**Final Project Phase-I**

ITIS 4P21: Introduction to Business Analytics

Group 4, Section 1

October 29th, 2021

James Boyle - 6196885

Naim Boufridi - 6357859

Shane Matthew Sodhi - 5362231

Maja Leonie Hermann - 7346299

Janvier Kurawige - 6361794

**Introduction & Problem Statement:** We begin the first phase of our analytics project with establishing our overall objective, and understanding the problem at hand. The Co-operators, a well known insurance brand, receives mass amounts of Web Quotes on a month-to-month basis. Based on their current system, there is little understanding of which quotes will become paying customers, which raises a large business problem for their sales and brokerage lead teams. This problem establishes our overall objective, which is to develop a predictive model by the end of the term that provides an estimate or prediction of which attributes are likely to indicate which quotes will become customers. This method is known as clustering. With the first phase, we aim to understand, clean, and prepare the data that will be used in this predictive model. Our main measure of success will be the bound rate (IS\_Bound), as this is a key measure of success for customer conversion. The goal is to narrow in on variables with greater purity than the dataset itself, which will yield information gain towards quote bound rate. In order to accomplish this, the data first was cleaned and processed by handling missing values and dealing with outliers/noise. Second, the overall purity of the data was determined and promising categories were discretized, reduced, and/or transformed. Third, variables’ importance in predicting the target were ranked using SAS’s function for determining attribute worth. This function uses the Gini, and Chi-Square methods. Last, the data was placed into pivot tables, and visualized to capture meaningful findings.

**Identifying and Resolving Missing values, Outliers and Noise:** The first stage of our process was identifying and resolving outliers, noise, and irrelevant data and understanding the data itself. To do this, we created summary statistics and Box-and-Whisker plots of each Numerical data category to understand the overall distribution and behaviour of the data, as well as count statistics for the amount of blank values (See tab Outliers&DescriptiveStats). From this analysis, we discovered outlying values that falsified and destabilized the numerical data through visualization and hard analysis. We have included full lists of the outliers removed in the Outliers&DescriptiveStats tab, under each respective category that they were removed in. An outlier or noise value was only removed in scenarios where the data provided was impossible, unexplainable, or uninterpretable. Some examples of outliers removed were “annual KM” distances of +2M KM/Year, as we felt this was an unjust and highly unlikely occurrence. The removal also included “0” values for this attribute, as having no distance driven per year was unjustifiable, and skewed the data lower. For Age/Year of Birth, we removed data values that were below the legal driving age in their respective domain, and deleted a handful of values with a birth year of “9999”.

For Commute Distance, we removed values that had a 1-way commute distance of +1000km, as this threshold was highly unlikely and unachievable on a daily basis. With respect to Commute distance, we also removed commute distance values if the quote had a “VehicleUse” specified as something other than Commute Distance, as the specified commute distance held no value if the vehicle use was not intended to be a Commuter Vehicle. For “Years Licensed”, we removed 3 Values with +117 years licensed. We felt this was an appropriate threshold, as drivers would need to be a minimum of 133 Years old to achieve these values. For Vehicle Age, the outliers that were removed were Vehicles with a model date in 2018 or 2019. Since this dataset is dated in 2016, these vehicles would not be available for consumer purchase. Vehicles with a model year of 2017 were permissible, as dealers sell vehicles for the next year in the current calendar year. For Vehicle Make it was possible to correct most of the spelling mistakes, get rid of additions like TRUCK/VAN and interchange the model with make to clean the data and minimize the amount of categories. A few had to be deleted as they were referring to fictional characters or were not interpretable.

This process also pointed out specific attributes within the data set that we believe hold no value on the Bound rate of a quote. For example, this included “Assigned Losses PD 5 YRS” and “Suspension Count” being excluded from the analysis, as they had 95,640 and 100,755 “0” observations respectively. This means that over 95% of the data in these categories is the same. It is reasonable to assume that these categories have little to no impact on the Bound Rate of the quote, which is our main objective. All the conviction categories were also excluded with similar rationale. This methodology was also applied to the “Years as Principal Driver” as the column contained more than 101,000 blank observations. This action was also proved in our Gini and Chi-Square tests as these attributes had little to no determination on our target variable. Regarding missing values, It has been observed that 63.50% of values in the Vehicle\_Value field are missing. Due to this magnitude of missing values, any imputation method will likely generate biased and unrealistic values. Furthermore, It has also been observed that 34.80% of values in the Commute\_Distance field are missing. However, Commute\_Distance is not applicable to all records. The overall process of outlier and Noise removal in this phase of our project resulted in significantly more stable summary statistics in our analysis (Before and After Results can be seen in Outliers&DescriptiveStats tab). This is desired, as it will lead to a more accurate and stable model when we begin our predictive analytics for the Bound rate.

