NOVA IMS Information Management School

MAA

Master Program in Advanced Analytics

Portfolio Rule-based Clustering at Automobile Insurance in Portugal

Octaviani Devi

Internship report presented as partial requirement for obtaining the Master's degree in Statistics and Information Management Proposal

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PORTFOLIO RULE-BASED CLUSTERING AT AUTOMOBILE INSURANCE IN PORTUGAL

by

Octaviani Devi

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics Proposal.

Advisor / Co Advisor: Roberto Henriques

June 2016

ACKNOWLEDGEMENTS

I wish to express my gratitude to my supervisor, Professor Roberto Henriques, for all his time, effort, understanding, generous guidance and assistance that made it possible to work in this project. Also, I would like to extend my gratitude to my supervisors, Paula Santos and Andre Rufino for their vital comments, supports and valuable discussion. It is a pleasure working with them. I am pleased to thank Sjoerd Smeets for the opportunity and Luis Cardoso who prepare this protocol. Also I would like to thank the office team, Tiago Cavaleiro, Diana Duarte and Filipe Santos for their help and support during the project.

I am pleased to thank Professor Leonardo Vanneschi as the head of Master Advanced Analytics who also support me with his advice. His lectures are amazing. I am also thankful to friends in Master Advanced Analytics, especially to Thies Bucker, Bruno António, Minang Suh Franklin and Ali Godjali for their support and help. Also, I would like to express my gratitude to all my teachers in Master Advanced Analytics.

Finally, I am thankful to my beloved friends, Maria Tjahjadi and Iput Pudji Astuti for their endless love, help and care. Also, I dedicate this thesis to my beloved parents and families.

ABSTRACT

Defining pricing strategy is a challenge for every insurance company. Competition makes insurers need to be more careful to adjust the premium since it may affect the reaction of the existing customer or the new ones. Correspondingly, it may impact the relationship with customer, also the profitability of the company. Moreover, the increment of number of policies will lead to the diversity of policy's risk profile and characteristics which becomes a challenge for insurer to manage their portfolio. Therefore, a deep understanding of portfolio segmentation is important for the company to fine tune the pricing strategy and gain more profit. The project aims to discover portfolio clusters by using k-means clustering algorithm and extract the rules of each cluster by developing classification model using Decision Tree algorithm. The result of the model shows that the clusters give different characteristics and behavior. Complement with KPI metrics, the company is able to monitor the performance of each clusters. So that, the company may use the analyses to optimize the strategy of growth and profitability.

Keywords: automobile insurance, rule-based clustering, k-means, clustering, classification, decision tree.

.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CBR	Clustering Based on Rules
CCC	Cubic Clustering Criterion
CRISP-DM	Cross-Industry Standard Process for Data Mining
GWP	Gross Written Premiums
KPI	Key Performance Indicator
LOB	Line of Business
MAD	Median Absolute Deviation
SSE	Sum of Squared Error
TPL	Third Party Liability

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Once a policy holder is signing the contract and or is renewing it, the insurer is exposed to the risk. Insurer classifies each risk and calculates the applicable premium associated with that risk. This premium is the price that a policy holder has to pay for insurance coverage. It can vary significantly based on difference characteristic of the risk (Werner and Modlin, 2010). Insurance premiums need to cover not only the expected claims but also certain level of profitability of the company. Therefore, the challenge of pricing strategy is to set premium prices not too high so that market share is not impacted as customers have right to exercise the price and to choose their insurers (Yeo, et all, 2001b).

Competition is forcing insurers to adjust rates to retain existing customers and to attract new ones. Insurance companies such as automobile insurances operate in a more competitive environment. A policy holder easily switches from one insurer to another because of the price sensitivity. The price changes do not need to be the same across all customers. For instance, the increment by 5% will give different reaction to the policy holder (Guelman, et al, 2014a). Because of this price change, some policy holders are likely to switch to an alternative insurer, while others will remain with the same company. Insurers that implement pricing strategy may take benefit from customers' preferences (Dolgui and Proth, 2010). Customers have different purchasing behavior over time and their willingness to pay. They are attracted by different benefits offered by the same type of products. For instance, the same car model may be proposed in different versions and may attract a particular type of customer (Dolgui and Proth, 2010).

One of the automobile insurance in Portugal faces the same challenge in defining the pricing strategy. The policy holders are very sensitive to price changes as well as claim management. In August 2015, the insurer's loss ratio is in line with the market which the market's is 66.8%. However, this loss ratio is higher about 6.5 pp compare to one of the other insurance company in Portugal. In addition, it is important to the company to manage cancellation rate and to see if it is needed to adjust pricing strategy without damaging the relationship with policy holders and to stay competitive in the market. Customer-oriented pricing can make a significant contribution to performance enhancement and strengthening the customer relationship (Murdock and McGrail, 1994, cited from Stomer, 2013). Stomer (2013). It presents evidence regarding the importance of adjusting pricing models by considering consumers' perception. According to Stomer (2013), in automobile insurance, vehicle specific attributes are also taken into consideration in premium calculation as well as customer characteristics. Therefore, it is crucial to understand the customer and vehicle characteristics. In this case, clustering is used to discover the characteristic similarity by grouping a heterogeneous population into a number of more homogeneous clusters (Berry and Gordon, 2004). Furthermore, rule extraction will be explored in order to know the characteristics of each cluster. This project is aligned with the strategic initiative 2015 of the company which is "Capture the non-life stand-alone opportunity in bancassurance" and "Foster Customer Centricity", also strategic initiative 2020 which are Customer and Innovation.

1.2. STUDY OBJECTIVES

From the stated problems, the project aims to discover cluster model that has better split to the existing portfolio in homogeneous groups. The cluster model is extended by building a set of rules that extracted from the defined clusters. So that, the rule-based cluster model is defined and can be used to assess the pricing strategy. The main objectives of rule-based cluster model development are as follow:

a. Create portfolio segments with technical characteristics such as customer characteristics, vehicle characteristics or any important characteristics.

b. Increase knowledge of Motor portfolios on the technical Key Performance Indicators (KPIs) such as Loss Ratio, Claim Frequency, Average Claim Cost, Renewal Rate and Customer Growth.

c. Able to be used to fine-tune motor pricing strategy to achieve an optimized portfolio mix based on discovered cluster model.

Based on the objectives above, the scope of works of this project is defined. The portfolio data in this project is limited to the personal portfolio as the majority of the data. Following is the scope of works:

a. Identify variables used in clustering model and identify parameters such as number of clusters.

b. Develop portfolio clusters of insurance portfolios based on their characteristics by using k-means clustering algorithm. Choose the best cluster based on meaningful interpretation and mathematical decision.

c. Extract the rules within the clusters by developing classification model. The model is developed using Decision Tree Algorithm with the respect to the cluster id.

d. Design and create Key Performance Indicator (KPI) dashboards to enable a better knowledge of the cluster characteristics. Provide analyses to understand the characteristics for each cluster.

2. LITERATURE REVIEW

2.1. INTRODUCTION

Berry and Linoff (2014, p 11, 350) mention that clustering provides a way to learn about the structure of complex data and break up into similar pieces. In clustering, there are no pre-classified classes and no distinction between explanatory and response variables. The algorithms discover similarities and group the data into clusters. Clustering methods have been widely used in many domains, including marketing, pattern recognition, biology and many areas. In business, clustering help marketer to characterize customer segmentation (Brito et al, 2015) and then target the marketing efforts to the most attractive segment. In biology, it can be used to categorize gene (Hasan and Duan, 2015) and to derive plant and animal taxonomies (Han and Kamber, 2006). Clustering is not only used as segmentation but can be used for outlier detection, such as credit card fraud (He, Xu, Huang & Deng, 2004). However, according to Mirkin (2005), the implementation of clustering is not straightforward because of two factors. First, similarity distance measurement and clustering techniques frequently give different results. Even though the same technique is implemented, it may lead to different cluster solutions dependent on the choice of parameters such as initial setting or number of clusters. Second, the interpretation of cluster structures is not straightforward. In this case, the clusters to be found are not depending only to data but also the user's goal and degree of granulation. That is why, clustering method often considered as art rather than science. Although, Mirkin (2005) mentions that discovered clusters must be treated as an "ideal" representation of the data that could be used for recovering the original data from aggregate clusters.

After in depth study about available clustering techniques (Han and Kamber, 2006), this project will focus on partitioning methods. Partitioning methods, k -Means, determine the clusters in such a way that objects within a cluster are "similar", whereas the objects of different clusters are "dissimilar". The clusters have to satisfy the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group.

According to Han and Kamber (2006), one of the most well-known partitioning methods is k-means algorithm. The k-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting *intra cluster* similarity is high but the *inter cluster* similarity is low. Cluster similarity is measure in regard to the mean value of the objects in cluster. The algorithm attempts to determine *k* partitions that minimize the square error function. It works well when the clusters are compact cloud separated from one another. Also, Han and Kamber say that the method is relatively scalable and efficient in processing data sets. Correspondingly, Madhulatha (2012) says the reason behind choosing the k-means algorithm is because of its popularity for the following advantages: (1) Time complexity of the algorithm is *O(nkl)*, where *n* is the number of points, *k* is the number of cluster, and *l* is the number of iterations taken by the algorithm to converge, it shows that the k-means is linear to number of points, so that k-means is scalable and efficient in processing large dataset (2) Space complexity is *O(k+n)*, the space complexity is modest because only depend on number of data points and number of clusters (3) Order-independent, for a given initial seed set of clusters, it generates the same partition of the data regardless the order of presented data to the algorithm. A k-means algorithm often used as customer segmentation technique among the other clustering techniques (Wang & Keogh, 2008; Liao, Chen & Lin, 2011; Ghoreyshi, 2015).

The major drawback of k-means that mentioned by Hand and Kamber is that it often terminates at a local optima and the result largely depends on the initial clusters centers. Furthermore, the k-means method can be applied only when the mean of a cluster is defined. This may not be the case in some applications, such as when the data have categorical attributes. The necessity for users to specify *k* in advance can be seen as a disadvantage. The k-means method is not suitable for covering clusters with non-convex shapes of clusters of very different size. Moreover, it is sensitive to noise and outlier data points because a small number of such data can substantially influence the mean value.

2.2. CLUSTERING BASED ON RULES

Describing the structure and obtaining the knowledge of a complex system are difficult tasks. As mentioned by Berry and Linoff (2014), the clustering is a way to understand about it, however according to Mirkin (2005), the interpretation of clusters is not straightforward. Gibert, Aluja and Cortes (1998) introduce Clustering Based on Rules (CBR) as a methodology that is developed to find the structure of complex domains. This approach performs better than traditional clustering algorithms or knowledge based system approaches. Gibert, et al explains that the CBR methods combine clustering algorithm and inductive learning or supervised learning method that focus to the problem of finding, interpreting special patterns and extracting useful knowledge. The CBR can be seen as a process of building knowledge model of a set of rules.

Similar approach is introduced by Williams and Huang (1997) which combine clustering method and supervised learning method. The motivation of this approach is driven by the fact that datasets become more complex and extremely large nowadays. The key knowledge or discoveries may be lost in this wealth of knowledge clusters and difficult to be analyzed. Therefore, it is needed to adopt a strategy to better focus on the most precious nugget. In this paper, it is presented hot spots methodology by adopting multi strategies to find the important nugget. Two cases are demonstrated in this study, which are insurance premium setting and fraud detection. The hot spot methodology has three steps as follows: (1) Develop unsupervised clustering; in this case, k-means clustering algorithm is used. (2) Perform supervised learning to build a symbolic description of the clusters. In this project, decision tree is used to produce a rule set. The result of rule is called nugget which each nugget corresponds to a subset of original dataset. (3) The final step is to evaluate each nugget in the nugget set to find particular importance based on key variables. Statistical summaries are performed to evaluate the nuggets. Lastly, visualization tools are used to provide effective presentations.

Moreover, Han and Kamber (2006) mention that classification as supervised learning is an effective means for distinguish groups or classes of objects. However, it requires labeling and costly collection to model each group. It is often more desirable to proceed in the reverse direction. First, separate the set of data into groups based on similarity by clustering and assign labels to the relatively small number of groups. Berry and Linoff (2014, p11) also suggest to do clustering as the first step in a market segmentation effort instead of trying to come up with a one-size-fits-all rule. Cluster analysis can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. A number of studies have introduced clustering model as a first step before doing

classification or prediction model. For instance, study from Peng et al (2005) explored unsupervised learning method for credit card accounts classification. Jurek and Zakrzewska (2008) use cluster analysis to improve Naïve Bayes Models of Insurance Risk. Lastly, Yeo et al (2001a), use clustering to classify policy holders into homogeneous risk groups. Within each cluster, a prediction model is developed to predict the effect of premium changes on motor insurance.

