International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN 2964-3139 e-ISSN (on Process)
Vol. 1, No. 1, page 20 - 34
Submitted 30/12/2022; Accepted 01/01/2023; Published 14/01/2023

Customer Value and Data Mining in Segmentation Analysis

[1*]Ahmed Gunandi, [2]Heba Awang, [3]Eman Alhawad, and [4]Lotfy Shabaan
[1,2,3,4] School of Computing & Informatics, Universiti Teknologi Bruner, Brunei Darussalam
e-mail : [1]gunandi548@gmail.com, [2]heba.barbie12@gmail.com, [3]emanhawaddi@hotmail.com, [4]shabahfyi@hotmail.com

Corresponding Autor: gunandi548@gmail.com

Abstract
Customer Value is the accessed value that a customer has to an organization. In Business, the customer is always right. This statement gives us the impression that all customers are viewed as equal in terms of potential value. Each customer is treated differently according to how much profit they can bring to a company. We use various Data Mining techniques to determine who are these customers and how we can acquire more customers like them who can bring more profit. A loyal customer will be treated differently than a customer that rarely do business with the organization. These customers are usually given bonus gifts and special offers as a form of thanks for their loyalty thus further strengthening that bond. Companies need a way to determine which of their hundreds of thousands of customers are deserving of this attention. Customer Value Segments are used in this specific situation.

Keywords— Data Mining, Customer Retention, Decision Tree, Segmentation, Regression, Predictive Model, Customer Value.

1. Introduction
The dataset that we have chosen is the SEGMENTATION dataset. It contains 10,000 lines with 26 different variables. We will call the company represented in this dataset as Teguh Inc. From my previous assignment, we have found that this company focuses on offering its customers consulting services as well as offering training program services to other companies to train their employee's. The dataset contains information regarding Teguh's client base. The company wants to make more profit on the sales of their products. The idea is to target specific segments from their customer base and focus on the ones that are more likely to purchase their product. They sent out test mailing to about 10,000 of its customers picked randomly.

Customers which used the coupon code included in the mailing with their purchase will be classified as a responder to this campaign. At the end, the company will better understand which segment of their customer will be more likely to buy from their catalog and others who are less interested. This is known as purchase propensity [1]. All the customers in this dataset have bought at least 60 dollars' worth of product in the last 24 months. The total amount for the purchases made by the customers will be summed up at the end. The ones that has a total that is not zero will be classified as a responder to the test mailing campaign. Teguh has categorized its customers to certain categories such as the demographic, geographic, purchase history and response rate.

2. Business Issues
Teguh Inc. is facing several issues that prompted their marketing team to launch this test mailing program. Their customer base was segmented into three profiles. The first issue was one segment of their customers do not purchase the consulting services at all. They also rarely purchase any of the teaching products offered but when they do, it is mostly from your resellers which will reduce the organizations profit margin. There is also the problem of disloyal customers. These types of customers are those who will not hesitate to purchase similar products by your rivals or purchase more services with other companies aside from Teguh. A large number of

customers were found to not have purchased anything from the company in more than a year. This might lead reduced customer retention if not addressed immediately. Another issue is the potential wastage of company resources in promoting their products to customers who are more likely to not buy any of your product anyway. This resource could have been used to target other customers that will be interested [5].

According to authors [2], which discussed the different types of data mining techniques used in customer segmentation, Value-based segmentation is the most useful type of customer segmentation. The customers are evaluated and ranked in terms to the value he or she gives to a company. They argued that Customer Relationship Management (CRM) objectives can be solved using the right techniques in data mining. They found that we can gain a much more useful understanding of our customers when we group them in certain groups with similar characteristics. Each of this groups will have their own specialized strategy for marketing. Segmenting you customers will also help divide the customers in terms of their demographic, behavior, and their loyalty for example. Author [3] stated that there has been a major increase in the growth of information in every organization today. This amount of data may be too much for the average analyst to use as a tool to increase their companies' competitive advantage. The main takeaway is that Data Mining can be used to bridge the gap and allow for a more efficient way of discovering patterns and relationships. In the case of customer retention. Author [3] argues that a customer that has a potential to defect can be discovered before they completely stop doing business with them. The loss of these customers will inevitably result in the loss of revenue from the company.

The loss is worse if the company is in an industry with multiple rival businesses. This usually means that there will be more customer defections thus increasing marketing costs. Author [4] proclaims that in today's world of intense competition between firms and the increase of complexity, businesses should strive to create innovative ways to fulfil the needs of your customers to help improve customer retention. CRM aims to create customers that will have a long-lasting relationship with the company and maximize profit. Customer Lifetime Value or CLV potential profit you can gain from a customer over their lifetime. One of the applications of CLV is Customer Segmentation. For example, we can cluster our customers into groups that can help decision makers in your company to make recognize segments in the market more efficiently. This can help improve customer retention as better marketing strategies can be developed [6].

3. Methodology

The SEGMENTATION dataset contains all sorts of information on Teguh Inc's customer base. This database was created in order for the creation of a model of the purchase propensity of the customers. This paper aims to help shed some light on how to perform this and predict which customer is more likely to make a purchase from the company's catalog. Various Data Mining techniques such as Clustering and Regressions for example will be explored. The RESPOND variable will be the main analysis target while the others will serve as the input for the model. SAS Enterprise Miner will be used to ease the data mining activities. This software contains useful tools grouped as SEMMA which contains the Sample, Explore, Modify, Model, and Assess tools [7]. The Partition Node will be used to divide the customer data into two and three parts which are the Training, Validation and the Test data sets before we perform Logistic Regression [8]. That will be performed using the Regression Node [8]. We will also use a SAS Code Node to specify which segment we are going to predict going forward. This prediction will be based on the cluster analysis we perform to determine which customer segment should be predicted to solve a specific issue. Other nodes used include the Metadata node and the Cutoff node and the usage of these will be explored further in the paper.

4. Results and Discussion

4.1. Preprocessing Tasks

The first step to training a regression model is to set the role of the model and change its measurement scale if necessary.

