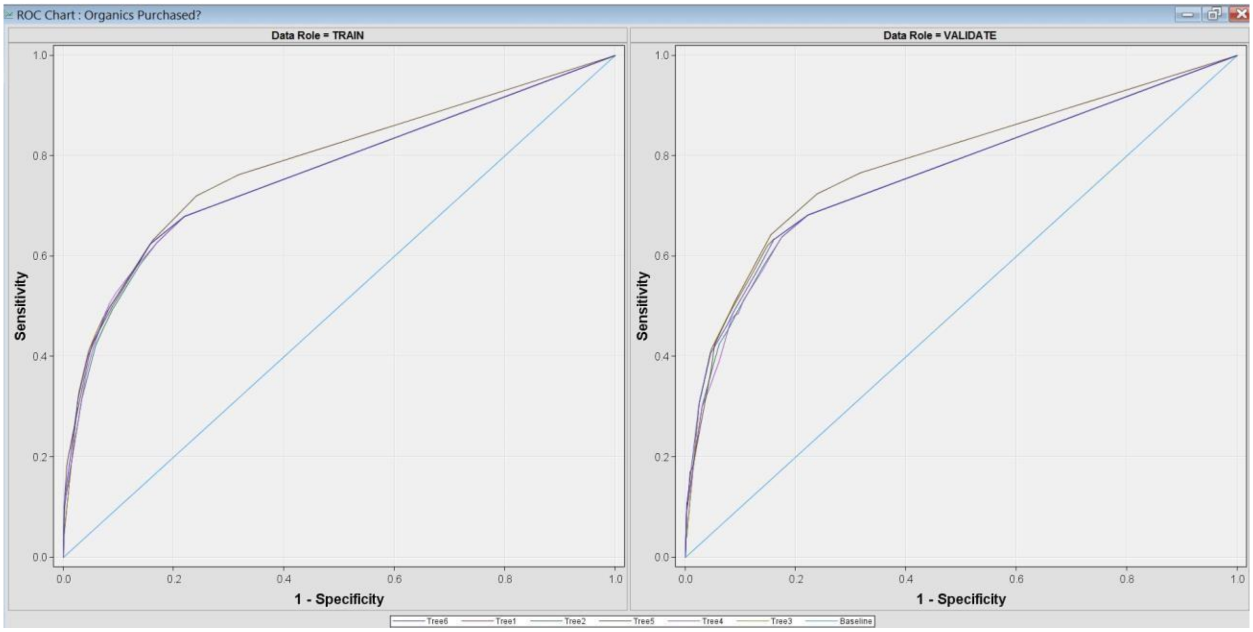
Tree4 - 'Nominal Target Criteria' was changed to 'Gini' and 'Missing Values' to 'Largest Branch'.

Tree5 - 'Nominal Target Criteria' was changed to 'Entropy' and no other changes were made.

Tree6 - 'Missing Values' was changed to 'Most correlated branch'

After making these models with above mentioned settings, these models were then compare to find out which tree can be considered the best. The results obtained were as follows:

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic
Tree3	ORGYN	Organics...	0.181262	11112	0.180526	0.921569	3061.432	0.137753	0.371151	22224	11112	11111	0.181262	0.921569	3065.686	0.137957	0.371426	22222	0.765	0.531	0.465
Tree1	ORGYN	Organics...	0.181262	11112	0.180526	0.921569	3061.432	0.137753	0.371151	22224	11112	11111	0.181262	0.921569	3065.686	0.137957	0.371426	22222	0.765	0.531	0.465
Tree6	ORGYN	Organics...	0.182252	11112	0.182595	0.92674	3057.776	0.137589	0.37093	22224	11112	11111	0.182252	0.92674	3064.299	0.137895	0.371342	22222	0.766	0.531	0.465
Tree5	ORGYN	Organics...	0.182792	11112	0.182145	0.92674	3017.778	0.135789	0.368496	22224	11112	11111	0.182792	1	3031.244	0.136407	0.369334	22222	0.784	0.568	0.479
Tree2	ORGYN	Organics...	0.189272	11112	0.187545	0.92674	3117	0.140254	0.374505	22224	11112	11111	0.189272	0.92674	3133.062	0.140989	0.375485	22222	0.762	0.525	0.459
Tree4	ORGYN	Organics...	0.193502	11112	0.184755	0.92674	3088.995	0.138994	0.372819	22224	11112	11111	0.193502	0.92674	3168.565	0.142587	0.377607	22222	0.764	0.528	0.459



From the fit statistics and ROC chart, it can be said that Tree1 and Tree3 gave lowest misclassification rate but the exact same results. This means that interval target criteria has no effect on the overall model. Tree5 gives good result as can be seen from ROC chart, which can mean that nominal target criteria has some positive effect on the tree. However, the misclassification rate of it is more than Tree1 and Tree3.

