**Co-operators General Insurance Company Report**

Project member: Weini Lei, Zhenyue Qian, Lu Sun, Yueyue Wang, Jinghui Wu

1. **Executive Summary**

The Co-operators General Insurance Company is a leading Canadian multi-line insurance and financial services company. Auto insurance is one of its businesses. Every month, tens of thousands of customers log on to CGIC's website to fill in the information and obtain relevant quotations for auto insurance. After comparing different prices, customers will decide whether they are bound to CGIC or not. To improve the company’s work efficiency, we will help CGIC understand the characteristics of those customers who have bound quotations in the past. Then, create a predictive model, that is, based on the information provided by the new customer, the probability of the new customer's success bound can be effectively predicted. The predictive model can help CGIC determine which quotations should be focused on. Decision tree, logistic regression, and artificial neural network technique will be applied to set up the model. In addition, clustering will be used to capture the natural structure of the data, helping CGIC to dig out customer profiles and target groups.

In the first phase, we conducted data exploration, visualization and preprocessing on 27 attributes and 101981 auto insurance quotes in 2016. We dropped 12 attributes because some attributes have missing values exceeding 90%, and some attributes are the threshold for whether the company provides insurance. Impute missing values with median value and global average level. Detect and deal with outliers and noise values accordingly. Finally, the data used for prediction includes 7 numerical variables, 2 interval variables, and 6 categorical data with 101,427 quotation data. Summary, steps, and key findings from data exploration as following.

1. **Summary of Key Data**

When pre-processing the data, we obtained the ranking of the importance of the data through ranking. The first five items are Years\_Licensend, Multi\_Product, Province, Customer\_age, Annual\_KM.

* Years\_Licensend, we detect 46 records out of control that is larger than 73. The statistics info: Mean:20.75, Median:17, Mode:3, S.E.:0.05, Min:0, Max:135. In general, the length of customers who holds a driver's license is concentrated between 6 and 33 years.
* Multi\_Product, there is no missing data. When bound, the proportion of Multiple Products is much greater than that of No Multiple Products.
* Province, customers concentrated in Ontario is much more than the sum of other provinces. This will be a more practical help for our subsequent forecasting model.
* Customer\_age, is a new data we added. The statistics info: Mean:12.97, Median:41, Mode: 25, S.E.:1.52, Min:-7983, Max:98. For different age groups, the amount of non-bound is larger than bound.
* Annual\_KM, for the original data, the amount of missing data is quite large, and also more outliers.
  1. **Attributes to Delete**
* Marking system & Tracking system, delete this column because more than 99.98% of the data shows that none of the systems is installed.
* Area code, delete because it corresponds to the cell phone number area. The cell phone is not restricted by region and cannot reflect correctly where the customer is located.
* Years as principal driver, more than 99% of the data is missing.
* Occupation, not-known or the blank occupies more than 90%.
* Vehicle model, we decide not to use because there are too many types of car and very difficult to reclassification.
* Conviction & Assigned & Suspend columns, according to the client’s policy, we do use these columns for predictive modelling.
  1. **Deal with Outliers, Noise & Missing values**
* Vehicle year: To calculate **Usage year** by using the 2016 mins record year, we transform this variable by using a log function before we apply the logistic regression technique.
* Annual KM: There are 488 outliers because about 99.52% of commute distances are below 40,000 km, so we replace data according to the median of the inliers of the corresponding vehicle use, 20,000 for business use, 10,000 for pleasure use, and 15,000 for commute use and other categories.
* VEHICLE\_OWNERSHIP：There are 75304 blanks, we fill the cell of Blanks randomly according to the equivalence of the ratio “Owned: 73%”, “Lease: 18%”, “No-owned: 9%”, this proportion data come from Statista's statistics.
* YEAR\_OF\_BIRTH: To calculate Customer age through using 2016 mins record year. According to the laws of different provinces in Canada, we concluded that the minimum age that can get a driver's license in Canada is 14, so the data under 14 are outliers and delete. Then we categorize customer age into 5 groups, “Teenager: 15-20”, “Young-age：21-40”,” Middle-age: 41-60”, ” Elderly: over 60”.
* YEARS\_LICENSED：According to the box plot, we can detect 46 records out of control that is larger than 73，so we drop these data.
* GENDER: There are 5 blanks and 1 unknown data, so we delete these 6-noise data.
  1. **Category attributes**
* Vehicle maker: Group them into few categories by country of the make.
* VEHICLE\_USE: The attribute has 8 categories. We regroup it into 4 categories, Business, Commute, Pleasure, and merge the remaining 4 types into Other categories.
* POSTAL\_CODE: We sort the initial letter of values and replace them to Province by correspondent province name. Then category the record into 2 groups: “Ontario”, “Other province”.