Furthermore, data mining technique for application to insurance is introduced by Guo (2003). Two important data mining techniques: cluster discovery methods and decision tree analysis are presented to address the issues and techniques for casualty actuaries. Cluster analysis is implemented to discover groups and to identify interesting distributions and patterns in the underlying data, such as segmenting existing policy holders into groups and associating a distinct profile with each group. This analysis helps insurance company to make rate making strategies and to select accurately which policies and services to offer to which customers. The example that is showed by Guo is clustering automobile drivers based on geographic and demographic factors. These segmented drivers will be used to estimate the claim frequency that allows actuaries to evaluate the potential profit for specific segments. After clustering process, claim frequency model is created by using combination of decision tree techniques and logistic regression. First, decision tree algorithm is used to identify the factors that influence claim frequency. After the factors are identified, the logistic regression technique is used to quantify the claim frequency and the effect of each risk factor.

Rule-based cluster model will be proposed in this project. Firstly, cluster analysis will be implemented in the interest of discovering the knowledge about insurance portfolios. This knowledge is important for the company to understand portfolio patterns in order to implement the appropriate strategy for each cluster. A k-means clustering is used to perform segmentation tasks in this study. It is necessary to exercise the result of clustering to achieve a reliable and good result. Secondly, after getting the final clusters, further analysis of the characteristics of clusters are performed. An empirical tree represents a segmentation of the data is created by applying a series of simple rules. Decision trees as part of the Induction class of data mining technique will be used in the respect of cluster id. Statistical analysis and reports based on key performance indicator will be created to understand more the characteristics of the clusters. Finally, the analyses of the clusters will be created.

2.3. K-MEANS

The k-means is clustering algorithm that attempts to divide the entity set in K non-overlapping clusters. These clusters are represented by lists of entities and by their centroids. The cluster centroid is the vectors of means characteristics across the clusters members as a result of iterative procedure. Formally, the cluster structure is represented by S_k (subset) I and M-dimensional, $k = 1, \dots, K$. Subsets S_k form partition $S = \{S_1, \dots, S_K\}$ with a set of centroids $C = \{C_1, \dots, C_K\}$.

The clustering process of k-means is as follows. First, K initial centroids are selected, where K is pre-specified by the user and indicate the desired number of clusters. The algorithm then selects cluster centers and each of the observations in the data is assigned to a cluster based on the shortest distance of the data point from the cluster centroid. New cluster centers are created by averaging

the observations assigned to a cluster. The centroid of each cluster is then updated based on the points assigned to that cluster. This process is repeated until a convergence criterion is satisfied (Wu, 2012). Suppose $D = \{x_1, \dots, x_n\}$ is the dataset. The k-means can be expressed by a function that depends on the proximities of the data points to the cluster centroids as follows:

$$\min_{\{m_k\}, 1 \le k \le K} \sum_{k=1}^{K} \sum_{x \in C_k} \pi_x \, dist(x, m_k)$$

Where π_x is the weight of x, n_k is the number of data objects assigned to cluster C_k, $m_k = \sum_{x \in C_k} \frac{\pi_x x}{n_k}$ is the centroid of cluster C_k, K is the number of clusters set by the user and the function "$dist$" computes the distance between object x and centroid m_k, $1 \le k \le K$. The selection of the distance function is based on the squared Euclidean distance, i.e. $\|x - m\|^2$.

2.4. DECISION TREE

Decision tree models are exploratory models that produce a series of decision rules, which can be expressed in English so they are easy to understand and to interpret by people (Berry and Linoff, 1997). The decision tree model is an algorithm that uses various ways to split large of data into smaller set of records that represents segments with the respect of the target. These segments are called as node which originates with a root node at the top of the tree that consists of entire data. The set of rules that is represented as a branch is applied until the tree is pruned or there is no possible split and become a leaf node. Each leaf node of the decision tree has a group of observations as a result of unique rule that can be represented as IF-THEN logic. The algorithms for building decision trees uses a sample for the split search by computing the worth and by observing the splitting criteria such as number of splits, level of the tree, input variables type, and target variables type (Guo, 2003). Decision trees have two techniques of splits into branches, which are binary splits and multi splits.

The decision trees are classified into two models based on the target variable type. The model is called a classification tree, in the case that target variable type is categorical. In the other hand, if the target variable is continuous, the decision tree model is called regression tree. The decision tree algorithm decides which input variable makes the best split that can separate the records so that a single class predominates in each group. The measure used to evaluate a potential split is purity. The purity measures evaluate the splits using gini, entropy, information gain ratio or chi square test for categorical target variables, while for continuous variable, the algorithm use reduction in variance or F test. The ability to handle missing value in splitting data becomes one of the advantages of using decision tree. The effectiveness of a decision tree is determined by observing the percentage classified correctly. Decision tree could be combined with the clustering algorithm to find rules and pattern (Berry and Linoff, 1997).

2.5. BASIC INSURANCE TERMS AND RATIOS

As part of the business understanding, it is important to know basic insurance terms and ratios. Given that some terms and ratios will be used in this project for example in the input variables, analyses and reporting. This section provides some basic concepts of fundamental insurance terms

and ratios, based on the interview with Actuary and based on summary of basic ratemaking by Werner and Modlin (2010). Basically, insurance has two fundamental characteristics according to Vaughn (2008), those are shifting risk from individual (called as insured) to another party (called as insurers) and sharing losses on some equitable basis by substitutes the premium to the uncertainty loss. The process to predict future losses and expenses becomes one of the main activities in insurance that is called as ratemaking or insurance pricing. This process will determine the price per unit of insurance for each exposure unit.

Risk exposure is the basic unit of risk that underlies the insurance premium. The exposure level is measured in years "insurance". If the contract is valid for the whole year, then the exposure is 1-year insurance. If the insurance only has half year contract (canceled after 6 months or only have 6 months contract) that the insurance will exhibit 0.5 (half year insurance). While ***pure premium*** is a measure of the average loss per exposure of cost per unit of time. This premium must be sufficient to cover losses and expenses. To obtain the premium, the insurer must predict the claim cost. The pure premium is calculated as:

$$\text{Pure Premium} = \text{Frequency} \times \text{Severity}$$

$$\text{Pure Premium} = \left(\frac{\text{Num.of Claims}}{\text{Num. of Exposure}}\right) \times \left(\frac{\text{Tot.cost of Claims}}{\text{Num.of Claims}}\right)$$

$$\text{Pure Premium} = \text{Tot. cost of Claims}/\text{Num. of Exposure}$$

Pure premium shows industry trends in overall loss costs because of the changes in both frequency and severity. Pure premium only covers cost of claims. The company has expenses for example cost of operation and maintenance, cost reinvestment, profit margin, inflation and commissions. These expenses will be added into pure premium and it is called ***commercial premium***. The difference between them is usually big.

As can be seen above, frequency and severity of loss is used in determining the premium. Frequency is important feature of an insurance portfolio. The ***frequency*** is not the number of accident that occur or number of a contract claim. It is a measure of claims ratio over risk exposure. Below is the formula:

$$\text{Frequency} = \text{Num. of Claims}/\text{Num. of Exposure}$$

Intuitively, the frequency is zero if number of claims is zero and there is no maximum value of the frequency. It is useful to compare between portfolios with different characteristics. For example, let portfolio A has 2,000,000 exposures and 100,000 numbers of claims and portfolio B has 1,000,000 exposures and 100,000 numbers of claims.

$$\text{The frequency of Portfolio A} = 100{,}000/2{,}000{,}000 = 5\%$$

$$\text{The frequency of Portfolio B} = 100{,}000/1{,}000{,}000 = 10\%$$

It can be seen that even though the number of claims are the same, the frequencies are different. Based on example above, it can be said that portfolios B is more likely to have accidents. Analysis of changes in claims frequency can identify general industry trends associated with the incidence of claims or the utilization of the insurance coverage. It can also help measure the effectiveness of specific underwriting actions.

Furthermore, ***severity or average claim cost*** is average value of the cost of claims. It is also called as intensity or severity. For example, in highway, the frequency of accidents is very low but in general the costs associated with accidents at high speeds are high. In this case, it can be categorized as high intensity. Analyzing changes in severity provides information about loss trends and highlights the impact of any changes in claims handling procedures. Below is the formula:

$$\text{Severity} = \text{Total amount of Claim Costs} / \text{Num. of Claims}$$

Claim cost itself is sum of all costs associated with the claim such as cost of investigation, compensation and repairs. In the beginning, claims are not known by the insurer because there is time lag between initiation of claim process and establishment of final cost. The date when the loss happens is called accident date or date of loss. Claims not currently known by the insurer are referred to as unreported claims or incurred but not reported (IBNR) claims. In this case, the insurer estimates of the cost of the claim (called reserve) and will be adjusted later. The report date is the date when the claim is known by the insurer and is called as a reported claim. Before the claim is settled, the reported claim is considered as an open claim. Once the claim is settled, this claim is categorized as a closed claim. ***Loss*** is the amount of compensation paid under the terms of the insurance policy. The difference between claim and loss is term claim to refer to the demand of compensation, while loss to refer to the amount of compensation.

Lastly, to assess the profitability of the business, the insurers uses loss ratio. ***Loss Ratio*** is the portion of total losses that is paid by each premium dollar. It is calculated as:

$$\text{Loss ratio} = \text{Pure Premium} / \text{Average Premium}$$

$$\text{Loss ratio} = \text{Tot. cost of Claims} / \text{Earned Premium}$$

, which Earned Premium is the amount of the premium that the insurers has already earned in relation to the exposure. It can be expressed as Premium * Exposure. Loss ratio is used also to determine the appropriate pricing and the creation of business strategies.

3. STUDY RELEVANCE AND IMPORTANCE

The insurer is the largest bancassurance operator in Portugal. It offers a wide range of life and non-life insurance. Currently, Life insurance of the Insurer is in the third, while Non-Life insurance is in the sixth position in the market. The automobile insurance as part of Non-Life insurance is in the eleventh position with only 2% of market share. Since 2011, a set of company strategies for 2011 – 2015 has been formulated not only to expand the opportunity and to do penetration in the market but also to focus on maintaining the customers. As automobile insurance is a very competitive sector in the market, several actions for automobile insurance have been done to execute the strategies such as creating new product. Additionally, this project becomes the most recent of strategy execution.

Several analyses have been done using data of August 2015 regarding to the automobile performances to give insight about the current condition. The analyses are done based on year on year basis. Total amount of Gross Written Premiums (GWP) in increases about 13.4% compared to August 2014 or 1.4% compared to the market growth. This increment expands the market share from 1.8% to 2%. However, total GWP is still below to the company ambition. Also, the cancellation rate can be improved by understanding the portfolio characteristics. Number of new policies also increases about 26.6% and number of renewal policies increases about 9.3% compared to same month of the last year. However, the increment of number of polices is followed by the increment of gross claim for about 25% year on year. Moreover, the gross loss ratio which is proportion of claim cost and premium is also very high with range between 65% and 72% in 2015. Those analyses become the importance of this study.

Looking at the current condition of automobile figures, one of the initiatives in accord with company strategic is to understand the portfolio characteristics. For instance, the increment of gross claims will impact to the pricing strategies (Werner & Modlin, 2010). While adjusting the pricing should be done carefully since the reaction of each customer can be different (Guelman, et al, 2014a). A rate increase has a direct impact on the premium customer payment, but there is also an indirect impact as a result of the “causal” effect of the rate change on the customer’s decision to renew the policy term (Guelman and Guillen, 2014b). As a result, understanding the nature of price sensitivities at the insurance policy level is valuable for insurers. Moreover, the increment of number of policies will lead to the diversity of policy’s risk profile and characteristics which becomes a challenge for insurer to manage their portfolios. Therefore, a deep understanding of portfolio segmentation is important for the company to maintain the relationships and gain more profit.

For this reason, in this study, portfolio data will be explored by using clustering technique. The clustering algorithm, k-means, will be demonstrated to get the best clustering result. Hence, insurers can gain insights from the discovered clusters and natural patterns can be revealed (Wang & Keogh 2008). For further analysis, will be developed classification model using Decision Tree algorithm in order to build a set of rules to extract the knowledge regarding to the cluster. KPI reports will be created to explore further the behavior of the portfolio inside the clusters.

4. METHODOLOGY

Several methodologies are available in data mining field. Azeved and Santos (2008) do parallel comparison between KDD, SEMMA and CRISP-DM. Based on the study, it is mentioned that CRISP-DM is more complete methodologies compared to others. In CRISP-DM methodology, it includes business understanding phase which is an important phase in data mining project while others do not include it. By understanding the problems, goals and resources in the business understanding phase, it can minimize the project risk (IBM SPSS Modeler CRISP-DM Guide, Chapman, 2000). Some adjustments are made in the methodology in order to adapt to the project. There are four phases in the project: (1) Business Understanding, (2) Data Understanding and Data Preparation, (3) Modeling, and (4) KPI Reports and Analyses. Below is the brief description of the phases based on IBM SPSS Modeler CRISP-DM Guide and Chapman (2000) with some adaption to the project.

