CGIC INSURANCE FORECASTING MODEL

Xiao, Yiwen #6604052

Fan, Lin #6056881

Lyu, Xinxin #6802888

Wang, Binbin #6573547

July 14, 2020

Table of Contents

1.Executive Summary

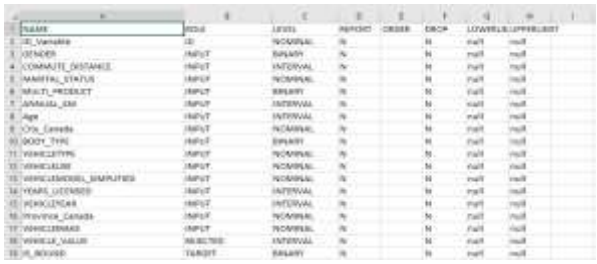
The Cooperators General Insurance Company has a large volume of web quotes, the company is not able to reach each one of them due to the time and cost constraints. Thus, CGIC is asking us to help them decide which quotes should be followed up. To assist the company, we have completed the work for carrying out three predictive models, which are decision tree, logistic regression, and neural network, as well as Twostep clustering analysis to give our final recommendations based on our analysis.

To give a better estimate, we use several configurations for our models and create 19 decision trees, 9 logistic regression models, and 6 neural network models in total. Among all the models we created, we choose the neural network as our final model for CGIC future scoring. Since it has the highest number in the ROC index and cumulative lift value, as well as the lowest number for both misclassification rate and average square errors for its validation data set.

After finalizing the model, we give several recommendations accordingly. First, geographically, CIGC should focus on people who live in ON, especially in southwestern, northern, and eastern areas. Second, demographically, CIGC should pay more attention to customers who are young adults aged between 25-39, and create promotions towards this group accordingly. Third CIGC needs to request the information of vehicle types and age, as different vehicle types and age combinations will affect the bound rate dramatically. Fourth, comping up with a new loyalty program is necessary. People who have multiple products with CIGC has a higher bound rate of 25%, which is nearly 10% higher than people who do not have (16%).

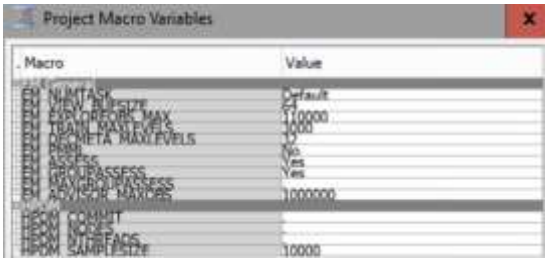
2.SAS Enterprise Miner

After we pre-processed our raw data in the Phrase I, we decide to input the variables below. In order to prevent unnecessarily impute mistakes, we replace all the missing values of nominal data by mode in our excel file manually. Since there are more than 65% and 90% of missing data in both the vehicle value variable and the occupations variable, we decide to reject these two variables into SAS Enterprise Miner.



After imported our file, there are two major things we changed in the setting, "Explorers Max" and "Train Max Levels". This enables us to fetch all the records and increase the maximum level of class level for our future modeling.

Table 1 Setting Changes



Data Processing

For all our models, because in our binary target variable, there is only 30% of the records are bound with CIGC, we have a significant class imbalance. A "Sample node" was introduced to achieve a 50%, 50% class balance for all our models to solve this problem. Second, because the machine did not know the missing value, we have to identify them for the machine. We used a replacement node to label missing values. After conducted the replacement node, there were 16609 records of commute distance that were labeled as a missing value in our data set. Third, we split our data into 50% training and 50% validation data sets by using the "Data partition" node, so that there will be two sets of data for the machine to measure. Based on that, because logistic regression and neural network can not

