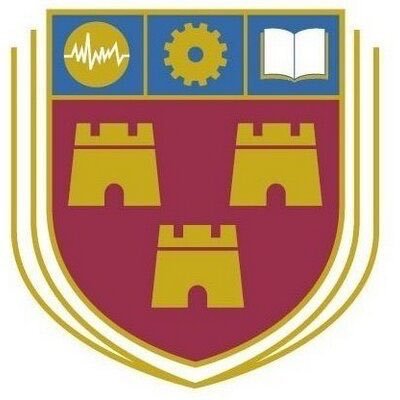
**DATA ANALYSIS ON ORGANICS DATASET USING MACHINE LEARNING ALGORITHMS**

****

**INSTITUTE OF TECHNOLOGY CARLOW**

**By**

**Kavya Naveen Kumar**

**C00223350**

**MSc. Data Science**

**Institute of Technology Carlow**

**Ireland**

**Supervisor – Greg Doyle Submission Date – 28th April 2018**

**CONTENTS**

1. Introduction…………………………………………………………………. 3

Research Questions …………………………………….............. 3

2. Literature Review…………………………………………………………. 3

3. Data Analysis……………………………………………………………… 7

3.1. Data Preparation …………………………………………………. 7

3.2. Handling missing value…………………………………………… 8

3.3. Data partition ……………………………………………………… 9

3.4. Impute missing values ……………………………………………… 9

3.5. Modeling…………………………………………………. ……... 10

3.1. Decision Trees…………………………………………. 11

3.2. Naive Bayes Classifier…………………………………. 15

3.3. Deep Learning…………………………………………... 16

3.4. Generalized linear regression…………………………… 17

3.5. Logistic regression………………………………………… 18

4. Evaluation………………………………………………………………… 19

5. Conclusions………………………………………………………………… 21

References……………………………………………………………………... 21

**1.Introduction**

The primary purpose of this report is to implement the detailed analysis of the organics dataset which helps the supermarket to understand the demand for its recently launched line of organic products. Since the number of supermarkets is growing day by day, the supermarkets are striving to draw and retain increased buyers as their profit depends on them and they do so by providing promotional coupons, loyalty points, etc. to the consumers. The data is collected from the initial days of the launch of products by issuing the vouchers to the customers. Understanding customer behavior and shopping patterns are the keys to do this.

The collected data set for the analysis is about a scenario wherein a supermarket is going to begin a new line of organic products, and the management of supermarket would like to understand, what group of people are most likely to buy these recently launched organic products to improve their sales. The supermarket also provides customer loyalty program to its customers which is an initial buyer incentive plan. In this plan, coupons are provided to the customers with the help of which data is collected whether those customers have purchased the organics products or not.

The primary measure in data analysis focuses on the understanding of the fundamental objective and requirements from a business view, and once clear with the understanding stage; we have to employ this knowledge to implement the data analytics solution to the problem. Numerous models can be developed in Rapid Miner based on the business requirement of supermarket management. These models will help to target the set of customers who are more likely to purchase the organic products.

**Research Questions**

1. Can we Predict Customers' Behavior?

2. How do the purchase orders differ by gender?

3. Does age play a part in purchasing organic products?

**2. Literature Review**

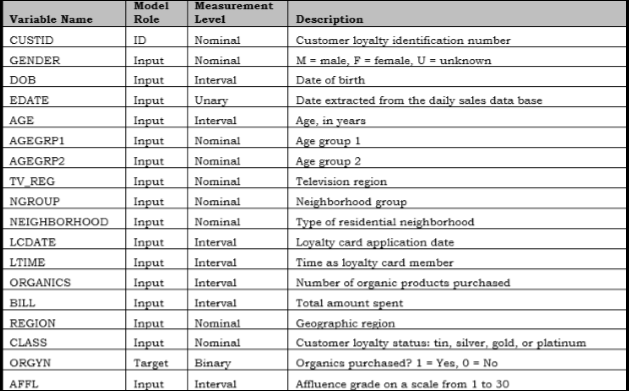
1. **Factor affecting customer behavior:**

To identify the multiple factors influencing consumer behavior, the study looked at various queries. How does a consumer work? What are the urges and purposes that lead him/her? What are the different factors that affect it? What will make him/her prefer one product or brand over another? According to multiple studies on consumer behavior, there are four basic kinds of elements that influence customer conduct; social variables, social elements, unique elements, and emotional elements.

Purchasing decisions and behaviors are apparently also inspired by the lifestyle of each consumer. A customer does not purchase similar items or conveniences at age 20 or 70. Their lifestyle, their values, their environment, their activities, leisure activities and their propensities for utilization advance for the duration of their life. For instance, in the course of their life, a consumer may switch from unbalanced products (fast food, ready meals, canned food, etc.) to a healthier diet. As a family, in the middle of his life before having to follow, a little later, to a low cholesterol diet to avoid health problems. The factors that influence its purchasing decision process can also evolve. For example, the social esteem a brand will, for the most part, assume a more critical component in the choice for a 25-year-old consumer than 65 years old.

**2. Data understanding:**

This stage concentrates on a preparative investigation of the data and processing this data concerning the Business Requirements. The ORGANICS dataset contains 22,223 observations and 18 variables. The variables in the data set are displayed below with the relevant roles and levels.