As the models with other settings failed to generate better results than what the machine automatically suggested, hence it would be better to go ahead with Tree1.

Regression Model

Another model considered for the classification is regression model. Regression model is sensitive to missing values. Hence, it is important to impute the missing values present in the data. For this, interval variables were imputed with Median, as when the values are spread out largely, then median can prove o be less sensitive to it and in replacing values in skewed distribution.

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
AFFL	MEDIAN	IMP AFFL	M AFFL	8	INPUT	INTERVAL	Affluence grade	525
AGE	MEDIAN	IMP AGE	M AGE	54	INPUT	INTERVAL	Age	751
LTIME	MEDIAN	IMP LTIME	M LTIME	5	INPUT	INTERVAL	Years as Loyalty Card Member	136
REP REP AGEGRP1	TREESURR	IMP REP REP AGEGRP1	M REP REP AGEGRP1	.	INPUT	NOMINAL	Replacement: Replacement: ...	751
REP REP AGEGRP2	TREESURR	IMP REP REP AGEGRP2	M REP REP AGEGRP2	.	INPUT	NOMINAL	Replacement: Replacement: ...	751
REP REP GENDER	TREESURR	IMP REP REP GENDER	M REP REP GENDER	.	INPUT	NOMINAL	Replacement: Replacement: ...	1215
REP REP NEIGHBORHOOD	TREESURR	IMP REP REP NEIGHBOR...	M REP REP NEIGHBORHO...	.	INPUT	NOMINAL	Replacement: Replacement: ...	328
REP REP NGROUP	TREESURR	IMP REP REP NGROUP	M REP REP NGROUP	.	INPUT	NOMINAL	Replacement: Replacement: ...	328
REP REP REGION	TREESURR	IMP REP REP REGION	M REP REP REGION	.	INPUT	NOMINAL	Replacement: Replacement: ...	236
REP REP TV REG	TREESURR	IMP REP REP TV REG	M REP REP TV REG	.	INPUT	NOMINAL	Replacement: Replacement: ...	236

After imputing the values, data was then transformed to stabilize skewness, improve model accuracy and non-normality. For this Log10 function was applied to interval variables.

```
Computed Transformations
(maximum 500 observations printed)

Input                   Input
Name        Role        Level      Name             Level         Formula

BILL        INPUT       INTERVAL   LG10_BILL        INTERVAL      log10(BILL  + 1)
DOB         INPUT       INTERVAL   LG10_DOB         INTERVAL      log10(DOB  + 15248)
IMP_AFFL    INPUT       INTERVAL   LG10_IMP_AFFL    INTERVAL      log10(IMP_AFFL  + 1)
IMP_AGE     INPUT       INTERVAL   LG10_IMP_AGE     INTERVAL      log10(IMP_AGE  + 1)
IMP_LTIME   INPUT       INTERVAL   LG10_IMP_LTIME   INTERVAL      log10(IMP_LTIME  + 1)
```

After imputation and transformation, this data is then supplied to regression model.

Model1 - Model selection under properties was selected as stepwise and rest of the settings were left unchanged.

Model2 - Optimization technique was chosen as 'Congra', 'Link function' is selected as 'Cloglog'. Model selection under properties was selected as stepwise.

Model3 - Link function is selected to 'Probit', 'Use selection defaults' was set to 'No', to not choose default values for the model selection technique. Model selection under properties was selected as stepwise.

Model4 - 'Uses Defaults' under 'Convergence Criteria' was chosen 'No', 'Selection Model' is not chosen and left as None for this model as to see the effects.

Model5 - Optimization technique 'Newrap' was chosen, 'Selection Model' as stepwise.