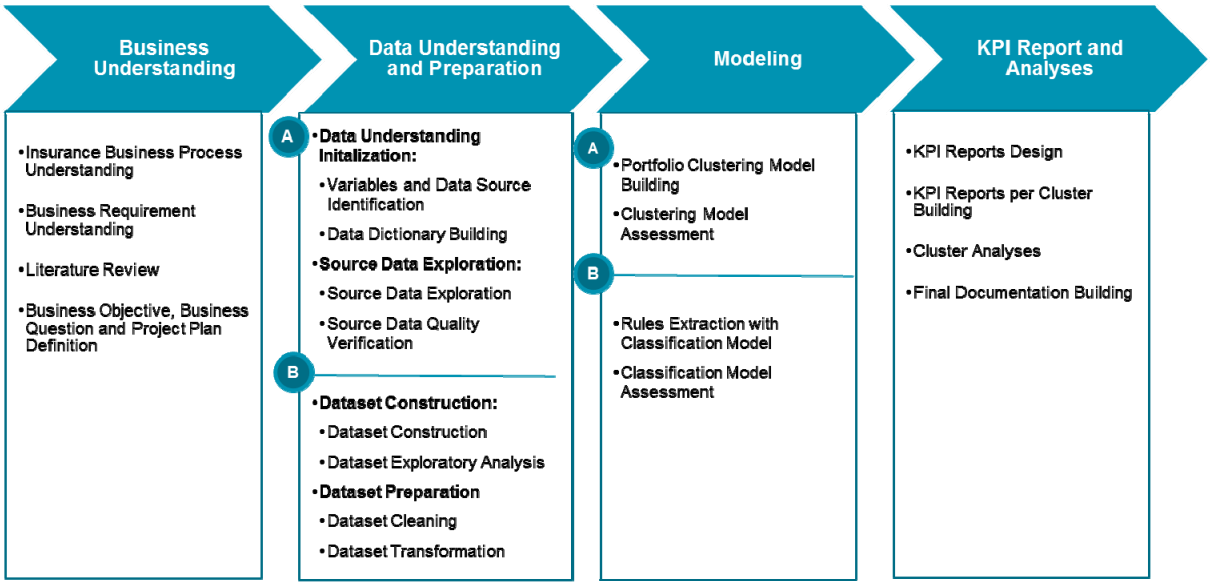


Figure 1: Project Methodologies

4.1. BUSINESS UNDERSTANDING

According to CRISP-DM Guide, it is important to explore what the organization expects to accomplish from the data mining project. Therefore, in the initial phase, the first task is to gain as much insight as possible to understand business objectives by clarifying problems and goals and discussing with as many key people. However, in order to understand the business requirement, it is important to know insurance business concept such as business process, insurance terms and ratios. It is useful not only to understand the context of the problem but also to work on data mining process. According to Ahlemeyer and Coleman (2014), the domain knowledge is a major part of successful data mining, because taking advantage of the knowledge in data mining decision is done most of the time, for example in data preparation on how to treat zero and missing values. The next step is assessing the situation regarding the available resources to the project such as business expert, data experts, dataset and software tools. Lastly, converts the knowledge of business objectives into data mining goals and preliminary plan designed to achieve the objectives.

4.2. 0DATA UNDERSTANDING AND PREPARATION

The data understanding phase starts with data acquisition by performing any analysis and discussion with business experts. These activities give benefit because specific information regarding data problem can be provided, so unnecessary process can be avoided (Wang & Keogh, 2008). In this step, the possible input variables and data sources are identified and collected. Also, data is selected according to the relevancy of data mining goal. The next process is describing the data including format, quantity and sampling of the records by building data dictionary. In this step, simple statistical analysis and some graphs or plots are made in order to find some interesting data that may feed into the transformation phase. The last step is verifying data quality to discover of data completeness, errors and missing data.

The data preparation phase covers all activities to produce the final dataset (data that will be used by modeling). This process normally takes a lot of time and effort in the project. Basic tasks in data preparation phase are as follows:

- **Data Construction and Integration**

  This task is creating new data which operations such as creating derived attributes, generating records, discretization of data by reducing the number of levels of attribute and any transforming data activities. There are two basic methods of integrating data which are merging data set and appending data. Merging data refers to joining together two or more datasets that have different information but have the same object information. In the opposite, appending data refers to joining two or more datasets that have same information but have different object information. Deriving variable building and simplification of the data such as by aggregating are done in this step. The result of these processes is dataset that is used by the project.

- **Data Cleaning and Data Transformation**

  Data cleaning involves several techniques to solve the data problem such as filling in missing values, smoothing out noise, handling outliers, detecting and removing redundant data. In this step, it is possible to drop some variables. Last step is data formatting which refer to data transformation. Sometimes, certain modeling technique requires particular format or order to the data.

4.3. CLUSTERING AND CLASSIFICATION MODELING

In this phase, various modeling techniques are selected and applied. It is possible that this step has been done in business understanding which refers to the specific modeling technique. Modeling is usually conducted in multiple iterations because of the needs to fine-tune the parameters or revert to the data preparation phase for manipulation required by the model. Some experiment will be done before making final conclusions. In order to track the progress with a variety model, it is necessary to keep notes of the parameter setting and data used for each model and description of model result. This phase will be divided into two which are portfolio clustering modeling and classification modeling based on cluster data.

4.4. KPI REPORT AND ANALYSES

At this stage in the project, the data mining models have been built. To understand the result of clustering model in more detail, KPI reports will be developed. Further analyses per cluster will be done. Before proceeding to final step, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. Final documentation is created to record all the activities and results from this project.

5. DATA UNDERSTANDING AND PREPARATION PROCESS

5.1. INTRODUCTION

Real world data are generally noisy, incomplete and inconsistent that may impact the result of data mining (Han and Kamber, 2006). Analyzing data that has such problems can produce misleading or wrong results. Thus, the representation and quality of data is first and foremost before running an analysis. Data understanding and preparation process will be divided into four groups which are Data Understanding Initialization, Source Data Exploration, Dataset Construction and, lastly, Dataset Preparation. The detail of the process is follow:

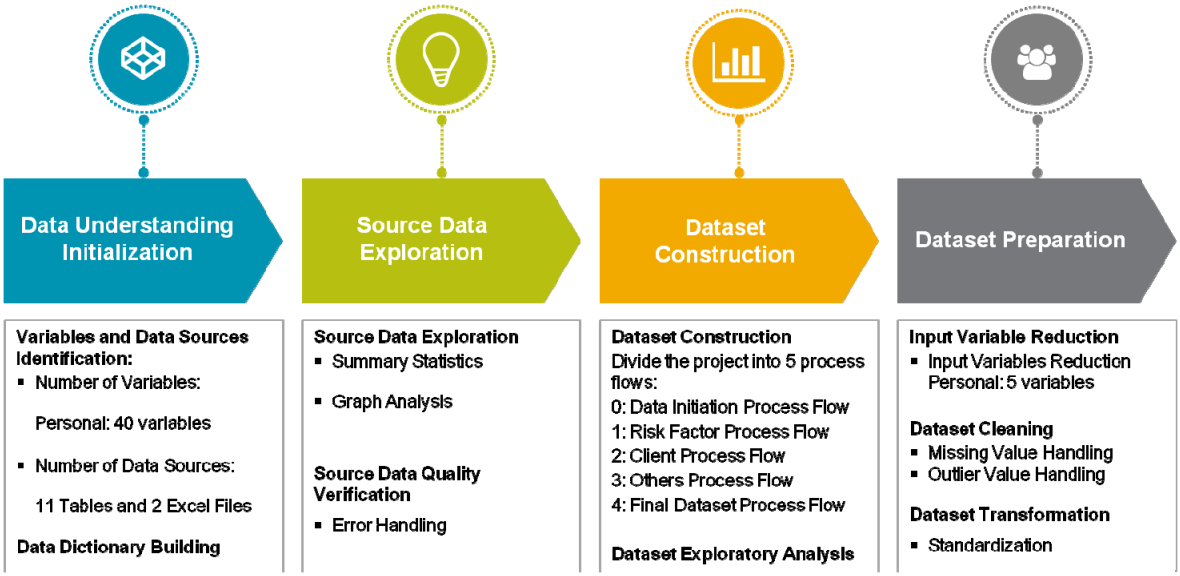


Figure 2: Data Understanding and Preparation Process

5.2. DATA UNDERSTANDING INITIALIZATION

The data understanding initialization is started with the identification of variables and data sources based on Business Understanding phase and Interview with data expert. After the variables and data sources are defined, the next step is building data dictionary to document the metadata and to understand the characteristics of the data.

5.2.1. Variables and Data Sources Identification

The main objective of creating clusters is to determine pricing strategies based on the characteristics of the clusters. The pricing itself is a function of some observable variables describing the risk. In motor vehicle insurance, the pricing can be determined as a function of vehicle characteristics such as power, weight and capacity and customer's characteristics such as age of the policy holder (Pelessoni and Picech, 1998). Therefore, the input variables that are defined will consider risk of portfolio that is used in pricing. The interview with an actuary has been conducted to decide which variables that possible to be used to create the clustering model.

There are two types of customer who buy insurance portfolio which are Individual and Commercial Line. These customer types have different data characteristics. Thus, it is necessary to

split the data into two. The individual data has the majority of the data, in this case, the project only focus on individual data and filter out the company data.

For the initial step, there are 39 possible input variables that has been defined which are grouped into five business perspectives. The business perspectives are driver characteristics, vehicle characteristics, policy characteristics, insurance customer characteristics and demographic characteristics. More detail variables are listed below:

No	Variable Name	Granularity Level
Driver Characteristics		
1	Driver's Age	Object
2	Number of Years of Driving License	Object
3	Driver's Gender	Object
4	Driver's Marital Status	Object
Vehicle Characteristics		
5	Vehicle Type	Object
6	Capital Car (The current price of the vehicles or sum of insured)	Object
7	Vehicle's Age	Object
8	Vehicle Weight	Object
9	Vehicle Power (hp)	Object
10	Vehicle Weight Power Ratio	Object
11	Vehicle Capacity (cc)	Object
12	Vehicle Fuel	Object
13	Vehicle Brand	Object
14	Vehicle Number of Seats	Object
15	Vehicle Number of Doors	Object
Policy Characteristics		
16	Policy Tenure	Policy
17	Payment Type	Policy
18	Claim Frequency	Policy
19	Claim Frequency based on profile	Policy
20	Average Claim Cost of Contract	Policy
21	Average Percentage Premium changes at renewal	Policy
22	Number of average changes of data	Policy
23	Number of Coverage	Policy
24	Collision Coverage Indicator (with or without collision)	Policy
25	Discount per contract	Policy
Insurance Customer Characteristics		
26	Customer Tenure	Customer
27	Number of Active LOB product	Customer
28	Number of Active Non-Life Ex. Health LOB product of clients	Customer
29	Number of Active Motor policies	Customer
30	Number of Active Life policies	Customer
31	Number of Active Health policies	Customer
32	Number of Active Accident policies	Customer
33	Number of Active Multi Risk policies	Customer
34	Customer Risk Scoring	Customer
35	Customer Segmentation	Customer

No	Variable Name	Granularity Level
	Demographic Characteristics	
36	Family Score	Object
37	Education Score	Object
38	Income Focus	Object
39	Rural / Urban	Object

Table 1: Possible Input Variables

After, the possible input variables are identified, the next step is to do the mapping between input variables and data source. The mapping document is called as codebook, but it cannot be presented here because of the confidentiality. In summary, there are 10 tables and 2 excel files of data sources that will be used in this project, they are:

No	Source Name	Description
1	Non-life Policy Table	This table provides information of all non-life policies. In this project this table is used to get the status of policies.
2	Underwriting Table	This table is snapshot values per month based on underwriting point of view. Underwriting View shows the accumulation value annuity. In this view, the claim cost will be reported on the transaction date. This table only contains Motor Policies.
3	Coverage Table	This table is information about portfolio per underwriting period. In this table, information about coverages is used. This table only contains Motor Policies.
4	Transaction Table	This is table transaction. This table is used to get number of changes per policies.
5	Policy Table	This is table from Marketing Data Mart that provides information for all policies in the company.
6	Discount Table	This is table for discount of Motor Policies
7	Risk Score Table	This is table for Risk Score Information
8	Customer Table	This is table from Marketing Data Mart that provides information about clients in the company.
9	Demographic Table	This is table from Marketing Data Mart that provides information about demographics data.
10	Object Table	This is table that contains detail information of objects.
11	Excluded Policies Excel	This is an excel file that contains information of policies that belong to the company. These policies need to be excluded.
12	Urban Rural Excel	This is an excel file that contains mapping between postal code and information of Urban, Rural or Mix.

Table 2: Data Source

5.2.2. Data Dictionary Building

One of the most valuable tools before performing data understanding is building data dictionary. Data dictionary can give idea about what the data contains. The data dictionary includes information about data definitions, as follows:

1. Location of source table
2. The description of the source table
3. Key column(s) of source table
4. Variable Names, Types, Length, English description, example data and some notes.