Table 1: SEGMENTATION Variables

Variables - Clus

(none) not Equal to

Columns: Label

Name	Use	Report	Role	Level
LOC	Default	No	Input	Nominal
INCOME	Default	No	Input	Interval
MARRIED	Default	No	Input	Binary
DISCBUY	Default	No	Input	Binary
COA6	Default	No	Input	Binary
FICO	Default	No	Input	Interval
SEX	Default	No	Input	Binary
RETURN24	Default	No	Input	Binary
VALUE24	Default	No	Input	Interval
OWNHOME	Default	No	Input	Binary
ORGSRC	Default	No	Input	Nominal
PURCHTOT	Default	No	Input	Interval
AGE	Default	No	Input	Interval
BUY6	Default	No	Input	Nominal
CLIMATE	Default	No	Input	Nominal
C1	No	No	Input	Interval
C2	No	No	Input	Interval
BUY18	No	No	Input	Nominal
BUY12	No	No	Input	Nominal
RESPOND	No	No	Input	Binary
C6	No	No	Input	Interval
C5	No	No	Input	Interval
C7	No	No	Input	Interval
C4	No	No	Input	Interval
C3	No	No	Input	Interval
ID	Yes	No	ID	Nominal

The first thing that was made sure was that the role of the variable ID was given the role "id". Some variables are highly correlated to one another thus we only need to pick the most important one. Similarly, in [9], the variables BUY6, BUY12 and BUY18 are highly correlated thus we need only pick BUY18 as it represents the total amount of purchases in the last 18 months. All the variables from "C1" to "C7" was rejected. Internal Standardization of the cluster node was set to "Range" changes from the default "Standardization" to make SAS EM divide the variables by their range. The number of clusters will be set to 5. We are using K-Means clustering by specifying the number of clusters.

Figure 1 Segment Plot

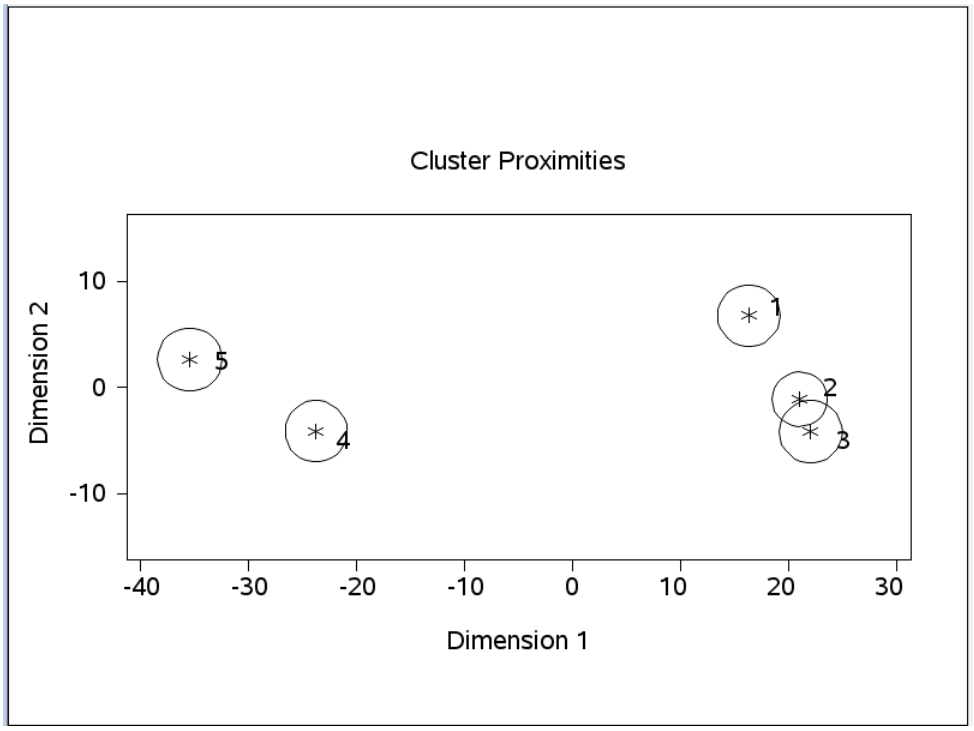

Figure 2 Plot for Cluster Distances

After we run the cluster node, we can view the cluster distance plot to access the clusters further. We can see that Cluster 1 is closer to Clusters 2 and 3. Despite this, there is a big difference in this group as cluster 1 is entirely made up of Females while Clusters 1 and 2 are made up of males. On the other hand, Clusters 5 and 4 are further away from the other 3 while being close to each other. Next, we will look at the ranking of the variables in terms of importance.

Table 2: Variable Importance Table

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
LOC	Location of ...	1	6	1
VALUE24	Total value ...	0	10	0.864479
INCOME	Yr Income i...	0	6	0.846406
OWNHOME	1 if own ho...	6	0	0.772101
PURCHTOT	Test mailin...	0	4	0.742229
CLIMATE	Climate co...	4	0	0.734116
SEX	F or M	1	0	0.649197
AGE	Age in years	0	6	0.605777
FICO	Credit Score	0	4	0.531327
BUY18	# of purcha...	1	3	0.497759
MARRIED	1 if Married,...	2	0	0.455787
RETURN24	1 if product ...	0	1	0.056667
DISCBUY	1 if discoun...	0	1	0.055788
ORGSRC	Original cu...	0	0	0

Based on Table 2, we can see that the variables that holds the most importance is location, total value of purchases made in the last 2 years and Income in thousands which make up the Top 3.

Table 3: Mean Statistics

Mean Statistics

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Age in years	Credit Score	Yr Income in thous.	Test mailing purchase total by product category	Total value of purchases last 24mo	BUY18=0
0.310818	.0002919	.	5	2007	0.328533	2.987464	2	1.451066	43.63532	693.3798	48.07262	7.238166	288.0992	0.710513
0.310818	.0002919	.	4	2368	0.321207	2.876486	3	1.28139	46.69611	695.138	42.39983	7.625	275.8184	0.675676
0.310818	.0002919	.	1	1656	0.299884	2.887831	2	1.481499	45.7673	695.063	56.26101	4.423309	236.942	0.702899
0.310818	.0002919	.	2	1981	0.28669	2.547223	5	1.451066	44.2491	693.5251	49.55151	6.418476	231.5952	0.69258
0.310818	.0002919	.	3	1988	0.313784	2.925051	4	1.28139	42.18303	694.499	46.19126	5.161469	230.3481	0.723843

The cluster with the highest mean value of the variable VALUE24 is cluster 5 as seen in Table 3. This makes it unique compared to the other clusters. We can classify this cluster as the group where we should target more and allow it to grow and further increase this value. It is a general knowledge that customers that has a history of doing business with you will buy from you again thus making this approach a smart one. Before we go about creating a predictive model based on our dataset which is a customer data set and predict prospective customers, we need to make sure that the data of the customer looks similar to the prospect database. This is to make sure that the model will provide us with results that provide fantastic business model. We have decided to predict cluster 5 as it seems to be the best choice. I will create a new Target variable that will represent the response we are going to predict.

