**Developing Predictive Models for The Co-operators**

ITIS 4P21: Introduction to Business Analytics

Group 4, Section 1

December 3rd, 2021

James Boyle - 6196885

Naim Boufridi - 6357859

Shane Matthew Sodhi - 5362231

Maja Leonie Hermann - 7346299

Janvier Kurawige - 6361794

**Executive Summary**

The Co-operators General Insurance Company receives a large volume of requests for auto insurance quotes each month. Many of these quotes are not actioned as customers tend to ‘shop around’ and CGIC does not have the capacity to address all of the quotes they receive. CGIC provided historical data from 101,981 auto insurance quotes which includes details about the customer, their vehicle, their location and whether they were bound. This paper is the summary of a data analysis project conducted by five Brock University students with the goal of identifying which customers are most likely to be bound, thereby increasing CGIC’s customer bound rate significantly. The project utilized data mining techniques such as data visualization and predictive modeling to identify key attributes of the most likely bound customers which resulted in a substantial increase to the bound rate above the baseline rate of 22%. We propose prioritizing customers who live in Ontario, are 37 or younger, already have an insurance product from CGIC, and have been licensed for over 7.5 years. This segment makes up approximately 11% of customers and bound rates are predicted to be 37.65%.

**Pre-processing**

With any data analysis, it was an integral component of our success to thoroughly clean, explore, and prepare the data to generate predictive models. This step was split into multiple components including outlier and noise reduction, discretization, and variable selection for our models. To summarize our phase I work, we removed quotes with conflicting information, unusable information, and extreme outliers that reduced the overall quality of the data, rather than using imputation. Imputation was used in certain cases where missing variables made up a relatively small proportion of the data, in order to maintain full model accuracy.

For the sake of our final models, we needed to define and configure our Metadata to determine which input variables were going to be used. We will walk through our rationale for the variables that were rejected below: Quote date was rejected as a predictor because there is only 1 year of data, so there is no way to detect seasonality or cyclicism within the data. The GINI index confirms that quote date will have negligble impact on bound rate. Vehicle age Classification was rejected because we are limited by the bins we created in our discretization. If our bins are inaccurate/skewed, it will misrepresent the data classification, and SAS will automatically bin the observations by equal depth/width. Vehicle year was rejected, and Vehicle Age is used instead to avoid redundancy. Vehicle Model was rejected due to there being too many vehicle model categories to model effectively (+128). Vehicle Usage intensity was rejected as it is captured in a non-descritized manner through “annual KM”. Work Travel distance is rejected from the model, instead we opted to use the numerical “commute” classification, due to the possibility of our discretization bins being inaccurate/skewed. Vehicle ownership was rejected due to a high amount of missing values (76,021 Blanks) - This value exceeds a reasonable level of missing values for imputation. Vehicle value was rejected due to a large amount of missing values (64,673 Blanks). Year of birth and age range is rejected due to it not having a significant GINI score in comparison to the continuous classification of age. FSA was rejected since it has too many categories. FSA\_POPULATION\_SIZE and Province will be used as geographic indicators instead. Area code was rejected as it doesn’t provide true location data.

Driver experience level was rejected, as “years licensed” will be a much more effective measure, as it is not limited by the category discretization we created in driver experience level. Years of principal driver was rejected due to there being over 101,000 blank observations. Assigned Losses and Suspension count are rejected due to them being the lowest ranked attributes of importance, and there being very few values that deviate from 0. For occupation, we created a new category of occupation and converted it into either “known” or “not known” since each individual profession has little worth on it’s own due to large quantities of classification. For the full metadata breakdown, see Figure 4.

**Predictive Modelling**

**Decision Trees**

The purpose of developing this model is to concisely determine input variables which can be used to predict with confidence whether a customer is likely to be bound for a quote or not. After importing the data using a file import node , the metadata is checked and edited accordingly to align with decisions made in the preprocessing phase. Due to the high disproportion between bound and not bound records observed in the preprocessing phase, we will use sampling to create a fair distribution of records for the target variable. Our target variable will now have 60% (33,361) as “not bound” and 40% (22,241) as “bound”. This will ensure proportionate distribution of records which will facilitate effective learning of the algorithm.Now that our dataset has been sampled with a fair distribution of records, let's partition our data using a data partitioning node. After a deliberate experimentation with different partitioning sets, we have decided to go with 55% of records for training and 45% for validation. This data partitioning set has been chosen because the root node starts with a Multiproduct variable which is among highly ranked predicting variables .This variable is also binary hence a good starting point. There will be three main phases during this modelling. The first phase will use Average Squared Error as a subtree assessment measure. During this phase, several systematic experiments will be conducted to determine the best model which minimizes complexity and average squared error. The second phase will use misclassification error rate as a subtree assessment measure. Just like phase1 of modelling, there will be systematic experiments to determine the best model which minimizes model complexity and misclassification error rate. Lastly, the best model in phase 1 will be compared to the best model in phase2 using cumulative rift and ROC.