**Discretization and Variable Transformation:** A number of transformations were made to the data to improve its quality. Methods such as discretization, generalization, and attribute construction were used. Our primary goal was to increase information gain on variables with high-worth (as per SAS Gini & Chi-square statistics) and to improve interpretability for visualization and multi-dimensional analysis in all other variables.

**Province & FSA Population**Using the forward sortation area data provided, we were able to improve interpretability and create a more real-world applicable categorization by cross referencing the data against a 2016 StasCan census (Appendix:Sources). This census included forward-sortation areas, their respective population and their province. Using excel’s Match and Index functions (similar to Vlookup), the attribute “Province” was constructed as well as a population attribute. The population attribute was discretized into categories of tiny (<1,000), small (<30,000), medium (<65,000), large (<99999) and huge (>100,000). This decision was based off of a 2011 StatsCan article which defines preferred classifications for population centre sizes (Appendix: Sources). The census defined large population centres as those with 100,000 or more residents, however, larger cities commonly have two or more postal codes, whereas medium or smaller cities have just one. For this reason, a large FSA population was redefined as one with 65,000 or more residents and huge as one with over 100,000. This also helped to equalize bin width.

**Vehicle Model**An iterative process took place to clean the Vehicle model to remove unnecessary information such as the number of doors, or the specific engine the model was using. Additionally, the models were simplified if they belonged to one family model and any unnecessary add-ons such as ‘sport’, or ‘luxury’ were removed. This was completed only on the condition that the base model is identical to other ‘children’ models. The process's steps are as follows: we trimmed the first word from the vehicle model category, imputed that information into a new column then we trimmed the 2nd word of the vehicle model and inputted that information into a new column. Afterwards we combined both columns and compared it to a VLookUp Table ensuring that the combination of first and second words were logical, and if they were not we would adjust back to only one word. This was done manually. In this process we found models that did not exist and we removed those from our dataset. Lastly, we created the final column using the VLookUp formula for input.

**Vehicle Type**As vehicle models still had well over 22,000 categories, an initiative was made to generalize the data into vehicleType categories. Categories chosen to fit the data were that of Sedan, SUV, Wagon, Van, Truck, Hatchback and Sports. These categories are the most common car categories according to a number of popular car enthusiast websites including carandriver.com, and jdpower.com. As the data contains over 22,000 different vehicle model categories, a Python algorithm was used to categorize the large majority of vehicle model records. Using the Pandas library, the final project file was read into a Pandas DataFrame using the read\_excel function. A number of conditional statements (one for each vehicleType) were then written which each contained a str.contains function. This function tests if a specified pattern or string is located within a series. To determine a vehicle’s type, key-words which were unique to each type were extracted from automotive websites using Copy-pasting web scraping. The excel data was also scanned for common unique keywords indicating a specific vehicleType, such as “Crew Cab” for a Truck. When the algorithm is run using these key-words, it locates the specific matching strings and then inputs the corresponding result, the vehicleType.

**Numerical Descritized Attributes:** On top of making modifications to the existing Non-Numerical data, there was an inherent need to discretize the Numerical attributes in order to understand the overall data, and facilitate the creation of our pivot tables for multi-dimensional analysis. This was one of the larger tasks in our process, and a comprehensive list of the newly created attributes can be found in the data dictionary within the excel document. To summarize our “binning”, we will walk through some of the categories created to group the large amounts of numerical data. The first category we created was an age classification for Vehicles using their model year. We also performed a similar method of discretization for Customer Age based on Stats Canada age classification. We also created a descritized column for the Annual KM driven, classified into “Light Intensity”, “Medium Intensity”, and “High Intensity”. For the sake of space, the criteria of binning has been specified in our data dictionary (See appendix 1). All of our discrete variable ranges were decided based on two factors: 1. An external source of data classification or 2. Based on the distribution of the data found in the post-outlier removal descriptive statistics. The next Descritized category we created was “Work\_Travel\_Distance”, which specifies a ranking of commute distance based on the distribution of the data. We also created a category that highlights Driver experience levels, based on the years licensed. The experience level of the driver increases with the amount of time licensed, and is classified into “Inexperienced Driver”, “Moderately Experienced Driver”, “Experienced Driver” and “Long-term Driver”

**Data Insights & Critical Findings:**After our comprehensive data cleaning and exploration stage, the next logical analytical step was to capture useful insights through visualization. By generating useful, organized visuals, we could further familiarize ourselves with the data, and create a compelling case for our future modeling. These insights also form a baseline for The Co-operators to better understand quote habits, and gain business intelligence with respect to their online quotes and their bound rate. In order to start generating interaction insights, we opted to perform single and multi-dimensional analysis to better understand how our data impacts the target variable. In our search for insights, close attention was made to attributes of high-worth based on both SAS StatExplore results and our own single category analysis. Attributes such as Multi-product, Province, Driver Experience, and Vehicle Type were of critical interest as some of their categorical properties held bound-rates which varied significantly from that of the dataset.