**Fig. 1: Data variables with a summary.**

Nominal- categorical

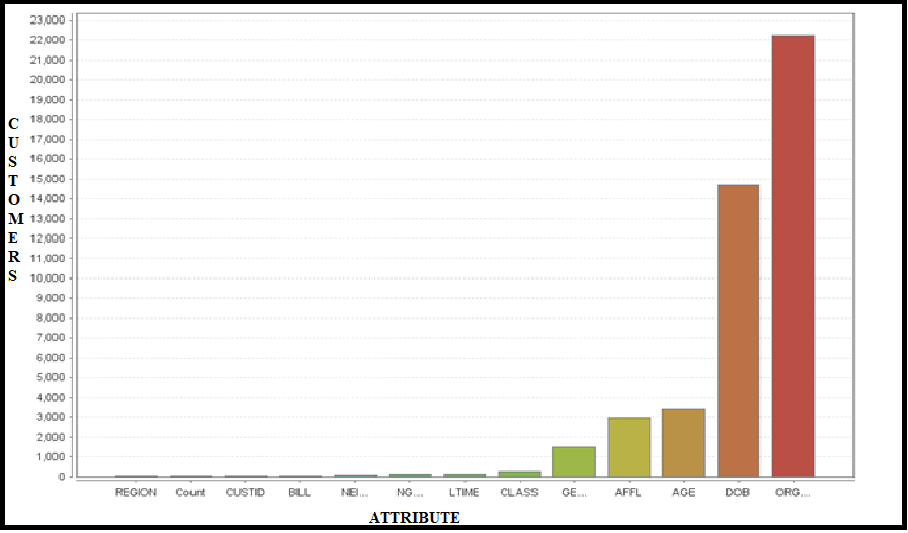
Interval- continuous

Target- data of interest.

As shown in above table, ORGYN is Target variable. It is represented in binary format (1 = Yes and 0= No) to decide if Organics is purchased or not.

These data presume that the majority of respondents are familiar with the main characteristics of organic products and as such, can make informed decisions in filling the response questionnaires with regarding having or having not purchased organic products. This information is encouraging because it validates the efforts made in recent years to understand the populace's acquisition of organic products better.

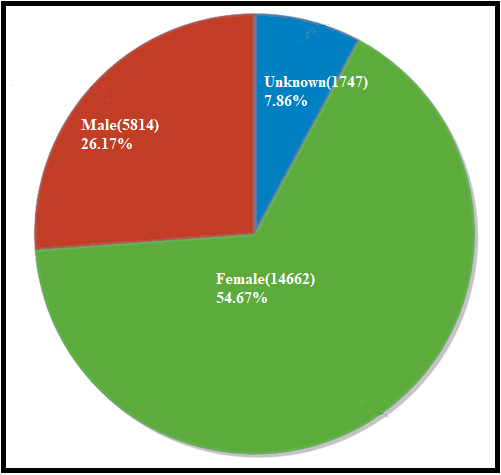
Exploratory data analysis is conducted to understand the dataset. From the figure below, it can be seen that ORGANICS variable has the highest chi-square value telling that the relation between the target variable and itself is very high followed by DOB, AGE, AFFL, and GENDER and so on.



**Figure.2: Chi-square plot**

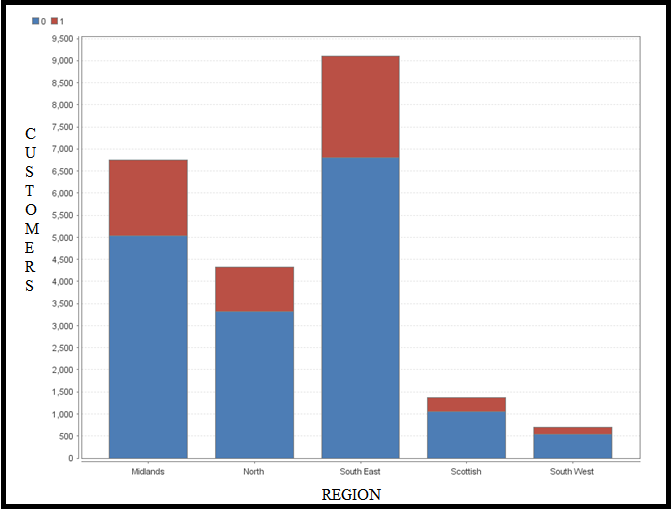
People aged 18 and above responded to the examination on the acquisition of organic products. No reflection was taken whether or not, the respondent had previously consumed organic food, and preferably emphasis is on a purchase of the organics. 54.67% of respondents were female, 26.17% were male, while 7.86% were unknown. The above stats showed in the pie chart in fig.3.

Regarding gender, out of total person interviewed, it can be easily recognized that woman have confessed to having bought organic products more compared to male respondents. Although this could imply a lower acceptance by males, it may also reflect the fact that females are often more concerned with purchase often foods as compared to their male counterparts. Nonetheless, this no doubt depicts female as a prime market target for the organic foodstuffs.

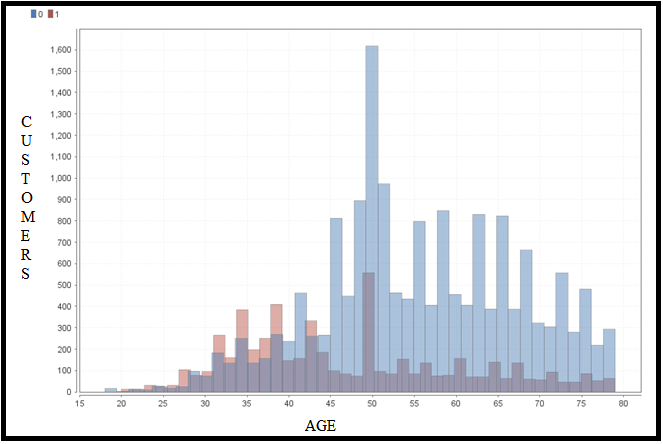


**Figure.3: Pie chart of gender distribution.**

From a regional view, as portrayed in the diagram, the highest number of organic products (ORGYN) were purchased in the South East. Although, this could be attributed to the large population in the area and hence a large market. Overall though, the organic product's update in the market remains below standard constituting only a small part of the population as a majority, in all the regions considered still do not buy organic products.