1. **Key Findings and Visualization**

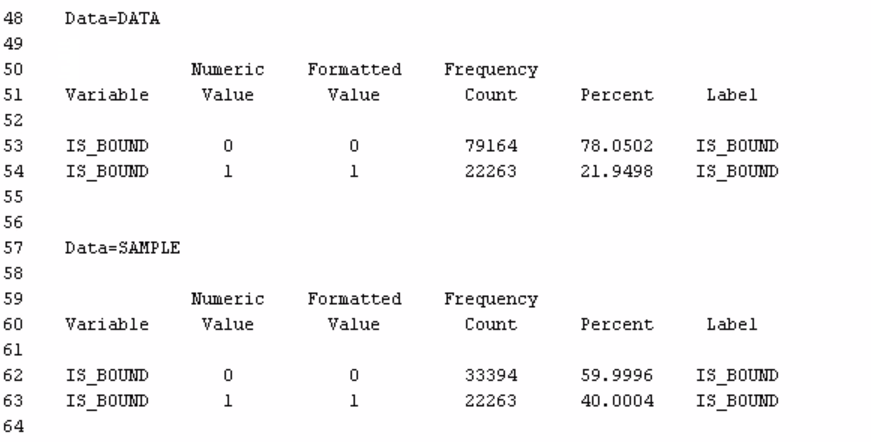
|  |  |  |
| --- | --- | --- |
| **Important Attribute** | **Findings** | **Exhibit** |
| **YEARS\_**  **LICENSED** | The more years licenses own, the less likely bound customers are. Thus, the clients with less driving experience are the company’s target. |  |
| **MULTI\_**  **PRODUCT** | we can see that when bound, the proportion of Multiple Products is much greater than that of No Multiple Products. |  |
| **Province** | The group from Ontario is our major customer. |  |
| **Customer\_**  **age** | The group of 21-40 years old and 41-60 years old are mainly ones bound to quote. Thus, the customer group targets 21-60 years old. |  |
| **Vehicle\_**  **country** | Vehicles made from America, Japan, Italy, Germany, and South Korea are a majority of car types. And American and Japanese cars are most popular and more likely to bound. |  |

1. **Prediction Techniques Used and Key Results** 
   1. **Preparation for prediction**

Below attributes are used for prediction.

|  |  |  |
| --- | --- | --- |
| **Name** | **Measurement  Level** | **Description** |
|  |
| ID\_VARIABLE | Interval | Quote Identification Number |  |
| Month | Interval | The month that the client was quoted |  |
| Usage\_Year | Interval | The number of years from the vehicle was produced to 2016 |  |
| Vehiclecountry | Nominal | The country where the vehicle maker is located |  |
| ANNUAL\_KM | Interval | Number of kilometres the client drives on an annual basis provided by the client |  |
| COMMUTE\_DISTANCE | Interval | Distance of the client’s commute from their home to their place of work provided by the client (in kilometres) |  |
| VEHICLE\_OWNERSHIP | Nominal | Is the vehicle owned, non-owned, or leased |  |
| VEHICLE USE | Nominal | The main use of the vehicle |  |
| GENDER | Nominal | Client gender |  |
| Customer\_age | Interval | The age of Clients until 2016 |  |
| Province | Nominal | The province where the client is located |  |
| YEARS\_LICENSED | Interval | The number of years the client obtained the license until 2016 |  |
| MULTI\_PRODUCT | Binary | Have other insurance products with us (Home, Life, Farm, etc.) (results in a 10% discount in premium) |  |
| MARITAL\_STATUS | Nominal | Marital Status of the client |  |
| Is\_Bound | Binary | Target: The quote became bound (1 = Yes; 0 = No) |  |

Since 78% of the original data is no bound and 22% bound, it is easy to produce no bound deviations from the results, so we sampled the data for prediction, which no bound is 60%, and bound is 40%, as follows,



We choose the ROC index to measure the performance of models because the purpose of this predictive model is to help CGIC identify which new businesses are most likely to be bounded. ROC index is mainly used to test the accuracy of the model. The higher the index, the higher the accuracy of the model.

* 1. **Decision Tree**

Firstly, we use the StatExplore node to understand the relative importance of the variables for the prediction task. We choose to use the Gini index to rank inputs in terms of prediction importance. The top 10 inputs are YEARS\_LICENSED, MULTI\_PRODUCT, Province, Customer\_age, ANNUAL\_KM, COMMUTE\_DISTANCE, useage\_year, Vehicle\_country, Martial\_Status, VEHICLE\_USE.

Then, we add a Data Partition node and assign several settings:

a). 75% of the data for training and 25% for validation. Under this setting, we found that only 8,000 data were recorded out of the 30,000 data sampled, and the value of the ROC index was small under different Decision Tree, so this setting was not the best. Then, we have tried several other settings.

b). 65% of the data for training and 35% for validation.

c). 60% of the data for training and 40% for validation.

d). 55% of the data for training and 45% for validation.

e). 50% of the data for training and 50% for validation.

We find that the value of the ROC Index is "bell-shaped" at different ratios, and the value of the ROC Index is largest when 60% of the data for training and 40% for validation, and then as the percentage of training changes, the value of ROC Index decreases gradually. Therefore, 60% of the data for training and 40% for validation is the best model.

Next, we explore the decision tree for different options. Firstly, we use the Average Square Error as the model assessment statistic. Run a decision tree model.