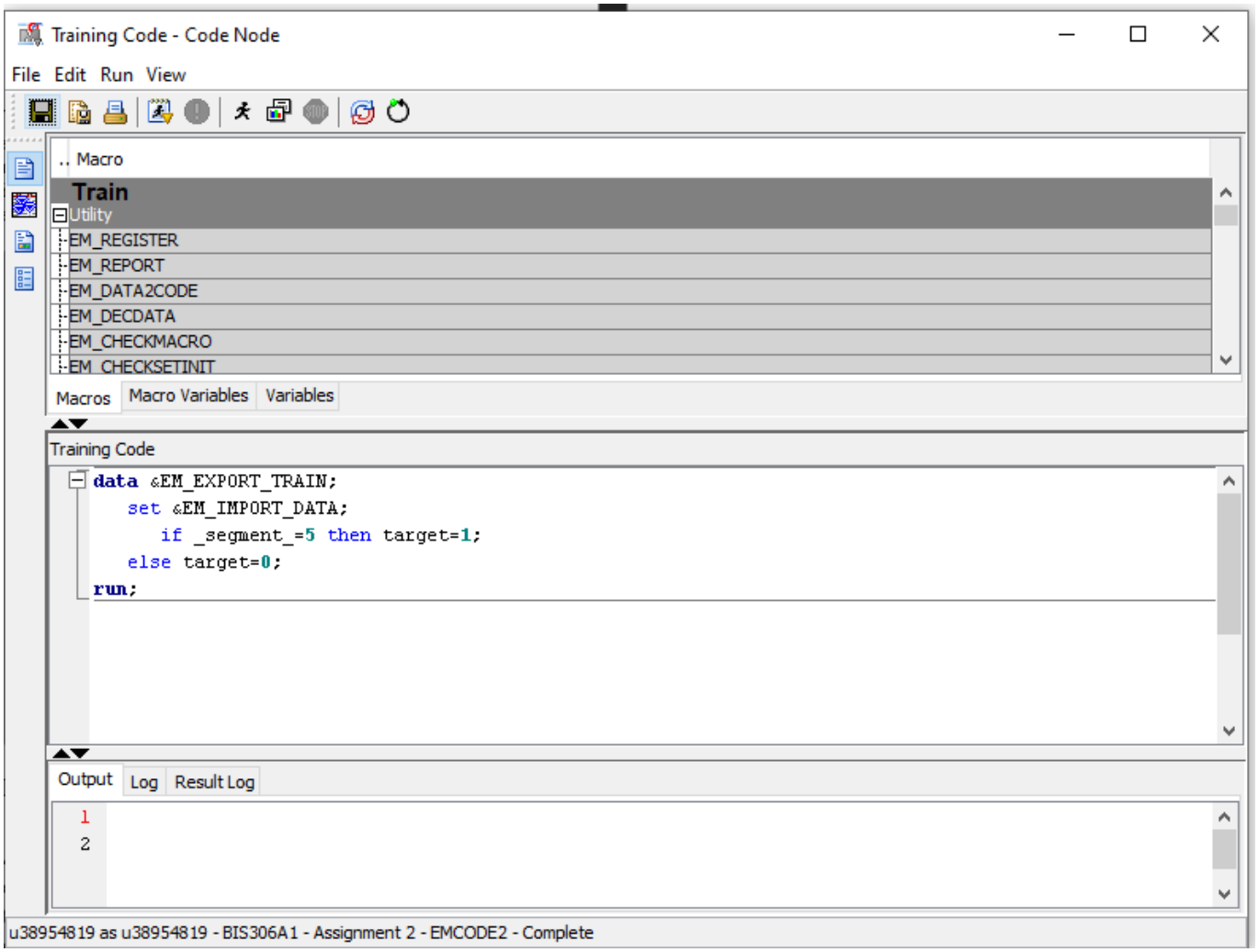

Figure 3 SAS Code Node

We used the macros provided which are &EM_EXPORT_TRAIN and &EM_IMPORT_DATA. The purpose of the first macro is to retrieve the training data we created previously. The second macro will retrieve the results

from the cluster node that we ran. After we run this node, we should have a new Target variable. This variable is automatically set to Input thus we need to change it to Target as we want to predict this variable in the model. To perform this, we use the Metadata node.

Table 4: Metadata Variables Table

Variables - Meta

(none) not Equal to

Columns: Label Mining Basic

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
AGE	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
BUY12	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**
BUY18	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**
BUY6	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**
C1	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
C2	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
C3	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
C4	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
C5	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
C6	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
C7	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
CLIMATE	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**
COA6	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
DISCBUY	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
Distance	N	**Default**	Rejected	**Default**	Interval	**Default**	**Default**	**Default**
FICO	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
ID	N	**Default**	ID	**Default**	Nominal	**Default**	**Default**	**Default**
INCOME	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
LOC	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**
MARRIED	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
ORGSRC	N	**Default**	Input	**Default**	Nominal	**Default**	**Default**	**Default**
OWNHOME	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
PURCHTOT	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
RESPOND	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
RETURN24	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
SEX	N	**Default**	Input	**Default**	Binary	**Default**	**Default**	**Default**
VALUE24	N	**Default**	Input	**Default**	Interval	**Default**	**Default**	**Default**
SEGMENT	N	**Default**	Segment	**Default**	Nominal	**Default**	**Default**	**Default**
_SEGMENT_LAB	N	**Default**	Rejected	**Default**	Nominal	**Default**	**Default**	**Default**
target	N	**Default**	Target	**Target**	Interval	**Binary**	**Default**	**Default**

In the table above, we changed the role of "target" to Target like we mentioned earlier and changed the level to Binary from the original. Next, we will drag a Partition node to the diagram and connect it to Metadata node. This node will split the data into 3 different categories. These are Training, Validation, and the Test data. The training partition will be used to make the model we are predicting and will have its characteristics examined while the Validation data will be used to fine tune the predicted model to better fit. The partitioning method was also changed to Stratified from the default and the variable "target" had its role changed to stratification. What this does is allowing the training, validation and test data will have the same percentage for the variable "target" as the initial data set which helps keep the partitions have the correct percentages.