The full data dictionary document cannot be presented because of the confidentiality. Below is the example of data dictionary which column 'Example Data' uses dummy values:

SAS Folder: **Seguros/Seguros User Data/Marketing/Tabelas**
Description: Table from Marketing Data Mart about all policies
Key Column: Id_Apolice,
ValidoDe

Name	Type	Label	Description	Example Data
Cod_Nif	Character	Id Customer		123456789
Id_Apolice	Character	Policy Number		AU12345678
Cod_RamoSimples	Character	LOB	Line of Business such as health, life, and motor.	AUX-Protecção Automóvel
Cod_RamoAgregado	Character	LOB Group		AU-Automóvel
Dt_EmissaoApolice	Date	Create Date		1/16/13
Dt_InicioApolice	Date	Active Date		1/16/13
Dt_EfeitoAnulacaoApolice	Date	Cancel Date		6/16/15
Nif_EntPagadora	Character	Payer Insurance Id Customer	Life insurance use this customer id	123456789
Nif_EntTomadora	Character	Insured Id Customer	Non-Life insurance use this customer id	123456789
ValidoDe	Date	Start Valid Date		8/31/15
ValidoAte	Date	End Valid Date		12/31/99

Table 3: Example of Data Dictionary

5.3. SOURCE DATA EXPLORATION

5.3.1. Source Data Exploration

The second step after data understanding initialization is starting to explore the source data to get some insights of the data. Following actions are done to analyze the source data:

1. Checking the data integrity
2. Checking key column uniqueness for each table
3. Creating basic information including total records, number of non-missing value (N), Number of missing value, Distinct Value
4. Doing univariate analysis, which are divided into two based on the type of variables:
   - Continuous Variables: Summary Statistics / Histogram Chart / Box Plot
   - Categorical Variables: One-way Frequency / Bar Chart / Pie Chart (see the distribution)

The detail data understanding result cannot be presented because of confidentiality.

5.3.2. Data Quality Verification

While doing exploration of source data, the verification of data quality is analyzed. Several problems such as data error and inconsistencies are found during understanding data process. The summary of error data handling is follows:

Problem	Explanation and Solution
There are individual ages that higher than 100. UnderwritingTable: - IDD_COND_VALOR - ANO_CARTA - CONDSEXO - CONDESTC	This data means that there is more than 1 driver, so in this case the ages of the drivers are unknown. There are rules to change the value into missing. This rule will be applied for below variables for all data: - Age - Marital Status - Gender - Number of years driving license The solution is setting variables to missing for all objects that fit certain rule condition.
There are values of E and T at Marital Status. UnderwritingTable.CONDESTC	This is error data. Need to be changed into missing value.
There are values of E and A at Gender. UnderwritingTable.CONDSEXO	This is error data. Need to be changed into missing value.
There are negative values of total claim cost. UnderwritingTable (I_FECHADO + I_CURSO -I_IDS_credor)	One of the reasons is because of the rejected claim. Set to zero for claim cost below than 0
There are zero weight. UnderwritingTable - PESOBRUT	In reality, it doesn't make sense to have zero weight. In this case, the zero weight values are replaced with the median weight (exclude zero values) based on cross join between TP_VIAT (Vehicle Type) and group of VMARCA (Vehicle Brand). It is applied also for variable WeightPowerRatio.
Zero power of vehicles. UnderwritingTable - POTENCIA	Normally zero power is for trailer which is only available at OTHERS vehicle type. In this case, zero power values are replaced with median of power based on cross join between TP_VIAT (Vehicle Type) and group of VMARCA (Vehicle Brand). It is applied also for variable WeightPowerRatio.
There are differences of Client Type in same policy but different object. UnderwritingTable.CLIENT_T	Use the latest CLIENT_T at policy level and use max CLIENT_T at customer level.
There is difference payment type between UnderwritingTable. tipfracc and NonLifePolicyTable. tipfracc	Based on sample policies, the payment type from UnderwritingTable are correct. Use information from NonLifePolicyTable
Discrepancy status between NonLifePolicyTable.sitapol and UnderwritingTable. I_CANCELATION_POLICE.	Use information from NonLifePolicyTable.

There are policies that have number of coverages equal to zero. Coverage Table	Replace the value with value 1
Mismatch status between NonLifePolicyTable and PolicyTable.	Use information from NonLifePolicyTable
There are customers that do not have customer identification number. In this case it can't be connected to PolicyTable and CustomerTable	Set into missing value
The issue date is not correct based on investigation from Marketing Team. If use this variable the result of customer tenure will be not correct. PolicyTable	To get the first time customer join this company, use minimum date of issue date and start date
Currently, default missing value is -99.99. DemographicTable.IncomeFocus	Need to replace -99.99 into missing value

Table 4: Data Error Handling

5.4. DATASET CONSTRUCTION

5.4.1. Dataset Construction

After process of data source exploration and data quality verification, the next step is building dataset. In this process, the cleaner dataset can be created. The dataset is built using SAS Enterprise Guide. Based on analysis in the previous step, the project was designed into five process flows, as follows:

No	Project Name	Project Description	Table Output
0.	Data Initiation	This is used to create active or in-force policies that become base dataset for the other process flows.	CLS_ACTIVE_OBJECTS CLS_ACTIVE_POLICIES
1.	Risk Factor Process Flow	This is used to create variables related to risk factor	CLS_RISK_FACTOR
2.	Customer Process Flow	This is used to create variables at customer level	CLS_CUSTOMER
3.	Others Process Flow	This is used to create other variables	CLS_OTHER
4.	Final Dataset Process Flow	This is used to create final dataset for individual after filtering Company Policies	CLS_PERSONAL_DATASET

Table 5: Dataset Process Flow

Table CLS_PERSONAL_DATASET is the final dataset that is used for modeling. The dataset uses inforce policies of November 2015 and consists about 120,000 records after excluding commercial line data. There are 39 variables that have been developed based on data understanding initialization phase. The business rules and error handling are also applied in this dataset. Detail process flow diagram from source data into final datasets can be seen at Appendix A.1.

5.4.2. Dataset Exploratory Analysis

Dataset exploratory analysis is also called as descriptive statistics normally based on fundamental statistical analysis. The dataset is examined to get prior information on variables and their correlation before they are selected as input variables in the modeling process. The dataset exploratory analyses are very useful not only for understanding the variables' characteristics but also for giving fundamental statistics information before they are selected for further steps in clustering analysis. In this project, three analyses have been conducted including univariate analysis, summary statistics and variable correlations. The analysis is done using SAS Enterprise Guide.

SAS Guide Tools	Description
Characterize Data	To analyze the data including frequency and univariate analysis
Summary Statistics	To analyze the data including basic statistics and percentiles
Correlation	To analyze the correlation between Interval Variables

Table 6: Exploratory Analysis

Characterize data feature in SAS is a simple approach to quickly getting a summary information of all variables in dataset. In summary, the result of characterize data shows:

a. Frequency table of categorical variables. This summary displays 10 most frequent distinct values per variable, frequent count and percent of total frequency.
b. Descriptive statistics of interval variables. This summary displays basic statistics such as N (Number of non-missing values), NMiss (Number of missing values), Total, Min, Mean, Median, Max and StdMean.
c. Frequency Chart of categorical variables. This chart displays frequent count of each value per variables in bar chart.
d. Histogram Graph of interval variables. This graph displays frequent count of binning value per variables.

For Summary Statistics, there are 3 categories statistics has been analyzed, those are central tendency, dispersion and shape of distribution. These summary statistics are complement from previous analysis, which includes shape of distribution such as percentile and skewness.

Category	Common Measures
Central Tendency	Mean and Median
Dispersion	Standard Deviation and Range
Shape of Distribution	Maximum, Minimum, P1, P5, P25, P50, P75, P95, P99, and skewness

Table 7: Summary Statistics

Correlation analysis is used to detect input variable redundancy in the dataset (Han and Kamber, 2004). The analysis uses correlation coefficient (also known as Pearson's coefficient) to measure how strongly one attribute implies the other. However, the correlation does not imply causality. Below are the input variables that are correlated (more than 80%), those are:

a. Customer Tenure and Policy Tenure have a correlation of 92.3%
b. Driver Age and Years Driving License have a correlation of 83%
c. Number of Coverage and Coverage Collision Indicator have a correlation of 83.96%

The detail information about correlation result can be seen in Appendix A.2. Based on this result, later in dataset preparation, the action will be taken by dropping one of variables that are high correlated. Beside those analyses, some analyses are done by ad hoc based on the need for analysis, such as using One Way Frequencies to show more complete frequency result than characterize data.

5.5. DATASET PREPARATION

From the previous step, the data sources have been integrated. Also, the dataset has been created and ready to be used for the next step. For data preparation to be successful, it is essential to have an overall picture of the dataset. Descriptive data summarization technique that has been done in previous step can be used to identify the typical properties and highlight which data values that have more missing value and should be treated as outliers. After dataset exploratory analysis, further data preparation is done by treating missing data, outlier data and lastly transforming the dataset.

5.5.1. Input Variable Reduction

Based on dataset exploratory analysis, it can be seen that some of variables have high correlation (Appendix A.2). In this case, the input variables can be reduced based on below criteria:

Reduction Type	Criteria	Action	Reduction
High Correlation	Correlation > 80%	Choose One	CustomerTenure vs PolicyTenure (92.3%) DriverAge vs YearsDrivingLicense (83%) NumOfCoverage vs CovCollision (83.96%)

Table 8: Input Variable Reduction

As mentioned above, there are three set variables that have high correlation. In this case it is enough to use only one of them. The variables that are chosen based on high correlation are PolicyTenure, Driver Age and NumOfCoverage. Based on this step, in total there are three variables that are not dropped. However, in extraction rule model, input variable CovCollision is used instead of NumOfCoverage. It is because CovCollision gives clear split in business rule.

5.5.2. Dataset Cleaning

5.5.2.1. Missing Value Handling

During data preparation, the important aspect is treating missing data. Collica (2006) does the experiment on the effect of missing value in the clustering result. It concludes that the missing value gives different result on variable selection. The Variable Selection is attempting to analyze which variable has the largest impact on the target variable penetration. This causes clustering algorithm to cluster observations differently. Furthermore, Matignon (2007) says that ignoring the incomplete observation and analyze only those records that consist of complete data may lead to discarding information that is quite useful because of the other non-missing values.

Refaat (2007) says that there are three basic strategies to treat missing value, those are: ignoring the record, substituting value based on general characteristics of the variable such as mode or mean or user defined value, and imputing the value by attempting to recreate the value. The last strategy is trying to generate the value using the result of the model by predicting the value that is missing. SAS has provided features to handle missing value by several method imputation and

substitution. Below is the list of imputation strategies (see 'Strategies' column, part 'Substitute or Impute the Value').

Strategies	Continuous	Categorical
Ignore the record	Only for data error and small data	Only for data error and small data
Drop the variable	If missing > 20%	-
Substitute / Impute The value	Mean Median Zero or any user-defined value Maximum value Minimum value	Mode User-defined value New category indicates missing value Impute Tree

Table 9: Missing Value Handling Strategies

After analysis and discussion with business expert, below is the decision of strategies for the missing value treatment for Individual Portfolio Dataset.

Variable Name	Percentage of Missing Value	Strategies
DriverGender	0.90%	Impute using Tree Algorithm
DriverMaritalStatus	1.52%	Impute using Tree Algorithm
VehicleFuel	1.34%	Impute using Tree Algorithm
CustomerSegmentation	14.19%	New Category of Missing Value
CustomerRiskScoring	2.72%	Median
CustomerTenure	4.34%	Median
DriverAge	0.90%	Median
PaymentType	7.97%	Median
WeightPowerRatio	0.81%	None
YearsDrivingLicense	0.90%	Median
EducationScore	11.20%	Median
FamilyScore	11.20%	Median
IncomeFocus	11.20%	Median
UrbanRural	11.20%	Median

Table 10: Missing Value Handling

According to Refaat, substitute value can create bias in the data. In the case of continuous variable, using median usually creates less bias than does using mean. More, Refaat mentions that this method is simple and straightforward. The justification for selecting specific replacement option is often easy. For the same reason, the substitute value with median is done for interval variables except for WeightPowerRatio variable. The missing value of WeightPowerRatio variable categorized as an undefined value resulting from mathematical operation. This happens because trailer vehicle has zero power values. The replacement with new values such as zero or median values will produce misleading and wrong imputation. Considering small amount of missing data which is 0.81%, there is no taken treatment, so that SAS algorithm will automatically ignore these values. However, in scoring data, missing values will be handled automatically by decision tree as rule extraction.

For categorical variable, the strategy is creating new category for missing value if the missing value is higher than 10% and using imputation strategy if the missing value is not too many. The

imputation method is a strategy to recreate the data by simulating model to predict the values that is missing. The reason of using imputation strategy instead of substitute of mode value is because the proportion of mode values are not significantly high. In this case, tree algorithm is used to do the imputation.

5.5.2.2. Outlier Handling

Outliers are extreme values that appear in the dataset and can be extremely small or extremely large. Outliers give impact to the clustering result as the clustering algorithm uses distance measurement to group the data. The k-means as clustering algorithm is sensitive to the outliers because they can cause cluster center shifted. SAS Miner has provided the feature to detect the outliers. The identification of outliers can be divided into two groups based on the type of variables whether the variable is interval or categorical.