**PHASE1: AVERAGE SQUARED ERROR AS A SUBTREE ASSESSMENT MEASURE.**

**First experiment: Using default settings:** This experiment uses default settings to train the model. As a result, there are 24 optimal leaves generated. The average squared error for training is 0.2250 and 0.2259 for Validation.  
**Second experiment:** **increasing maximum branches while keeping other settings at default**: This experiment increases the number of maximum branches to 3 and then to 4. The aim is to find out if an increase in the number of branches improves the model. As a result, there are 29 optimal leaves for 3 branches and 31 optimal leaves for 4 branches. The average squared error for 3 branches is 0.2249 for training and 0.2263 for Validation. As for 4 branches, the average squared error is 0.2248 for training and 0.2260 for validation. At this point, increasing maximum branches is making the model more complex while failing to reduce average squared error for validation.  
**Third experiment: reducing and increasing maximum depth while keeping other settings at default:** This experiment uses a lower depth of 5 and then a high depth of 20. The aim is to find out if changing the number of depths improves the model. As a result, 20 optimal leaves are generated at 5 depth and 24 at 20 depth. The average squared errors at 5 depth are 0.2253 for training and 0.2261 for validation. Clearly, the average squared is not improving with smaller depth. As for 20 depth, the average squared error is 0.2250 for training and 0.2259, therefore there is no difference to the default experiment done earlier.  
**Fourth experiment: Increasing the leaf size to 100 while keeping other setting at defaul**t This experiment tries to increase the minimum records that a leaf can have from 5 to 100. This is done after observing that some splitting involves less than 100 records. In total, 28 optimal leaves are generated while the average squared error for training is 0.2246 and 0.2258 for Validation. Clearly, increasing the leaf size to 100 has minimized the average squared error.  
**Comparing experiments in phase 1 using average squared error to find the best:** After running a model comparison node using average squared as selection criterion with emphasis on validation data, it is not surprising that the model in the fourth experiment is the best. It has the lowest average squared error. Please refer to the decision tree appendix (Figure 2.2) to see the subtree assessment plot for this model.

**PHASE2: MISCLASSIFICATION RATE AS A SUBTREE ASSESSMENT MEASURE.**

**First experiment:** **Using default setting**: This experiment uses all default settings except changing the subtree assessment measure from average squared error to misclassification rate. As a result, there are 5 optimal leaves generated. The misclassification rate is 0.3732 for training and 0.3779 for validation.  
**Second experiment: reducing and increasing maximum depth while keeping other settings at default**: This experiment uses a lower depth of 5 and then a high depth of 20. The aim is to find out if changing the number of depths improves the model. As a result, the number of optimal leaves is 5 at both 5 depth and 20 depth. The misclassification error for validation is also similar at 0.3779. Clearly, there is an issue of underfitting at any maximum depth while other settings are at default.  
**Third Experiment: increasing the leaf size to 100 while keeping other settings at default**: This experiment tries to increase the number of observations a leaf can have from 5 to 100. This is done after observing that some splitting involves fewer than 100 records. After running the model, The misclassification rate for training is 0.3732 and 0.3779 for validation. At this point, increasing leaf size is not improving the model.  
**Fourth experiment: increasing maximum branches while keeping other settings at default**: This experiment increases the number of maximum branches to 3, 4 and then to 10. The aim is to find out if an increase in the number of branches improves the model. As a result, there are 10 optimal leaves for 3 branches and 12 optimal leaves for 4 branches. The misclassification rate for 3 branches is 0.3717 for training and 0.3759 for Validation. As for 4 branches, the misclassification rate is 0.3723 for training and 0.3757 for validation. Clearly, the misclassification error rate decreases as we increase the total number of branches.   
**Firth experiment: Increasing maximum branches to 10 and leaf size to 100 while keeping other settings at default.** Considering that one variable called province has 10 categories, let's increase the number of maximum branches to 10 with a pre-pruning technique of 100 leaf size. The results are even much better with 18 optimal leaves and a misclassification error rate of 0.3710 for training and 0.3746 for validation.   
**Comparing experiments in phase2:** After running a model comparison using misclassification error rate as a selection criterion with emphasis on validation data, it is not surprising that the model in the firth experiment with 10 branches and 100 leaf size is the best. It has a misclassification rate of 0.3710 for training and 0.3746. Please refer to the decision tree appendix (Figure 2.3) to see the subtree assessment plot for this model.

**Comparing the best model in Phase1 against the best model in Phase2 of Decision tree using ROC and cumulative lift at a selection depth of 20%:**Now that we know the best models using misclassification error rate and average squared error, let's compare the best model from each phase of the decision tree using cumulative lift. Based on the ratings of cumulative lift and ROC, the best model comes from phase1 of decision tree modeling where average squared error was used as a subtree assessment measure. This model uses 100 leafs while keeping other settings at default. It has a ROC rating of 0.639 and cumulative lift rating of 1.3977. Please refer to the decision tree appendix (Figure 2.1) to see the table and graph of cumulative lift.

**Regression Modelling**

From a perspective of predictive modelling, it’s important to utilize multiple tools to determine the best predictive approach. For the next component of our predictive model, we tested out different configurations of a Logistic Regression model to determine key indicators, and benchmark it to our final chosen model. To begin, we must discuss the model setup employed to build our regression models. First, we employed the use of a sampling node to better represent the proportion of Bound quotes to a 60-40 split to prevent imbalanced target outcomes, as discussed in our Decision Tree component of this paper. The next component of our regression model was the use of a data partitioning node. This step required a great deal of experimentation in order to build a model with the least possible amount of over/underfitting on our validation data. In order to test this, we experimented with each of our model types (Forward, Backward, and Stepwise) with different combinations of data partitioning. Throughout experimentation, we found that a partition of 65-35 (Train-Validate) delivered the lowest Misclassification rate on the validation data, a key signifier of model importance. To see the results of our partitioning experiment, refer to Regression results (Figure 1.1). After we determined the best partition of our data, the next required step was performing variable transformation on highly skewed variables. Since regression models are highly sensitive to skewed variables, we needed to transform these variables to a normal distribution in order to combat overall model skew. From visual exploration of our numerical values, we determined that VehicleAge, AnnualKM, CommuteDistance, and Vehicle age showed high levels of variable skew. In order to normalize skew, we applied a Log-based transformation on these variables to improve model performance. In our Phase 1 analysis, we opted to consider quotes with conflicting and/or unusable information due to some categories having over 60% of data missing, as well as massive outliers, or noise contributors, as “dead quotes” and they were removed within Phase 1 of the assignment. Due to this course of action, there are very few categories left to impute (since values with over 60% missing were rejected, with exception to commute distance which will be imputed). There are a few numerical categories that required small amounts of imputation due to rogue missing values that were missed in initial data cleaning, but this was limited to 2-3 missing variables for categories such as Years\_licensed. A few categorical data types were excluded from imputation, due to the potential for generating false values. For example, blank values within categories including VehicleType and VehicleMake when imputed will generate based on the most frequent classification, and this has a strong possibility of misclassification of the Vehicle type. To elaborate, this has the possibility of classifying a truck, as defined in the data row, as a Sedan (Assuming Sedan is the most common category). For this reason, we will exclude these two that are dependent on the results of other data fields from imputation. This was the key component in our model setup, and from here, we experimented with deriving the best possible result. A key component of this stage was experimenting with the data partitioning, and the impact of changing the entry and stay P-value, as well as max steps for our models. For modelling purposes, we employed the use of a backwards, forwards, and stepwise model using an entry significance of 0.5 and a final stay significance of 0.1. We found this to be an optimal tradeoff between maintaining model quality, and using a richer level of variables for our regression formula. For a full breakdown of the experimentation of P-value settings, see Figure 1.2. Our final, and best performing model was a forward regression model with a stay significance of 0.1, 40 max steps with a 65-35 partition. This model resulted in a total of 8 steps, and contained the variables: FSA\_POP, Years licensed, Province, Marital Status, Vehicle Age, MultiProduct, and Vehicle Use. This model netted a Misclassification rate of 0.38 and an ASE of 0.226. From an interpretation standpoint, this model was not as accurate or impactful as our ANN or decision tree results from a performance measure (This model netted a higher MCR, ASE, and lower Cumulative lift statistic in comparison). Because of this, we will use our more accurate models for final business interpretation.