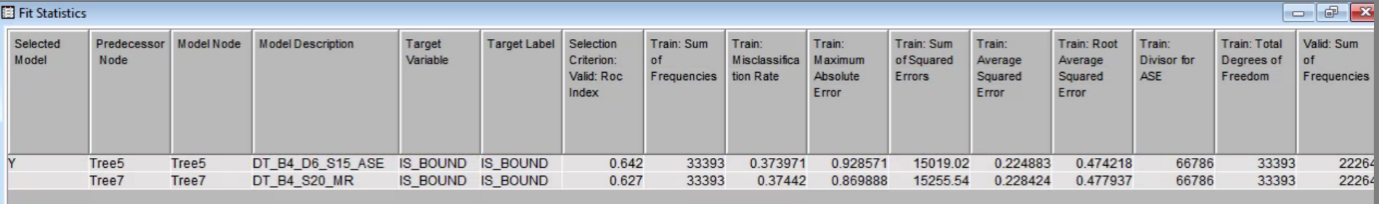
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Selection Method** | **Regression Name** | **Parameter setting** | **Best model (within)** | **Model Comparison (between)ROC** |
| Average Square Error  (different Branch) | DT \_Default | Default | DT\_B4\_ASE | The final Decision Tree model is DT\_B4\_D6\_S15\_ASE. The ROC index is 0.642 |
| DT\_B5\_ASE | Branch 5, Tree depth 6, Leaf size 5 |
| DT\_B4\_ASE | Branch 4, Tree depth 6, Leaf size 5 |
| DT\_B3\_ASE | Branch 3, Tree depth 6, Leaf size 5 |
| Average Square Error  (different Depth) | DT\_B4\_D6\_ASE | Branch 4, Tree depth 6, Leaf size 5 | DT\_B4\_D6\_ASE |
| DT\_B4\_D5\_ASE | Branch 4, Tree depth 5, Leaf size 5 |
| DT\_B4\_D4\_ASE | Branch 4, Tree depth 4, Leaf size 5 |
| DT\_B4\_D3\_ASE | Branch 4, Tree depth 3, Leaf size 5 |
| Average Square Error  (different Leaf size) | DT\_B4\_D6\_ASE | Branch 4, Tree depth 6, Leaf size 5 | DT\_B4\_D6\_S15\_ASE |
| DT\_B4\_D6\_S10\_ASE | Branch 4, Tree depth 6,  Leaf size 10 |
| DT\_B4\_D6\_S15\_ASE | Branch 4, Tree depth 6,  Leaf size 15 |
| DT\_B4\_D6\_S20\_ASE | Branch 4, Tree depth 6,  Leaf size 20 |
| DT\_B4\_D6\_S30\_ASE | Branch 4, Tree depth 6,  Leaf size 30 |

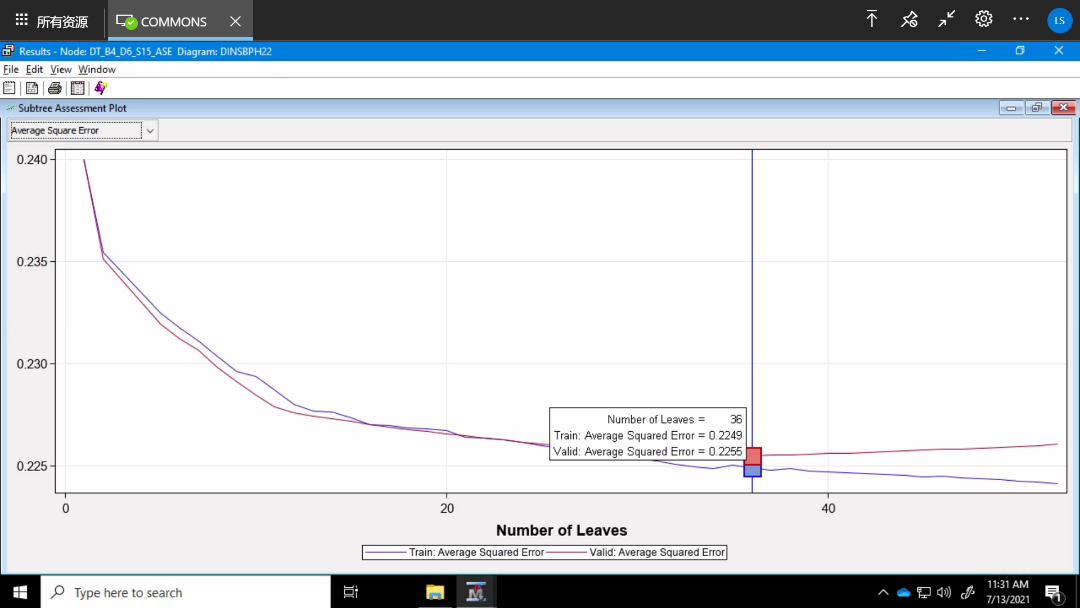
Secondly, we use Misclassification Error as the model assessment statistic. Run a decision tree model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Selection Method** | **Regression Name** | **Parameter setting** | **Best model (within)** | **Model Comparison (between)ROC** |
| Misclassification Error  (different Branch) | DT \_Default | Default | DT\_B4\_MR | The final Decision Tree model is DT\_B4\_S20\_MR, The ROC index is 0.627 |
| DT\_B5\_MR | Branch 5, Tree depth 6, Leaf size 5 |
| DT\_B4\_MR | Branch 4, Tree depth 6, Leaf size 5 |
| DT\_B3\_MR | Branch 3, Tree depth 6, Leaf size 5 |
| Misclassification Error  (different Depth) | DT\_B4\_MR | Branch 4, Tree depth 6, Leaf size 5 | DT\_B4\_MR |
| DT\_B4\_D5\_MR | Branch 4, Tree depth 5, Leaf size 5 |
| DT\_B4\_D4\_MR | Branch 4, Tree depth 4, Leaf size 5 |
| DT\_B4\_D3\_MR | Branch 4, Tree depth 3, Leaf size 5 |
| Average Square Error  (different Leaf size) | DT\_B4\_MR | Branch 4, Tree depth 6, Leaf size 5 | DT\_B4\_S20\_MR |
| DT\_B4\_S10\_MR | Branch 4, Tree depth 6, Leaf size 10 |
| DT\_B4\_S15\_MR | Branch 4, Tree depth 6, Leaf size 15 |
| DT\_B4\_S20\_MR | Branch 4, Tree depth 6, Leaf size 20 |
| DT\_B4\_S30\_MR | Branch 4, Tree depth 6, Leaf size 30 |