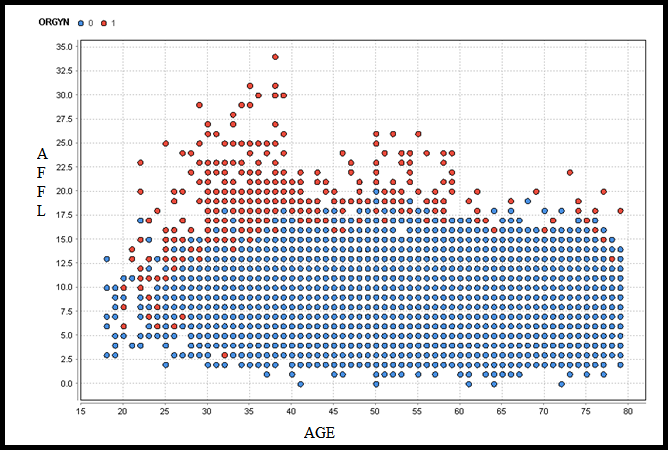


**Fig. 4: Bar chart for regional distribution.**



**Fig. 5: Histogram plot for age distribution.**

Regarding age group, the most significant uptake (Orange blocks, i.e., ORGYN = 1) is seen amongst those aged between 30 and 40. Nonetheless, this is not the exciting aspect of the study. Instead, it is a fact that amongst the respondents aged 20-30 and 30-40 more people reported purchasing organic products as compared to other segments of the population where a majority did not buy the products. This could be attributed to this generation being a well-informed and health conscious generation that understands the importance of consuming organic foods.



**Fig. 6: Scatter plot for age v/s affl value.**

From the above scatter, we can observe that maximum number of the times the organic products were bought by people with affluence level above 12.5 and the age above 30 years.

**3. Data Analytics**

**3.1. Data Preparation:**

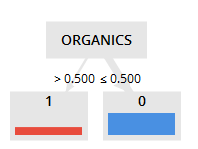
Data preparation stage focuses on retaining and discarding the variables towards the analysis. In particular, to perform the required activities to set up the dataset like variable selection, role selection, measurement level selection depending on the business requirement. Sometimes, it may also need additional data cleaning and transformation actions.

The variables AGE, AGEGRP1 and AGEGRP2 represents age details and therefore use of only one variable AGE is sufficient thus we have to reject AGEGRP1 and AGEGRP2 variables. Similarly, LCDATE and LTIME variables represent the nearly similar bit of information. As a result, LCDATE variable is rejected. On the other hand, EDATE is extracted date and is holding a single value which is of no use for doing additional analysis. As a result, EDATE is also rejected. ORGYN is Target variable and is represented in binary format (1 = Yes and 0= No).

There are few other variables mainly; AFFL is an affluence grade having Interval Level and role as Input with values ranging from 0 to 34. CLASS represents loyalty membership, and it is classified into four groups namely tin, silver, gold, or platinum. REGION represents mainly 6 Geographic regions. NEIGHBORHOOD is having numeric values ranging from 1 to 55, and these neighborhood values are grouped into 6 NRGROUP namely A to F, so it necessary to get rid of any one of the variables. As a consequence, rejecting NEIGHBORHOOD variable. There is another variable called BILL having the total amount spent values on shopping, and 6487 entries out of 22223 have the spending amount of 0.01 which is too low, and chi-square test also confirmed that this variable does not play a significant role in our analysis hence rejecting BILL variable as well. DOB variable is also eliminated from use in some of the models, as the AGE variable gives the information related to it.

The ORGANICS variable explained a highly significant relationship with the target variable, yet when used in different models, resulted in several issues. For instance:

1. In a decision tree, the tree had only one level of depth with misclassification rate equal to zero as shown in fig.7 below.



**Figure.7: Decision Tree is built considering all the variables.**

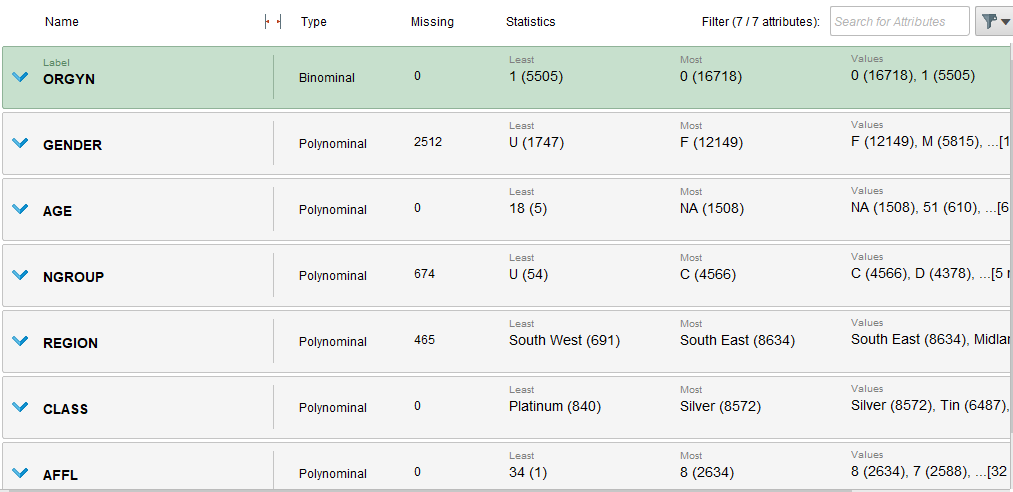
The root node is split by variable having highest logworth value telling the variable worth in predicting the target variable. The logworth value is achieved from the chi-square test which was discussed in Data Understanding section.