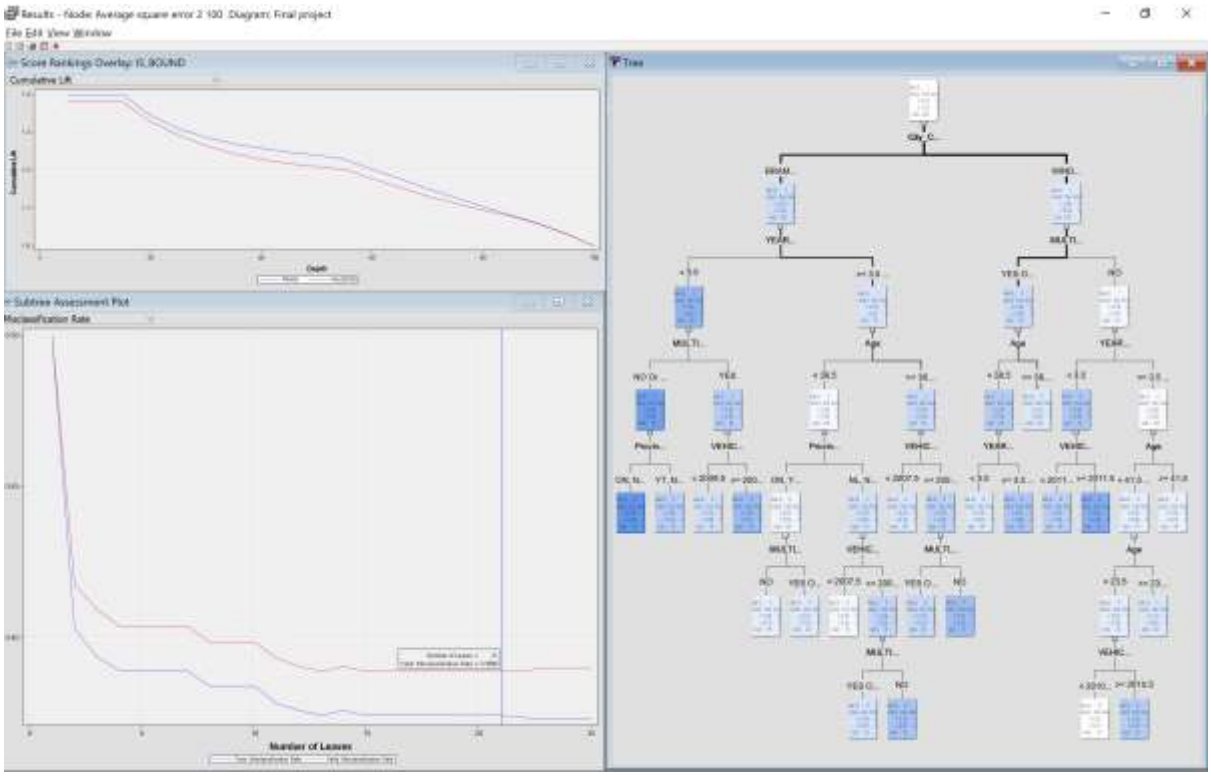
handle data with skewness and data with missing value, we applied a “Transform” node to change the skewed variable of “REP_COMMUTE_DISTANCE” and “YEARS_LICENSED” into a normal distribution form by log, and a impute node to replace all the missing value of “LOG_REP_COMMUTE_DISTANCE” by its mean.

Decision Tree

In decision tree, in order to see a different result, there are 2 assessment methods, and 3 attributes we mainly manipulated with. “Maximum branch size”, “Maximum depth” and “Leaf size”. After we tried many runs, we notice that the average depth of the decision tree is 5, so we set all the decision tree “Maximum depth” to 10 instead of the default value. Moreover, we tried a different combination of the setting “Maximum branch” with “Leaf size”. Since we have around 60,000 records in total, we decide to pick “100”, “500”, and “1000” as our leaf size. (Comparing with the total records, too small of a leaf size may not generate any useful results, as they will form small leaf with only a couple of few records in a class). The result of their misclassification rate and average square error of the validation data sets are showed below.

Assessment type	Maximum branches	Leaf size	Valid: Missclassification rate	Valid: Average square error
Average square error	2	100	0.388777	0.233123
Average square error	2	500	0.393173	0.233802
Average square error	2	1000	0.391917	0.234915
Average square error	5	100	0.392724	0.233647
Average square error	5	500	0.397479	0.234965
Average square error	5	1000	0.392634	0.235345
Average square error	10	100	0.395416	0.234486
Average square error	10	500	0.400709	0.236113
Average square error	10	1000	0.411474	0.239449
Misclassification rate	2	100	0.388777	0.2363
Misclassification rate	2	500	0.93173	0.236873
Misclassification rate	2	1000	0.391917	0.236923
Misclassification rate	5	100	0.391109	0.236819
Misclassification rate	5	500	0.397479	0.237234
Misclassification rate	5	1000	0.392634	0.235445
Misclassification rate	10	100	0.391872	0.23633
Misclassification rate	10	500	0.400709	0.238266
Misclassification rate	10	1000	0.411474	0.239866

We found out that when we set the “Maximum branch” as 2 and the “Leaf size” as 100, it generates the best model amount of our decision trees. The misclassification rate is 0.3887 and the average square error is 0.233123. We decide to choose this decision tree as the best model to compare with logistic regression and the neural network later.

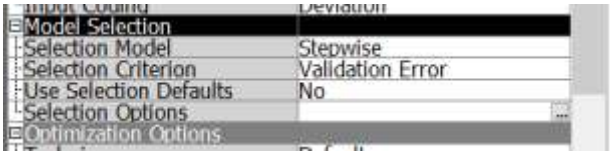


Logistic Regression

Model building

We know there are three different model selections in the regression model, “forward”, “backward” and “stepwise”, we decided to run three of them and to choose one for our final output and also as the input for our neural network model.