Interval	Description
Business Consideration	Use specify limit
± 3 Standard Deviation	Specifies the number of standard deviations from the mean to be used as cutoff value. That is, values that many standard deviations away from the mean will be filtered out.
± 9 Median Absolute Deviation (MAD)	Specifies the number of deviations from the median to be used as cutoff value. That is, values that are that many median absolute deviations away from the median will be filtered out
± 0.5 Extreme Percentile	Cutoff percentiles for extreme percentiles
± 9 Modal Center	Specifies the number of spacing from the modal center to be used as cutoff value. That is, values that are that many spacing away from the modal center will be filtered out

Table 11: Interval Outlier Handling Strategies

Nominal	Description
Business Consideration	Use user specify limit
Rare Percentages (1%)	Specifies the percentage of rare values
Rare Values (1)	Specifies the number of rare values

Table 12: Nominal Outlier Handling Strategies

Based on discussion with the business expert, below is the decision of outlier handling per variables.

a. Interval Variables

Variable Name	Outlier Handling Rules	Variable Name	Outlier Handling Rules
DriverAge	None	PaymentType	None
VehicleAge	Extreme Percentile	TotalDiscount	Extreme Percentile
VehicleCapacity	Extreme Percentile	AvgPremiumChg	Modal Center
VehicleNumOfDoors	None	NumOfAccidentPolicies	Extreme Percentile
VehicleNumOfSeats	Extreme Percentile	NumOfHealthPolicies	Extreme Percentile
VehiclePower	Extreme Percentile	NumOfLOB	None
VehicleWeight	Median Absolute Deviation	NumOfLifePolicies	Extreme Percentile
WeightPowerRatio	Modal Center	NumOfMotorPolicies	Extreme Percentile
YearsDrivingLicense	None	NumOfMultiRiskPolicies	Extreme Percentile
CapitalCar	Extreme Percentile	NumOfNonLifeLOB	Extreme Percentile
ClaimFreq	Extreme Percentile	NumOfcoverage	None
ClaimFreqByProfile	Extreme Percentile	RiskScore	None
CustomerTenure	Extreme Percentile	EducationScore	None
PolicyTenure	Extreme Percentile	FamilyScore	None
AvgChgFreq	Extreme Percentile	IncomeFocus	None
AvgClaimCost	UserSpecify >5000		

Table 13: Interval Outlier Handling

b. Nominal Variables:

Variable Name	Outlier Handling Rules
Cov_Collision	Rare Percentages (1%)
UrbanRural	Rare Percentages (1%)
DriverGender	Rare Percentages (1%)
DriverMaritalStatus	Rare Percentages (1%)
VehicleBrand	Rare Percentages (1%)
VehicleFuel	Rare Percentages (1%)
VehicleType	Rare Percentages (1%)

Table 14: Nominal Outlier Handling

5.5.3. Dataset Transformation

As mentioned previously, the cluster algorithm depends on the measuring the distance. Different length measurements will give impact on the result. Variables with large values contribute more to the distance measure than variables with small values. For this reason, before performing cluster analysis, it is necessary to scale or to transform the variables. In this project, the dataset is transformed by adjusting the values using standardization. Standardization is used to scale the value such that the dataset has zero mean and unit standard deviation. This makes variables have same scales and are able to compare.

6. MODELING

6.1. MODELING FRAMEWORK

As stated in the study objective, this section describes the model framework that covers clustering model and rule extraction model development. Below is the model framework for this project:

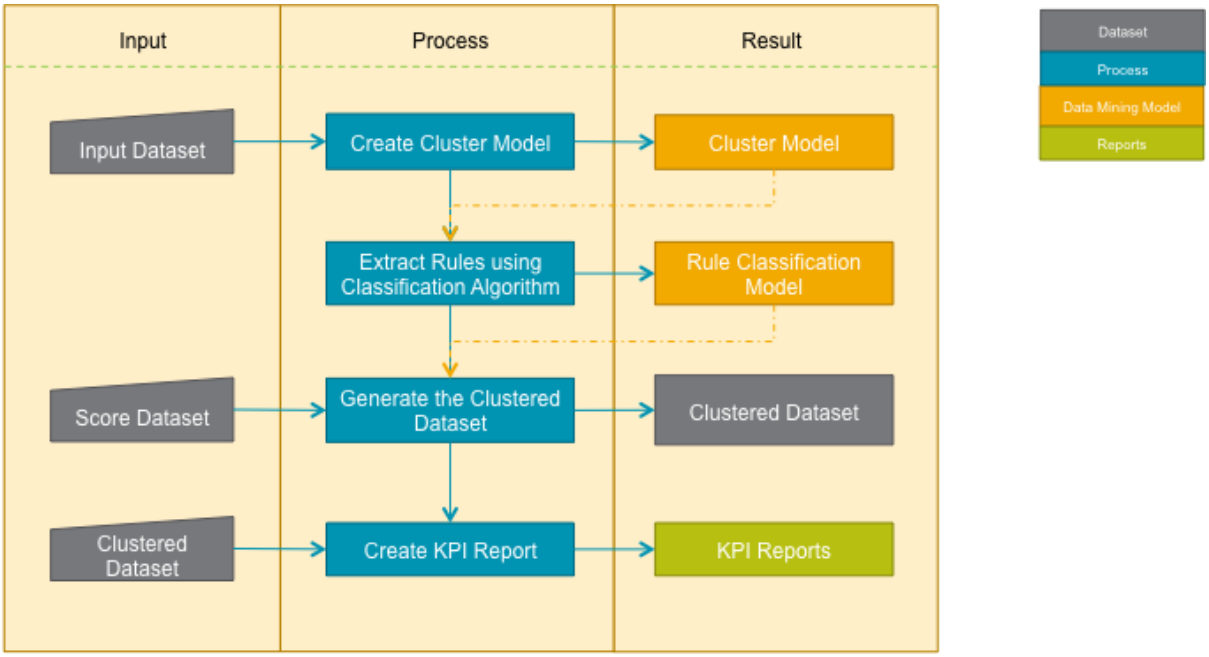


Figure 3: Model Framework

The dataset prepared in the previous step is used to create cluster model. Several experiments will be taken using variation of input variables and parameters such as number of clusters and seed initialization method. The result of this process is to create the best clustering model that has meaningful clusters based on business justification. After the cluster model is created, the extraction rules using classification model is built on the respect of cluster id. The input variables that are used in the classification model are the same input variables but using original variables before to do imputation and standardization. The purpose of extraction rules is not to get the best extraction rule classification model, but to understand how the cluster model is formed. After the rule extraction model is created than score dataset is applied using the rule extraction model instead of clustering model. Lastly, to learn the characteristics of the cluster, KPI Report per cluster will be built that covers information such as number of new businesses, loss ratio, average severity and claim frequency.

Ghoson (2010) carried out comparative analysis of decision tree induction and clustering techniques of three popular data mining software tools, those are SAS Enterprise Miner, SPSS Clementine and IBM Intelligent Miner. The analysis includes four main criteria of performance, functionality, usability and auxiliary Task Support. Using those criteria, weighted average was calculated. According to the results, SAS Enterprise Miner is the best data mining software among them for both decision tree induction and clustering techniques. Therefore, this study supports the use of SAS Enterprise Miner as a tool to create predictive and descriptive model in this project.

6.2. CLUSTER MODEL

6.2.1. Cluster Development Scenario

Conceptually, clustering algorithms use input variables to distinguish each observation and to form concepts that characterize each created cluster based on subset of the input variables (Guo, 2003). Hence, cluster development scenario in this study is divided into two steps, as following:

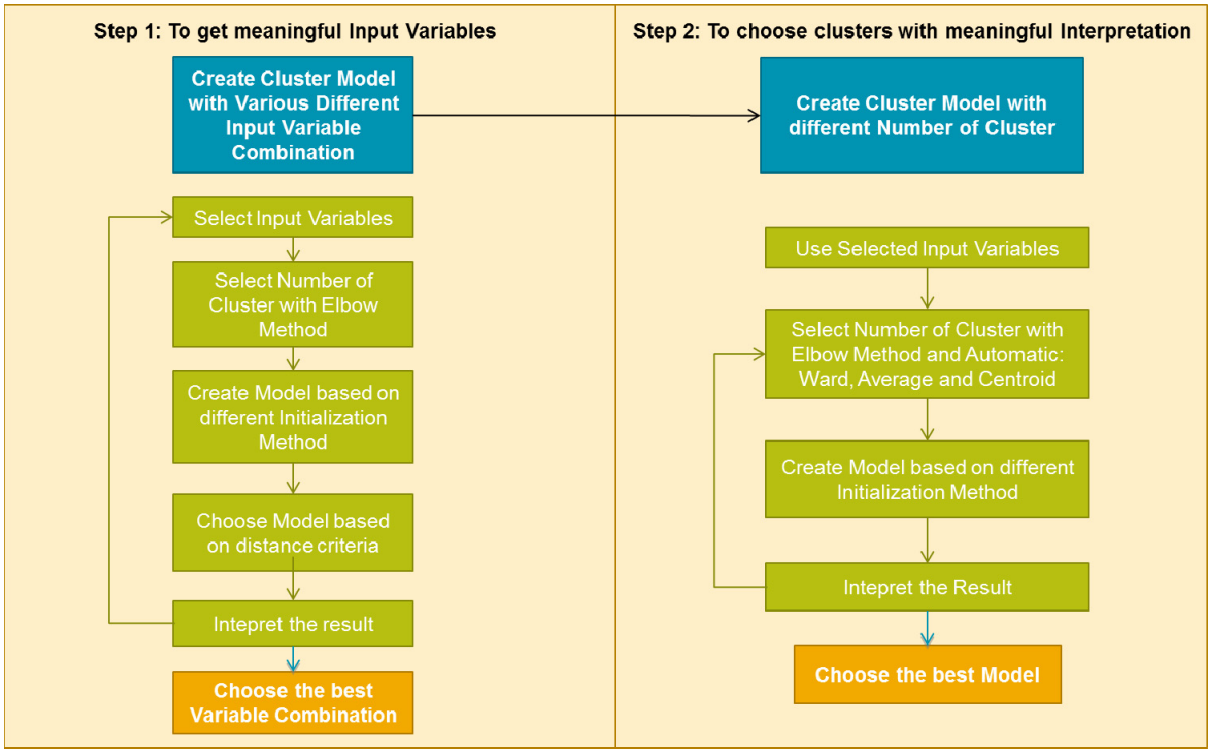


Figure 4: Cluster Development Scenario

Step 1. To get meaningful input variables

Mirkin states that the cluster finding stage involves not only clustering algorithm but also interpretation in terms of the variables. In this stage, the business expert may analyze the clustering results and suggest a modification of the input variables by adding or removing the variables. Supported by Advanced Business Analytics Notes from SAS, one of the criteria to choose input variables is the variable that meaningful to the analysis objective because it is important for the interpretation and explanation of the generated clusters.

Similar process, in this project, there will be several trials of building cluster model with different input variable combinations. First of all, input variable combination need to be decided. Then, a number of clusters is chosen based on elbow method. There are three cluster models that are created using difference seed initialization methods. Those are MacQueen, First and Princomp. According to SAS Help, MacQueen method uses MacQueen's k-means algorithm to compute initial seed, First method will select the first complete cases as initial seed and Princomp method uses principal component to set seed initialization on evenly grid in the plane of the first two principal components. While other methods, those are Full Replacement and Partial Replacement, are not used because these methods are for outlier detection purpose. From these three models, only one

will be chosen based on distance criteria that maximize dissimilarity between clusters and similarity inside the cluster. Lastly, interpretation of cluster is done. Based on an expert judgment and interpretation, in term of variables, the expert may not see no relevance in the result and suggest new input variable combination. Therefore, the same process will be done repeatedly until the most meaningful input variable combination is found.

Step 2: To choose clusters with meaningful interpretation

The second step is to choose cluster model with meaningful interpretation. With selected input variables from step 1, a number of clusters need to be decided. SAS Miner provides two options for specifying the number of clusters. They are fixed number of clusters option and automatic option. To get fixed number of cluster, elbow method is used. Elbow method shows information of sum of squared errors (SSE) for each value of k. The goal is to choose a small value of k that has low SSE. Second option is using automatic mode from SAS Enterprise Miner. According to SAS Miner references, there are two processes. First of all, SAS Miner do preliminary clustering using the value from 'Preliminary Maximum' property used as beginning of the number of clusters. Second process uses multivariate means of the clusters in the first process as inputs. Hierarchical algorithms using agglomerative are performed to reduce the number of clusters. Then the smallest number of clusters that meets following condition is selected:

- The number of clusters must be greater than or equal to the number that is specified in the Minimum value of Selection Criterion properties.
- The number of clusters must have cubic clustering statistic values that are greater than the CCC threshold as specified in the Selection Criterion properties.
- If there is no matched number of clusters, then the number of clusters will be set to the first local peak.
- The final clustering produced using the smallest number of clusters that meet those criteria as inputs.

There are three methods to calculate the distance of hierarchical clustering; those are Average, Centroid and Ward. Average method calculates based on average distance between pairs of observations, one in each cluster. The Centroid method calculates distance between two clusters based on (squared) Euclidean distance between their centroids or means. Ward method calculates using ANOVA sum of squares between the two clusters summed over all the variables. Finally, similar like previous step, seed initialization methods need to be specified, those are MacQueen, First or Princomp.

6.2.2. Selected Meaningful Input Variables

Based on business understanding phase, it is decided to have 39 possible input variables that reduced into 36 because of the high correlated variables. The clustering models in step 1 use several variable combinations from these 36 input variables. There are 16 variable combinations that have been exercised to get the meaningful result. The table of variable combination can be found in Appendix B.1. The interpretation process has been done during this phase to get meaningful clusters. The major decision of input variable choices will be discussed here.