Thirdly, we compared the final best Average Square Error decision tree model and Misclassification Error decision tree model. Because the DT\_B4\_D6\_S15\_ASE tree has the schedule model of Y, the ROC index for the Average Square Error decision tree model is higher than the Misclassification Error decision tree model.

Therefore, the best Decision Tree model is DT\_B4\_D6\_S15\_ASE, the average squared error on validation data and training data are 0.2255 and 0.2249 separately. The number of leaves is 36. Following are the result,





**Analysis of Final Decision Tree**

First of all, because we selected 60% Train and 40% Validation, we believe that more than 40% of Bound's data are good data, and we will extend the analysis and conclusions to the superior node in turn. In addition, we believe that the data whose Bound Rate difference between Train and Validation is greater than 2 is not good, so it is not presented as a conclusion. On this basis, we carefully looked at all the nodes and came to the following conclusions:

**1.**Customer outside Ontario

In the customer group outside Ontario, only customers meet the following three conditions at the same time: (a). No multiple products or missing. (b). The driving license is between 7.5 years and 18.5 years or missing. (c). The vehicle service life is greater than or equal to 6.5, then the Bound rate of the annual customer base is greater than 40% which equals 43.45%, and the gap between Train data and Validation data (42.75%) is not large.

Diagram, timeline

Description automatically generated with medium confidence

**2.** Customers in Ontario or missing

Among the customers in Ontario or the province that is missing, the group with a Bound rate greater than 40% accounts for the majority.

* For people who have multiple products, the driver’s license is ≥7.5 and <20.5 or missing, and customer age is ≥24.5 and <26.5 or missing has the highest Bound rate, which is Train 68.23% and Validation 67.12%
* At the same time, for people who have multiple products, the driver’s license is ≥7.5 and <20.5 or missing, and customer age is ≥26.5 and <36.5 or missing, the proportion of train data and validation data in the Bound rate of the customer group is the same as 58.59%.

Combining the above two situations, we concluded that for people who have multiple products, the driver’s license is ≥7.5 and <20.5 or missing, and customer age is ≥24.5 and <36.5 or missing, the Bound rate of this kind of customer group is better. And it can be used as the company's main target group, and it will be mainly targeted at this group of people in subsequent activities and publicity.

* People who have the license year ≥11.5 and <20.5, and the age ≥36.5 and <42.5 have the Bound rate greater than 40%, which is Train 48.30% and Validation 48.82%. In this node, whether or not the group has multiple products will not affect their Bound rate.

Diagram

Description automatically generated

* For those whose Year\_Licensed ≥20.5 and also has multiple products, regardless of any age group, the Bound rate is always greater than 40%. For those people, the customer whose age is less than 42.5 has the highest Bound rate, which is Train 51.6%, Validation 50.15%, followed by people who is ≥54.5 or missing, with the Bound rate as Train 46.77% and Validation 45.44%, the least Bound rate is Train 41.36% Validation 42.42% for people who aged from 42.5 to 54.5.

Diagram

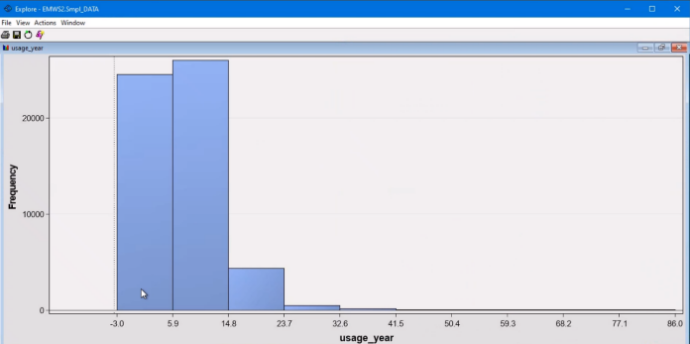
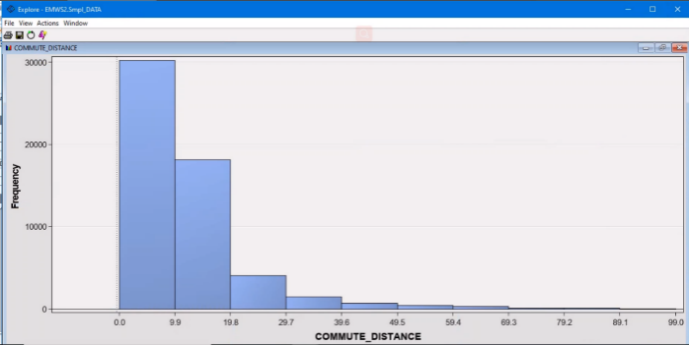
Description automatically generated