We first set the selection criteria property as the validation error to specify the criterion for choosing the final model, as

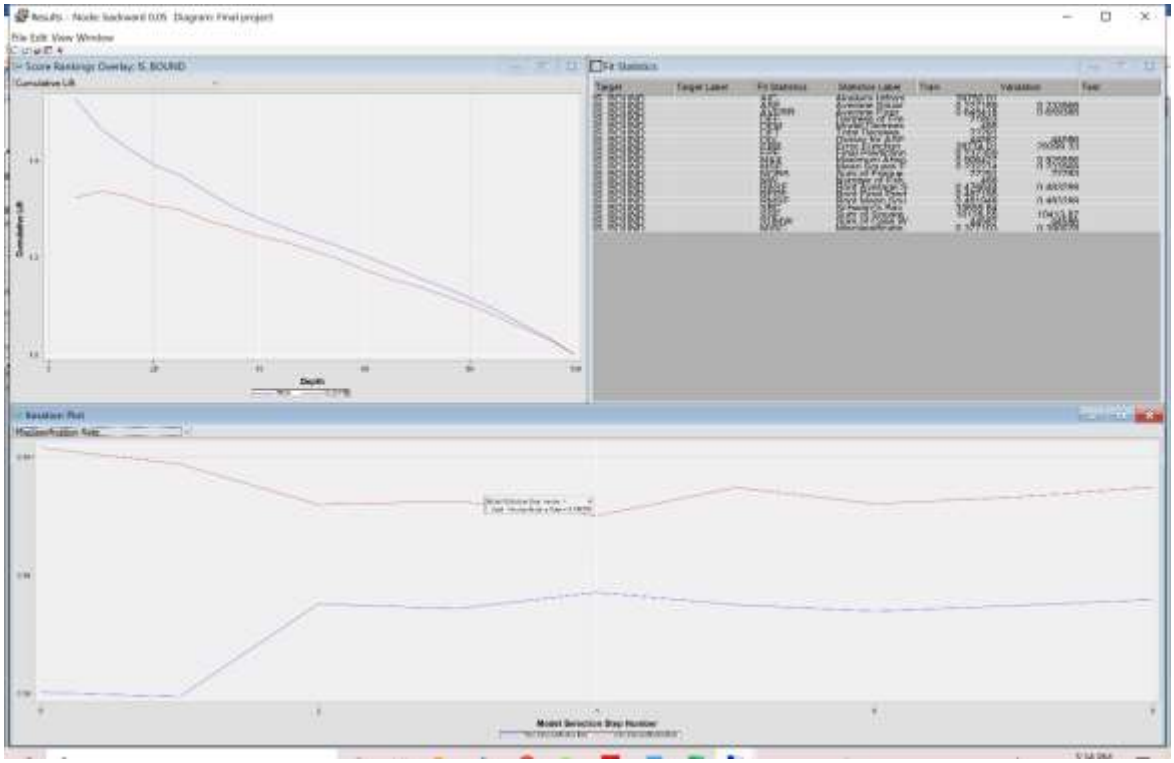


the model with the smallest error rate will be chosen. Then we set the "Use selection default" property to NO to use non-default values such as the maximum number of steps, entry, and stay significant level. We set the maximum number of steps to 30, so that machine will not stop too early before getting the optimum step result. Meanwhile, we also set the entry significance level as 1.0 as we want the model to include as many variables as possible for the input, and the stay significance level as 0.8, 0.5, and 0.005 to test the results.

After running all 9 different regression node with different combinations of stay significance level 0.8, 0.5, 0.05 and assessment methods "forward", "backward" and "stepwise", we found that backward model selection has the lowest misclassification rate and average square error in their validation data set, no matter the stay significance level. We think it is because some of the unimportant variables do not affect the result that much.

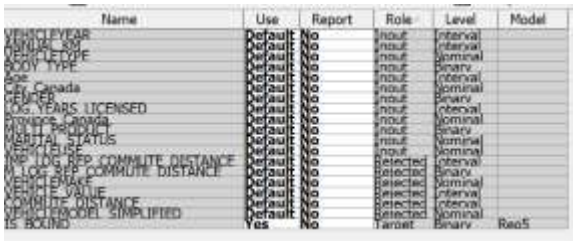
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg2	backward 0.5	0.39008	0.22719	0.37710	0.23357
	Reg5	backward 0.05	0.39008	0.22719	0.37710	0.23357
	Reg8	backward 0.8	0.39008	0.22719	0.37710	0.23357
	Reg	Stepwise 0.5	0.39461	0.22688	0.37679	0.23380
	Reg3	Stepwise 0.05	0.39461	0.22688	0.37679	0.23380
	Reg4	FORWARD 0.5	0.39461	0.22688	0.37679	0.23380
	Reg6	FORWARD 0.05	0.39461	0.22688	0.37679	0.23380
	Reg7	Stepwise 0.8	0.39461	0.22688	0.37679	0.23380
	Reg9	FORWARD 0.8	0.39461	0.22688	0.37679	0.23380

Since there is no difference, we will pick the regression model with backward model selection and a stay significance level of 0.005 with a misclassification rate of 0.39008 and an average square error of 0.23357 for the validation data set, as the model to compare with decision tree and neural network.