In the beginning, the variable combinations included insurance customer characteristics such as number of motor policies and number of LOB. After the analysis, it was found that the cluster

model was not relevance with the objective which is for pricing strategy. The model was more useful for marketing because it showed a cluster that customer buy more LOB products which related more to loyalty program for example. So that it was decided to not include insurance customer variables in the cluster model. Next, it was also tried to include claim frequency and average claim cost as input variables. It was expected that the data could be split based on risk characteristics. However, after several trials, it was realized that claim frequency and average claim cost variables cannot be used to group new portfolios, since those variables are unknown in the beginning of contract. Because of this, risk input variables were excluded. However, the cluster model without risk input variables need to be checked whether it is able to produce distinguish cluster based on risk or not. The result shows that the cluster model can produce significantly different claim frequency even though without risk variables (Figure 7). Further, gender as input variable is also discussed. According to Vaughan (2007), traditionally, an automobile insurance uses customer criteria such as age, gender and marital status as risk rating. However, gender neutral rating becomes a debate. Some jurisdictions do not allow to use gender as rating variables (Modlin, 2010). In European Union legislation requires insurer to demonstrate that the gender has correlation to underlying risk. To avoid controversy, it is decided that gender variable is excluded. Lastly, it was also decided not to include demographic characteristics. The family score and education score seem not relevance to the objective. More, rural/urban and income variables were not used by the model to distinguish the cluster. It shows that the variable importance equal to zero.

Finally, there are 13 input variables chosen to build clustering model. These selected input variables are important factors for pricing modeling. Those are driver characteristics including age and marital status, vehicle characteristics including vehicle types, capital car, vehicle age, vehicle weight, vehicle horse power, vehicle weight power ratio, vehicle capacity and vehicle fuel. Other than those, input variable of policy tenure, risk scoring and number of coverage are also included. Number of coverage indicates whether the drivers buy only third party liability coverage or also own damage coverage.

No	Variable Name	Description	Granularity Level
1	Driver's Age	The age of the Driver	Object
2	Driver's Marital Status	Marital Status of the Driver	Object
3	Vehicle Type (PrivateVehicleInd)	Vehicle Type to indicate whether the vehicle is private or not.	Object
4	Capital Car	Sum insured of vehicle	Object
5	Vehicle's Age	The age of vehicle	Object
6	Vehicle Weight (kg)	The Weight of Vehicle	Object
7	Vehicle Power (hp)	Horse Power of Vehicle	Object
8	Vehicle Weight Power Ratio	The weight of vehicle divided by power (hp)	Object
9	Vehicle Capacity (cc)	Motor capacity of Vehicle	Object
10	Vehicle Fuel (GasolineInd)	Fuel type of Vehicle	Object
11	Number of coverage	Number of coverage per policies that is bought by policy holder. The maximum number of coverage is 14	Policy
12	Policy Tenure	Policy Tenure	Policy
13	Customer Risk Scoring	The risk scoring	Customer

Table 15: Selected Input Variables

6.2.3. Cluster Model Parameters

To initialize k-means algorithm, it needs to specify the number of clusters and initial centroid. The number of clusters should not be too few because this could result in a lack of discriminating information for adequate segmentation. In the meantime, the total number of clusters cannot be too many which cause to less number of observations in each cluster. As mentioned in the cluster development scenario, there are two methods on choosing number of clusters. First is elbow method and second is SAS automatic option.

According to Kodinariya and Makwana (2013), elbow method is a visual method shows, that at some point, the value for certain k will drop dramatically and become flat when the k increase further. However, there is possibility that the graph does not show the elbow clearly and sometimes it shows several elbows. Berry and Linof mention there is no a priori reason to select a particular value. Further, they mention that finding a number of clusters is an iterative process instead of computer program. After defining one value of k, the result needs to be evaluated and then try again with another k value for several trials. The cluster result needs evaluation on a more subjective basis to determine their usefulness for a given cluster.

The elbow method is generated using R program. Several trials have been performed to get stable number of k. For the overall elbow graphs can be seen at Appendix B.2. Based on the elbow graphs, it seems the number of clusters are not defined very clearly. However, there are some points that give drop value before it increases again. Regarding the elbow charts, number of clusters equal to 9 and 12 are used to create a cluster model. Below is one of the example of elbow chart.

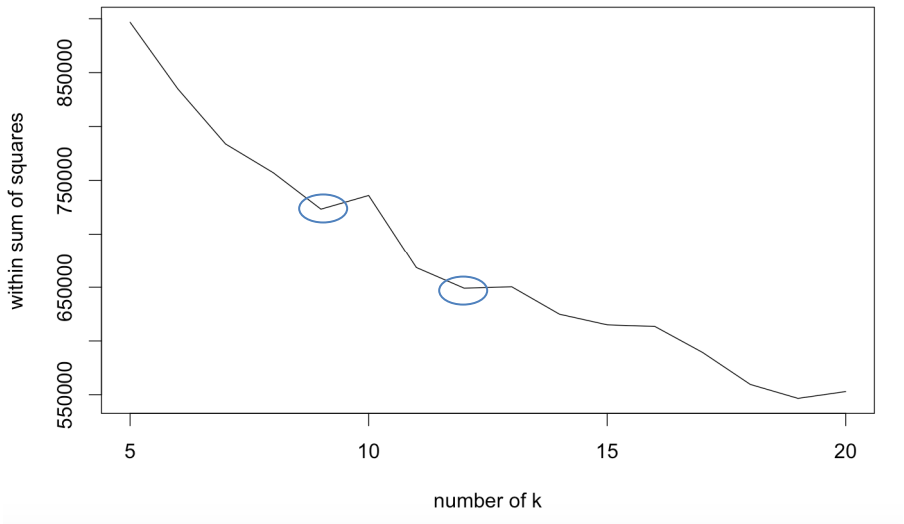


Figure 5 Elbow Method

As mention previously, there are three types of distance calculation of cluster model which are Ward, Average and Centroid. Also, there are three types of seed initialization method that are used those are MacQueen, First and Princomp. Based on combination of clustering method and seed initialization method, automatic option produces 9 cluster models. The result of automatic option is presented below. A negative CCC is not meaningful (Sas Institute, 2012). Therefore, as seen at table above, the suggested number of clusters are 18, 19, 21 and 36.

Cluster Method	Seed Initialization	Number of Cluster	CCC Value
Ward	MacQueen	21	450.752
Ward	First	36	523.324
Ward	Princomp	20	-
Average	MacQueen	5	-493.411
Average	First	5	-383.642
Average	Princomp	19	153.444
Centroid	MacQueen	6	-282.755
Centroid	First	5	-383.642
Centroid	Princomp	18	190.914

Table 16: Number of Cluster of Automatic Option

Considering business objective, number of clusters less than 15 is preferable. Accordingly, number of clusters equal to 9 and 12 are selected. The next step is to run the model within number of clusters equal to 9 and 12 using different seed initialization methods. The generated cluster models for each number of clusters are evaluated by comparing the average of distance between records (Maximum Distance from Cluster Seed) and average of distance to nearest cluster (Distance to Nearest Cluster are observed). The evaluation criteria are the lower Maximum Distance from Cluster Seed the better the model and the higher Maximum Distance from Cluster Seed the better the model. Second criteria to choose the model is that the more data evenly distributed among clusters the better the model. Lastly, it is also checked how well the misclassification rate for decision tree to extract the rules. This decision tree is not final model for rule extraction. The purpose of creating decision tree model is to give an idea the complexity of the rule extraction model. If the misclassification is high, it means that the model has difficulty to extract the rules. For this purpose, the decision tree models are created using the same parameter setup and using whole data, so that they are comparable. Below is the comparison table:

Number of Cluster	Seed Initialization	Avg of Maximum Distance from Cluster Seed	Avg of Distance to Nearest Cluster	Number of Cluster < 5%	Misclassification Rate
12	MacQueen	6.965294585	2.861719384	6	N/A
12	First	7.299467122	3.034373771	3	7.76 %
12	Princomp	6.992300525	2.775963448	2	12.74 %
9	MacQueen	7.789305845	2.803474485	3	N/A
9	First	7.439175434	3.101177878	1	N/A
9	Princomp	7.557609172	2.822550654	1	N/A

Table 17: First Stage Cluster Model Comparison

For Number of Clusters equal to 12, MacQueen model gives the best result in term of intra cluster distance, but a lot of small clusters below than 5%. There are 6 out of 12 clusters below 5%. This result is not preferable. Therefore, the choice is either First model or Princomp Model. First model gives better result on the maximization of inter-cluster distance, while Princomp Model gives better result on the minimization of intra-cluster. For misclassification rate criteria, the Princomp Model gives very bad result, it is 12,74%. It means that how this cluster is formed is difficult to understand. In conclusion, The First model is selected for number of clusters equal to 12. For Number of clusters equal to 9, it can be seen that First model is better than the other models in term of distance criteria and distribution. In this case, First model is selected.

The next step is evaluating the characteristics of cluster model results and the usefulness of the clusters. There are two cluster models which both are using First model as seed initialization. Those are First 9 Cluster Model and First 12 Cluster Model. After analyzing those two models, it is decided to use First 12 Cluster Model, because it gives more detail result such as, it produces policies with own damage coverage group into two clusters instead of one.

6.2.4. Final Cluster Model

Twelve clusters have been selected as the final results using First 12 Cluster Model. The number of observations in the smallest cluster is 1.93% (Cluster 1) and the highest cluster is 23.38% (Cluster 10) of overall data. Further analysis is taken to see risk profile of average claim cost and claim frequency. For Cluster 1, despite the number of objects is small, it gives unique characteristics. It has very low claim frequency compare to others. Because of this, cluster 1 is preserved. Below is the original proportion of cluster model.

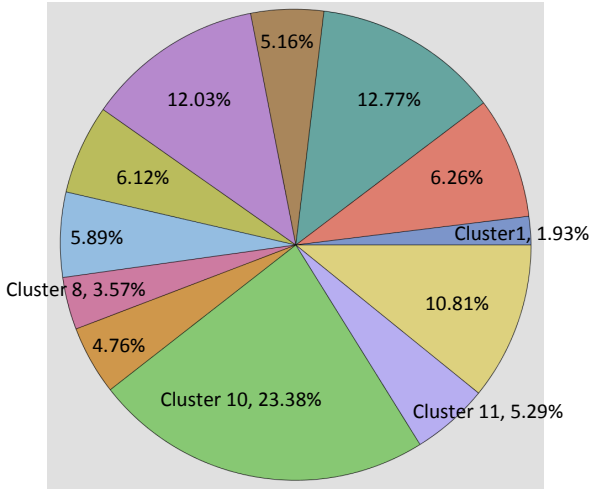


Figure 6: Proportion of First 12 Cluster Model

As seen in graph above, cluster 10 is a big cluster; in this case, second clustering model is done to split this cluster. It uses the same process like the previous one, however the input variables are adjusted since cluster 10 have specific characteristics such as married customers and use gasoline. Because of that, some input variables will not give any advantages on splitting the clusters. The input variables are reduced into driver age, capital car, vehicle age, vehicle weight, vehicle horse power, vehicle weight power ratio and vehicle capacity. Furthermore, the automatic option produces the same number of clusters which are 2 clusters. With number of cluster equal to 2, cluster models are created using different seed initialization. The distance of these models are very similar, however First model produces very slightly better than the others and has better misclassification rate. Therefore, the second clustering of Cluster 10 uses First Cluster Model.

Number of Cluster	Seed Initialization	Avg of Max. Distance from Cluster Seed	Avg of Distance to Nearest Cluster	Misclassification Rate
2	MacQueen	164.8679792	1.835295176	5.78%
2	First	164.8541244	1.836721481	4.69 %
2	Princomp	164.8576496	1.835356629	4.90 %

Table 18: Second Stage Cluster Model Comparison

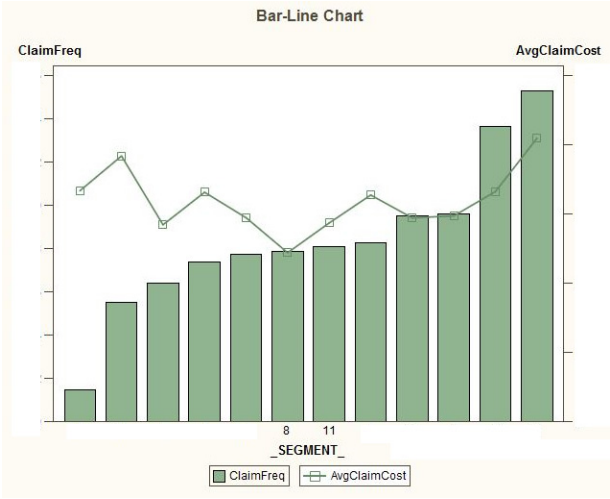


Figure 7 Claim Frequency of 12 First Cluster Model

Looking at the risk profiles (Figure 6), cluster 8 ad cluster 11 have similar risk profiles. Also based on the analysis both clusters have similar characteristics. Therefore, it is decided to combine these two clusters. Final Clusters are presented below:

Original Cluster Id	Percentage	New Cluster Id	Percentage	Notes
8	**3.6%**	8	**8.9%**	Combination of original Cluster 8 and 11
11	5.3%			
10	**23.4%**	10	**12.1%**	Splitting of Cluster 10 into Cluster 10 and 11
		11	**11.2%**	

Table 19: Final Cluster Model

After final cluster model is formed, the next thing to do is to extract the rule using decision tree algorithm with the respect of cluster id. The rule extraction will be run twice, the first is to extract rule of the original clusters and the second one is to extract rule of tenth cluster model.