The tree split stops after one level because the input data has been split to its purest form which is the goal of the decision tree, i.e., The data belonging to target variable 0 is in the left node whereas with the target variable is in the right node. The misclassification rate is also zero. It has zero false positives and zero false negatives. All the output values seem too good to be true. The model is not considering the other factors which might be contributing towards the organic purchase. That is why other decision tree models are built by rejecting the ORGANICS input variable.

This issue is not just occurring in decision tree model, but also in other models like regression, neural network, naive Bayes, etc., the misclassification rate displays zero if ORGANICS variable is used. Thus, it is eliminated from all models as it is not giving any pertinent information.

**3.2. Handling missing value**

There are missing values in most of the variables, and this is handled by using Replace missing values operator in rapidminer tool. As shown in below fig. 8, GENDER, NGROUP, REGION has missing values, and those are replaced successfully by using Replace missing values operator. Missing values do not build problems while working with Decision Tree. However, Regression and Neural Network model does create a problem by ignoring the observations having missing values. Therefore, it decreases the size of training data set which can influence the predictive power of the model. To overcome this issue, it is necessary to impute the missing values before proceeding.



**Fig.8: Missing data.**

The count is selected as default input method for class variables whereas, the average is selected for interval variables.

**3.3. Data partition**

Once done with replacing missing values we have to move forward to distribute the dataset in Training and Validation part. When splitting dataset, we have to be sure that validation data set is large enough. If this is not the case, then it might lead to erroneous results while evaluating the decision tree. Therefore, data has been divided in 80-20% for performing training and validation on the decision tree.

**3.4. Impute missing values**

The impute operation is performed to handle missing values in the dataset. Missing values do not create problems while working with Decision Tree. Regression and Neural Network model ignore the observations having missing values. Therefore, it reduces the size of training data set which can impact the predictive power of the model. To overcome this issue, it is necessary to impute the missing values before proceeding.

**3.5. Modeling**

In machine learning, classification is a supervised learning approach in which the computer program determines from the data input provided to it and then uses this learning to classify new observation. This data set may be bi-class, or it may be multi-class too. While in this case, it is ORGYN variable which is a binary class with 0 and 1’s (0= did not buy organic products, 1 = bought organic products)

The types of classification algorithms used in this analysis are:

1. Regression models like:

a. Generalized Linear Classifiers.

b. Logistic Regression.

2. Naive Bayes Classifier

3. Decision Trees

4. Neural networks:

a. Deep learning

The following methods are chosen for evaluation of the models:

1. Misclassification Rate using confusion matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. By computing the statistical measures namely the true positives (TP), True Negatives (TN), False Positive (FP), and False Negatives (FN).

Classification error = (Incorrect predictions) / (Number of Examples) = (FP + FN) / (TP + FP + FN + TN)

1. Lift chart: A lift chart shows how much better a machine learning model performs compared with a random guess. It also shows the point at which the predictions become less useful.

The lift chart shows ten bins for test data. Each bin is filled with decreasing confidence of the model for the target class. That means that the examples with highest confidence values are in the first bin, then in the second, and so on. The chart consists of two parts. The bars show the correct percentage of the target class.

The following part of the chart is a line which shows the cumulative coverage of the target class if one would consider only samples of at least the confidence of the corresponding bar or higher. A value of 60% at the third bar, for example, means that you covered 60% of the desired target class at that point. However, the third bar only represents 30% of your total population. That means that this model would correctly identify 60% of the target with only using 30% of the total population (the 30% with the highest confidence for this class). In contrast to this, a random model would only achieve 30% of the target class.

1. ROC Comparisons: shows the ROC curves for all models, together on one chart. The closer a curve is to the top left corner, the better the model is.

**3.5.1. Decision tree:**

A decision tree is a tree-like arrangement of nodes assigned to create a decision on values connection to a class or an estimate of absolute target value which is ORGYN in this case. Each node represents a splitting rule for one specific Attribute which is decided by different criteria. For classification, this rule separates values belonging to different classes.

The structure of new nodes repeats until the stopping criteria are not achieved. The prediction for the class label Attribute is determined depending on the majority of Examples which reached this leaf during generation. The Target Attribute must be nominal for classification.

After the decision tree is generated, the model can be applied to the testing dataset by using Apply model operator. Each observation follows the branches of the tree according to the splitting condition until a leaf node is reached.

While building a Decision tree, few parameters needs to be changed to perform the analysis. The parameter setting for the splitting rule, leaf node. The value of the maximum depth of tree should be changed to 7. This allows rapid miner to train tree up to 7 steps of splits. Usually, values from 6 to 9 help to build more complex tree. Leaf size is changed to 8. Therefore, it requires a minimum of eight training observations for performing the exploration.

There are various criteria upon which the splitting is decided like information gain, gain ratio, Gini index, etc.

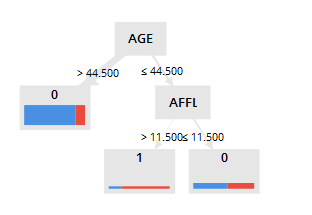
**3.5.1.1. Decision tree using information gain criteria:**

In this case, the entropies of all the Attributes are calculated, and the one with least entropy is chosen as a criterion for the split. Entropy is the measure of the lack of predictability. The entropy is highest when uncertainty is maximum. This method has a bias towards selecting Attributes with a large number of values.