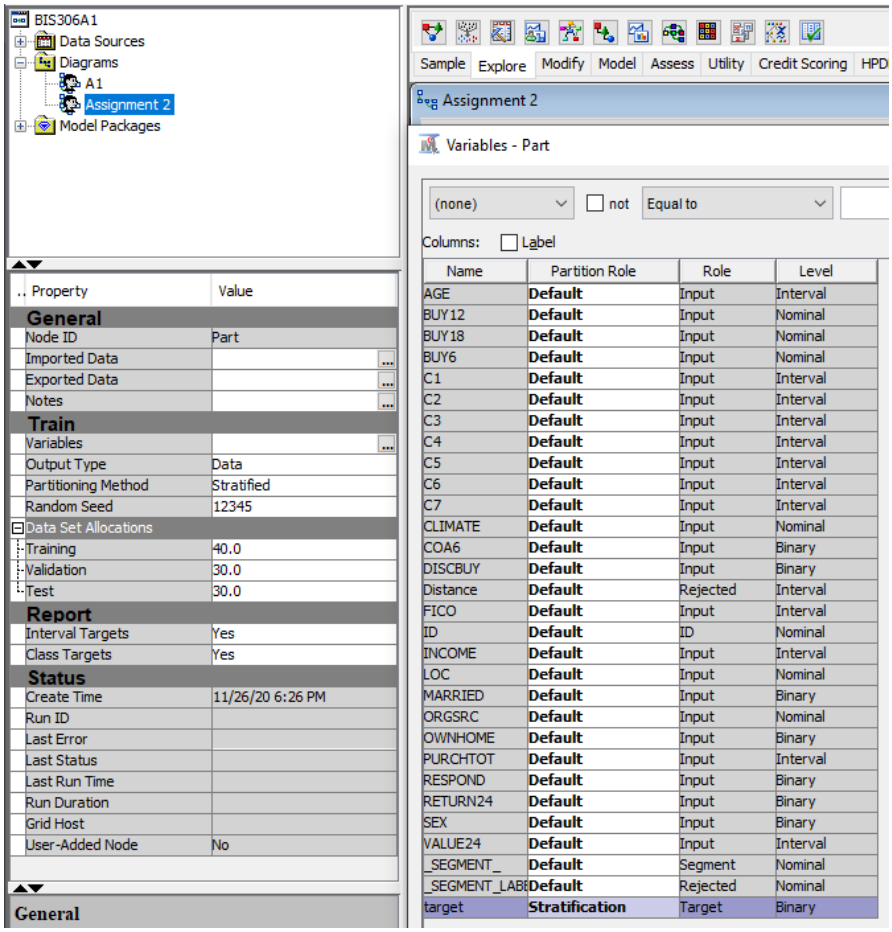

Figure 4 Partition Node Setting

After we have partitioned the data, we can connect it with a Regression Node to perform logistic regression. The diagram now will look something like Figure 5 below.

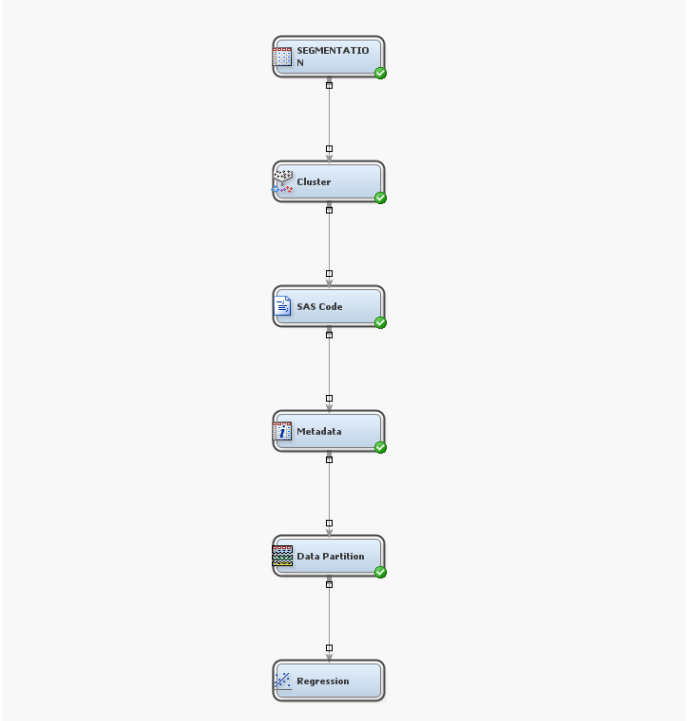

Figure 5 Overall Diagram

Table 5: Regression Node Variables

Variables - Reg

(none) | not | Equal to | Columns: | Label | Mining

Name	Use	Report	Role	Level
INCOME	Default	No	Input	Interval
MARRIED	Default	No	Input	Binary
BUY18	Default	No	Input	Nominal
DISCBUY	Default	No	Input	Binary
VALUE24	Default	No	Input	Interval
OWNHOME	Default	No	Input	Binary
RETURN24	Default	No	Input	Binary
ORGSRC	Default	No	Input	Nominal
PURCHTOT	No	No	Input	Interval
CLIMATE	No	No	Input	Nominal
RESPOND	No	No	Input	Binary
FICO	No	No	Input	Interval
COA6	No	No	Input	Binary
C3	No	No	Input	Interval
C1	No	No	Input	Interval
AGE	No	No	Input	Interval
SEX	No	No	Input	Binary
_SEGMENT_LAB	No	No	Rejected	Nominal
BUY12	No	No	Input	Nominal
BUY6	No	No	Input	Nominal
C6	No	No	Input	Interval
C7	No	No	Input	Interval
Distance	No	No	Rejected	Interval
C5	No	No	Input	Interval
C2	No	No	Input	Interval
C4	No	No	Input	Interval
LOC	No	No	Input	Nominal
target	Yes	No	Target	Binary

The initial variables chosen for the logistic regression is shown in table 5. The variables are selected according to my considerations on which variables would be more important from a business point of view rather than selecting it automatically.