6.3. RULE EXTRACTION MODEL

Decision tree algorithm is chosen for rule extraction because it uses English rules that is easy to understand by human. The rule extractions are found in the leaves of decision tree model. They use IF-THEN rules and have mutually exclusive where there is no rule conflicts (Han and Kamber, 2006). The experiments are divided into two stages. The first stage is to get the composition of training dataset and validation dataset. Second stage is to get final rule extraction using dataset composition as a result in the first stage. According to Advanced Business Analytics from SAS Institute, there are different summary statistics correspond to prediction types. Those are decisions, rankings and estimates. Based on the nature of the prediction model in this study, decision statistic type is used to judge the model in the first stage. A decision prediction is rated by their accuracy. The proportion of disagreement between prediction and outcome is calculated to get misclassification rate. Berry and Linoff (2004) mention that when a decision tree is used for predicting the score, having a large number of leaves is advantageous because it can predict more accurately. In contrary, when the objective is to generate rules, the fewer the rules the easier to understand. Therefore, this criterion is used to create final rule extraction model in the second stage.

Training dataset is used to train the initial model. Validation dataset is used to adjust the initial model so that it more general and to avoid over fitting model. There are several methods to partition the dataset into training and validation dataset which are simple random, stratified and cluster. According to SAS documentation, simple random method gives every observation to have the same probability of being chosen to one of the partitioned datasets. Stratified method gives all observations within each stratum to have an equal probability to be chosen to one of the partitioned datasets. The stratum is formed based on variables from input data. Cluster method is not used since it needs to set partition role of the cluster variable which is not available. After the experiments, it shows that stratified gives very slightly better misclassification rate. In this case, stratified method is used.

The input variables that are used for rule extraction are the same as input variables for cluster model except for NumberOfCoverages variable. Unlike in cluster model that the input variables are standardized, the decision tree uses original value of the input variables in order to get informative rule. Recall that NumberOfCoverages and CovCollision Indicator have high correlation. The higher number of coverages the higher probability that a policy holder buys collision coverage and the opposite way. Because of this, only one variable is chosen in the cluster model. As seen in the variable combination list, both variables are tried to get meaningful cluster model. Based on the result, it is concluded that NumberOfCoverages variable gives preferable results instead of CovCollision variable in the same variable combination. It is also supported that k-means algorithm works better in interval variable which is NumberOfCoverages variable. However, in business, it is required to have clear split whether the policy holders have collision coverage or not. Given that, in rule extraction model, it will be used CovCollision variable instead of NumberOfCoverages variable.

The first stage is to compare models using 100% dataset that is called as DT100 model, 80% training dataset and 20% validation dataset that is called as DT80 model, 70% training dataset and 30% validation dataset that is called as DT70 model and 60% training dataset and 40% validation dataset that is called as DT60 model. The full trees are created with the same pruning condition, which sets minimum leaf size equal to 300. Based on the trials, the models show that binary tree with maximum branch equal to 2 give the best result. Below is the summary of the models:

Rule Extraction Model	Number of Rules	Misclassification Rate	Misclassification Validation Rate	Score Misclassification Rate
DT100	52	7.76%		8.07%
DT80	38	8.22%	8.10%	8.45%
DT70	33	8.53%	8.89%	8.95%
DT60	31	8.36%	8.77%	8.84%

Table 20: Decision Tree Comparison

According to Berry and Linoff (2004), a large difference of misclassification rate between training and validation dataset is a symptom of an unstable model. It can be seen that the difference of misclassification rate of DT80 model is the lowest difference compare to DT70 and DT60 model. Therefore, DT100 is proven as the best model, followed by DT80, DT60 and the worst is DT70 model. Model of DT100 is part of consideration because to extract the rule, it is not necessary to split the data between training and validation dataset, especially when the objective is only want to extract the knowledge on how the clusters are formed. However, it is arguable because the rule extraction

model will be used not only for current dataset but also the future dataset. Therefore, considering generalization concept, it is selected to use DT80 instead of DT100 as base for rule extraction. The second stage is to adjust the rule extraction model using 80:20 compositions.

As Han and Kamber (2006) mention that decision trees may suffer from sub tree repetition and replication. It can be large and difficult to interpret, that is why pruning the resulting rule set is important. Moreover, the tree pruning methods also address over fitting problem. The Decision Tree tool from SAS provides feature to create the model autonomously or interactively. By using interactive decision tree, the rule extraction can be more controllable such as when it wants to be pruned and which variables that is used to split the data. In addition, building models interactively can be informative (Advanced Business Analytics, SAS Institute). There are two models that are created using interactive model. The difference is only on the splitting of policy tenure variable. Model Int80-1 is using 8.5 as a splitting point and Model Int80-2 is using 7.5 as splitting point. Even though Model Int80-1 produces better misclassification model, model Int80-2 is chosen as base model based on business decision. Further analysis of rule extraction is done. From 16 rules, it is reduced into 14 rules that is presented as Int80-Simplified model.

Rule Extraction Model	Number of Rules	Misclassification Rate	Misclassification Validation Rate	Score Misclassification Rate
Int80 - 1	16	12.45%	12.66%	12.80%
Int80 - 2	16	12.59%	12.76%	12.93%
Int80 Simplified	14	13.27%	13.29%	13.55%

Table 21: Interactive Rule Extraction Model

Model of Int80-simplified is the final rule extraction model of the original cluster model. The second rule extraction of Cluster 10 Model is simpler. It produces 4 rules, but then it is simplified into only 2 rules with misclassification rate becomes 11.6% from 4.69%. The rules divide the original tenth cluster into cluster 10 and 11 that become low vehicle power and high vehicle power. Then, they are combined with the rule set from Int80-simplified model. Detail rule extraction can be seen at Appendix B.3. Final cluster labels are obtained by using the rules that were semantically interpretable. Below are the cluster labels of Final Ruled Cluster Model (the order is not the same with cluster proportion at Figure 6):

Cluster Label
Non Private Vehicles, Low Weight and Capacity (trailers, motorbikes, small tractors)
High Weight or High Capacity Vehicles, such as Vans, Pick-ups, Truck, Heavy Machinery and powerful Motorbikes
Private Vehicles, Single/Divorced Drivers, Gasoline
Private Vehicles, Married Drivers, Gasoline, Low Vehicle Power
Private Vehicles, Married Drivers, Gasoline, High Vehicle Power
Private Vehicles, Single/Divorced Drivers, Diesel, Low/Medium Capacity
Private Vehicles, Single/Divorced Drivers, Diesel, High Capacity,
Private Vehicles, Married Drivers, Diesel
Private Vehicles, The worst risk score
Private Vehicles, Own Damage, Low Power and Low Capital Car
Private Vehicles, Own Damage, High Power or High Capital Car with Low power (new vehicles)
Private Vehicles, Old policies

Table 22: Final Ruled-based Cluster Model

7. KEY PERFORMANCE INDICATOR REPORTS

Further analysis of Rule-based Cluster Model is taken by analyzing the key performance indicator. Basically, most of the KPIs measured are related to the insurance ratio. These metrics include risk analysis, for instance exposure, earned premium, average premium, claim frequency, average severity and loss ratio. Some of the metrics are analyzed based on type of coverage such as collision and TPL coverage and policy status such as new business or renewal. Below is the list of KPI:

Number of Policies

These KPIs are used to monitor number of policies in term of proportion and growth per cluster. The proportion of new business in policies can be used to monitor when the insurer wants to grow the number of policies at certain clusters. The growth number of policies is designed to provide comparison between current values to previous month values. The growth is accounted from the number of new policy and renewal policy.

- Cluster Proportion of Number of New Business Policies
- Growth Number of Policies per Cluster

Exposure and Earned Premium

Exposure is a prorate theoretical risk based on the total exposure. The analysis of proportion exposure of clusters often ranks risks according to their probability of occurring. While earned premium is total premium that collected by insurer. Analysis of earned premium per cluster shows which cluster gives more income to the company. It also analyzes the earned premium from new business.

- Percentage of Cluster Exposure in total of Portfolio
- Cluster Proportion of Earned Premium
- Cluster Proportion of New Business Earned Premium

Average Premium

Average premium metric is to know how much premium paid in average by policy holders. The original average value and the variation of average premium will be analyzed. Variation is calculated by comparing the average premium of 12 month rolling from current month and average premium of 12 months rolling from previous month. The average premiums in total and new business are monitored per cluster.

- 12 Month Rolling of Average Premium per Cluster
- 12 Month Rolling of Average Premium Variation per Cluster
- 12 Month Rolling of New Business Average Premium per Cluster
- 12 Month Rolling of New Business Average Premium Variation per Cluster

Average Severity

Average severity measures claim cost filed by the policy holders. The purpose of this KPI is to help to assess the risk associated per cluster characteristics and to adjust policy pricing together with claim frequency. It is useful to analyze the behavior of the policies with difference coverage type.

- 12 Month Rolling of Average Severity of Total per Cluster
- 12 Month Rolling of Average Severity of Collision Coverage per Cluster
- 12 Month Rolling of Average Severity of TPL Coverage per Cluster

Claim Frequency

Claim frequency shows the likelihood that a loss will occur. A low frequency means the loss event is possible, but the event is rarely happened in the past and it is not likely to occur in the future. The opposite way, a high frequency means the lost event happens regularly in the past and expected to occur regularly in the future. The analyses of claim frequency total and per coverage type are presented.

- 12 Month Rolling of Claim Frequency of Total per Cluster
- 12 Month Rolling of Claim Frequency of Collision Coverage per Cluster
- 12 Month Rolling of Claim Frequency of TPL Coverage per Cluster

Loss Ratio

Insurance is the business of managing risk, so that the insurer needs a thorough understanding of the incurred loss ratio. If the value is higher than expected, then further investigation is required to figure out what has happened. If it is lower than expected, it could indicate irrelevant products or difficulties in claiming, possibly affecting customer satisfaction. Therefore, loss ratio is an important KPI. Loss Ratio is the relation between company cost and collected premium. In this report, Loss Ratio will be analyzed per coverage types such as collision coverage and TPL coverage and per customer status such as new business, renewal and total and combination of both.

- 12 Month Rolling of Loss Ratio of total per Cluster
- 12 Month Rolling of Loss Ratio of Collision Coverage per Cluster
- 12 Months Rolling of Loss Ratio of TPL Coverage per Cluster
- 12 Months Rolling of Loss Ratio of New Business Total per Cluster
- 12 Month Rolling of Loss Ratio of New Business Collision Coverage per Cluster
- 12 Months Rolling of Loss Ratio of New Business TPL Coverage per Cluster
- 12 Months Rolling of Loss Ratio of Renewal Total per Cluster
- 12 Month Rolling of Loss Ratio of Renewal Collision Coverage per Cluster
- 12 Months Rolling of Loss Ratio of Renewal TPL Coverage per Cluster

Average Number of Policies per Customer

These KPIs are used to analyze the average number of policies per line of business for each customer. The line of business is divided into four groups; Motor (Motor is part of Non-Life LOB), Non-Life, Life Risk and Health. One of the purposes of analyzing this KPI is to know whether customer loyalty may relate to the certain cluster characteristics.

- Average Number of Motor Policies per Customer per Cluster
- Average Number of Non-Life Policies per Customer per Cluster
- Average Number of Life Risk Policies per Customer per Cluster
- Average Number of Health Policies per Customer per Cluster

The KPI reports are presented in excel dashboard. There are four main dashboards that are created, those are:

1. Cluster Summary Dashboard, shows summary characteristics of 12 clusters.
2. Key Performance Indicator Summary Dashboard. On this dashboard, the KPI summary of all clusters will be showed. It is also completed with the scattered plot of cluster ids with the

following metrics:

a. Renewal Loss Ratio vs New Business Loss Ratio
b. Renewal Claim Frequency vs New Business Claim Frequency
c. Renewal Average Premium vs New Business Average Premium
d. Loss Ratio vs % Growth Policies
e. Loss Ratio vs Claim Frequency

3. Key Performance Indicator Trend Dashboard. On this dashboard, the 13-month-trend of each KPI per cluster will be showed. It also provides the features to do comparison trend among cluster id and allows the user to select the value based on product or tariff.
4. Cluster Characteristics Dashboard. On this dashboard, the characteristics per cluster will be showed. The information such as proportion of vehicle type, vehicle fuel, vehicle age, driver age and marital status proportion are presented. Also, breakdown KPI metrics per product can be analyzed in this dashboard.