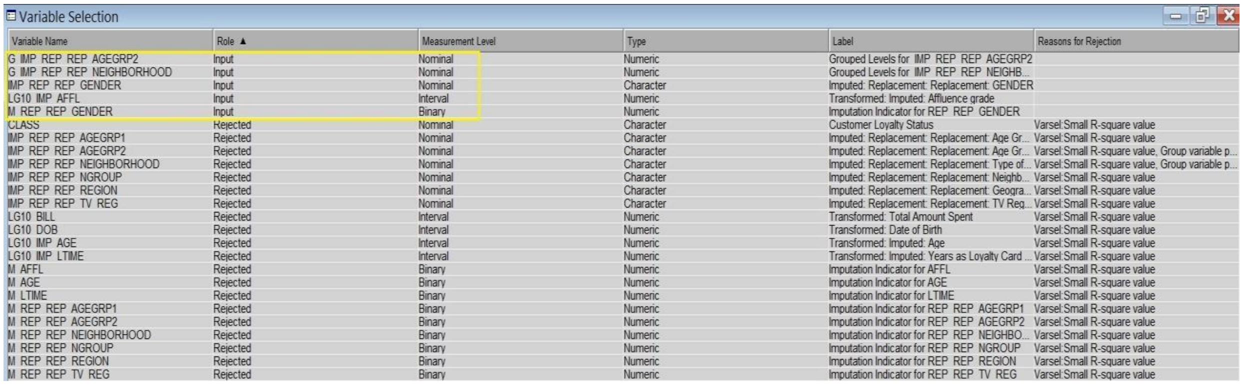
All the models were compared and analyzed based on following statistics:

Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error
Reg	Regressi...	ORGYN	Organics...	0.184052	9430.942	0.134493	0.423189	11099	13	11112	22224	9404.942	0.134808	0.98862	0.13465
Reg5	Regressi...	ORGYN	Organics...	0.184052	9430.942	0.134493	0.423189	11099	13	11112	22224	9404.942	0.134808	0.98862	0.13465
Reg3	Regressi...	ORGYN	Organics...	0.184592	9450.308	0.134758	0.42406	11099	13	11112	22224	9424.308	0.135074	0.994486	0.134916
Reg4	Regressi...	ORGYN	Organics...	0.186122	9511.599	0.13335	0.419438	11017	95	11112	22224	9321.599	0.13565	0.988145	0.1345
Reg2	Regressi...	ORGYN	Organics...	0.247682	12444.11	0.18637	0.55985	11111	1	11112	22224	12442.11	0.186404	0.75225	0.186387

Regression model 1(Reg) and 5 yielded the best results as compared to other models. It should also be noted that with optimization technique 'Congra' and link function as 'Cloglog' (Regression model 2), the misclassification rate and average square error is the maximum, hence regression model 2 should be rejected. Also, regression models 3 and 4 gives misclassification rate higher than what models 1 and 5 offers. So, either of the model 1 or 5 can be chosen.

Neural Networks

Neural networks are also sensitive to missing values. So the values imputed and transformed above can be used by neural networks as well. Also, it is a good idea to provide less number of inputs to this type of model, to improve complexity and results. To achieve this, Variable Selection node was connected to neural networks. This node will take the transformed data and only pass those variables to neural networks whose value will be greater than R^2 value, other variables will be rejected. After variable selection, following attributes were passed to neural networks.