Neural network

Since the neural network is very sensitive towards the variables that we input, we decide to connect the neural network node with our selected backward logistic regression, as the backward logistic regression has the lowest misclassification rate and average square error, it also helped us to filter unnecessary variables.



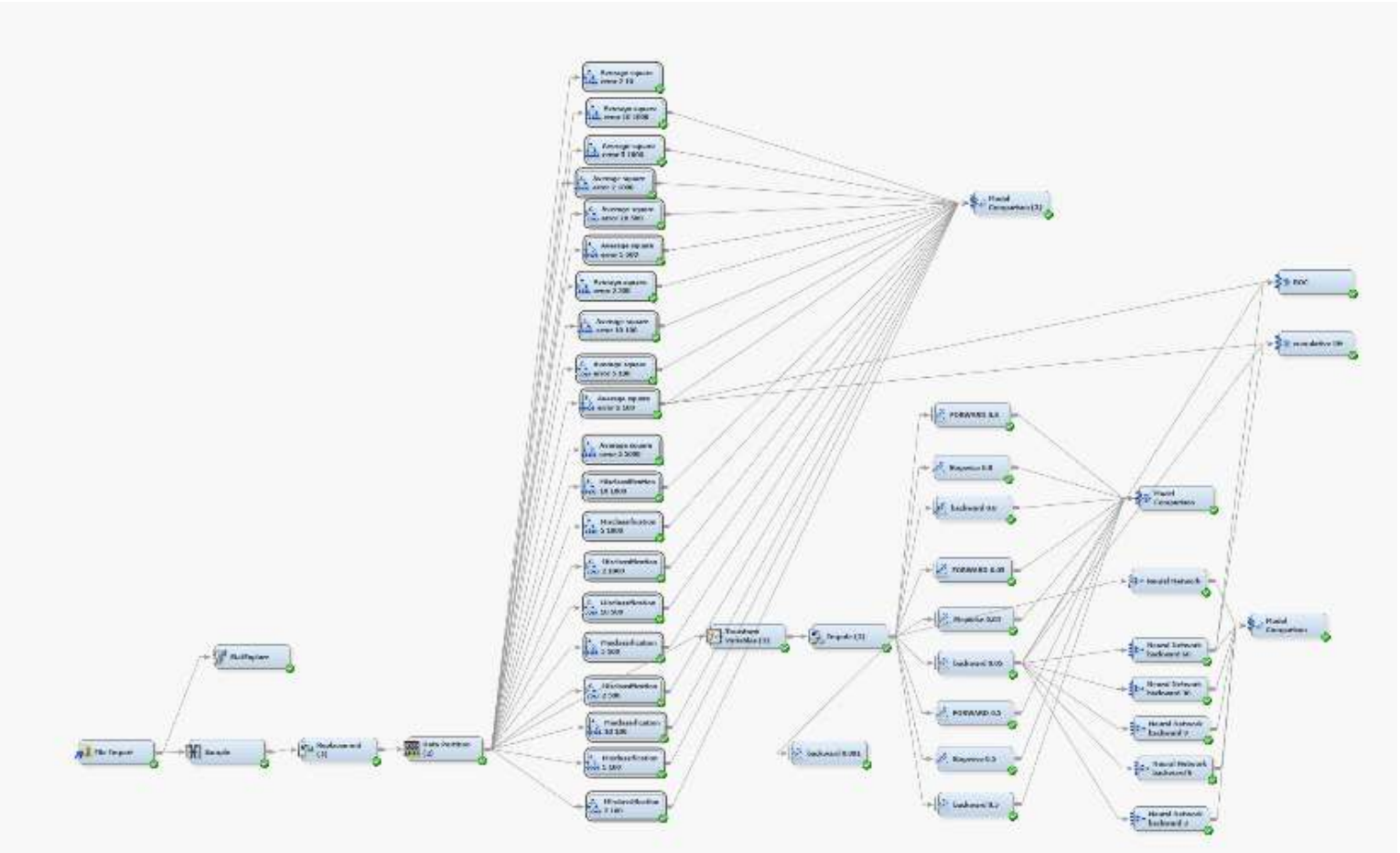
Name	Use	Report	Role	Level	Model
VEHICLEYEAR	Default	No	input	interval	
VEHICLE_KM	Default	No	input	interval	
VEHICLETYPE	Default	No	input	nominal	
BODY TYPE	Default	No	input	binary	
Age	Default	No	input	interval	
City Canada	Default	No	input	nominal	
GENDER	Default	No	input	binary	
LOG_YEARS_LICENSED	Default	No	input	interval	
PROVINCE_ROAD	Default	No	input	nominal	
MARITAL_PERIOD	Default	No	input	binary	
MARITAL_STATUS	Default	No	input	nominal	
VEHICLEUSE	Default	No	input	nominal	
TIME_LOG_REP_COMMUTE_DISTANCE	Default	No	rejected	interval	
M_LOG_REP_COMMUTE_DISTANCE	Default	No	rejected	binary	
VEHICLEMAKE	Default	No	rejected	nominal	
VEHICLE_VALUE	Default	No	rejected	interval	
COMMUTE_DISTANCE	Default	No	rejected	interval	
VEHICLEMODEL_SIMPLIFIED	Default	No	rejected	nominal	
IS_BOUND	Yes	No	target	binary	Reg5