**1 Dimensional Analysis  
Success Rate:** We have defined success rate as the ratio of yes bounds when compared to overall bounds. This ratio is 21.97 percent. The most relevant deviations from this rate are shown here.   
Quotes by Province: When comparing the four provinces with the highest sample sizes, Ontario, Newfoundland & Labrador, Nova Scotia, and New Brunswick, Ontario have the highest percentage of yes bounds by 10% or more. Refer to the Non-numerical category stats sheet. This analysis means that the order of yes bounds is Ontario, Nova Scotia, New Brunswick, and Newfoundland and Labrador.  
**Quotes by Age Range:** Most of the quotes are from the 25-54 age range. This means that this age range is the most popular one. The 24 and under age range are less likely by 5% to have bound rates while 55 and over are less likely by 2%.  
Quotes by FSA Population Size: If a quote comes from a huge FSA location there is a minor advantage of almost 3 % in the yes bound rate. Greater population size may indicate an increased bound rate.  
**Quotes by Driver Experience Level:** Inexperienced drivers are less likely to have yes bounds when compared to any other experience level in its table. Their rate is 16.54%. The opposite is the case for moderately experienced drivers who have a 27.21% rate. In terms of experienced drivers and long term drivers their rates are 24.98% and 20.75% respectively.   
**Quotes by Multiproduct:** When customers stated no for Multiproduct they had a chance of 15.89% of being bound, while those who already purchased a product had a higher chance of 25.15% to be bound.  
**Quotes by VehicleAgeClassification:** Even though the difference is small it can be observed that the older a vehicle is, the more likely it is that the quote is being bound. This can be seen as new vehicles have a 21.3% of being bound while old vehicles have a 24.19% yes bound rate.  
**Quotes by VehicleUsage:** Quotes with the input of Farm Pleasure have a high rate of being bound with a rate of 36.03%. A small indication can be seen for vehicles which are used for business purposes with a yes bound rate of 24.12%.

**2 Dimensional Analysis**

**Driver Experience & Vehicle Age Classification & Multi-Product:** This table gives further insight into which customers with which cars regarding their age have the highest rate of already owning a different product from the company. This would mean that a long-term driver who wants to insure a new vehicle has a rate of 77.99% to have a multi-product. Similarly, a moderately experienced driver who asks for a quote for an old vehicle has only a 37.65% rate of having a multi-product.

**Driver Experience & Age Range & Is Bound:** It shows the effect on the quote being bound by the experience level and age range of the customer. It can be observed that a customer who is 24 and under and an inexperienced driver has a lower rate of 16.45% to be bound. Whereas somebody who falls under the age range 25-54 and is a moderately-experienced driver has a high rate of 27.82% to be bound, similar numbers can be seen for an experienced driver in that age range.

**Bound Rate - With vs Without Multi-Product by Driver Experience:** The likelihood of a customer who has a multi-product versus one who does not changes the likelihood of a bound quote. In figure , it is evident that drivers who have multi-products are much more likely to be bound than those who are not. This is especially important for experienced drivers and long-term drivers. Refer to figure 4.

**Vehicle Usage Intensity & Vehicle Type:** For the high intensity column it is very unlikely for a sports driver to have a bound quote. Their percentage is 15.29% whereas every other percentage is very close to 20% or higher. Under light-intensity driving Sedans are less likely to have yes bound quotes when compared to every other vehicle type. They are 18.35% whereas all the others are over 20%. Most drivers appear to be medium intensive drivers who drive sedans.

**Targeting**

**Targets:** By using characteristics that define drivers with high-bound rates we can target specific segments of the population. These characteristics consisted of the drivers being in Ontario, experienced and moderately experienced drivers. They belong to FSA population sizes that are medium, or large. These drivers also typically have high-rates of multi-product.

**Non-Targets:** On the other hand there are characteristics which define drivers with low-rates. These characteristics are the drivers that do not belong to the Ontario province, they do not have multi-products, and their experience levels are inexperienced or long-term. They belong to FSA population sizes that are tiny, small and huge.