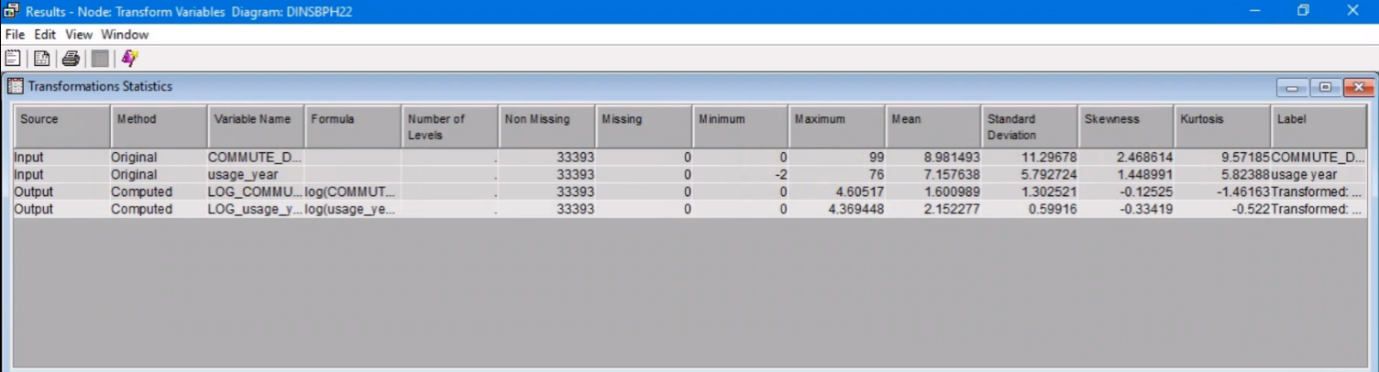
In general, in the decision tree, the main split variables are province, customer age, Year\_Licensed, and Multi\_Products, which are basically consistent with the rank conclusion we got when we first used StatExplore.

For practical applications, the bound rate of people with multiple products will generally be higher, the age distribution is distributed from 24.5 years old, and the Year\_Licensed distribution is also distributed from 7.5 years later. This is also consistent with the actual situation. In the correlation, we have already learned that Year\_Licensed is positively correlated with the age of the customer. Young drivers will cherish the use of vehicles more and will use more insurance to ensure the safety of their vehicles. For older customers, insurance is a protection for their life safety.

* 1. **Logistic Regression**

Before we take the logistic regression, we need to examine the distribution of inputs to avoid an exaggeration of the result. We explore all the interval variables and then find that the input of COMMUTE\_DISTANCE and usage\_year is left-skewed. Thus, we plan to change from Default to Log for the two variables in the Transform Variables node.