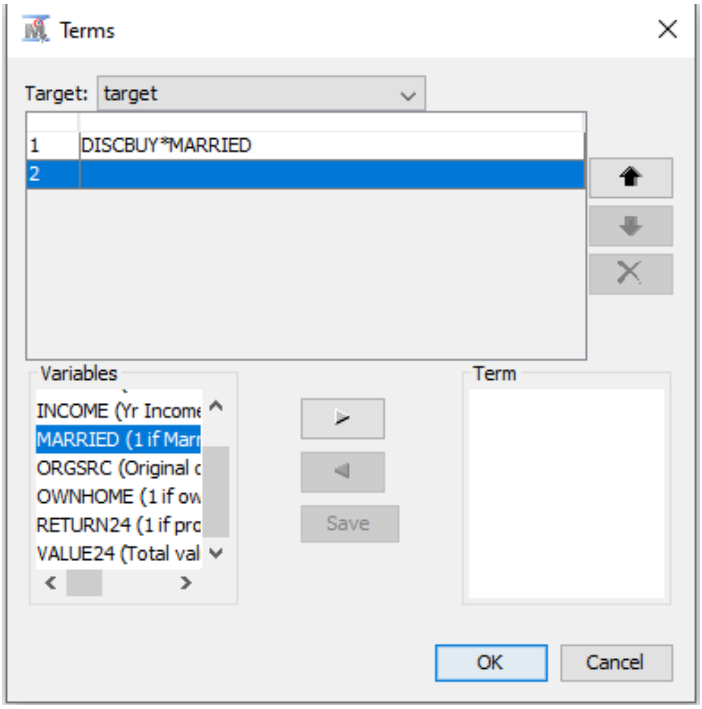

Figure 6 Terms Editor

A term was added to the model with the term's editor. We will check if the interaction between the two variables is significant, if not, we can choose another combination that might work better.

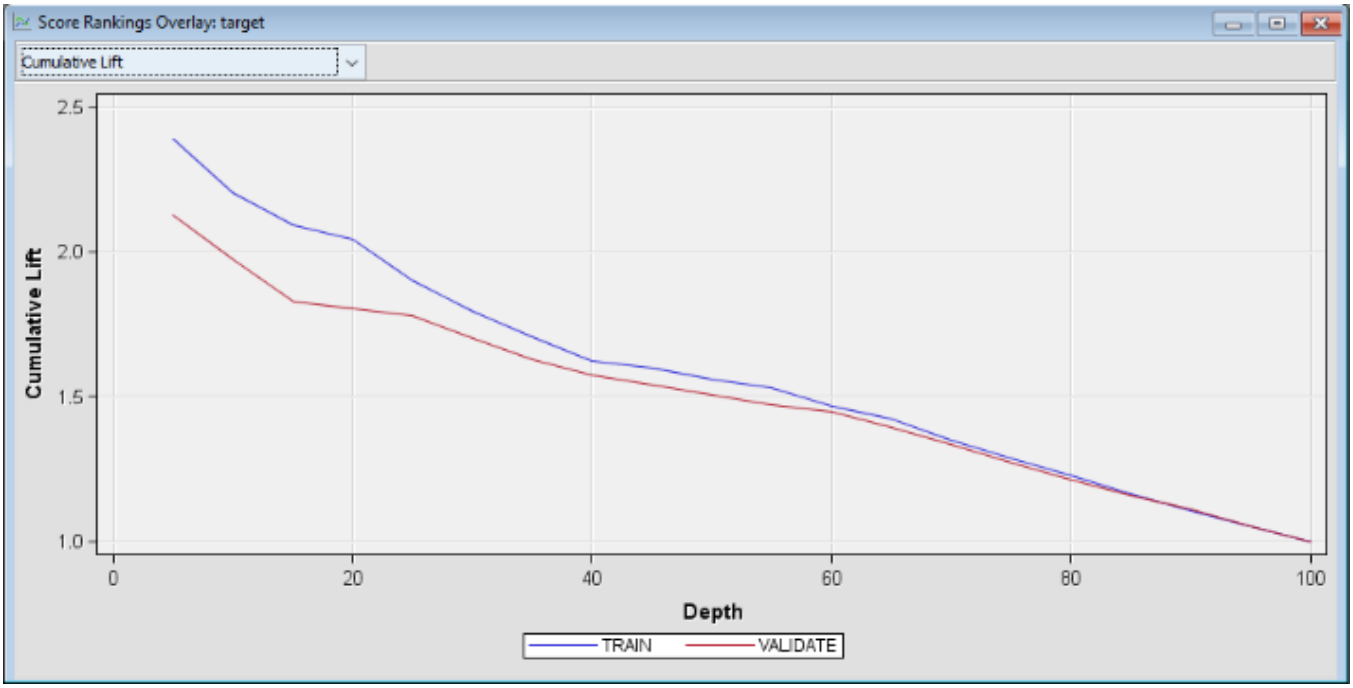

Figure 7 Score Rankings

The figure above represents the score rankings as a lift chart. The training data is represented in blue while the validation data is represented in red. The curves of the two datasets seem to not be similar in the lower depth values. This might be because there was a problem with some values being very correlated with each other.