Refer to figure 8 to see the visualization of targets and non-targets.

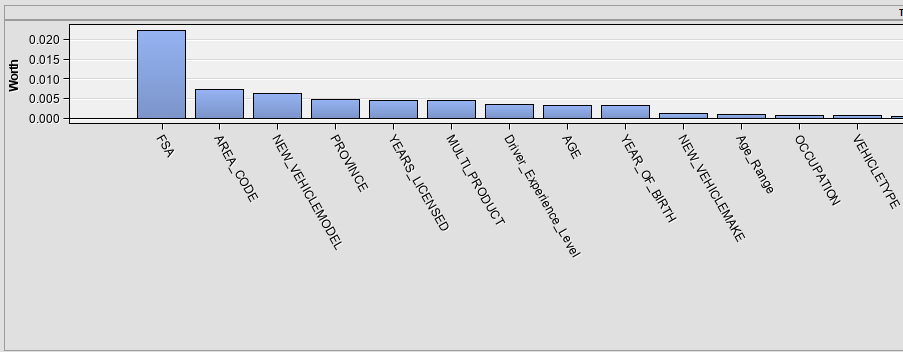
**Conclusion:** In short, this phase of our analysis provided massive amounts of insights and data preparation which has created an opportunity for The Co-operators to better understand their data. Throughout our process, there has been an emphasis on creating a data file that is easy to understand and develop insights from, even for a user that may not be experienced in data interpretation or analysis. Our multidimensional analyses, pivot tables, and editable charts provide digestible information and insights as discussed throughout this paper.

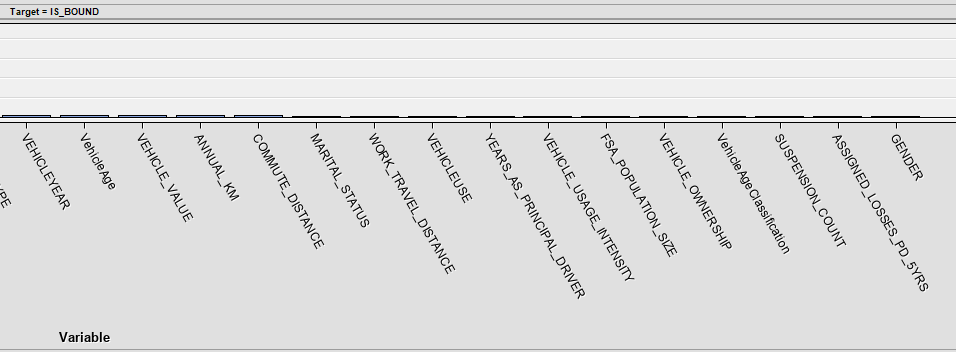
As we continue, our emphasis will now shift from developing insights and data exploration to building a model that will help with future quote prediction and cluster analysis. This differentiates from what we have already accomplished as we have now provided basic insights for the Co-operators to work off of, but our upcoming model will be able to be applied to newer data sources and provide a reasonably accurate result.

**Appendix**

1. Data Summary/Dictionary

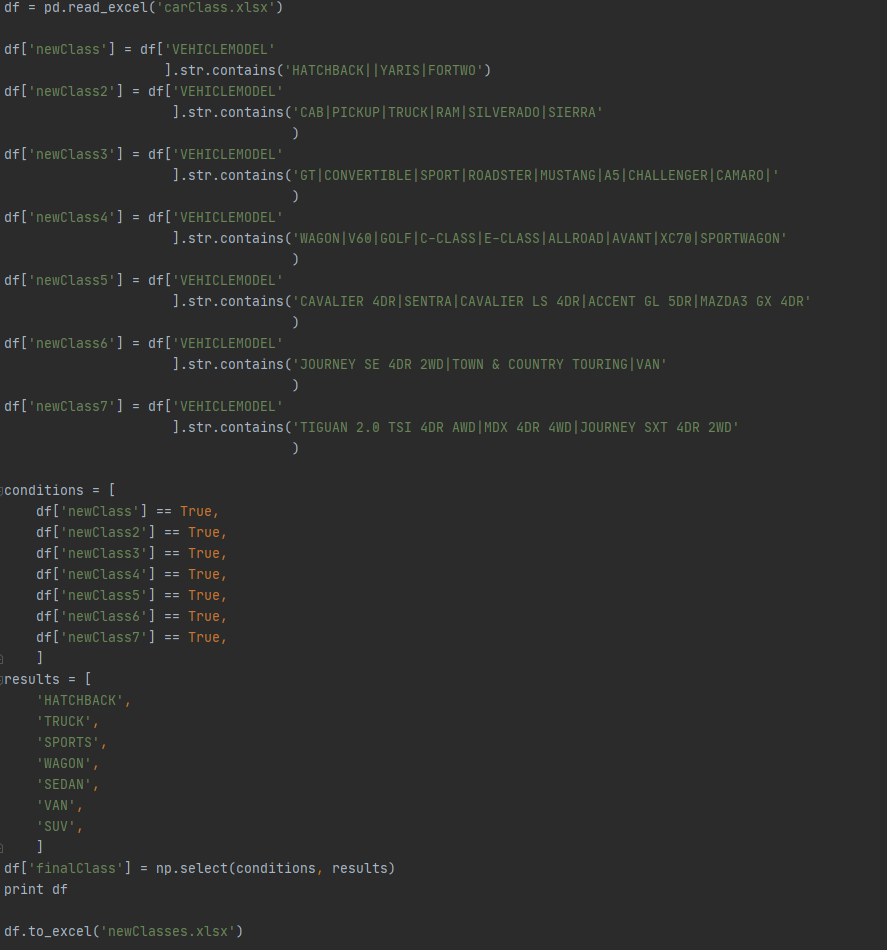
**Attribute Rankings - Variable Worth (A Function of GINI Statistic)**

