|  |
| --- |
| Co-operators General Insurance Company |
| Phase 2 Report |
|  |

|  |
| --- |
| Wahaj Siddiqui  12-3-2021 |

**Table of** **Contents**

[**Executive Summary 2**](#_Toc89442752)

[**Problem Statement 3**](#_Toc89442753)

[**Data Preprocessing 3**](#_Toc89442754)

[**Predictive Modelling 7**](#_Toc89442755)

[**Logistics Regression 7**](#_Toc89442756)

[Default Stepwise Model 7](#_Toc89442757)

[Relaxed Stepwise Model 8](#_Toc89442758)

[Relaxed ASE Stepwise Model 8](#_Toc89442759)

[**Decision Tree 9**](#_Toc89442760)

[**Neural Networks 11**](#_Toc89442761)

[Misclassification Rate 11](#_Toc89442762)

[Average Error Function 12](#_Toc89442763)

[**Segmentation Analysis 12**](#_Toc89442764)

[**Table of Figures 17**](#_Toc89442765)

# **Executive Summary**

As the Cooperators General Insurance Company, or CGIC, begins to embrace their online platform, they have begun offering auto insurance quotes on their website in exchange for information. They hope to use their proprietary data for business development purposes. One opportunity of interest is within the data mining field. In the online platform, CGIC has been prone to logjams and quote processing delays. Consequently, the company has hired business analysts to unravel the data and recommend a predictive model which would forecast which individuals have a high propensity for entering into an insurance agreement based on the information they provide.

There was preprocessing needed for the data, as it was in a raw state. Columns were removed or recalculated; rows were imputed on; and entries were corrected. The clean dataset held 18 columns, 101,411 values, and held minimal noise and granularity. After generating simple insights on the cleaned datasets, relationships were uncovered, and a modelling roadmap was generated.

The data underwent a 70/30 training and validation split. The modelling process accounted for three different machine learning algorithms; Decision Trees, Logistic Regression and Artificial Neural Networks. Rigorous systematic experimentation was conducted on each type of model around assessment criteria, discretization, filtering, and model design. From each type, a comparison study was conducted using the highest cumulative lift and cumulative % response rate. Once the top models from each type were selected, there was a model comparison node consisting of ROC and average square error-based assessment of the validation data. The final recommendation was based on our overall understanding of the model, as well as the scoring distribution as compared to others. In this experiment, the maximal decision tree model was the best scoring experiment in several tests. The advantages of this model included its consistency and productivity in terms of classification.

For the last step, Clustering analysis was conducted to validate our initial insights and to discover any associations or patterns within the data which our predictive models overlooked. Through this step, we found several insights which proved to be key considerations in the project’s success. The clustering technique used is called K-means clustering, which is an iterative approach to finding centroids.

The final recommendations and key discoveries will be discussed throughout this report.

# **Problem Statement**

The Co-operators General Insurance Company, or CGIC, is a provider of home, auto, life, business, travel, and farm insurances. As of 2016, they had 101,981 auto insurance quotes that they have collected from their web services. CGIC receives a large volume of web quote requests monthly. Currently, their request processing is prone to logjams due to the need for each quote request to be individually reviewed. This, in turn, results in many ‘to be bound’ quotes to become defunct before the request is reviewed. CGIC has a database filled with the quote information which has been submitted and processed, point -- “IS\_BOUND.” This is a data mining opportunity. The company wants to learn more about its client pool and use its proprietary data to estimate the propensity of successful business interactions. This will help the company prioritize its insurance request process towards individuals who, according to the model, are more likely to form an agreement with CGIC.

# **Data Preprocessing**

In this section, we will summarize our preprocessing and make any assumptions about the relevance of the variables within our project. This step is critical to building good predictive models as they alleviate pain points and incorrect data. Currently there are 26 variables as well as one identifier attribute. We individually dissected each variable to ensure that it is clean and relevant, the key changes are listed below.