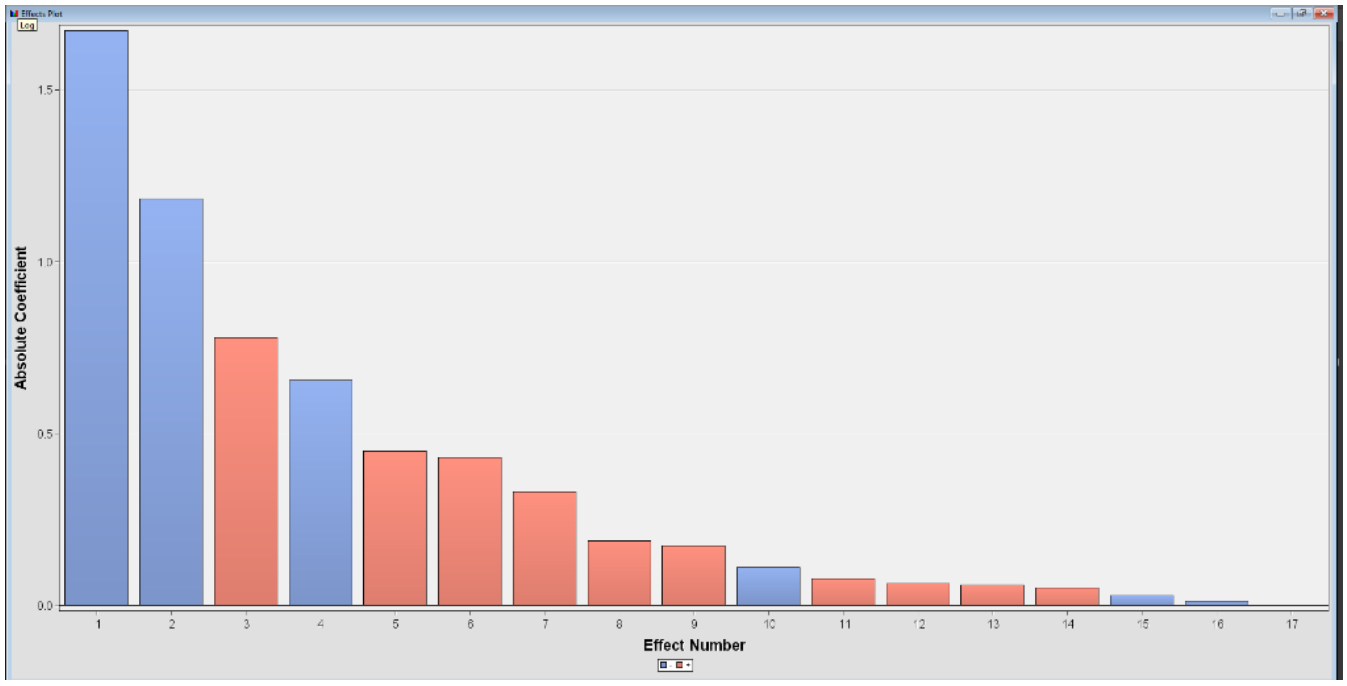

Figure 8 Effects Plot

Figure 8 above tells us how big of an impact each variable has when predicting segment levels.

Table 6: Type 3 Analysis of Effects Table

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DISCBUY*MARRIED	1	1.0569	0.3039
BUY18	3	36.8737	<.0001
DISCBUY	1	0.0000	0.9961
INCOME	1	14.1507	0.0002
MARRIED	1	32.5359	<.0001
ORGSRC	6	12.8664	0.0452
OWNHOME	1	214.5667	<.0001
RETURN24	1	0.4095	0.5222
VALUE24	1	87.0784	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-4.1973	0.3841	119.41	<.0001		0.015
DISCBUY*MARRIED	0 0	1	0.0507	0.0494	1.06	0.3039		1.052
BUY18	0	1	0.9894	0.3016	10.76	0.0010		2.690
BUY18	1	1	0.2382	0.2917	0.67	0.4142		1.269
BUY18	2	1	0.1363	0.3199	0.18	0.6702		1.146
DISCBUY	0	1	0.000240	0.0494	0.00	0.9961		1.000
INCOME		1	0.0104	0.00276	14.15	0.0002	0.0929	1.010
MARRIED	0	1	0.2832	0.0497	32.54	<.0001		1.327
ORGSRC	C	1	-0.2978	0.1306	5.20	0.0226		0.742
ORGSRC	D	1	0.2372	0.1098	4.67	0.0307		1.268
ORGSRC	I	1	0.1363	0.3769	0.13	0.7177		1.146
ORGSRC	O	1	-0.0294	0.1100	0.07	0.7893		0.971
ORGSRC	P	1	-0.1118	0.1224	0.83	0.3611		0.894
ORGSRC	R	1	0.0182	0.1284	0.02	0.8872		1.018
OWNHOME	0	1	1.1358	0.0775	214.57	<.0001		3.114
RETURN24	0	1	-0.0548	0.0857	0.41	0.5222		0.947
VALUE24		1	0.00343	0.000368	87.08	<.0001	0.2939	1.003

The output window can tell us much more than the other sections of the results will. The table above shows us the variables ordered by their statistical significance as regards to our target variable. We are looking for variables with p-values of less than 0.05 which can indicate a significance in its statistical value. The variables that are significant are Income, Married, Ownhome, Orgsrc and Value24. The DISCBUY variable had a p-value of 0.9961 thus we can say that it is not significant. When we allow DDISCBUY to interact with MARRIED, the p-value improves to 0.3039. This might tell us that this interaction might be significant in predicting our target variable, which is Segment 5. We can improve this model by removing some variables deemed insignificant. After removing DISCBUY and RETURN24 and rerun the Regression node, we get the following output.

Table 7: Updated Type 3 Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DISCBUY*MARRIED	1	1.0589	0.3035
BUY18	3	36.8779	<.0001
INCOME	1	14.1507	0.0002
MARRIED	1	32.5674	<.0001
ORGSRC	6	12.8758	0.0451
OWNHOME	1	214.5910	<.0001
RETURN24	1	0.4095	0.5222
VALUE24	1	87.1002	<.0001

Table 8: Odds Ratio Estimate

Odds Ratio Estimates

Effect		Point Estimate
BUY18	0 vs 3	10.520
BUY18	1 vs 3	4.964
BUY18	2 vs 3	4.482
INCOME		1.010
ORGSRC	C vs U	0.708
ORGSRC	D vs U	1.209
ORGSRC	I vs U	1.093
ORGSRC	O vs U	0.926
ORGSRC	P vs U	0.853
ORGSRC	R vs U	0.971
OWNHOME	0 vs 1	9.696
RETURN24	0 vs 1	0.896
VALUE24		1.003

These values there are many variables that have p-values less than 0.05. this indicates they are all highly significant. We can a cutoff node to help us in accessing the model we created. This will be connected to our regression node. From Table 8, we can see the point estimates for each effect that can tell us the likelihood of the outcome. When the value is above 1.0, when there is a unit increase for the effect, the outcome variable will be increased by the number after the decimal. For example, an increase in income will increase the outcome variable by .010.