As we learned in the class, the more iteration we have the more accurate the model tends to be, we decide to set the maximum iteration to 200 and let the machine decide which step of the iteration is the best for our model. In addition**,** the number of hidden units is also an important factor that decides the accuracy of our neural network model. We tried different number of hidden units of “3”, “6”, “9”, “30” and “60” in our model.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate ▲	Valid: Average Squared Error
Y	Neural	Neural	Neural Network	backward 60	0.384112	0.230194
	Neur	Neur	Neural Network	backward 6	0.388238	0.230803
	Neur	Neur	Neural Network	backward 9	0.389988	0.231089
	Neur	Neur	Neural Network	backward 3	0.390212	0.231935
	Neur	Neur	Neural Network		0.394608	0.233761

As a result, the neural network model with 60 hidden units performed the best. It has the lowest misclassification rate of 0.384112 and an average squared error of 0.230194 for its validation data set. We will pick this neural network model for further comparison.



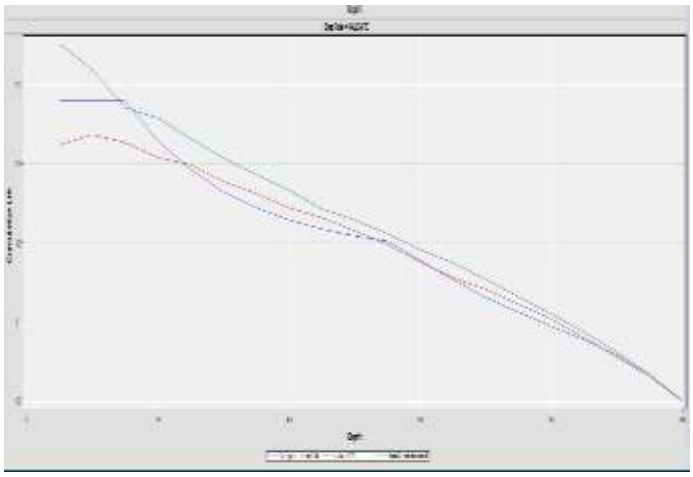
Model Comparison

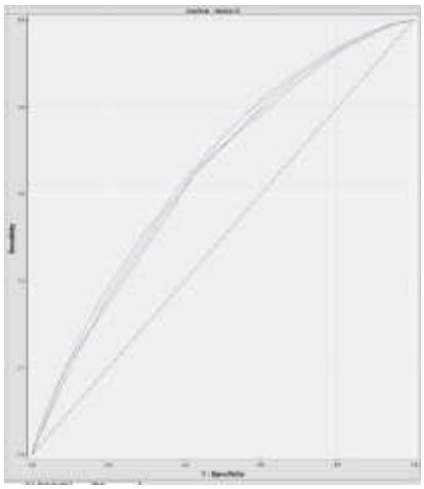
Final outcome

Finally, we need to compare all our models and choose the best one for CGIC's future scoring. We added two comparison node of "ROC" statistics and "Cumulative lift" statistic to compare their average square error and misclassification rate for the validation set as well as roc and cumulative lift index.

Selected Model	Model Node	Model Description	Valid: Roc Index	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Neural	Neural Network backward 60	0.661	0.22037	0.35884	0.23019	0.38411
	Reg5	backward 0.05	0.646	0.22719	0.37710	0.23357	0.39008
	Tree3	Average square error 2 100	0.643	0.22950	0.37419	0.23312	0.38878

Selected Model	Model Node	Model Description	Valid: Cumulative Lift	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Neural	Neural Network backward 60	1.41800	0.22037	0.35884	0.23019	0.38411
	Tree3	Average square error 2 100	1.37966	0.22950	0.37419	0.23312	0.38878
	Reg5	backward 0.05	1.33638	0.22719	0.37710	0.23357	0.39008





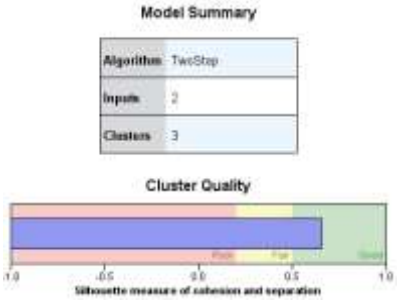
We can see that our neural network model has the highest number in roc index and cumulative lift (0.661 and 1.418), which means that they have better accuracy and predictability for our data. This can be also seen in both plots as its curve is above other models'. Besides, the neural network model also has the lowest number for both misclassification rate and average square errors for its validation data set (0.34811 and 0.23019). We decide to choose this model as our final outcome.