1. QUOTEDATE: We grouped each entry by month to improve the data’s ability to function in a predictive model. For this, we created a calculated column which categorizes each entry by month – 1-12.
2. VEHICLEYEAR: This variable shows the year of the vehicle for which the quote is being requested. Once we drill-down, we discover that there are two entries of vehicles from the year 2018 and 2019. This is an error as there are not any 2018 or 2019 models as of 2016 when the entries were created.
3. 2018 BMW X5 - ID 40014920 - to diagnose and fix the issue with this value, we filter some of the other variables (VEHICLEMAKE, VEHICLEMODEL, ANNUAL\_KM, VEHICLE\_VALUE). Through this process, we discovered an identical entry for vehicle, annual kilometers, and value, except the VEHICLEYEAR is 2015. The assumption of the vehicles being the same is made and the value 2018 is replaced.
4. 2019 VOLKSWAGEN JETTA - ID 40082272 - due to the lack of any distinctive entry (VEHICLE\_VALUE), we will remove this row from the dataset.

With that, the VEHICLEYEAR variable is clean.

1. VEHICLEMAKE: This variable shows the manufacturer of the vehicle. There was a lot of cleaning which needed to be done. Due to the page constraint, we will list one of the cleaning items we completed. There are several entries which categorize the VEHICLEMAKE with the suffix -TRUCK/WAGON and CAMION/WAGON. This is an added detail which complicates this variable. First, we separate the VEHICLEMAKE from the category by using the delimiting feature in excel. Now that a new column is created housing the categories, we must determine its value. Unfortunately, there was not enough clarity to assume whether the empty rows (not TRUCK/WAGON or CAMION/WAGON) meant that they could be categorized as sedans, or simply, cars. So, we decided to remove the new generated column. This column was now clean, as well.
2. VEHICLEMODEL: This variable shows the vehicle model for which the quote request was completed. There was also quite a bit of cleaning needed for this variable. Due to the sheer number of models and specifications that are available (5670 unique values), we decided to source just the first name of the model. For instance, a SILVERADO 2500 HD LT CREW CAB 4WD was converted into SILVERADO. As a result, the number of unique values decreased to 1261. The next step in the cleaning was to correct errors like ACC into ACCENT. After cleaning the little typos, the VEHICLEMODEL is ready for the next step.
3. ANNUAL\_KM: This variable tells the estimate of annual kilometers driven by the prospective client. First, we changed the data type from general to number, so we can run calculations on this variable. Additionally, we also run into an outlier in this variable -- 20,000,000 kilometers for ID 40094546. This entry is removed from the dataset because it is impossible. Next, we noticed that there may be other inaccurate or misplaced data within this category - this focuses on annual kilometer values under 100km. However, we must compare with the COMMUTE\_DISTANCE to gain a better understanding of what is going on. When we filter for values under a 100 km, we can see that the COMMUTE\_DISTANCE for each entry is significantly larger. The explanation that we are using is that the values were switched during any given step from the data entry to our team gaining access to it. So, we switch the values to show the correct order. Now we must replicate these changes into COMMUTE\_DISTANCE.
4. COMMUTE\_DISTANCE: Like the ANNUAL\_KM, we convert the data type to numbers. Then, we must create and enforce two rules
5. The only instance where a non-blank entry is acceptable is if the VEHICLEUSE entry says “Commute.”
6. Since the commute distance cannot exceed the total annual kilometers value, every instance where the commute distance exceeds the annual kilometers must result in a switch between the two items. Our assumption is that there must be ambiguity in the design of the web form which confuses individuals about the positioning of the two values. Once both rules are satisfied, the column is clean.
7. VEHICLE\_VALUE: This variable is an indicator of the value of the vehicle as estimated by the client. First, we change the data type to currency. Then, we discover that there are 64439 missing values in this column. Since this is a client estimate, we will not make any more changes to this column.
8. VEHICLEUSE: This indicates the main use of the vehicle. Two categories account for over 90% of all entries. We must group items, to decrease noise within the current entries – the changes made are as follows: Work and Leisure.