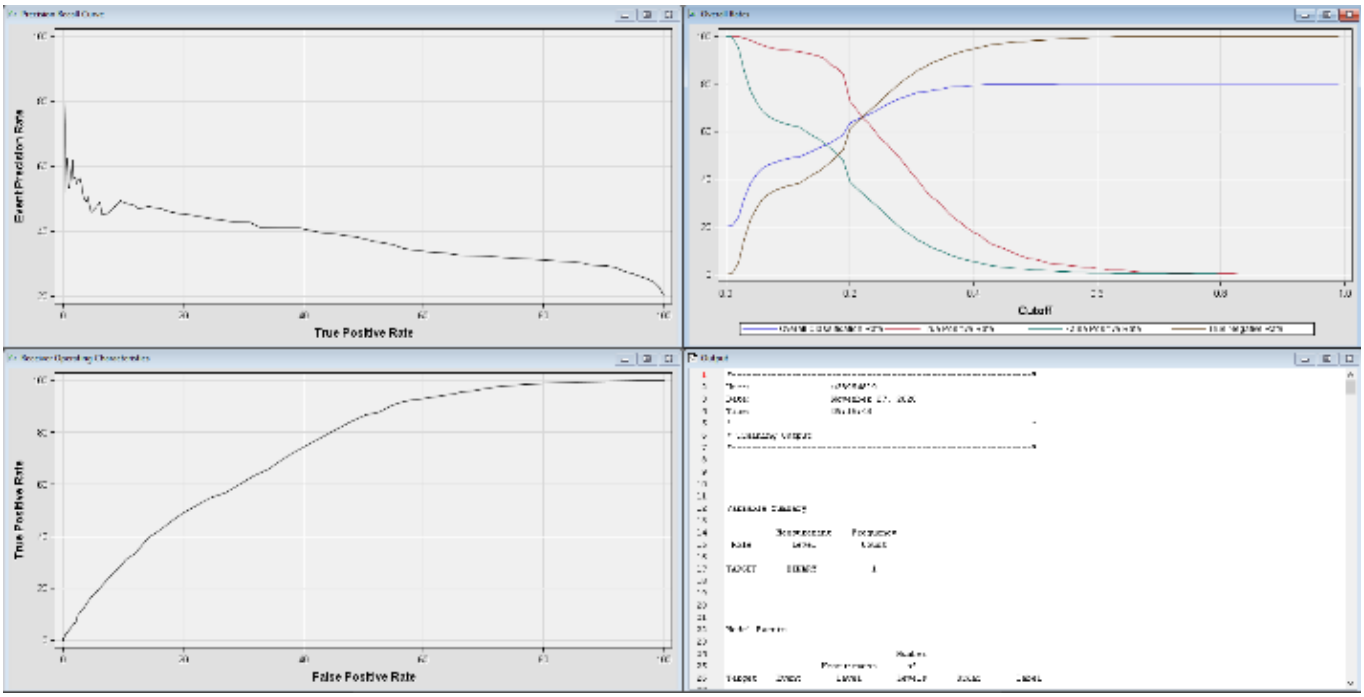

Figure 9 Cutoff Node Output

From the Receiver Operating Characteristics output, we can help find out if the if the true target level captured from our variable for Target is precise versus the false target variable. At the end, we now have a predictive model that predicted cluster 5 that came from the clustering we performed earlier. This model can be used to score data sets from other records. Cluster 5 was chosen as we presume it is the most valuable customer segment. This is because they have the highest average value for the variable VALUE24 which represents the total number of purchases made in the last 2 years. When we are predicting cluster 5, we should find customers that have similar average income as they have a higher likelihood of having similar purchasing habits as the customers in cluster 5. We found from the cluster analysis that cluster 5 is not the segment with the highest average income. The cluster with this characteristic is Cluster 1 according to the Mean Statistics from Table 3.

This shows an opportunity to possibly to grow this customer base and improve its customer value thus predicting Cluster 1 can also be beneficial. Teguh Inc can implement this model with a prospect database which will contain the variables of the original customer data but without the purchasing data as this is a database of non-customers you are trying to target. We can use this model to predict customers for Segment 5 using the variables from the prospect data. The probabilities for the potential customers in segment 5 can be scored using the prospect database variables. Teguh Inc can benefit by predicting customers that will be more likely to buy products from their catalog. This will increase the sales from the catalog and increase revenue. The customers in Segment 5 are the ones more likely to keep purchasing Teguh Inc products thus it is in their best interest to acquire more of these types of customers. They can be targeted better if they use this model to predict the customers in the segment.

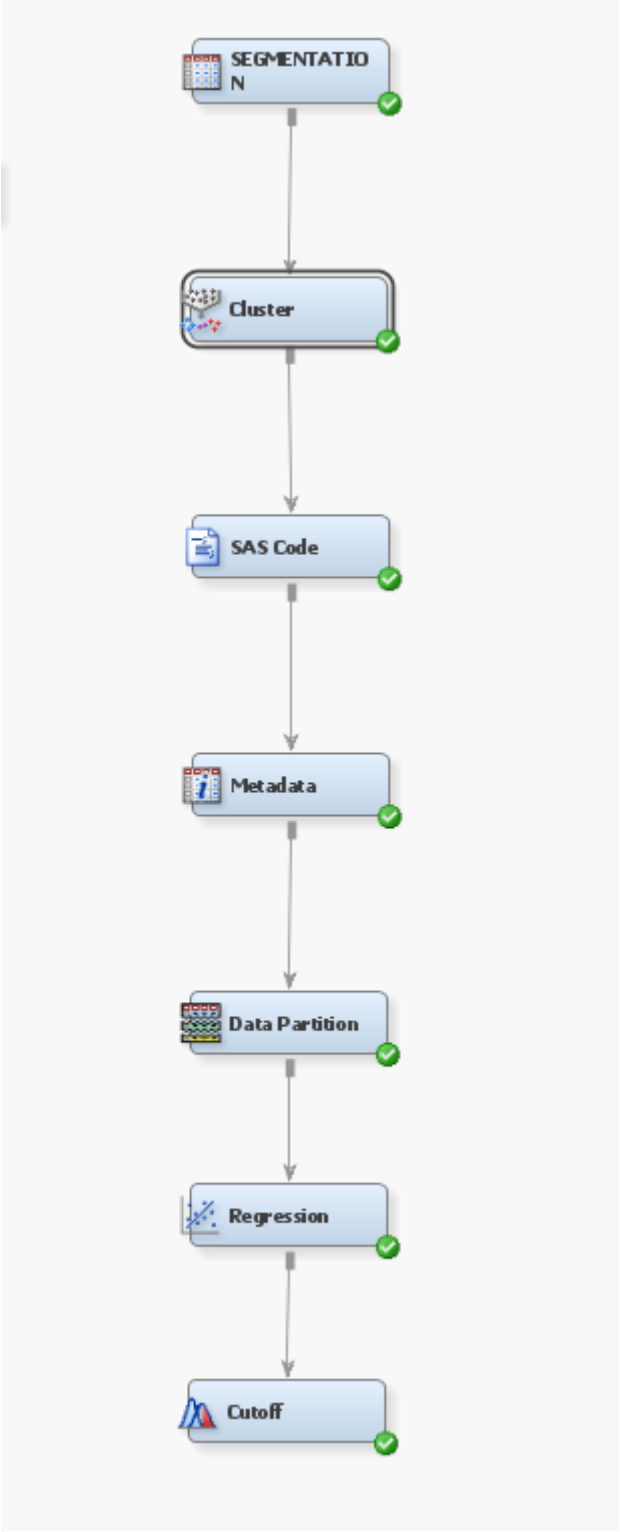

Figure 10 Updated Diagram

4.2. Challenges in Implementation

Teguh Inc might face many obstacles when trying to implement Data Mining to improve their decision making and review the marketing strategies currently in place and decide whether they should change it for the better. The main problems are as follows:

4.2.1. Lack of Quality of Data

An example of this is the presence of "Noisy Data". Noisy data is one of the most common problems that occurs when trying to implement Data Mining in a business context. The customer data that Teguh collects could

be unstructured and have a lot of meaningless data that might not really help in the data mining process. They might make errors during the collection of customer data that may be from human error or by instruments of measuring such as keying in the wrong address for a customer.