Variable Selection

Variable Name	Role ▲	Measurement Level	Type	Label	Reasons for Rejection
G IMP REP REP AGEGRP2	Input	Nominal	Numeric	Grouped Levels for IMP REP REP AGEGRP2	
G IMP REP REP NEIGHBORHOOD	Input	Nominal	Numeric	Grouped Levels for IMP REP REP NEIGHB...	
IMP REP REP GENDER	Input	Nominal	Character	Imputed: Replacement: Replacement: GENDER	
LG10 IMP AFFL	Input	Interval	Numeric	Transformed: Imputed: Affluence grade	
M REP REP GENDER	Input	Binary	Numeric	Imputation Indicator for REP REP GENDER	
CLASS	Rejected	Nominal	Character	Customer Loyalty Status	Varsel:Small R-square value
IMP REP REP AGEGRP1	Rejected	Nominal	Character	Imputed: Replacement: Replacement: Age Gr...	Varsel:Small R-square value
IMP REP REP AGEGRP2	Rejected	Nominal	Character	Imputed: Replacement: Replacement: Age Gr...	Varsel:Small R-square value, Group variable p...
IMP REP REP NEIGHBORHOOD	Rejected	Nominal	Character	Imputed: Replacement: Replacement: Type of...	Varsel:Small R-square value, Group variable p...
IMP REP REP NGROUP	Rejected	Nominal	Character	Imputed: Replacement: Replacement: Neighb...	Varsel:Small R-square value
IMP REP REP REGION	Rejected	Nominal	Character	Imputed: Replacement: Replacement: Geogra...	Varsel:Small R-square value
IMP REP REP TV REG	Rejected	Nominal	Character	Imputed: Replacement: Replacement: TV Reg...	Varsel:Small R-square value
LG10 BILL	Rejected	Interval	Numeric	Transformed: Total Amount Spent	Varsel:Small R-square value
LG10 DOB	Rejected	Interval	Numeric	Transformed: Date of Birth	Varsel:Small R-square value
LG10 IMP AGE	Rejected	Interval	Numeric	Transformed: Imputed: Age	Varsel:Small R-square value
LG10 IMP LTIME	Rejected	Interval	Numeric	Transformed: Imputed: Years as Loyalty Card ...	Varsel:Small R-square value
M AFFL	Rejected	Binary	Numeric	Imputation Indicator for AFFL	Varsel:Small R-square value
M AGE	Rejected	Binary	Numeric	Imputation Indicator for AGE	Varsel:Small R-square value
M LTIME	Rejected	Binary	Numeric	Imputation Indicator for LTIME	Varsel:Small R-square value
M REP REP AGEGRP1	Rejected	Binary	Numeric	Imputation Indicator for REP REP AGEGRP1	Varsel:Small R-square value
M REP REP AGEGRP2	Rejected	Binary	Numeric	Imputation Indicator for REP REP AGEGRP2	Varsel:Small R-square value
M REP REP NEIGHBORHOOD	Rejected	Binary	Numeric	Imputation Indicator for REP REP NEIGHBO...	Varsel:Small R-square value
M REP REP NGROUP	Rejected	Binary	Numeric	Imputation Indicator for REP REP NGROUP	Varsel:Small R-square value
M REP REP REGION	Rejected	Binary	Numeric	Imputation Indicator for REP REP REGION	Varsel:Small R-square value
M REP REP TV REG	Rejected	Binary	Numeric	Imputation Indicator for REP REP TV REG	Varsel:Small R-square value

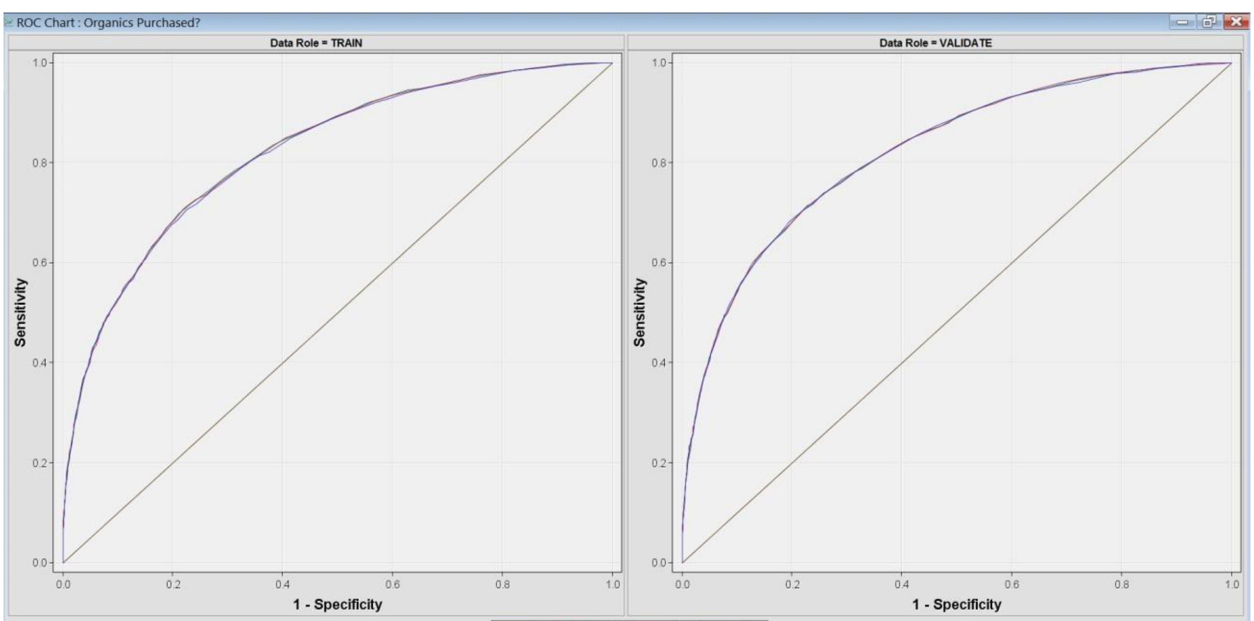
Several models were tested to achieve best results.