Work (66622) – (Business, Commute, Courtesy Car, Driver Training, Fire Dept, Fisherman, Police)

Leisure (34870) – (Farm Pleasure, Pleasure, Rally)

1. GENDER: This categorizes the individual’s gender by Male, Female or Unknown. There was only one “Unknown” Value, so it is removed to decrease noise. Only Male and Female values remain.
2. YEAR\_OF\_BIRTH: This is the year the individual was born in. This column was used to create a calculated column which imputes the individual’s age into a new column called AGE. This column was also converted into a number data type for computation. Upon creating the AGE column, we discover that there are a few quotes filed by individuals under the age of majority for attaining the driver’s license - 16 years. These rows (3 total) are dropped because they are flawed entries.
3. POSTAL\_CODE: This indicates the FSA to which the prospective client belongs to. This is the primary location attribute. We use the first character in the code to group the values by region. When conducting visualization, we will use an external FSA dataset to define more specific geographical values to this column. A calculated column ‘Region’ is created and serves as the primary location attribute. This column is removed. For a few items, AREA\_CODE is used to impute values and the replaced by the region column.
4. YEARS\_LICENSED: Number of years for which the individual has been licensed. There are 8 missing values; these values are removed to decrease overall noise as we do not have the ability to estimate the values. There are also more incorrect values that are too large. The rule we employ to remove these is: if the YEARS\_LICENSED exceeds age, remove the row.
5. OCCUPATION: The column is removed as there are not enough values to conduct valuable analysis.
6. VEHICLE\_OWNERSHIP, YEARS\_AS\_PRINCIPAL\_DRIVER, CONVICTION\_COUNT\_MINOR\_3YRS, CONVICTION\_COUNT\_MAJOR\_3YRS, CONVICTION\_COUNT\_CRIMINAL\_3YRS, ASSIGNED\_LOSSES\_PD\_5YRS, SUSPENSION\_COUNT, MARKING\_SYSTEM & TRACKING\_SYSTEM:  
   These columns are removed due to lack of relevance.
7. IS\_BOUND: This is the dependent variable of this dataset. There is one missing value in this column at ID 40026648 - we remove that row.

The final cleaned dataset looks quite different from the original file. There are only 18 columns remaining for which the data summary is recorded in *Figure 1*.

# **Predictive Modelling**

Our predictive modeling and analyses step was conducted using *SAS EM*. There were a few preprocessing steps conducted in that as well. We conducted final variable selection and dropped three columns – YEAR\_OF\_BIRTH, VEHICLEMODEL and VEHICLEVALUE columns due to redundancy and high granularity, respectively.

For Data Partitioning, we performed a 70-30 default split. This is the standard partitioning rule in business analytics. Once we run the DP node, we receive the partition summary as the data is partitioned into 70% training and 30% validation split. Out of the 101,411 remaining entries, 70,988 were allocated to training, and 30,423 were allocated to validation. The summary statistics for the partition is in *Figure 2*.

## **Logistics Regression**

Using the cleaned dataset from phase one, the inputs variables that were used create the logistic regression model are the following: AGE, ANNUAL\_KM, COMMUTE\_DISTANCE, GENDER, MARITAL\_STATUS, MONTHNUM, MULTI\_PRODUCT, REGION, VEHICLE\_MAKE, VEHICLEUSE, VEHICLEYEAR & YEARS\_LICENCED. The following logistic regression models were created to conduct analysis. Each of the regression models were created based on different properties.

### **Default Stepwise Model**

In the first stepwise model we used the default settings, which has the same entry significance level and stay level of significance, 5%, and default selection criterion After running the regression, the model in the first step was selected. This model consists of Intercept and MARITAL\_STATUS. The overall model is good since the overall probability, 0.0368, is less than the level of significance, 5% (*Figure 3*). Referring to the analysis of maximum likelihood estimates, single and married individuals seemed to have the most impact on the target as their probability is less than the significance level, 5%, and their odd ratio estimates is higher than odd ratio estimates of divorced, widow\_widower, and separated individuals (*Figure 4*).

According to the cumulative lift graph, If CGCI contacted 95% of the people applied for quotes using this model, they will be 22% more responsive than if they were contacted randomly. if they were contacted randomly. If CGCI contacted 40% of the people applied for quotes using this model, they will be 21.5% more responsive than if they were contacted randomly (*Figure 5*).

The model has 0.171342 average squared error on the training data, and 0.171399 average squared error on the validation data. It has 0.21959 misclassification rate on the training data, and 0.219629 misclassification rate on the validation data.