The maximal depth of the tree is set to ‘7’, criterion is set as ‘information gain’, minimal leaf size is set to ‘10’ as shown below:

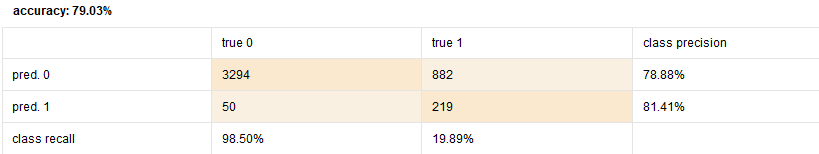


In this case, first split is done by AGE as it is having least entropy value followed by AFFL and GENDER as shown in figure. below:



**Figure. 9: Decision tree with information gain criteria.**

After Split 1, Split points are changed to AFFL to make it more understandable and further splitting is done on the basis of AFFL and so on based on interval splitting rule. The last level of splitting is done on the basis of AFFL (i.e Affluence grade). The resulting Interactive Decision Tree has 5 nodes with 3 leaf nodes

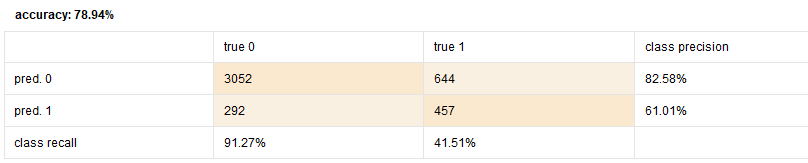


**Figure. 10: Confusion matrix for decision tree with information gain criteria.**

The accuracy of the decision tree with information gain is shown in above confusion matrix. Here, the split stops at second split itself with the misclassification rate of 0.21.

**3.5.1.2 Decision tree using gain ratio criteria:**

|  |  |
| --- | --- |
| |  | | --- | | Gain ratio is a variant of information gain that adjusts the information gain for each Attribute to allow the breadth and uniformity of the Attribute values. Entropy is the measure of the lack of predictability. The entropy is highest when uncertainty is highest. This method has a bias towards selecting Attributes with a large number of values.  The maximal depth of the tree is set as ‘7’, criterion is set gain ratio, minimal leaf size is set to 10 and pruning is applied to the tree as shown below:    In this case, first split is done by AFFL as it is having highest gain ratio value followed by AGE and GENDER as shown in the fig. 11 below:    **Figure.11: Decision tree with gain ratio criteria.**  After Split 2, Split points are changed to AFFL to make it more understandable and further splitting is done on the basis of AGE and so on based on interval splitting rule. The last level of splitting is done on the basis of AGE on left hand side of the tree. The resulting Decision Tree has 13 nodes with 7 leaf nodes | |
|  |



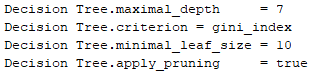
**Figure. 12: Confusion matrix for decision tree with information gain criteria.**

The misclassification rate can be seen from the above fig.12 for the pruned tree are similar to decision tree with information gain without pruning. However, the prune tree splits to its maximum and come back to the point where it found optimum split. Here, the split stops at leaf 5 thus giving a better model with less complexity having misclassification rate of 0.22.

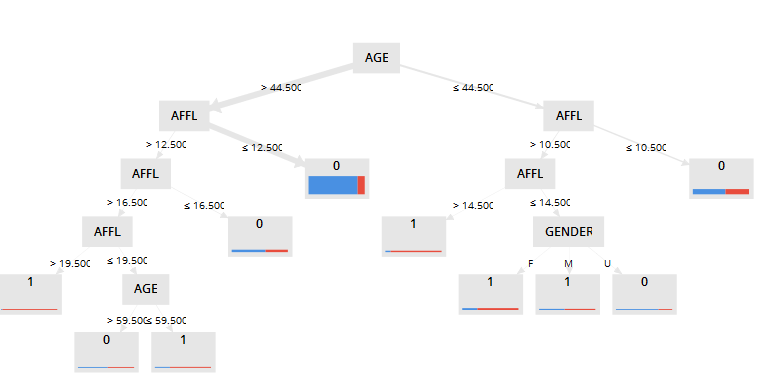
**3.5.1.3. Decision tree using Gini index criteria:**

A measure of inequality between the distributions of label characteristics. The Gini Index can be measured by deducting the sum of the squared probabilities of one class from another. The variable having the highest gini index will be selected for the first split.

The maximal depth of the tree is set as ‘7’, criterion is set Gini index, minimal leaf size is set as 10 and pruning is applied to the tree as shown below:

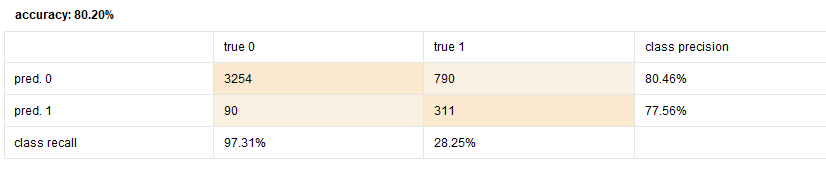


In this case, first split is done by AFFL as it is having highest Gini index value followed by AGE and GENDER as shown in the figure.13 below:



**Figure.13: Decision tree with Gini index criteria.**

After Split 1, Split points are changed to AGE to make it more understandable and further splitting is done on the basis of AFFL and so on based on interval splitting rule. The last level of splitting is done on the basis of AGE on left hand side of the tree and on the right hand-side it is based on GENDER variable. The resulting Decision Tree has 18 nodes with 10 leaf nodes. The pruned tree is built to get the optimum size of the tree.