4.2.2. Privacy and Protection Concerns

Another problem that might arise is the backlash from customers or the government over the privacy of data Teguh will collect. Data mining might lead to certain issues in data security such as the collection of customer behavioral data. This might be accessed by unauthorized individuals thus posing a security risk.

4.2.3. Lack of Skills in Consolidating Data

The data to be collected exists in many different forms. The customer data can range from in the form or words, numbers or even in videos and images. The right personnel are needed to compile all this data into one customer database. As data is inherently complex, a good data structure can help improve the data mining results thus Teguh can achieve better results. Scalability might also be an issue as when the number of customers a company has increases, so does the size of their database..

4.3. Best Practices

Many companies have adopted data mining in their operations and become very successful. Teguh can follow the steps of these companies and how they are using data mining technologies to make very beneficial business decisions.

4.3.1. Coca Cola

One of Teguh's goal as a company is to improve customer retention. Data mining allows us to look at customer's behaviors and patterns to offer better quality service. Coca Cola does this to improve quality of service and finding out what they want thus reducing the number of customers lost. They built a loyalty program back in 2015 which was digitally focused to improve customer retention and improve their data strategy [10]. They listen to the opinion their customers share with them via emails, phone calls or on social media platforms. This allows them to create advertisements that can relate to different types of people and aligned their brand into each person's characteristic.

4.3.2. Netflix

Netflix used targeted advertising to continue to retain their place at the top of the streaming industry. They have over one hundred million subscribers as of 2018 [10] and they collect data in a large scale so they can accurately target which movies or shows a user is more likely to watch and show them on their suggestions. This create a better overall experience for Netflix users. They managed to overcome the challenge of scalability thanks to the expertise of their team.

4.3.3. UOB Bank

UOB uses data mining technologies for Risk Management handling. As they may face big losses if their risk management is not sufficiently implemented, they pour a lot of effort into ensuring they develop the best system for Risk Management. Data mining has cut down the time to calculate risk from 18 hours manually to only a couple for minutes [10]. Teguh can take inspiration from this to prevent them from incurring huge losses due to an inadequate risk management plan.

4.4. Possible Extensions

In this paper, we only explored on creating a predictive model to only predict one segment level for Cluster 5 from the Cluster Analysis. This was choses as Teguh want to solve issues such as the lack of Customer Loyalty and the unnecessary resources used to target unsuitable customers that will not benefit them. We looked at variables such as the total number of purchases made in the last 2 years to determine which segment is the most valuable. Teguh can also look for segments which have a very high-income average. This will represent the segment with the most potential for growth thus predicting this segment can help with improving the effectiveness of targeted advertisements as targeting these customers is more worth it than some other segments with lower income averages. This could also be predicted using Neural Networks in future works which will have the benefit of us not having to describe how the model will perform or the significance of each variable to the model.

5. Conclusion

In this paper, we have discussed how to solve the problems faced by Teguh Inc. The main problem was how to predict which customers will give more value than others as it is very expensive to choose to target every single one of their customer bases. Some segments are clearly more profitable than others thus different strategies must be performed on these segments to maximize revenue potential. Teguh can identify these Customer Value Segments to give more attention to some segments according to their value. Some segments might contain more disloyal customers or customers that have not purchased from you in a while. Predicting these segments and providing them with the right offers or special treatment can help prevent customer attrition for example [11]. Determining which customers holds the most value can help solve the problems of Teguh Inc. One problem that we faced was determining which combination attributes will provide us with the best results in predicting the segment level. The better the segment level, the better the likelihood of predicting customers with the criteria of the most valuable [12].

BIBLIOGRAPHY

[1] . A. Gramegna and P. Giudici, “Why to buy insurance? an explainable artificial intelligence approach,” Risks, vol. 8, no. 4, p. 137, 2020.
[2] . S. Moradi and F. Mokhatab Rafiei, “A dynamic credit risk assessment model with data mining techniques: Evidence from Iranian banks,” Financial Innovation, vol. 5, no. 1, 2019.
[3] . K. S. Ng and H. Liu, Artificial Intelligence Review, vol. 14, no. 6, pp. 569–590, 2000.
[4] . M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, “Estimating customer lifetime value based on RFM analysis of Customer Purchase Behavior: Case Study,” Procedia Computer Science, vol. 3, pp. 57–63, 2011.
[5] . M. Anshari, M. N. Almunawar, S. A. Lim, and A. Al-Mudimigh, “Customer relationship management and Big Data enabled: Personalization & Customization Of Services,” Applied Computing and Informatics, vol. 15, no. 2, pp. 94–101, 2019.
[6] . H. Baber, “Fintech, crowdfunding and customer retention in Islamic Banks,” Vision: The Journal of Business Perspective, vol. 24, no. 3, pp. 260–268, 2019.
[7] . C. G. Thompson and B. Semma, “An alternative approach to frequentist meta-analysis: A demonstration of bayesian meta-analysis in adolescent development research,” Journal of Adolescence, vol. 82, no. 1, pp. 86–102, 2020.
[8] . E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” Journal of Clinical Epidemiology, vol. 110, pp. 12–22, 2019.
[9] . M. A. Syakur, B. K. Khotimah, E. M. Rochman, and B. D. Satoto, “Integration K-means Clustering method and elbow method for identification of the Best Customer Profile Cluster,” IOP Conference Series: Materials Science and Engineering, vol. 336, p. 012017, 2018.
[10] . M. Rong, D. Gong, and X. Gao, “Feature selection and its use in big data: Challenges, methods, and Trends,” IEEE Access, vol. 7, pp. 19709–19725, 2019.
[11] . E. Ascarza, S. A. Neslin, O. Netzer, Z. Anderson, P. S. Fader, S. Gupta, B. G. Hardie, A. Lemmens, B. Libai, D. Neal, F. Provost, and R. Schrift, “In pursuit of Enhanced Customer Retention Management: Review, key issues, and Future Directions,” Customer Needs and Solutions, vol. 5, no. 1-2, pp. 65–81, 2017.
[12] . D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, “Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda,” New Frontiers in Mining Complex Patterns, pp. 119–136, 2020.