### **Relaxed Stepwise Model**

In the second stepwise model we used the default selection criterion with entry significance level of 1 and stay significance level 0.5. We also changed the maximum number of steps to 30 steps. After running the model, the model in step 7 was selected. This model consists of Intercept, AGE, GENDER, MARITAL\_STATUS, REGION, and YEARS\_LICENCED. The overall model is good since its probability, 0.0388, is less than the level of significance, meaning some of the variable in the model affect the target (*Figure 6*).

According to the analysis of maximum likelihood estimates the following inputs (divorced, married, Metropolitan Toronto, British Columbia, and Western Quebec) seemed to have the most impact on the target as they have low probability, less than 7%, and their odd ratio estimates is higher than odd ratio estimates of the rest inputs (*Figure 7*).

According to the cumulative lift graph, If CGCI contacted 10% of the people applied for quotes using this model, they will be 22.5% more responsive than if they were contacted randomly. if they were contacted randomly. If CGCI contacted 40% of the people applied for quotes using this model, they will be 21.7% more responsive than if they were contacted randomly. The model has 0.171279 average squared error on the training data, and 0.171475 average squared error on the validation data. It has 0.21959 misclassification rate on the training data, and 0.219629 misclassification rate on the validation data (*Figure 8*).

### **Relaxed ASE Stepwise Model**

In the third stepwise model we used validation error as a selection criteria and entry significance level of 1 and stay significance level 0.5. We also changed the maximum number of steps to 30 steps. The model was selected in step 0. This model consists of the Intercept only. The overall model is good since the probability is less than 0.05 (*Figure 9*).

According to the cumulative lift graph, If CGCI contacted any percentage of the people applied for quotes using this model, they will all have the same response rate, 21.959% more responsive than if they were contacted randomly. The model has 0.17137 average squared error on the training data, and 0.171392 average squared error on the validation data. It has 0.21959 misclassification rate on the training data, and 0.219629 misclassification rate on the validation data.

## **Decision Tree**

The simplest type of prediction is the decision, also known as classification. Decision predictions often relate directly to a categorical target variable, and so they are identified as primary, secondary, and tertiary in correspondence with the levels of the target. By default, decision tree models in SAS Enterprise Miner assume decision predictions when the target variable has a categorical measurement level, which in our case, the target class “IS\_BOUND” has a binary level of measurement. Selecting useful inputs is vital as redundant input does not give any added information, therefore, it is effective that the modeling algorithm for decision tree models make input redundancy a minor issue.