3.Cluster Analysis

TwoStep Clustering Analysis

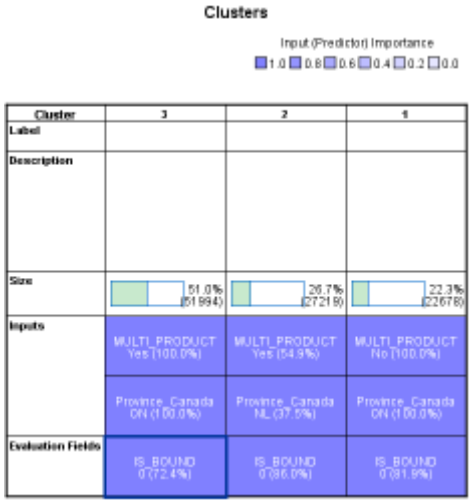
We use the TwoStep clustering method in SPSS to conduct cluster analysis. The input variables for carrying out TwoStep cluster analyses are focused on two sets of input variables: Province_Canada and Multiple_Product as well as VehicleType and VehicleAgeGroup.

1) Province_Canada and Multiple_Product



In the beginning, we started with six variables: Province_Canada, Multiple_Product, YearsLicencedGroup, VehicleAgeGroup, AgeGroup, and Marital_Status. After a couple of runs, we ended up picking Province_Canada and Multiple_Product for the TwoStep cluster analysis, as they showed a meaningful result in the analysis.

We set Province_Canada and Multiple_Product as Categorical Variables and IS_BOUND into the Evaluation Fields box, we set the maximum cluster as 8 and let the machine to determine the number of clusters automatically. As a result, we got a good clustering with 3 clusters and a ratio size of 2.29.



By default, clusters are sorted from left to right by size, as number 3 being the biggest one. Cluster 3 includes 51994 records, which accounts for 51.0% of the total population. Most of them come from ON and have multiple products with CGIC. Cluster 2 includes 27219 records and accounts for 26.7% of the total population. Among them, 54.9% have multiple products with CGIC. They live outside of ON and those who live in NL have the highest frequency in this cluster (37.7%). Cluster 1 includes 22678 records and accounts for 22.3% of the total population. They come from ON and do not have multiple products with CGIC.

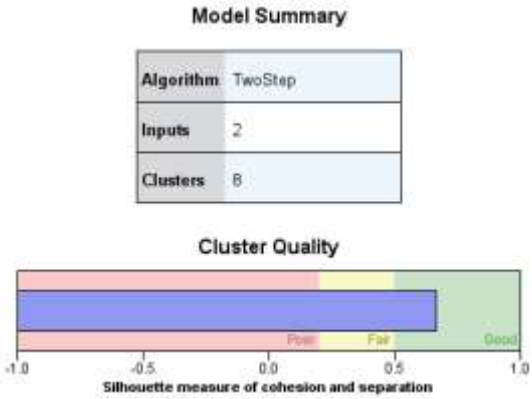
In terms of the historical data of IS_BOUND that we set for evaluation fields, the evaluation percentages of these four clusters are 27.6%, 14%, and 18.1% respectively. The evaluation percentages in the evaluation fields are used to evaluate the clusters, not bound rates.