**Artificial Neural Network**

To prepare our data for Artificial Neural Network (ANN) modeling, we used identical pre-processing settings for variable selection, sampling & variable transformation. However, our imputation & data partitioning settings were optimized differently. ANN’s do not handle missing values well and result in loss of entire rows of data, therefore, a number of methods were tested to address 9,181 missing values from the “Vehicle Type” attribute. First, a model was run without imputation. In this case, we essentially omitted entire rows with missing values. Next, a model was run with imputation using count on the vehicle type category. This resulted in more bias than the first model and yielded lower MCR & ASE. Finally, a model was run omitting the vehicle type category entirely, however, this also produced a less accurate model. With these results, it was determined that running the model with the missing values resulted in the best outcome.

Next, the model was run using a number of different training and validation data partitions including 70-30, 60-40, 50-50, & 45-50. These were run using default settings for both misclassification and ASE selection criterion on both gini-coefficient and chi-square ranked models. Ultimately, a 50-50 data partition was chosen as it appears to provide the best ASE and Misclassification results in comparison to other partitions.

**Data Partition Selection (Default Hidden Units 3)**

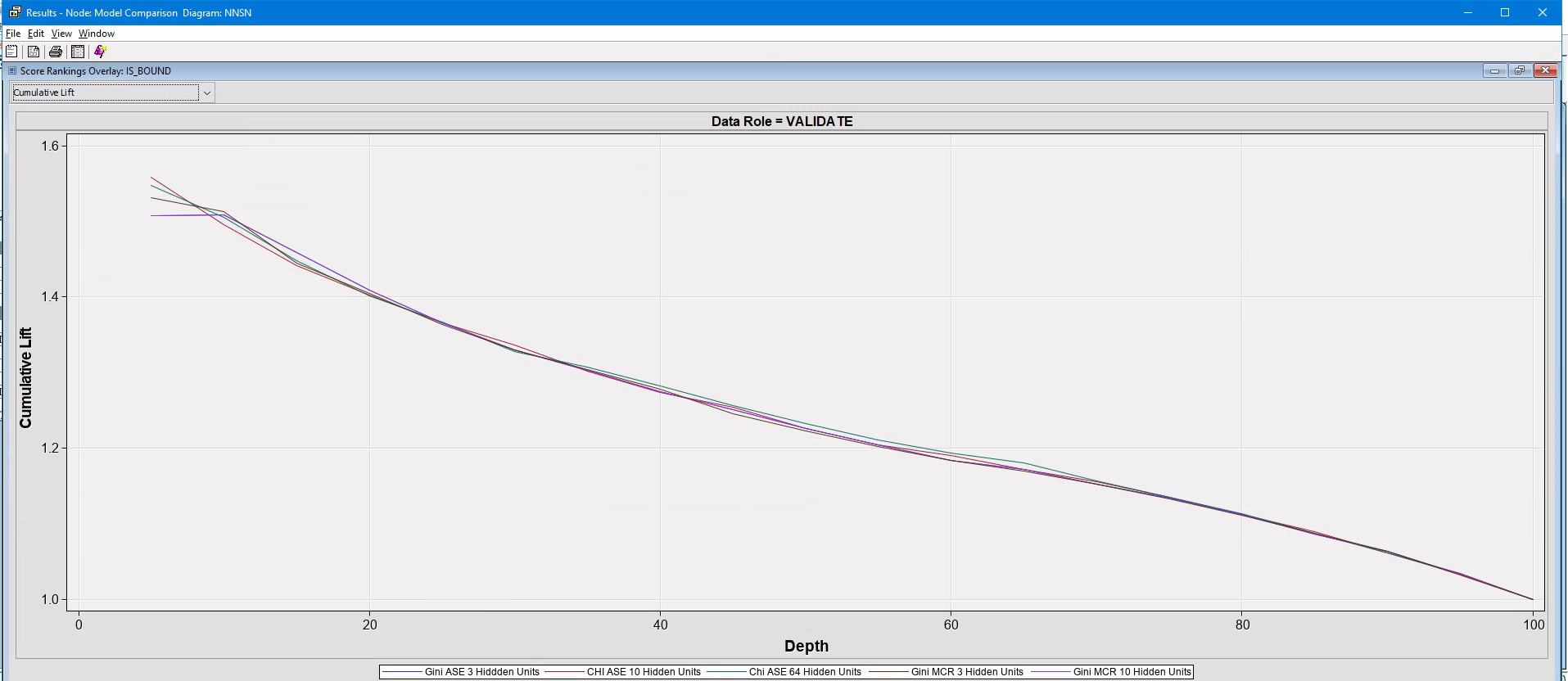
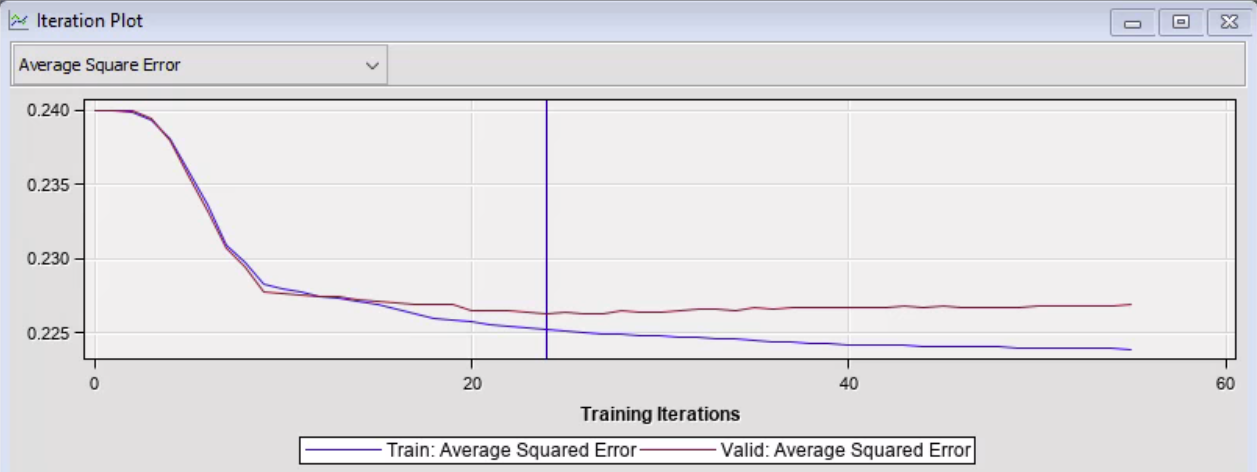
|  | **Gini-Ranked Model** | | **Chi-Ranked Model** | |
| --- | --- | --- | --- | --- |
| **Data Partition** | **MCR** | **ASE** | **MCR** | **ASE** |
| **70-30** | **.3754** | **0.2256** | **.3792** | **.2274** |
| **60-40** | **.3754** | **.227** | **.3782** | **.2271** |
| **50-50** | **0.3738** | **0.2262** | **.3752** | **.2264** |
| **45-55** | **0.3761** | **0.2270** | **.3752** | **.2268** |