The decision trees we will implement will address the model essentials which are predicting new cases, selecting useful inputs, and optimizing complexity through prediction rules, split search algorithms, and pruning. The initial step in developing the decision tree model is to implement the decision tree node in the diagram, click the interactive tab and split the node in the tree view. Examining the log worth in the dialog box, variable COMMUTE\_DISTANCE has the highest log worth (1.9476), and therefore, we develop an interval split rule with a specific branch of 2 where the first branch contains instances with a COMMUTE\_DISTANCE count less than 7.5 km, and the second branch contains instances with a COMMUTE\_DISTANCE count greater than or equal to 7.5 km. With the training data partitioned into two subsets, the first subset corresponding with cases for COMMUTE\_DISTANCE count less than 7.5 has a lower-than-average concentration of IS\_BOUND=0 cases. The second branch has a higher concentration than the first branch and the average for IS\_BOUND=0. We split the branch 1 node and create two more nodes for the VEHICLE\_USE variable and we split the branch 2 node and create two more nodes for the VEHICLEYEAR variable as these are the highest logworth values after COMMUTE\_DISTANCE. Work in VEHICLEUSE has a higher-than-average concentration for IS\_BOUND=1. Based off the final Maximal tree after further splitting the nodes, we can note down some observations made off the analysis for the commute distance node with less than 7500 km and greater than 7500 km. If we go down the tree from the right side (>= 7500 km), the node is split through the variable VEHICLEYEAR as a binary branch, which brings us to the two options, older than 2000 (<2000) or newer than 2000 (>=2000). Since most of the customer cars are newer than 2000, we can move down the branch on the right with greater than 2000 and split the node through the attribute VEHICLE\_USE which can be split into two options, ‘work’ or ‘leisure’. Since the initial node was split with the attribute COMMUTE\_DISTANCE, it does not make sense for leisure to be relevant and so we ignore that branch and work with the work or missing branch under ‘>=2000’. The count for clients who work under the training set is 8958, with 20.76% for IS\_BOUND=1 under the training data. If we split the node further through the MULTI\_PRODUCT attribute, we can analyze how many customers said yes or no to CGIC’s multi products in the previous years. With the node split, 3106 have said no in the previous years, whereas 5852 clients have a product with CGIC already. We can further split the node through the GENDER attribute for the “YES” branch, which gives us two new nodes, male with 3239 instances and female with 2613 instances. To get more accurate results we split both the male and female branches by age greater than 35 and less than 35. Since the count is still higher than 1000, we can further split the nodes to get a stronger observation for what kind of clients would be bound or not bound based on the different attributes and split points. If we are to utilize the maximal tree to the child nodes to develop assumptions, we can say that there are 702 instances that drive more than 7500 km for work purposes, who have existing products with CGIC, are classified as males over the age of 35, and are married and have been licensed for over 25 years (*Figure 10)*. With these observations, we can assume that this specific statistic count and concentration for IS\_BOUND=1 informs us that this many clients are willing to agree to a quote with CGIC based on the attributes used to split the nodes. We can conduct further analysis in the maximal tree by splitting nodes with different attributes to develop different observations. However, in our analysis we have made observations by following the thicker line in the maximal tree, as the thicker line shows us where most of the records have gone through the branch. The interval splitting rule dialog box in *Figure 11* highlights the variables used to develop the binary branches in the decision tree. The next step is to view the subtree assessment plot and analyze it. *Figure 12* shows 22 leaves with the plot for training data (0.2200) to be constantly above the validation data (0.2184). From this we can assume that the maximal 22-leaf tree generates a higher misclassification rate than any of its simple predecessors. By using the same model to assess model performance and variable usefulness, it can often lead to overfit models. However, looking at the assessment plot on the validation data brings us to the solution. Another alternative decision tree is developed with assessment measure as Average Square Error. Average square error for any decision tree in the sequence, have necessary inputs for calculating the statistics for the actual target value and the prediction. The results highlight a single node with IS\_BOUND=1 at 21.96 and IS\_BOUND=0 at 78.04. The reason there is only a single node is because we did not split the tree like we did with the maximal tree. Splitting the data was based on the logworth to isolate subregions of input space with how proportions of clients that are bound and are not bound. Due to having no splits, the average square error is affected and is inadequate for measuring the performance of the validation and training data. The decision tree model decides from the leaf or terminal node. By default, it is a classification based on the predicted target value.

## **Neural Networks**

Neural Networks are a complex algorithm which aims to generate a noise-free predictive output. Our neural network analysis consisted of 6 tests which consisted of different assessments, different optimization values and different preprocessing methods. After analysis, we compared them and selected the best option to compare with other predictive models. For each attempt, we changed one property each. For conciseness, we will group the analysis by model selection criterion. One of the criterions we will consider in our decision process is convergence (*Figure 13)*. The process as well as property changes are outlined below, and the entire diagram is listed in the appendix as *Figure 14*.

### **Misclassification Rate**

We conducted three neural networks using the misclassification. For the first node ‘NN\_MR,’ we set misclassification as the criterion (*Figure 15*). We ran the test with all default values. The results of the model were suspicious of error – the neural network satisfied the convergence, but the misclassification rate remains unchanged as the iterations increased – MR = 0.219604. So, we ran a misclassification node but reduced the input variables using chi-square threshold with a variable selection. We ran into the same result, so we decided to run subsequent tests with different network structures and optimization settings. Once again, the results were similar as the misclassification rate remained constant. We attempted to replace the kilometer and commute distance using a replacement node and same result. We discovered that misclassification rate criterion is not well-suited for our purpose (*Figure 16*). In our last test with misclassification, we changed the network manually to test whether multilayer perceptron characteristic was the culprit to poor result. We ran in to same issues; consequently, misclassification rate was deemed an incorrect criterion for our model.