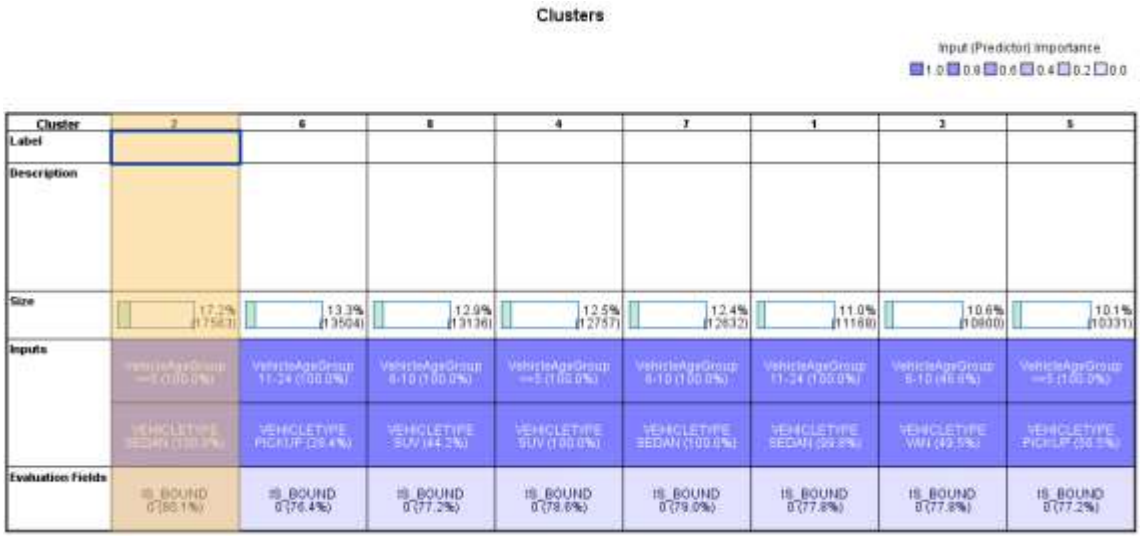
Based on the above analysis, we can see Cluster 3 has a higher bound rate. This group of customers is living in ON and have multiple products with CGIC.After carrying out the clustering analysis of Province_Canada and Multiple_Product combining with

VehicleAgeGroup, YearsLicencedGroup and AgeGroup individually. As a result, the clustering result was better when we combining it with VehicleAgeGroup.

2) VehicleType and VehicleAgeGroup

In this run, we set VehicleType and VehicleAgeGroup as Categorical Variables and IS_BOUND into the Evaluation Fields box, and let SPSS determine the number of clusters automatically with the maximum clusters of 15. As a result, we had an outcome of 8 clusters and a





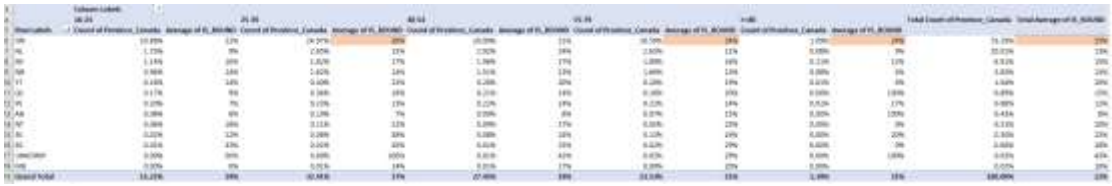
ratio size of 1.7. By default, clusters are sorted from left to right by size, so they are sequenced 2, 6, 8, 4, 7, 1, 3, and 5 in the clusters table.

Cluster 2 includes 17563 records and accounts for 17.2% of the total population. Those are all Sedans, which were within 5 years. Cluster 6 includes 13504 records and accounts for 13.3% of the total population. There are seven vehicle types (Pickup, SUV, Van, Coupe, Hatch, and Wagon), with Pickup being the highest amount (28.4%). Cluster 8 includes 13136 records and accounts for 12.9% of the total population. There are three vehicle types (SUV, Pickup, and Hatch), with SUV of 6-10 years being the highest amount (44.2%). Cluster 4 includes 12757 records and accounts for 12.5% of the total population who drive a relatively new SUV. Cluster 7 includes 12632 records and accounts for 12.4% of the total population. They are all Sedans, which have been bought for 6 to 10 years. The rest of the clusters only accounts for around 10% of the total population individually.

Comparison Analysis

1) Cluster Analysis vs. Pivot Table

In Phase I, we analyzed the bound rate by Province_Canada, and AgeGroup with Multiple_Product as a filter and concluded that among the customers in ON, those who are aged between 25 and 39 have the highest bounds rate at 30%, which is followed by the age group of 55-79 (24%) and above 80 (24%). With the factor of multi-products involved, we see a roughly 3% increase in all ages across all provinces.



In TwoStep clustering analysis, we suggest that CGIC should focus on the group of customers who are living in ON and have multiple products with CGIC. When combing with

the AgeGroup variable, we also suggest that CGIG should pay attention to those customers aged between 25 and 39 if they come from ON and have multiple products with CGIC.

The similar results also occur when we use VehicleType and VehicleAgeGroup to carry out pivot table analysis and cluster analysis, as pickup truck and SUV showed a relatively higher bound rate. However, compared with cluster analysis, the pivot table does not consider the weight of each category in the total record number, so the high bound rate from small classes may not be useful. Choosing cluster analysis might have a better outcome in this circumstance.