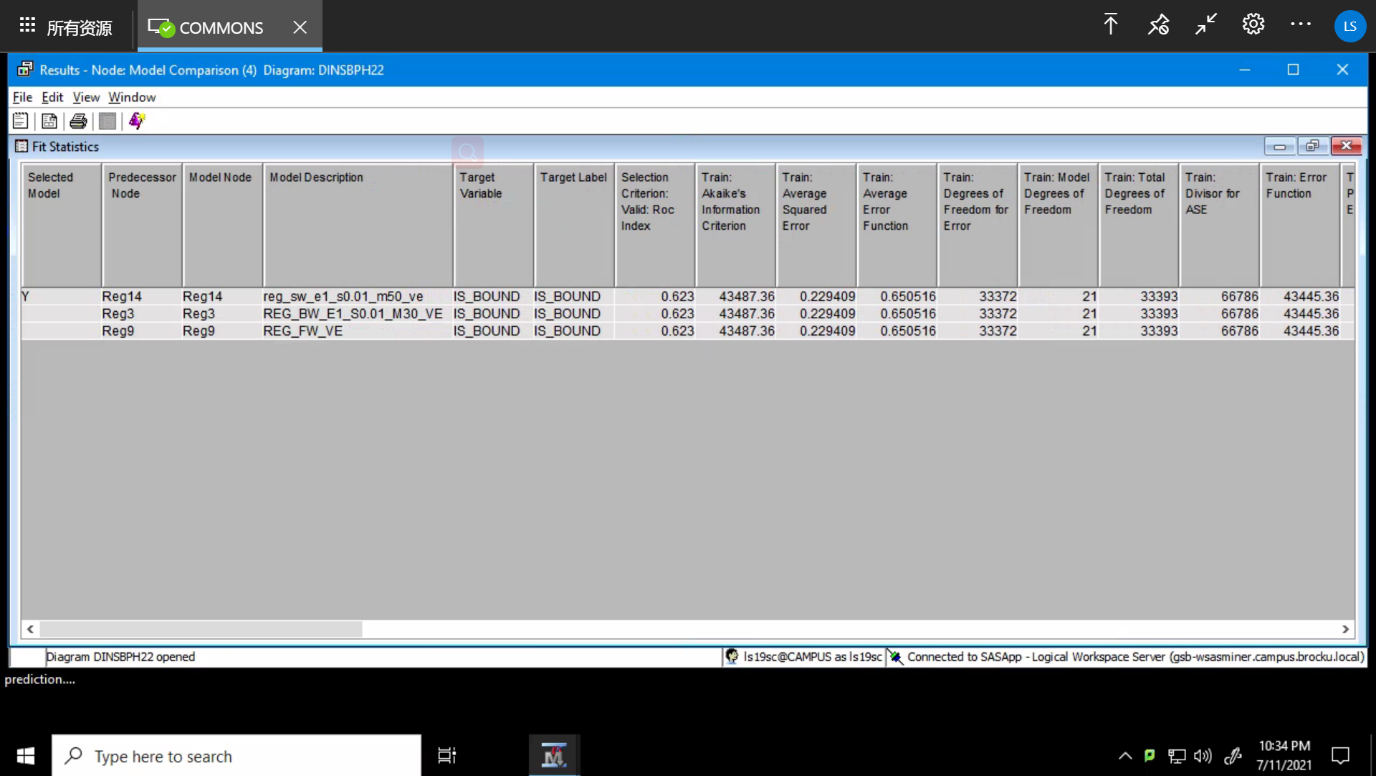


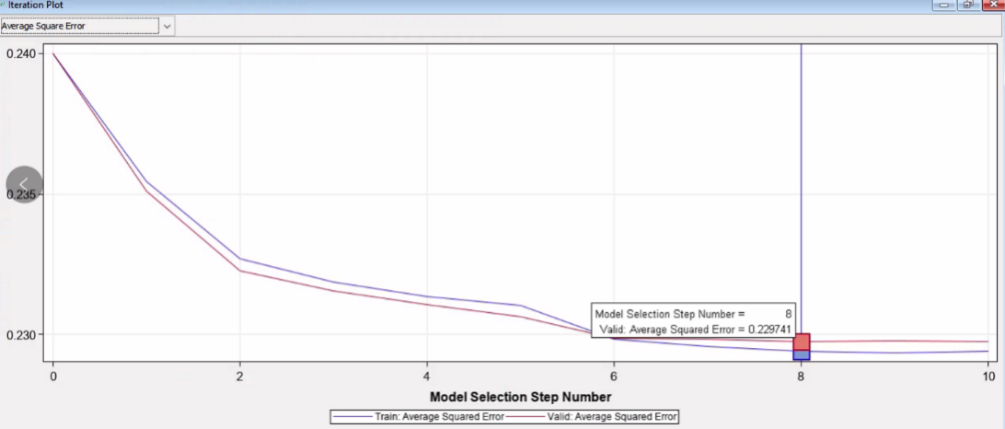
The following steps are building a series of models based on methods of **backward, forward, and stepwise** selections in various settings.

|  |  |  |  |
| --- | --- | --- | --- |
| **Selection Method** | **Regression Name** | **Parameter setting** | **Model Comparison (within)** |
| **Backward** | Reg | Default | Reg 3 |
| Reg2 | Entry1, stay0.5, Max\_step30, P-value |
| Reg5 | Entry1, stay0.05, Max\_step30, P-value |
| Reg3 | Entry1, stay0.01, Max\_step30, VE |
| Reg4 | Entry1, stay0.01, Max\_step30, MR |
| **Forward** | Reg6 | Default | Reg9 |
| Reg7 | Entry1, stay0.05, Max\_step30, P-value |
| Reg8 | Entry0.5, stay0.05, Max\_step30, P-value |
| Reg9 | Default, Validation Error |
| Reg10 | Default, Misclassification Rate |
| **Stepwise** | Reg11 | Default | Reg14 |
| Reg12 | Entry1, stay0.5, Max\_step30, P-value |
| Reg13 | Entry0.5, stay1, Max\_step30, P-value |
| Reg14 | Entry1, stay0.01, Max\_step50, VE |

After the number of trials, we firstly set different parameters in Backward, Forward, and Stepwise selection methods respectively. Then, we choose the best ones in three methods. Based on the best models won in the test, we do the model comparison in terms of the ROC curve, and we find the Reg 14 is the best one.

**Results:** The number of iteration is 8. The inputs selected are COMMUTE\_DISTANCE, Customer\_age, LOG\_usage\_year, MARITAL\_STATUS, MULTI\_PRODUCT, Province, Vehicle\_country and YEARS\_LICENSED. The Chi-Square of all variables is above 10. The Average Squared Error for the train and the valid data set is 0.22941 and 0.229741.