### **Average Error Function**

Our next model property change was to evaluate using Average Error as the model selection criterion. In the diagram, Test 6 and 7 represent these models. With default settings and average error as the model selection criterion, we ran Test 6 – AE = 0.526405 for the validation set (*Figure 17*). The results were improved, the convergence constraint was satisfied, and we saw a decrease in average error over the 5 training iterations. We also had a respectable cumulative lift ranking as depth increased. We did receive an error in optimization which we correct by increasing the number of maximum iterations from the default ‘5’ to ‘10’. Next, we used the replacement node as a test indicator for if “Test 5: NN\_MR.Replace” was a misstep. In this model, we see comparable results. The fit statistic ‘average error function’ value for the validation data is 0.53 for both tests (*Figure 18*). We will consider these tests when we are comparing all our neural networks for the choice for the final recommendation.

## **Segmentation Analysis**

Our goal is to subdivide our instances into subsets of customers where we can select a subset as a market target with a distinct marketing mix. Our next step will be to measure the clustering quality by observing the attributes that define their behavioral patterns in same and different clusters. With *Weka*, the data set is used as the clean dataset which includes 101,411 instances. As the initial step, we remove VEHICLE\_MODEL, QUOTEDATE, and ID\_Variable in the preprocessing tab. QUOTEDATE was removed because the MONTHNUM is sufficient as it will shows which month of the year the client requests for a quote. The VEHICLE\_MODEL is not as relevant as VEHICLEMAKE as it shows which car brand the average client has in each cluster. The ID\_Variable is removed as it does not provide sufficient granularity since each client has a specific ID, there would not be a mean value to represent the average client. We open the “cluster” tab and choose SimpleKMeans algorithm as our cluster to convert categorical attributes to binary so the preprocessing filter can handle the mixture of categorical and numerical attributes. The Euclidean distance measure will compute the distances between instances and clusters. The random seed option supports the clustering model by providing the same score for each split (*Figure 19* ).

We will leave the random seed as 10 and the number of clusters as 2 for our first observation. With 14 attributes, our cluster model is developed through the default training set cluster mode. The result window displays the centroids of each cluster along with their statistics and percentage of instances in each cluster. Each cluster centroid represents the mean vectors to characterize each cluster. The centroid for cluster 0 has 54,675 instances and shows that this is segment of cases are representing middle aged to old (approx. 50) males, married, living in Central, Ontario who have been licensed for around 27 years. This group has purchased products from CGIC in previous years and tend to drive for leisure, but for commuting purposes they drive around 16 km and an estimate of 14,119 km in a year. Furthermore, this group drives vehicles that are on average made in 2009 and tend to have a FORD vehicle make. Lastly, this group usually requests a quote in July and has a 21% chance of securing the quote with CGIC. The centroid for cluster 1 has 46,736 instances and shows that this is segment of cases are representing middle aged to young (approx. 34) females, single, living in Southwestern, Ontario who have been licensed for approximately 13 years. This group has not purchased products from CGIC in the previous years and tend to drive for working purposes. The commute distance is 15 kms, with an estimate of 15,872 kms annually. In addition, this group drives vehicles that are on average made in 2009 which tend to be Honda’s as the vehicle make. Lastly, this group commonly requests in July and has a 22% chance of securing a quote with CGIC. In the CGIC dataset, we treat the IS\_BOUND variable the same way as the other variables, rather than a classification attribute. A visualization of these two clusters is displayed in *Figure 20*.

We will now develop 4 cluster groups with random seed as 10 for our second cluster model analysis (*Figure 21*). With 14 attributes, our cluster model is developed with the default training set cluster mode. In this model, cluster 0 has 25,908 instances and shows us the segment of cases that represent middle aged to old (approx. 57) males, that are married and are living in Newfoundland and Labrador. This group has been licensed for 35 years and has purchased products from CGIC in previous years. This cluster tends to drive for leisure purposes but for work their commute distance tends to be 13 kms, and 12,485 kms on annual basis. The average vehicle driven are 2009 TOYOTA’s, with quotes being requested around July, and with a 22% of being bound to a quote. The centroid for cluster 1 has 15,865 instances and shows the segment of cases that represent middle to young aged (approx. 35) females, who are single, and live in Southwestern Ontario. This group tends to use their vehicles for work purposes, with an average of 14 kms for commute distance, and 15,454 kms on an annual basis. This group has been licensed for an average of 15 years and have bought products from CGIC in previous years. The centroid for VEHICLEMAKE is Mazda with an average year of 2009. This group also tends to look for quotes in July and has a 42% chance of being bound for a quote with CGIC. In comparison to the remaining 2 cluster groups, cluster 1 has the highest chance in having a successful interaction with CGIC for a quote. Therefore, it would be recommended that they use this cluster segment as their market target.