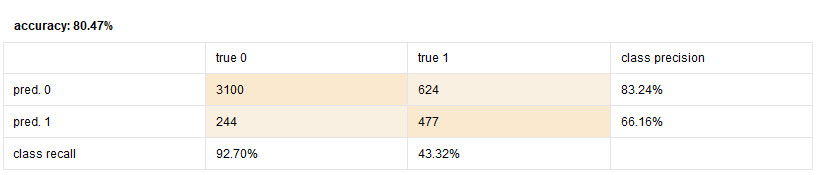
**Figure.14: Confusion matrix for decision tree with Gini index criteria.**

The misclassification from the above figure.14 for the pruned tree are similar to decision tree with information gain without pruning. However, the prune tree splits to its maximum and come back to the point where it found optimum split. Here, the split stops at leaf 8 thus giving a better model with less complexity having misclassification rate of 0.198. Since, this decision tree is having the least misclassification rate amongst all the 3 decision tree, this is the better model.

**3.5.2. Naive Bayes classifier:**

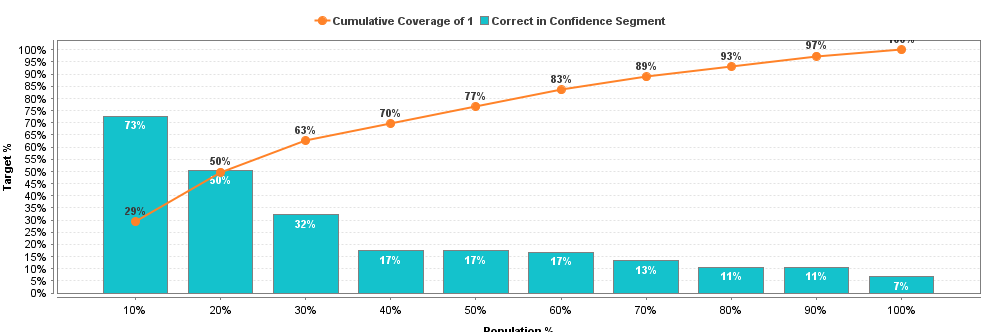
Naive Bayes is a high-bias, low-variance classifier, and it can build a good model even with a small data set. It is simple to use and computationally inexpensive. The fundamental assumption of Naive Bayes is that, given the value of the target (the class), the value of an Attribute is independent of the value of any other Attribute. Actually, this assumption is rarely true (it's "naive"!), but experience shows that the Naive Bayes classifier often works well. The independence assumption vastly simplifies the calculations needed to build the Naive Bayes probability model.

To complete the probability model, it is necessary to make some assumption about the conditional probability distributions for the specific Attributes, given the class. This Operator uses Gaussian probability densities to model the Attribute data.



**Figure.15: Confusion matrix for Naive Bayes classifier.**

The classification matrix on the test data (20%) above show that 244 observations were wrongly classified as target 0 even though the organics was purchased (target 1), which is a Type II error. The 624 observations were classified as organics purchased even though it was not purchased. The misclassification rate for this model is 0.195.



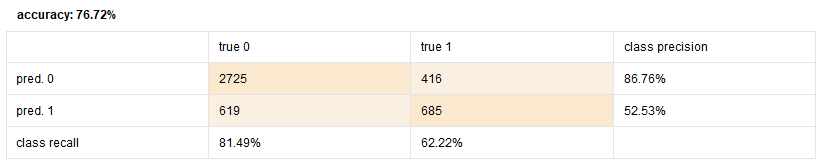
**Figure. 16: Lift chart for Naive Bayes classifier.**

A value of 63% at the third bar means that covered 63% of the desired target class at that point with only using 30% of the total population as shown in fig16, above. In contrast to this, a random model would only achieve 30% of the target class.

**3.5.3. Deep learning:**

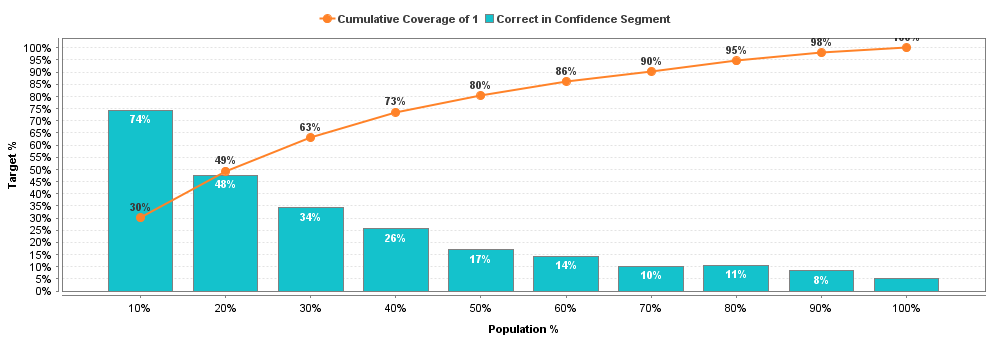
Deep Learning is another variant of neural network based on a multi-layer feed-forward artificial neural network that is trained using back-propagation. The network can contain a large number of hidden layers consisting of neurons with tanh, rectifier and maxout activation functions. Advanced features such as adaptive learning rate, rate annealing, momentum training, enable high predictive accuracy. Each compute node trains a copy of the global model parameters on its local data with multi-threading and contributes periodically to the global model via model averaging across the network.

The operator starts a 1-node cluster and runs the algorithm on it. Although it uses one node, the execution is parallel.. By default it uses the recommended number of threads for the system.