Additionally, changes to the number of hidden units were made to evaluate changes in model performance to increases and decreases in non-linear computations. Increasing the number of hidden units from the default of 3 appeared to have a very miniscule improvement to model performance for the chi-squared target model. In the gini model, ASE/MCR peaked when run with 10 hidden units then decreased thereafter. 64 units provided a worse result than 3. See below a summary of our results and best scores.

**Gini vs Chi ASE & MCR Predictive Models (50-50 Partition)**

|  | **Gini** | | **Chi** | |
| --- | --- | --- | --- | --- |
| **# Hidden Units** | **MCR** | **ASE** | **MCR** | **ASE** |
| **1** | **.3781** | **.2284** | **.3794** | **.2282** |
| **3** | **.3738** | **.2262** | **.3751** | **.2264** |
| **5** | **.3738** | **.2264** | **.3758** | **.2270** |
| **10** | **.3727** | **.2264** | **.3750** | **.2261** |
| **64** | **.3753** | **.2262** | **.3745** | **.2259** |

Next, a model comparison was run for cumulative lift. The chosen model was the Gini ASE model with 3 hidden units (Gini ASE 3 - iteration plot below). This model had the highest cumulative lift of 1.4089 at the depth decile of 20 (see first chart below). Higher performance in the lower deciles is particularly of interest as CGIC can only action a certain number of quotes each month. ROC charts did not show significant differences between the top 5 models.



**Clustering Analysis Setup**

We decided to use SPSS for our cluster analysis as the program has the advantage of scalability for mixed data and the ability to apply various user settings. Additionally, SPSS provides a user-friendly interface that gives a good overview and creates graphical solutions which help to interpret the result. Since our data set includes continuous and categorical variables, we decided to use the two-step analysis as it delivers a more accurate result through being more versatile than k-means clustering. For this method we chose the same variables that we used for the predictive modelling: Categorical: New\_VehicleMake, VehicleType, VehicleUse, Gender, FSA POP SIZE, Province, Marital\_Status, Multi\_Product, Occupation   
Continuous: VehicleAge, AnnualKM, CommuteDistance, Age, YearsLicensed  
We chose the continuous to be assumed as standardized due to our data pre-processing performed in stage 1 and applied no noise or outlier handling because we removed these within phase 1 of the project. Our goal is to detect 2 to 3 clusters that will give us more insight into groups that have a lower Is\_Bound rate of 0 than the average of 78% that was detected within our data that we received from the company. Therefore Is\_Bound is our target and will be the attribute on which different clusters will be compared with. This is a necessary step as the target is not being considered for the actual cluster results, it is only used to evaluate the effectiveness of the suggested clusters.  
Our first default set up led to a poor cluster quality as the 14 different inputs did not lead to any good clusters. The dominating values are all too close to each other, allowing barely any differentiation. We modified our setup by assigning 1% of outlier/noise handling to figure out if they are causing a problem. This showed no significant improvement to the analysis which was also the case after varying the number of pre-specified clusters as clusters got worse as the values were not distinct enough. The next step was reducing the number of inputs into the model by dropping variables that have low predictive importance according to attribute ranking from phase I. We removed Gender, VehicleUse and CommuteDistance as they have the lowest worth as variables while keeping our numerical attributes as assumed standardized and do not use outlier/noise handling. After generating poor results with this method we switched to splitting our variables into categories of tree descriptive criteria where we remove the least important from each descriptive criteria (Figure 3.1).   
After several attempts of dropping different attributes according to the just described procedure we ended up with the optimal cluster that includes the following attributes: Age, Marital Status, Multi Product, Vehicle Make, Vehicle Type, Years licensed, Occupation and FSA Population Size. Someone who falls into this cluster has a higher likelihood of being bound as the bound rate for this group is 22.3% instead of the 22% (Figure 3.2). Surprisingly the group with the highest bound rate is not based solely on the highest ranked attributes but also includes lower ranked ones and excludes high ranked such as annual kilometres or province. To find more insightful clusters regarding our target we focused on creating clusters that only display information about the driver, the vehicle or the geographic information. The best cluster that includes all attributes about the driver generates 2 clusters that have a higher bound rate with 22.9% and 24.1%. By removing the gender we were able to generate an even better outcome with a bound rate of 22.9% and 27.1%. The only cluster with good quality was achieved through missing marital status and gender which resulted in a bound rate of 24.8%. The cluster that included all vehicle related attributes besides commute distance, as this attribute had 0.1 impact, was able to generate a bound rate of 22.6%. The geographical attributes resulted in two clusters from which one had a bound rate of 24.8% and an opposite picture of 14.1%.