Our third cluster model involves a k input of 6 with the attributes we have used in the previous 2 models. The same default settings will be implemented, with random seed as 10 and the default training set cluster mode. From the 6 cluster groups, we will focus on the cluster segment that attains the highest percentage towards IS\_BOUND. In this model, cluster 3 has 10,181 instances and shows us that this segment of cases represents middle aged (approx. 45) females, who are married and are living in Eastern Ontario. The group tends to use their vehicle for work purposes with their commute distance being 18 kms and 15,520 kms on an annual basis. This group has been licensed for 22 years and has purchased a product with CGIC in the past. Furthermore, TOYOTA is the mean vector in this cluster with vehicles manufactured in 2010. This group tends to look for quotes in July and has a 55% chance of being bound for a quote with CGIC. For comparison purposes with other clusters, the results, and visualizations for the 6 clusters are shown in *Figure 22* and *Figure 23*.

The last cluster model involves a k input of 8 with the attributes we have used in the previous 3 models with the same default settings implemented. From the 8 cluster groups, we will focus on the cluster segments that attains the highest percentage towards IS\_BOUND. In this model, cluster 1 has 8286 instances and looks for quotes in July with a 51% chance of being bound for a quote with CGIC. The segment of cases in this cluster represents middle aged to young (approx. 31) females who are single and are living in Southwestern Ontario. This group uses their vehicle for work purposes, with a commute distance of 14 kms and 15,715 kms driven annually. The group has a license for an average of 11 years and has not purchased any products from CGIC in the past. The mean vector for VEHICLEMAKE is Mazda with the vehicle being manufactured around 2008. Cluster 3 has 3515 instances and shows us that this segment of cases represents middle aged (approx. 47) females, who are married and are living in New Brunswick. The group tends to use their vehicle for work purposes with their commute distance being 13 kms and 15,202 on an annual basis. This group has been licensed for 25 years and has not purchased any products from CGIC in previous years. In addition, CHEVROLET is the mean vector in this cluster with vehicles manufactured in 2010. This group tends to look for quotes in July and has a 45% chance of being bound for a quote with CGIC. For comparison purposes with other clusters, visualizations for the 8 clusters are shown in *Figure 24*.

We can now compare our results from the 4 cluster models we developed with our initial insights/patterns obtained in the data exploration stage. In the data exploration, we discovered an underlying trend between AGE and MULTI\_PRODUCT where older customers (=> 42) are more likely to have a ‘yes’ MULTI\_PRODUCT value with CGIC. We can confirm this trend through each of the cluster models developed, as individuals in model 1 (k=2) cluster 0; model 2 (k=4) cluster 0 and 3; model 3 (k=6) cluster 0, 1, and 3; and model 4 (k=8) cluster 0 and 7, are all older than the mean (42) and are existing clients of the company. Our second comparison is with the MARITAL\_STATUS and MULTI\_PRODUCT relationship we discovered in the initial data exploration. We can confirm this trend through each of the cluster models we developed, as individuals who are married in model 1 (k=2) cluster 0; model 2 (k=4) cluster 0 and 3; model 3 (k=6) cluster 0, 1, 3, and 5; and model 4 (k=8) cluster 0, 6, and 7 are all existing clients of CGIC.

Our third comparison is with the VEHICLEUSE and AGE model that was developed in our initial phase. The relationship between the two variables displays a trend where individuals who are younger are more likely to use their vehicle for work purposes while older individuals are more likely to use their vehicle for leisure. The split point to distinguish older individuals from younger is 50. Examples of this case are shown in model 1 (k=2), cluster 0 where a 50-year-old male would use their vehicle for leisure, meanwhile, in cluster 1, a 33-year-old female would use their vehicle for work. This trend can be shown in each model and each cluster, where individuals below the age of 50 will use their vehicle for work, and individuals above the age of 50 will use their vehicle for leisure. In the rare case, model 3 (k=6) cluster 2 has a mean vector of 37 for age and leisure for VEHICLE\_USE, and model 4 (k=8) cluster 2 has a mean vector of 40 for age and leisure for VEHICLE\_USE. These can be distinguished as potential outliers.