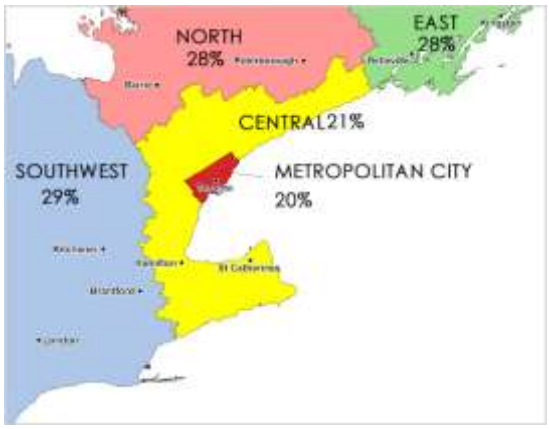
2	BODY TYPE	NORMAL				
3						
4	BOUND RATE	Column Labels				
5	VEHICLE TYPE	<=5	11-24	6-10	>=25	Grand Total
6	BUS	0%	0%		100%	25%
7	COUPE	19%	23%	20%	24%	21%
8	HATCH	23%	22%	23%	17%	23%
9	KIT				0%	0%
10	MOTOMOBILE	36%	20%	33%	33%	32%
11	PICKUP	22%	23%	23%	23%	23%
12	SEDAN	20%	22%	21%	23%	21%
13	SPORTS	23%	25%	18%	26%	23%
14	SUV	21%	25%	23%	29%	22%
15	TRAILER	50%	0%	0%		20%
16	TRUCK	67%	33%	50%	40%	42%
17	UTILITY VEHICLE	75%	14%	0%	0%	24%
18	VAN	22%	23%	23%	25%	22%
19	WAGON	26%	26%	22%	22%	24%
20	(blank)		0%			0%
21	Grand Total	21%	23%	22%	24%	22%

4.Recommendation

For our recommendation, we would like to make a suggestion based on our Decision Tree, Logistic Regression, Clustering Analysis, and pivot tables. According to the tables below, we will do our recommendations based on these variables.

Models	Recommended Variables
Decision Tree	Province_Canada, MULTI_PRODUCT, YEARS_LICENSED, Age, VEHICLEYEAR, City_Canada
Backward Regression	Province_Canada, MULTI_PRODUCT, Age, LOG_YEARS_LICENSED, MARITAL_STATUS, VEHICLEYEAR, City_Canada
Clustering	Province_Canada, Multi_Product, YearsLicencedGroup, VehicleYearGroup and AgeGroup, VehicleType

First, geographically, CGIC should focus on customers who live in ON, since it accounts for the majority source of its customers and has an average bound rate of 25%. When we break down the province of Ontario into areas, we find that the southwestern, northern, and eastern areas of Ontario have an even higher bound rate at around 28%. CIGC should focus on these areas, especially in cities like Guelph, Huron, Nepean, Ottawa, Peterborough, where the bound rate reaches almost 30%. When we analyze it with demographics, people aged 25-39 in Ontario tend to have a higher bound rate than any other age group. However, although NL province accounts for 10% of the population, which is the second biggest province group in our data, their bound rate is only 13%, to reduce the cost CIGC should try to pay less attention in this province.



Second, demographically, CIGC should focus on young adults who are aged between 25-39, analysis shows that this group has a higher bound rate of 27%. Amount them people who own wagon and sports cars showed a higher bound rate of around 30%. We suggest CIGC come up with more promotions towards young adults to increase their overall bound rate. Interestingly enough, when we add in the factor of gender, it shows a gap between different genders of people aged 80 and above, with a 25% bounds rate for females and 19% bounds rate for males.

Average of IS_BOUND	Column Labels					
Row Labels	16-24	25-39	40-54	55-79	>=80	Grand Total
COUPE	18%	24%	20%	23%	19%	21%
HATCH	19%	28%	19%	23%	21%	23%
MOTOMOBILE	0%	17%	44%	40%		32%
PICKUP	21%	28%	21%	20%	23%	23%
SEDAN	18%	26%	18%	20%	19%	21%
SPORTS	19%	29%	20%	25%	0%	24%
SUV	18%	28%	19%	21%	26%	22%
VAN	20%	27%	20%	21%	21%	22%
WAGON	22%	30%	19%	23%	20%	24%
Grand Total	**19%**	**27%**	**19%**	**21%**	**21%**	**22%**

Third, in terms of car types. Wagon, Sports, Pickup, and Hatchback types have a higher bound rate in general. When we break down by the vehicle age, CIGC tends to have

an above-average bound rate of 24% with cars above 25 years. We can see that New wagons, old sports cars, and old SUV are more likely to be bounded with CIGC with a roughly 27% of bound rate. In order to better target their clients with information about their cars and the age of the vehicle, we suggest CIGC makes vehicle type and age information mandatory.

Fourth, multi-products is another factor that will affect the bound rate significantly. In our analysis, we found that on average, people who have multiple products with CIGC have a higher bound rate of 25% which is nearly 10% higher than people who do not have (16%). CIGC should launch more loyalty programs and focus on their old customers, as it can lower their acquisition cost for new customers.