Neural1 - Model Selection Criteria is chosen as 'Misclassification' keeping rest of the settings unchanged.

Neural2 - Architecture for Network was chosen as 'Ordinary Radial - Equal Width', Direct connection is selected 'Yes' to directly connect inputs and outputs. In optimization, 'Training Technique' is chosen as Trust Region. Model Selection Criteria is chosen as 'Misclassification'.

Neural3 - Architecture for Network was chosen as 'Ordinary Radial - Unequal Width', Direct connection is selected 'Yes' to directly connect inputs and outputs. In optimization, 'Training Technique' is chosen as Levenberg-Marquardt.

The above mentioned networks were then compared and the results were assessed.



Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Average Squared Error	Train: Maximum Absolute Error
Neural2	ORGYN	Organics...	0.181802	11112	11047	65	65	9425.863	9901.389	0.13303	0.990668
Neural1	ORGYN	Organics...	0.182702	11112	11060	52	52	9395.72	9776.141	0.132973	0.984762
Neural3	ORGYN	Organics...	0.183512	11112	11060	52	52	9459.236	9839.657	0.133663	0.98178

Neural network 2 performs the best by giving the least misclassification rate. This model is chosen to go ahead with and compare with other models.

Model Comparisons and Evaluation

After coming up with the best decision tree, regression model or neural network model, these models are then compared to each other to determine the best model among these three. Following information was obtained after the comparisons:

Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error
Tree1	ORGYN	Organics...	0.181262	11112	0.180526	0.921569	3061.432	0.137753	0.371151	22224	11112	11111	0.181262	0.921569	3065.686	0.137957	0.371426
Neural2	ORGYN	Organics...	0.181802	11112	0.184575	0.990668	2956.463	0.13303	0.364733	22224	11112	11111	0.181802	0.994576	2955.711	0.133008	0.364703
Regressi...	ORGYN	Organics...	0.184052	11112	0.185025	0.98862	2988.962	0.134493	0.366732	22224	11112	11111	0.184052	0.994178	2944.849	0.13252	0.364032

Misclassification rate of decision tree is the least among all three models, the next model close to it is neural networks and regression model gives the largest misclassification rate among all. So, for this analysis regression model should not be considered.

Decision Tree and neural networks both has shown good results. However, implementation of decision tree is much more simpler than neural networks. Decision trees also provides an advantage that its performance does not get affected by missing values. However, this is not the case with Neural networks. Neural networks ignores observations with missing values. Due to this the size of training data gets reduced and this can affect the predictive power of the model. In this report, as the aim is to choose the best fit model for a retail shop, and in real world data set cannot be expected to be without missing values due to any reason, for example, customer refused to give details, certain detail about the customer is not known etcetera. Hence, a decision tree model will be the most suitable for the supermarket to decide which customers are likely to buy the organic product.

Recommendations to the retailer

The retail shop should try to avoid as many missing values as possible. As this data set consisted of a large amount of missing and NA values. There can be multiple ways of achieving this, for example, regular reviewing of data set, contacting the customer for any missing data, etc. This is important as ignoring missing values when preparing the model may not lead to the appropriate results. These missing values can be of importance but since the data is not available it will be either replaced or ignored. Also, the missing values can be manipulated or imputed by the analyst working on the data and this could introduce biasness and yet again not give the correct results.

The retailer can review the data set for duplicate rows or variables which provide the same information. For example, in the given data set, there are two variables TV_REG and REGION with some same values, now this can be confusing for the analyst working on it or the one who is reading the results. So, retailer should try to bring more clarity and quality in the data set.

Based on descriptive statistics and results of decision tree, retailer can target the correct group of customers based on information pointed out below and should try to make sure that the variables pointed are correctly obtained avoiding missing value:

- Most important factors to target the customers of interest are AFFL, DOB, GENDER, AGEGRP2
- Females spent more on organic products than Males
- Customers from affluence grade 7 to 9 should be targeted
- Most customers belonging to the age of 50 to 60 are more likely to spend