**Interpretation and Insights**

Based on our predictive model results, we found that our optimal model was a decision tree using ASE as the selection statistic, with a leaf size of 100. We know this is our best performing model as it delivered a Cumulative lift rating of 1.3977, and lowest ASE measure of 0.225. However, the most important component of this model is specifically the interpretation of the results, and what CGIC can improve on from a business perspective. From our optimal tree, we resulted in a few key descriptors of what is the most significant customer segment for CGIC sales agents and brokers to target. The root node of this model, and one of the key definers overall model, is a customer having a pre-existing multi-product. The next key customer descriptor was the overall province that the customer resides in. We found that the highest bound rate customers lived in Ontario, Nunavut, NWT, and Saskatchewan, in comparison to the other provinces. Following this, another key descriptor was the years licensed by the customer. Overall, the customer segment that follows the previous criterion and has been licensed for more than 7.5 years has the highest likelihood of gaining a bound quote. The final key descriptor in our optimal tree was the overall age of the customer, which was an age 38 or lower. To summarize the profile of the customer, our target customer is someone 38 years old or younger, lives in Ontario/NWT/Saskatchewan/Nunavut, already has an insurance product with CGIC, and has been licensed for 8 or more years. This customer segment contained over 11,000 observations with a predicted bound rate of 37.65%, which is significantly higher than the nearest customer segment in our model (which had a predicted bound rate of 33.26%). This is a proportionally larger breakdown between bound/not\_bound in comparison to the total data, which had an overall bound rate of roughly 22%. From an action perspective, brokers need to focus specifically on a customer segment that generates the highest return for the least effort. This customer segment is highly focused and extremely likely to become a bound quote. This theory is also confirmed through the use of Cumulative lift. We determined that our model is quite effective in guessing the first 20% of the bound quotes (Cumulative lift value of 1.397), and this has a significant impact on the actions taken by the broker, in the way that they can prioritize this customer segment with their limited capacity to address web quotes, as it falls within the first 20% of predicted quotes by our model. We also determined some key customer segments that paint a similar picture from our cluster analysis. Our best cluster (Figure 3.2) with attributes from all categories shows an average age of 32, the marital status of 95% of people within the group is single, half of them do not have a multi product, most people drive a Honda, even though this applies only for 15.5% of the group and the most occuring vehicle type is a Sedan. The average number of years licensed lies at 11.99%, most of the group does not provide information about their occupation (94.4%) and over half of the cluster lives in an area that has a small population size. This suggests that a main target group are younger adults that identify as single and rather live in smaller cities. Cluster analysis solely on the driver revealed that middle aged married people who have other insurance products with the company are 1.2x more likely to be bound than the just described customer, but 5% less of our customers fall into this group. Additionally, we observed that having a multiproduct has a positive impact on bound rate no matter whether the driver is a young adult or middle aged. The vehicle group that achieves a slightly higher bound rate shows an average of 14,600 annual km and are mostly used for commuting. Within that cluster most vehicles are SUVs from Ford and have been used for a while as their average age is 7. The geographical cluster with a higher bound rate suggests that a quote that is from Ontario and an area with a medium sized population has an almost 3% higher likelihood of being bound. A quote from a place with a small population size within Newfoundland and Labrador on the other hand should not be pursued because of its low bound rate of 14.1%.   
Overall using our suggested clusters can help to mark the quotes that have a higher chance of being bound. The cluster that considers attributes from all 3 different categories makes it possible to prefer these and ignore the others, yet the improvement from it is marginal. Using the clusters that focus only on one category creates more distinct groups regarding our target but leaves more room for variation as they only take a few variables into account. It could be used for specific marketing purposes whose success depends greatly on the type of customer that is targeted.

**Recommendations**

Our recommendations are focused on taking an extremely targeted approach to maximize the time of the sales and brokerage team. For our recommendations, we advise CGIC to pursue the customer segments discovered within our model interpretation section. We identified a few key customer segments in our analysis, both through predictive modelling, and cluster analysis. To review what we discovered, we will describe the ideal customer segment to CGIC for meeting business objectives. Overall, the most ideal customer segment is the quotes where the applicant already has a product with CGIC, lives in Ontario, NWT, Nunavut, or Saskatchewan, is aged 37 or younger, and has been licensed for 8 or more years. This customer segment has an overall bound rate of 37.65%, which is proportionally higher than the bound rate of the overall dataset (~22%). This is a very feasible target of just over 11,000 quotes that can be attended to by the sales department. From our clustering analysis, we discovered that there are three areas of focus for defining customer trends. 1. Geographical descriptors 2. Vehicle Descriptors 3. Customer descriptors. In terms of specific recommendations we advise from our findings, we advise that CGIC targets (from a geographical standpoint) the customer segment that resides in Ontario, within medium-sized population areas (FSA Areas with over 65,000 residents). From a vehicle standpoint, we found that Ford SUVs with an average yearly kilometer distance of 14,000km had the highest rate of bound from a vehicle description. From a customer description, we found that single young adults (<32 years old) with these descriptors had a proportionally higher rate of bound and made up 15% of our customer base.