**Figure.17: Confusion matrix for Deep learning neural network.**

The confusion matrix on the test data (20%) above show that 619 observations were wrongly classified as target 0 even though the organics was purchased (target 1), which is a Type II error. The 412 observations were classified as organics purchased even though it was not purchased. The misclassification rate for this model is 0.232 as shown in fig.17.

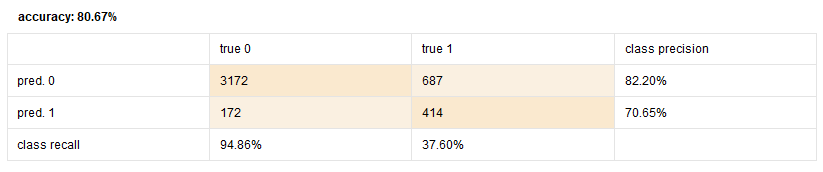


**Figure. 18: Lift chart for Deep learning neural network.**

A value of 63% at the third bar means that covered 63% of the desired target class at that point with only using 30% of the total population as shown in fig.18. In contrast to this, a random model would only achieve 30% of the target class.

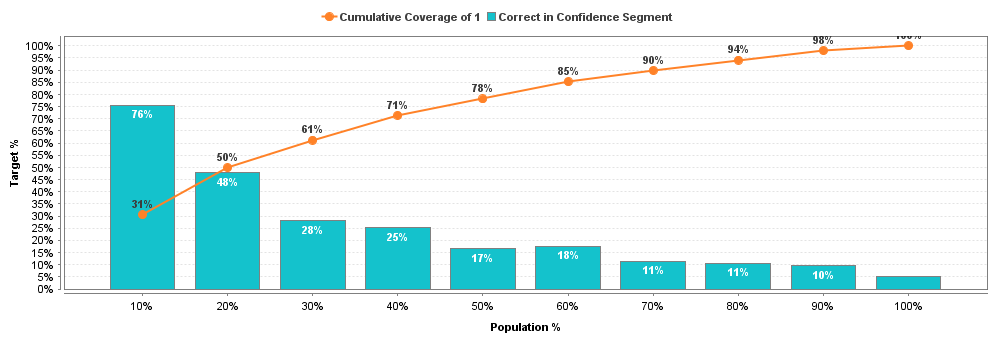
**3.5.4. Generalized linear regression:**

Generalized linear models (GLMs) are an extension of traditional linear models. This algorithm provides generalized linear models to the data by maximizing the log-likelihood. The flexible net penalty can be used for parameter regularization. The model fitting computation is parallel, very fast, and scales much well for models with a limited number of predictors with non-zero coefficients. The operator starts a 1-node cluster and runs the algorithm on it. Although it uses one node, the execution is parallel.



**Figure.19: Confusion matrix for generalized linear regression.**

The classification matrix on the test data (20%) above show that 172 observations were wrongly classified as target 0 even though the organics was purchased (target 1), which is a Type II error. The 687 observations were classified as organics purchased even though it was not purchased. The misclassification rate for this model is 0.193 as shown in fig.19.

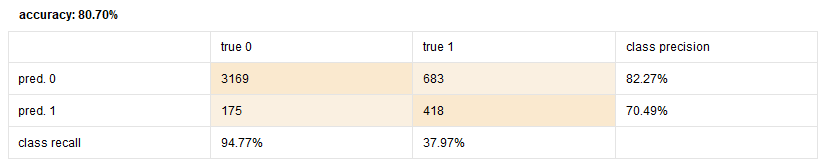


**Figure. 20: Lift chart for Deep learning neural network.**

A value of 61% at the third bar means that covered 61% of the desired target class at that point with only using 30% of the total population as shown in fig.20. In contrast to this, a random model would only achieve 30% of the target class.

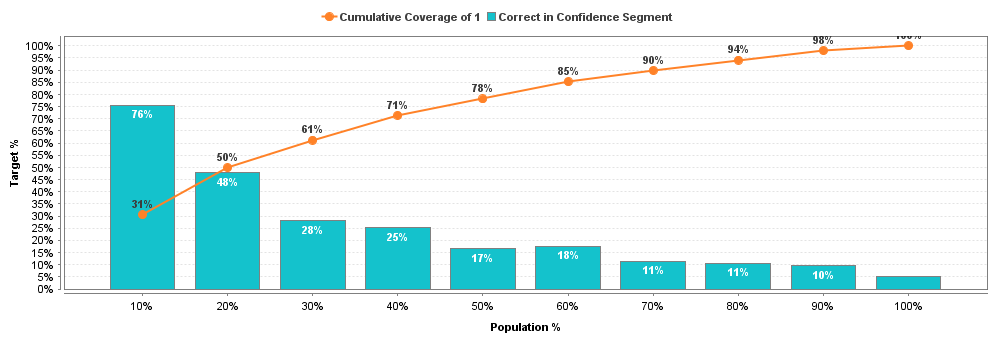
**3.5.5. Logistic regression:**

This operator is a simplified version of the Generalized Linear Model operator. It is a statistical process for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two probable outcomes). The purpose of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.



**Figure.21: Confusion matrix for Logistic regression.**

The classification matrix on the test data (20%) above show that 175 observations were wrongly classified as target 0 even though the organics was purchased (target 1), which is a Type II error. The 683 observations were classified as organics purchased even though it was not purchased. The misclassification rate for this model is 0.193 as shown in fig.21.