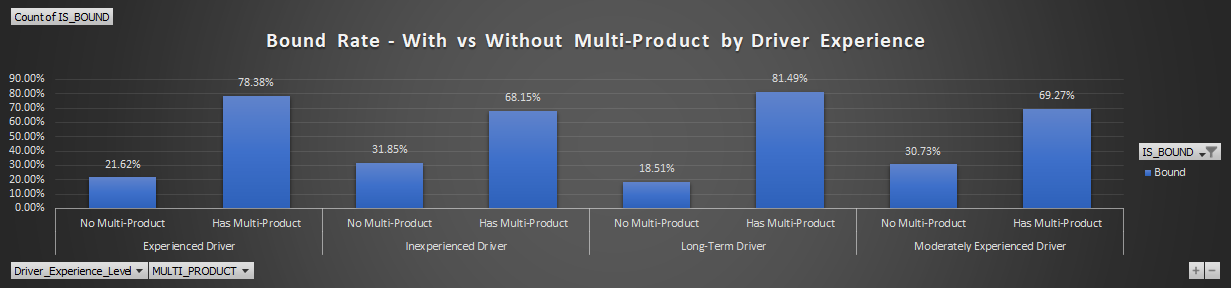




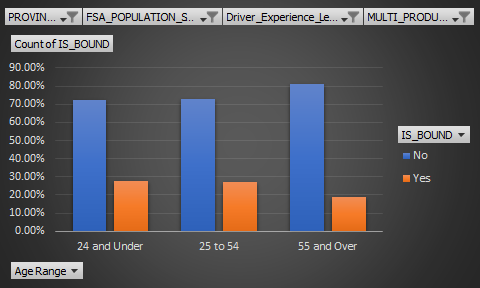
**VehicleType Classification Algorithm**

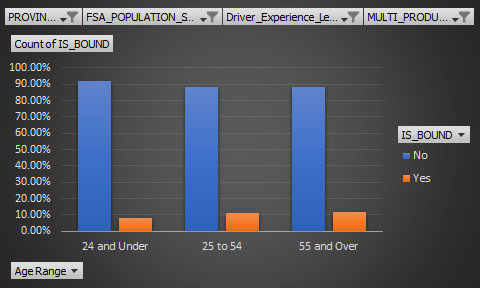
Note: A number of vehicle search terms were removed from within the conditionals for presentation.

****

**Figure 4) **

**Figure 5) Target \*& Non-Target in descending order.**

****



**Sources**

**Province & Population Size**

[**https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/comprehensive.cfm**](https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/comprehensive.cfm)

[**https://web.archive.org/web/20121213032942/http:/www.statcan.gc.ca/subjects-sujets/standard-norme/sgc-cgt/urban-urbain-eng.htm**](https://web.archive.org/web/20121213032942/http:/www.statcan.gc.ca/subjects-sujets/standard-norme/sgc-cgt/urban-urbain-eng.htm)

Population Sizes Sats Canada - h[ttps://www150.statcan.gc.ca/n1/en/pub/53f0007x/4224875-eng.pdf?st=mWVm\_YFy](https://www150.statcan.gc.ca/n1/en/pub/53f0007x/4224875-eng.pdf?st=mWVm_YFy)