**Recommendation**

Once each type of model and their corresponding experiments were completed, we selected three final predictive models to be compared using ROC scores, average square error, and cumulative lift scores. The three models were:

1. Maximal Decision Tree Model
2. Neural Network with Average Error
3. Relaxed ASE Stepwise Model

Each of these models had rigorous assessments conducted on them; the final selection experiment was completed on validation data through SAS EM’s model comparison node.

For the Relaxed-ASE-Stepwise Model, we had a misclassification rate of 0.2193, and an average square error of 0.17139. We also had a static cumulative lift at 1.00.

For the Neural Network with Average Error, we had a misclassification rate of 0.2196, and an average square error of 0.171377. Our cumulative lift value was 1.032542.

For the Maximal Decision Tree, we had a misclassification rate of 0.21959, and an average square error of 0.171150. Our cumulative lift for this model value was 1.00.

Through this assessment and considering the individual depth of the predictive models themselves, we determine that the Maximal Decision Tree Model is the most comprehensive model for the scope of this project. Although the cumulative lift value is low, the maximal decision tree is the least inaccurate (Misclassification) out of the options. The advantages of this model are that it is comprehensive, detailed, and easy to use. The neural network is another viable recommendation as it has the highest forecast added value of any predictive model (Cumulative Lift). However, the neural network is more complex and has a ‘Blackbox’ nature, which means we are unable to understand the process behind how a value was predicted. If this project had more data available, we assume that the neural network would become more comprehensive – and better than the maximal decision tree. But, for the solution or recommendation to the CGIC, we recommend the **Maximal Decision Tree Model.**

# **Table of Figures**

|  |  |
| --- | --- |
| Figure 1: Data Preparation Summary | Page 6 |
|  | |
| Figure 2: Data Partition Summary | Page 7 |
|  | |
| Figure 3: Likelihood Ratio Test for Default Stepwise Model | Page 7 |
| Table  Description automatically generated with medium confidence | |
| Figure 4: Maximum Likelihood Estimates Analysis for Default Stepwise Model | Page 7 |
| Table  Description automatically generated | |
| Figure 5: Cumulative % Response Iteration Plot for Default Stepwise Model | Page 7 |
| Chart, line chart  Description automatically generated | |
| Figure 6: Likelihood Ratio Test for Relaxed Stepwise Model | Page 8 |
| Table  Description automatically generated with low confidence | |
| Figure 7: Maximum Likelihood Estimates Analysis for Relaxed Stepwise Model | Page 8 |
| Table  Description automatically generated | |
| Figure 8: Average Square Error for Relaxed Stepwise Model | Page 8 |
| Chart, line chart  Description automatically generated | |
| Figure 9: Maximum Likelihood Estimates for Relaxed ASE Stepwise Model | Page 8 |
| Table  Description automatically generated | |
|  |  |
|  |  |
|  |  |
|  |  |
| Figure 10: Final Decision Tree View | Page 10 |
|  | |
| Figure 11: Logworth of Variables Under Split Node 1 | Page 10 |
| **Table  Description automatically generated** | |
| Figure 12: Subtree Assessment – Misclassification Rate | Page 10 |
|  | |
| Figure 13: Convergence Criterion – Average Error NN | Page 11 |
|  | |
| Figure 14: Neural Network Diagram | Page 11 |
|  | |
| Figure 15: Properties – Misclassification Rate | Page 11 |
|  | |
| Figure 16: Misclassification – Iteration Plot and Score Rankings | Page 11 |
|  | |
| Figure 17: Properties – Average Error | Page 12 |
|  | |
| Figure 18: Average Error – Iteration Plot and Score Rankings | Page 12 |
|  | |
| Figure 19 and 20: K-Means, Num 2 and Visualization | 12 and 13 |
|  | |
| Figure 21 and 22: K-Means, Num 4 and Num 6 | 13 and 14 |
| A screenshot of a computer  Description automatically generated with low confidence | |
| Figure 23 and 24: K-Means, Num 6 Visualization and Num 8 Visualization | 14 and 15 |
|  | |
|  |  |
|  | |