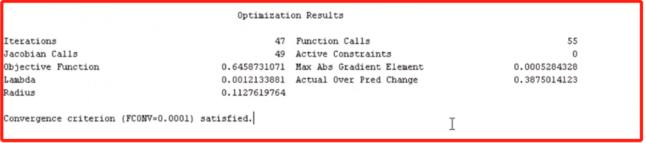


**Analysis of Odd Ratio:**

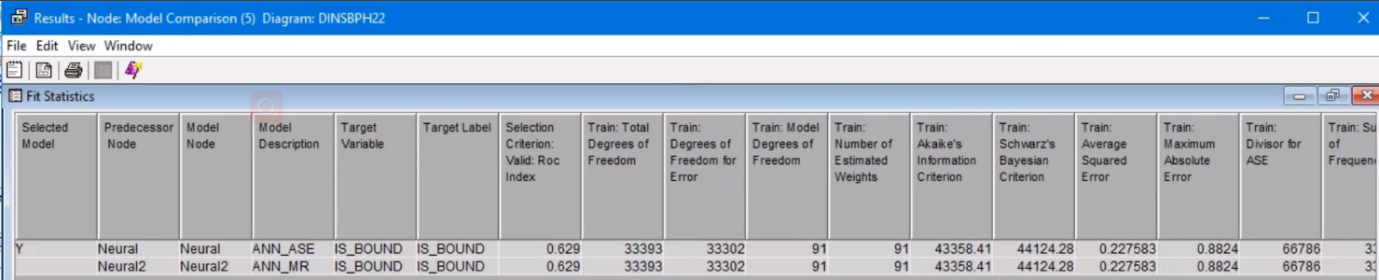
* COMMUTE\_DISTANCE, the odds ratio estimate equals 0.994. This means that for every kilometre increased for a commute from their home to the place of work, the odds of being bound change by a factor of 0.994, a 0.6% decrease.
* Customer\_age, the odds ratio estimate equals 0.976. This means that for each age of a driver increased, the odds of being bound change by a factor of 0.976, a 2.4% decrease.
* LOG\_usage\_year, the odds ratio estimate equals 1.110. This means that for each using year raised, the odds of being bound change by a factor of 0.110, an 11% increase.
* MARITAL\_STATUS, when the comparison group is Single and Widow, the odds ratio estimate equals 0.933. This means that cases with single status are 0.933 times less likely to be bound than cases with widow status. when the comparison group is Married and Widow, the odds ratio estimate equals 0.798. This means that cases with married status are 0.798 times less likely to be bound than cases with widow status. Divorced and Widow group has the similar result. Whereas, when the comparison group is Separated and Widow, the odds ratio estimate equals 1.107. This means that cases with separated status are 1.107 times more likely to be bound than cases with widow status.
* MULTI\_PRODUCT, the odds ratio estimate equals 0.563. This means that cases with a value of ‘N’ are 0.563 times less likely to be bound than cases with a value of ‘Y’.
* Province, when comparison group in Ontario and other provinces, the odds ratio estimate equals 1.969. This means that residents in Ontario are 1.969 times more likely to be bound than in other provinces.
* Vehicle\_country, when the comparison group is America and Swedish, the odds ratio estimate equals 1.176. This means that vehicles producing from America are 1.176 times more likely to be bound than vehicles producing from Swedish. Whereas, when the comparison group is Canada and Swedish, the odds ratio estimate equals 0.355. This means that vehicles producing from Canada are 0.355 times less likely to be bound than vehicles producing from Swedish.
* YEAR\_LICENSED, the odds ratio estimate equals 1.020. This means that each year of owing license raised, the odds of bound change by a factor of 0.020, a 2% increase.
  1. **Artificial Neural Network**

We created two neural models: **ANN\_ASE and ANN\_MR**, which compares with decision tree and regression models to figure out which model is best for this case.

**1. ANN\_ASE:** We set the model criterion as Average Square Error with the maximum iteration of 50 in this node. The primary training has also been set as No. Under this setting, the model could generate the fewest average error for the validation set. The optimization result shows iteration value is 47, which indicates the model will run at least 47 times to reach convergence.



**2.ANN\_MR:** The second ANN model was created and run under the criterion of Misclassification Rate. Same as Average Square Error, we set 50 as the maximum number of iterations and No to preliminary training. As the result, the ANN with an Average squared error has better performance. (Refer to the table below)

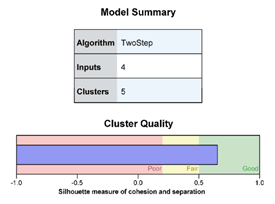
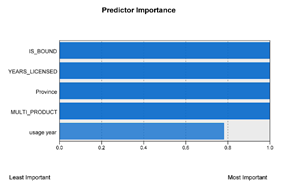


1. **Clustering and Results**

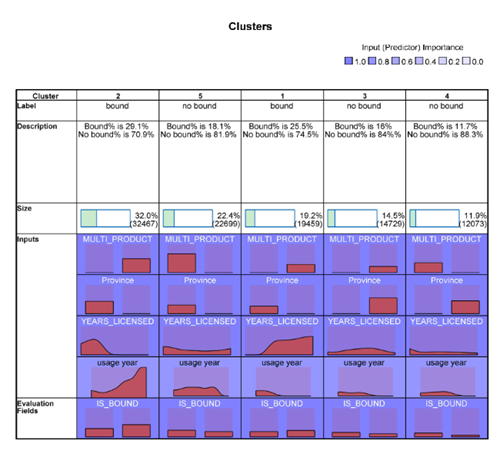
Through data exploration and pre-processing, we obtained 12 valid variables and reclassified them into three categories: Driver, Vehicle, Others, as follows,