Overall, our recommendation is to focus sales efforts specifically on these customer segments. Specifically, we found that our demographic seen in our decision tree has the highest impact on the overall bound rate, so this is the category of customer that should be prioritized the most. Under the assumption that the sales division can cover the first 20% of the quotes, this segment and the segment defined in our cluster analysis are the best way to maximize sales efforts and target the most likely demographics directly.

**Appendix**

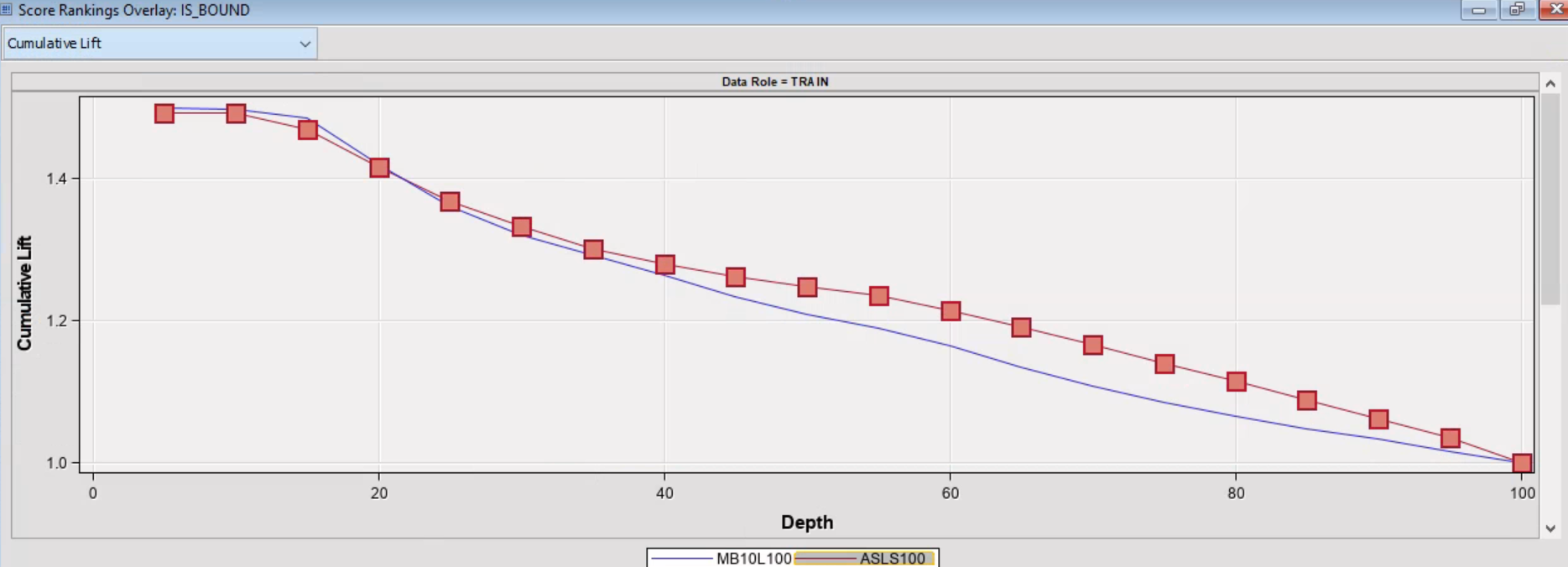
**Figure 1.1: Regression Results (0.5 Entry, 0,1 Stay, Validation, 40 Max Steps)**

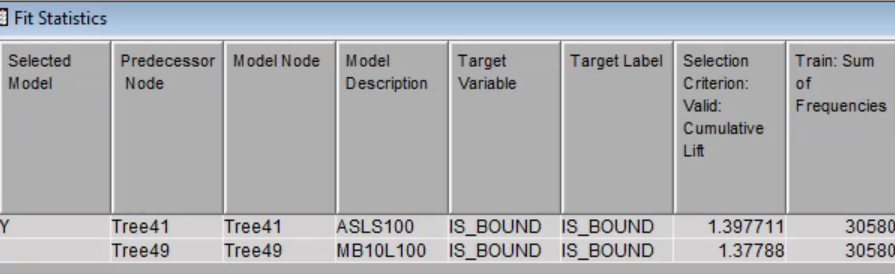
|  | **For-**  **ward** |  | | **Back-**  **ward** |  | | **Step-Wise** |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Partition** | **Steps In Model** | **MCR** | **ASE** | **Steps In Model** | **MCR** | **ASE** | **Steps in Model** | **MCR** | **ASE** |
| 45-55 | **7** | **0.382** | **0.227** | **6** | **0.382** | **0.227** | **7** | **0.382** | **0.227** |
| 55-45 | **7** | **0.383** | **0.227** | **2** | **0.383** | **0.228** | **7** | **0.383** | **0.227** |
| 65-35 | **8** | **0.380** | **0.226** | **5** | **0.381** | **0.227** | **8** | **0.380** | **0.226** |
| 35-65 | **7** | **0.38** | **0.227** | **8** | **0.380** | **0.227** | **7** | **0.38** | **0.227** |
| 50-50 | **8** | **0.38** | **0.227** | **5** | **0.382** | **0.227** | **7** | **0.380** | **0.227** |