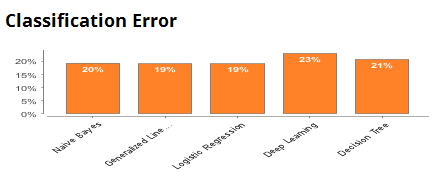
**Figure. 22: Lift chart for Logistic regression.**

A value of 61% at the third bar means that covered 61% of the desired target class at that point with only using 30% of the total population as shown in fig.22. In contrast to this, a random model would only achieve 30% of the target class.

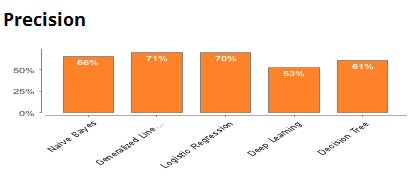
**4. Evaluation**

In the Modelling section, various models are reviewed and built considering different variables and properties. The evaluation measure is a decision for all the models, and therefore the selection criterion is Misclassification rate calculated on the validation data, which will be looked for to determine the best fit model. There are several decision trees created. The decision tree with information gain criteria, a tree with gain ratio criteria, tree with Gini index, deep learning neural network and Regression nodes are included in the final comparison model. Making minor changes in the properties makes difference in the performance measure of the models and this can be seen from all the models.

To find out the best fit model out of the 5 models built, Model comparison node is done based on misclassification rate.

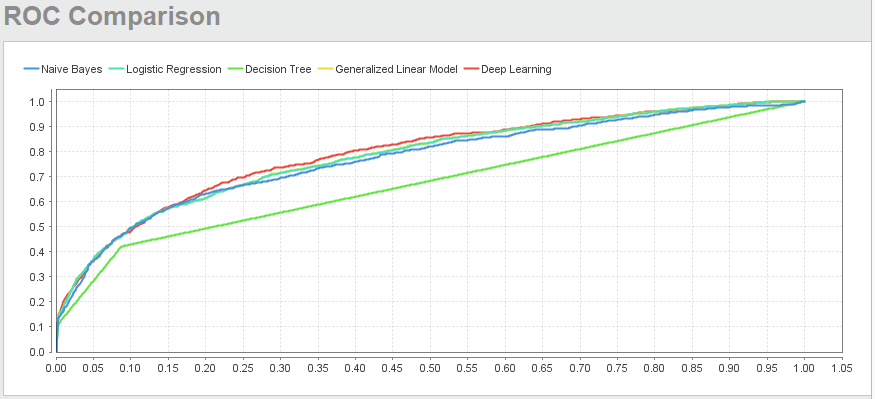


**Figure 23. Model comparison on basis of misclassification rate and precision.**



**Figure 24. Model comparison on basis of misclassification rate and precision.**

From the above figure 23 and 24, it is seen that logistic regression model is the best fit model for the input dataset as it has the least misclassification rate of 0.193. The Generalized Regression model is the second best with the 0.19 misclassification rate.



**Figure 31. ROC Curve**

The ROC curve for the Logistic regression covers the largest area under the curve, thus having more accurate classifier than other models.

**Deployment and Recommendations:**

The last step of Data Mining is Deployment, once the deployment is done and ready for use. The Supermarket management can perform analysis and create reports based on business requirements and can take effective business decisions by looking at trends based on different parameters and variables.

According to the Statistics, there is sufficient evidence to conclude that the prime market for organic products is majorly dependent on Affluence Grade, Age and Gender. Additionally, evidence suggests that some regions offer a more viable market than others, and as such emphasize the need to invest in more ads, TV ads inclusive in areas where uptake remains relatively low. Overall though, the uptake is positive and encouraging. To analyses and take business-related decisions based on data set, Logistic regression is best fit for this case. This model can be further improved by collecting organics related data because Logistic regression works better with large data sets. In this way, Supermarket Management should proceed with implementing Logistic Regression Model.

The Regression model can be used to target the customers likely to buy organic products as the model fits the best with necessary information which is readily available and achievable. The integration of the model to the existing one should also be taken into consideration while deploying it. After broad business understanding, automation of the model built can be carried on.

**Strengths:**

1. Since it is proved that logistic regression has the highest accuracy from the above analysis, this model can be implemented in real-life scenarios.
2. The Regression model can be used to target the customers likely to buy organic products as the model fits the best with necessary information which is readily available and achievable.
3. The integration of the model to the existing one should also be taken into consideration while deploying it. After broad business understanding, automation of the model built can be carried on which is an added advantage.

**Limitations:**

1. The obtained best model is providing the accuracy of 81%, which can be improvised using advanced feature engineering techniques.

**5. Conclusion**

The organics dataset study is conducted and based on the experiments performed, it is concluded that the Logistic regression is the best model followed by the generalized linear regression and naïve bayes model. It has the highest prediction accuracy. The assessment of the models is done on the basis of Misclassification Rate and ROC Plot.

**References:**

Building a Decision Tree : SAS/STAT(R) 14.1 User’s Guide: High-Performance Procedures. (n.d.). Retrieved November 26, 2017, from http://support.sas.com/documentation/cdl/en/stathpug/68163/HTML/default/viewer.htm#stathpu g\_hpsplit\_details01.htm

Ellison, B., Duff, B. R. L., Wang, Z., & White, T. B. (2016). Putting the organic label in context: Examining the interactions between the organic label, product type, and retail outlet. Food Quality and Preference, 49(Supplement C), 140–150. https://doi.org/10.1016/j.foodqual.2015.11.013

https://docs.rapidminer.com/latest/studio/how-to/

Organic Food Trends | Agricultural Marketing Resource Center. (n.d.). Retrieved November 26, 2017, from https://www.agmrc.org/markets-industries/food/organicfood-trends/

Organic Market Analysis | OTA. (n.d.). Retrieved November 26, 2017, from https://www.ota.com/resources/market-analysis