* Driver: Gender, Customer\_Age, Years\_Licensed, Marital\_Status
* Car: Vehicle\_Country, Vehicle\_Usage\_Year, Vehicle\_Ownership, Vechicle\_Use
* Others: Province, Annual\_km, Commute\_Distance, Multi\_Product

Then select one to two independent variables from each category, which can represent that category, and conduct clustering through the SPSS Two-Steps technique. After many multi-dimensional experiments, we found 5 clusters with 4 inputs (Years\_Licensed, Vehicle\_Usage\_Year, Province, Multi\_Product) are a better clustering with good quality. Four inputs are also ranked as the most important predictors.

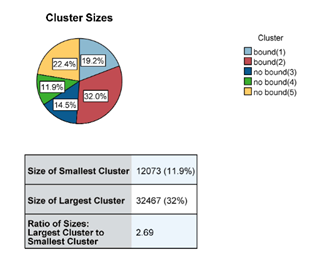
Since the raw data is 22% bound and 78% no bound, so when a cluster bound% exceeds 22%, we think that the customer of this segment has a higher probability of responding to the bound. The bound% of cluster 1 and 2 are over 22%, which are 25.5% and 29.1% separately. The remaining clusters are lower than 22%. It means that the customer segmentation of cluster 1 and 2 are more likely to be bound than other clusters. In addition, the bound % of cluster 2 is higher than cluster 1.



According to the pivot table, the common characteristics of Cluster 1 and 2 are that the vehicle is owned, other insurance products of CGIC have been purchased, and the location of Cluster 1 and 2 is Ontario. Unique characteristics of Cluster 1 and 2 as following,



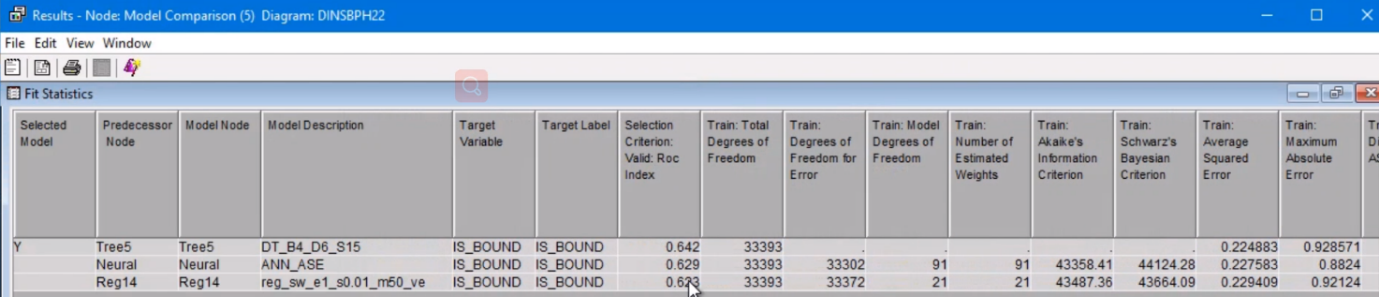
Overall, from the below bar chart, we found that cluster 1 and 2 accounted for 51.2% of the total, among which cluster 2 accounted for 32%. CGIC can first contact cluster 2 to increase the probability of bound.



The results of clustering are similar to what we found when we did data exploration. Clustering divides the customer's group into more details. For instance, clustering narrows the age range of customers to 21-40 years old. The vehicle is owned by the customer and the main purpose is commuting. This also makes us more convinced of the correctness and validity of the key findings from data exploration.

1. **Final Selected Techniques and Model Results**

Finally, we chose Decision Tree 5, Reg 14, and ANN\_ASE shown in the Node Model Comparison. Through the data indicated in the ROC index, we will use **Decision Tree 5**, which stands for as final model which shows ROC is 0.642.



**Results:** Customers who meet these characteristics are likely to bound the quotation: a) Live in Ontario, b) Age between 24 and 37 years old with a driver's license for 7 to 21 years, c) Own American or Japanese cars, mainly for commuting, d) Purchased other products of CGIC, e) Marital status is single or married. CGIC should focus on these customers’ quotations at first.

1. **Recommendations**
2. Younger consumers are more likely to buy insurance because they are safety conscious and have a high rate of car usage. Furthermore, they have some driving experience, have some understanding of driving insurance and are more aware of what kind of insurance they need. Therefore, companies can take advantage of these purchasing preferences to conduct marketing campaigns, such as posting advertisements near the finance street so that let more consumers in this category see the advertisement.
3. According to the prediction results, customers who have bought other insurance products are more willing to buy car insurance from the company. This indicates that customers have a high level of trust in the company and do not want to change insurance companies frequently. Also, these customers may consider the 10% discount in premium. Therefore, the company should focus on how to maintain the regular customers and meet their needs more often. For example, it can arrange a docking service, proactively ask if customers need to buy car insurance together when they buy other insurance products and offer more discount policies to these customers. Secondly, the company can try to improve service quality, such as simplifying the purchase process to make it easy for regular customers to buy.