**Figure 1.2: Regression Results (Experimentation on P-Value, using a Forward Model with a 65-35 partition, 0.5 Entry)**

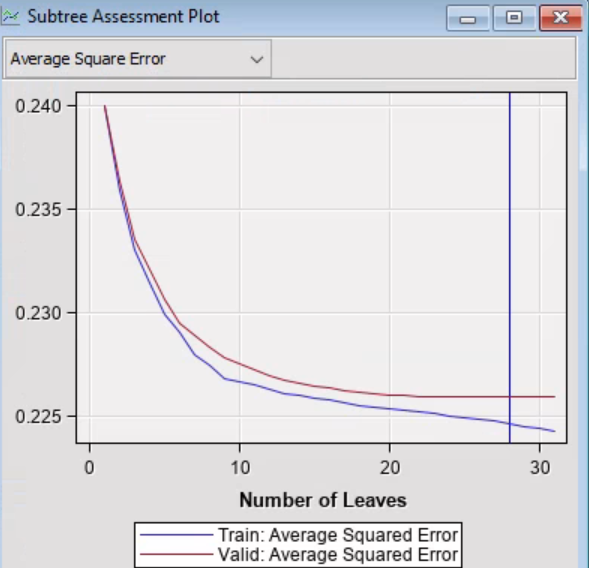
| Stay P-Value | **MCR** | **ASE** |
| --- | --- | --- |
| 0.02 | 0.380 | 0.226 |
| 0.05 | 0.380 | 0.226 |
| 0.1 | 0.380 | 0.226 |
| 0.2 | 0.380 | 0.227 |

**Figure 2.1: Cumulative lift output for the two best models in decision tree phases**

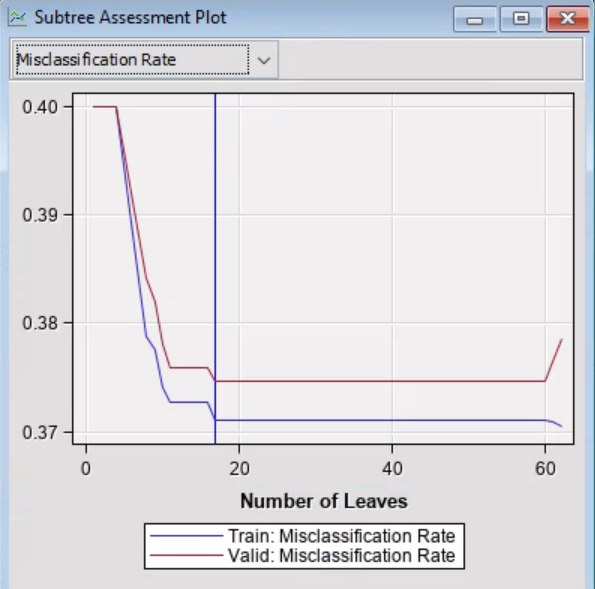
****

****

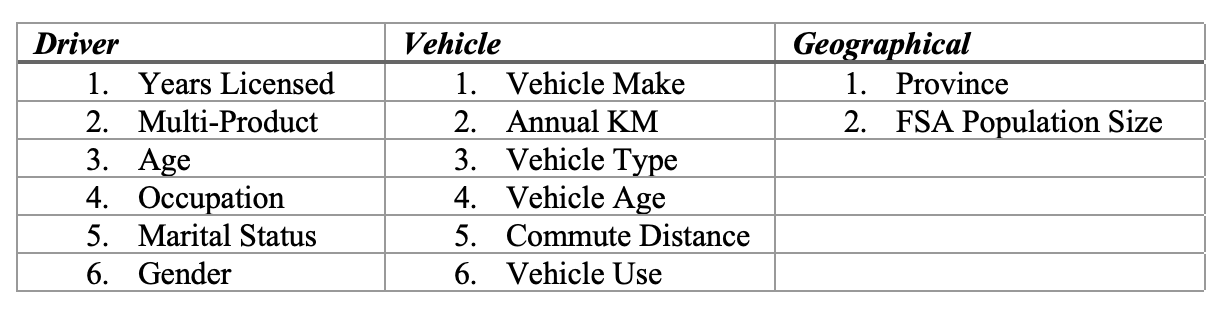
**Figure 2.2: Decision Tree Subtree Assessment Plot**

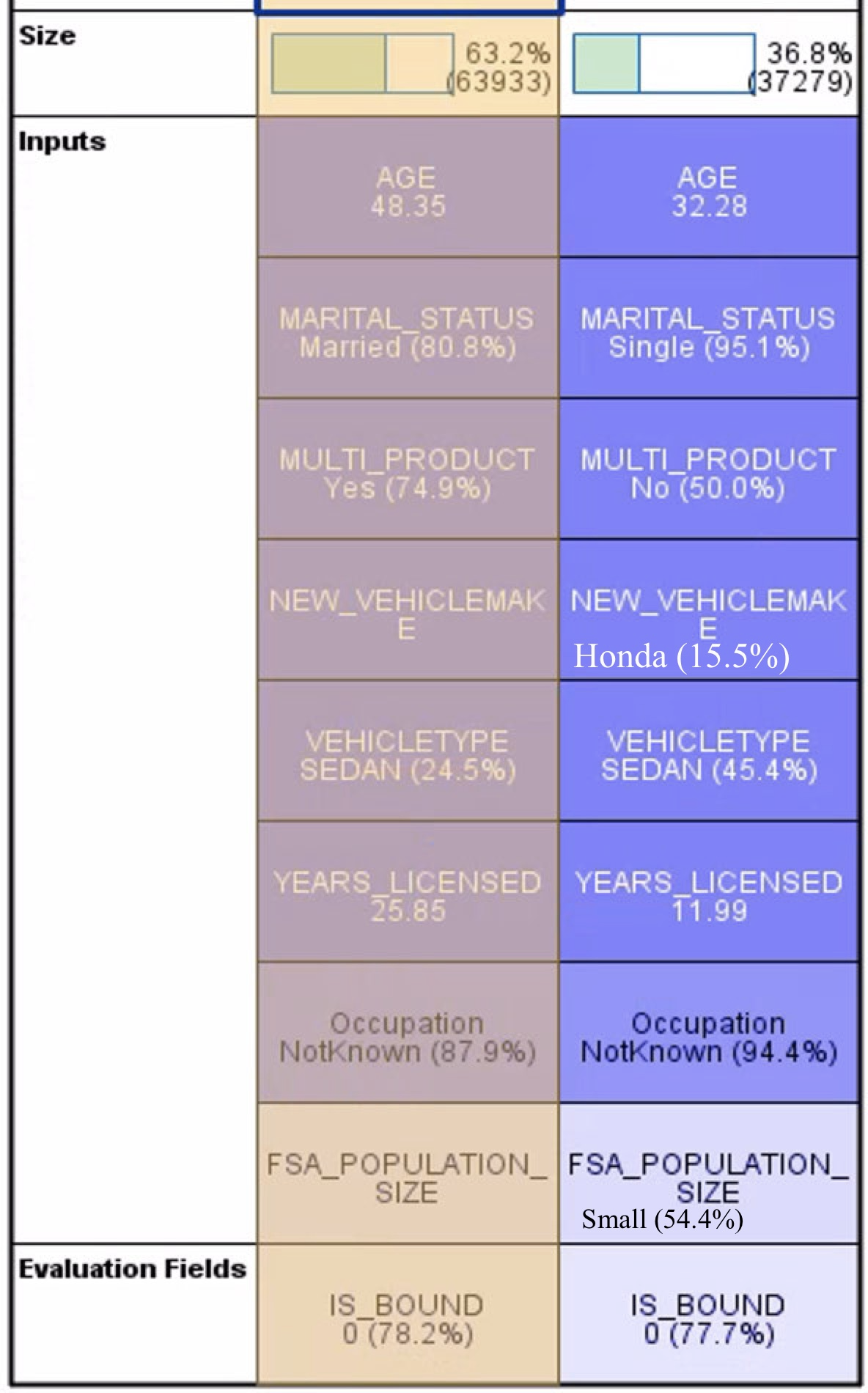
****

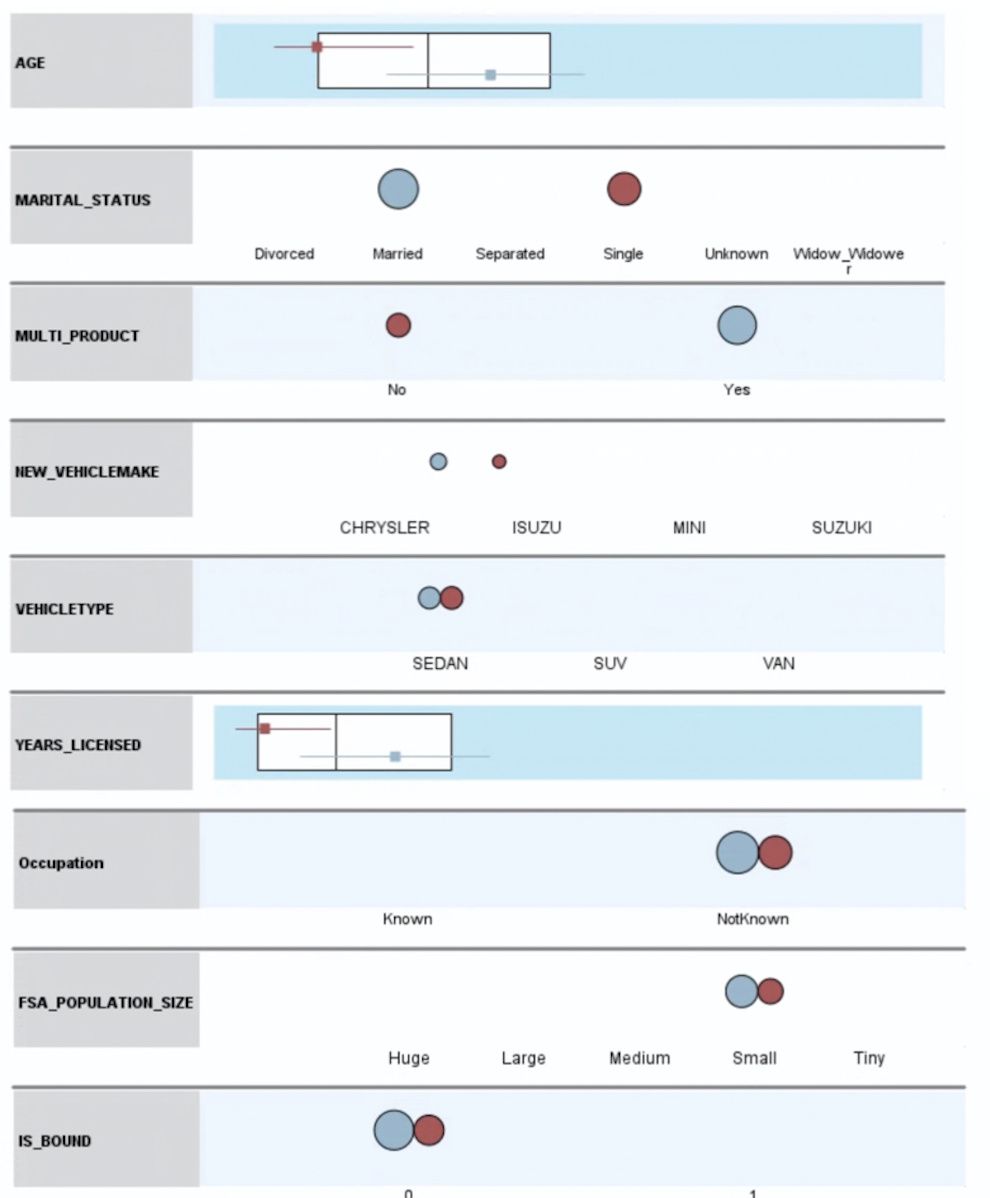
**Figure 2.3: Subtree Assessment Plot for best model in decision tree using misclassification error rate**

****

**Figure 3.1: Attributes sorted by importance into categories**

**Figure 3.2: Best Cluster all attributes and Cluster Comparison**





**Figure 4: Metadata Configuration**