8. CLUSTER CHARACTERISTIC ANALYSES AND DISCUSSION

8.1. CLUSTER CHARACTERISTIC ANALYSES

The cluster labels have been assigned based on automatic characterization from rule extraction model. The analyses of cluster characteristics are using score data of March 2016. Statistical analyses and KPI analyses are performed in order to identify characteristics of the clusters. Some insights could be gained into the nature pattern of the automobile insurance portfolio. Based on these analyses, interesting patterns can be derived per cluster that can help management to pay more attention to any potential problems or opportunities. Due to the confidentiality of the data, not all KPI metrics will be discussed in this report.

Single/Divorced Drivers who use Gasoline Private Vehicles

The policy holders in this cluster have marital status of 63.5% single, 26.4% divorce and 10% widow/widower. Vehicles in this cluster use gasoline fuel and has Private Vehicles type.

Married Drivers who use Gasoline Private Vehicles with Low Power

The policy holders in this cluster are married; about 96.3% aged more than 35 years old. The vehicles use gasoline fuel and have Private Vehicles type. In average, the vehicle power is low, about 57, because based on the rule extraction, the highest power is 70. About 97% of the vehicle aged over 5 years in which 56% of the vehicles in this cluster aged more than 16 years.

Married Drivers who use Gasoline Private Vehicles with Medium / High Power

The policy holder characteristics are similar like previous cluster. The policy holders in this cluster are married and about 96.6% people have age greater than 35 years old. The vehicles use gasoline fuel and have Private Vehicles type. About 95%, the vehicle ages are above 5 years which 36% vehicle age are between 11 to 15 years and 41.4% are above 16 years. The difference is on the vehicle power which is about 93 in average.

Single/Divorced Drivers who use Diesel Private Vehicles with Low/Medium Capacity

There are about 70% single policy holders, about 23% Divorce and about 7% are Widow / widower. This cluster has the biggest proportion of policy holder below 35 years old which is about 33.6% and about 74% are below 50 years old. The vehicles use Diesel and about 68% vehicle age are below 10 years. The average of vehicle power is 83 which is lower than cluster below (single/divorced drivers who use Diesel Private Vehicles with High Capacity). The vehicle capacity is below 1700.

Single/Divorced Drivers who use Diesel Private Vehicles with High Capacity

This cluster has similar characteristics as previous cluster. About 67% are Single, 24.7% are Divorce and about 7.5% are Widow/Widower. About 28.6% are below 35 years old and about 68% are below 50 years old. However, the vehicle age in average is older than cluster above (Single/Divorced Drivers who use Diesel Private Vehicles with Low/Medium Capacity). About 70% are above 10 years. Also, the vehicles in this cluster have higher capacity with minimum capacity is 1700 and higher average of vehicle power which is 122.

Married Drivers who use Diesel Private Vehicles

This cluster contains married drivers. About 94.5%, the drivers have age greater than 35 years old. The vehicles that they use are Diesel fuel and Private Vehicle Type.

Own Damage Coverage and Vehicles with Low Power and Low Capital Car

The main characteristic of this cluster is the policy holders buy own damage coverage. About 58.7% are married and 26.44% are single. Most of the vehicles are new. About 92%, the vehicle ages are below 10 years. The difference with the other own damage coverage cluster below is that this cluster has low sum insured value about 9800 Euro and low power which is in average about 92.

Own Damage Coverage and Vehicles with High Power or High Capital Car but Low Power

The main characteristic of this cluster is the policy holders buy own damage coverage. About 65% are married and 21.5% are single. Almost 50% the vehicle ages are below 4 years and about 92% are below 10 years. About 88.8% the vehicles use Diesel Fuel. It has the highest average weight and average capacity like next cluster. The vehicles in this cluster have high sum insured value, which is in average about 24.300 Euros. About 43% vehicles have Mercedes-BMW brand. The average of vehicle power is high, about 163.

High weight or high capacity vehicles

About 69% are married drivers and about 21% are single drivers. This cluster is mix vehicles between Non Private vehicles that have high weight or high capacity and Private vehicles that have high weight power ratio. It has the highest average weight and average capacity together with previous cluster. The different is previous cluster has higher average vehicle power than this cluster and most of the vehicles are new. About 97% of these vehicles use Diesel. The examples of vehicles in this cluster are heavy machinery, pick-ups, truck and powerful motorbikes.

Non Private vehicles with low weight and capacity

Inside this cluster, 58.6% are married and about 31.3% are single. This cluster consists only Non Private vehicle types; those are about 87.6% of Motorcycles and about 10.7% of other vehicle types such as trailers and small tractors. This cluster has the lowest vehicle weight, vehicle power and vehicle capacity compare to the other clusters.

The Worst Risk Score

The main characteristic of this cluster is the policy holders have high risk score. About 52% are married and 33% are single. Most of the vehicles have age more than 11 years, about 75%.

Old Policies

The main characteristic of this cluster is the policy tenure of the portfolio is greater than 7 years. This cluster may important for the company because of their loyalty. About 67% are married and 19.8 % are single. Only about 2.5% policy holders have age below 35 years old. As expected, the vehicle ages are 73% more than 11 years. About 47% use Gasoline fuel, 31.6% use Diesel and 20.9% use other fuel.

8.2. FURTHER CHARACTERISTICS DISCUSSION

The different portfolio cluster characteristics have been defined based on input variables. It can be seen that the clusters have unique characteristics among them. Further cluster analyses are explored in this section based on KPI metrics. Thorough analyses of portfolio clusters have been done to increase knowledge of cluster characteristics related to the KPI metrics. However, due to the confidential reason, disclose further specific or more detailed descriptions of the clusters identified cannot be presented. The KPI values cannot be shown and only key analyses of KPI metrics will be discussed.

Cluster Name	% # Written Policies	% Exposure Proportion	%Earned Premium Proportion	Avg Premium	Claim Freq	Loss Ratio	% Growth	%Renewal Rate
Non Private Vehicles, low weight and capacity								
High weight or high capacity (i.e. heavy machinery, powerful motorbikes and big vehicles								
Single/Divorced Drivers, Private Vehicles, Gasoline								
Married Drivers, Private Vehicles, Gasoline, Low Vehicle Power								
Married Drivers, Private Vehicles, Gasoline, Medium/High Power								
Single/Divorced Drivers, Private Vehicles, Diesel, Low/Medium Capacity								
Single/Divorced Drivers, Private Vehicles, Diesel, High Capacity								
Married Drivers, Private Vehicles, Diesel								
The worst risk score, Private Vehicles								
Own Damage Coverage, Private Vehicles, Low Power and Low Capital Car								
Own Damage Coverage, Private Vehicles, High Power or High Capital Car with Low Power								
Old Policies, Private Vehicles								

Table 23: Key Performance Indicator

First of all, the analysis of percentage exposure proportion is done. The comparison to the percentage of written policies proportion shows that there is not much differences between them. It means that the exposure is accordance to cluster proportions. However, the percentage of earned premium proportion shows significance differences for some clusters. It means that some clusters have higher premium compare to the others. It can be explained because there are clusters that buy more coverages so that the premium become higher. The comparison of average premium between new business and renewal portfolio per clusters can be seen at Figure 8. The graph shows that the average premium between them is linear that means there is no significance difference premium between new business and renewal portfolio. More, it clearly shows that there are two clusters that have high average premium. While others are concentrated in certain range of average premium.

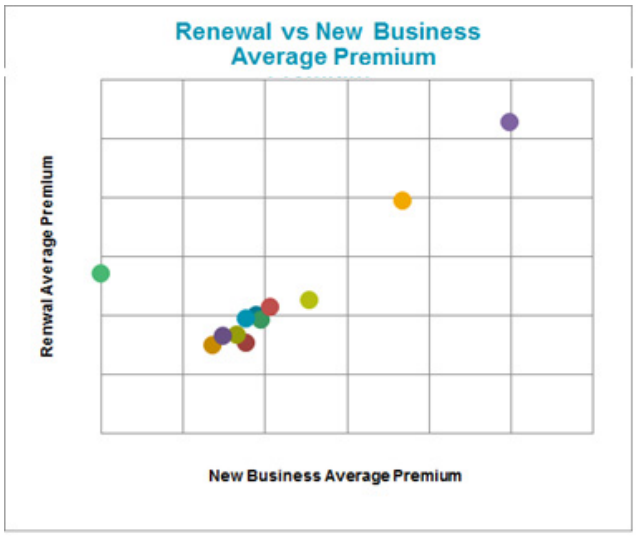


Figure 8: Renewal vs New Business Average Premium

The Figure 9 below shows the claim frequency and average severity per clusters that is ordered from the lowest claim frequency to the highest claim frequency. It can be seen that the cluster model is able to find clusters with significantly different claim frequency without including claim frequency as input variables. In other words, clustering is able to distinguish between low and high risk groups. The graph shows that there are two clusters that have significantly higher claim frequency compare to others. Those clusters are clusters with own damage coverage, so high claim frequencies are expected. Besides that, it can be revealed some knowledge that clusters with certain vehicle characteristics give higher risk profile no matter the marital status is. But with the same vehicle characteristics, it shows that married drivers have lower risk profile compare to single/divorced drivers. More, as seen (Figure 9), the claim frequency values fluctuate among the clusters. In contrast, the average severities of the clusters are not so fluctuating.

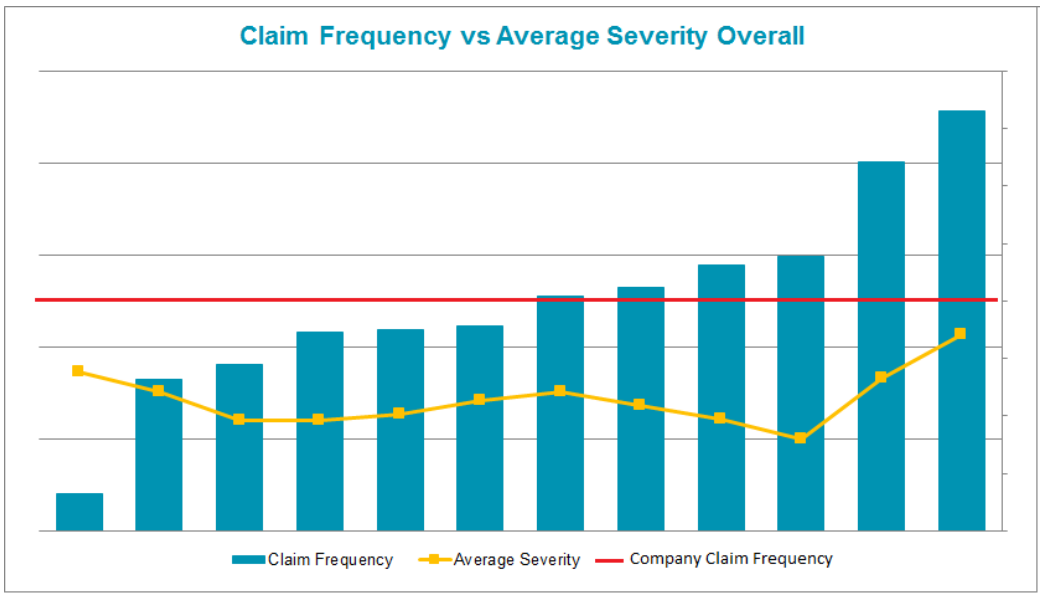


Figure 9: Claim Frequency vs Average Severity

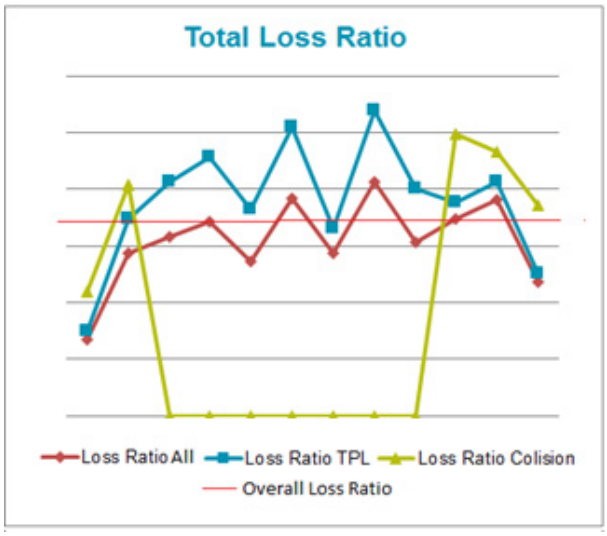


Figure 10: Loss Ratio per Cluster

The loss ratio graph acima presents that the cluster model is also able to show which clusters are profitable and which clusters are unprofitable. It is presented that there are three loss ratio spikes over overall loss ratio line. The insurer needs to pay attention to these unprofitable clusters by doing deeper analyses. Furthermore, together with risk analysis of claim frequency, it may give the company insight on how to adjust the pricing both new business and renewal pricing (Figure 11). According to Figure 11 (A), there are clusters that have high risk and high loss ratio. In this case, the insurer may increase the price to get lower loss ratio or more profits. More, it is presented that some high risk clusters have reasonable loss ratio (Figure 11-B) which means that the insurer apply appropriate price to these clusters. There are some clusters that show low loss ratio and low risk (Figure 11-C). In this case the insurer may decrease the price of these clusters to increase renewal rate and to attract more new business coming.

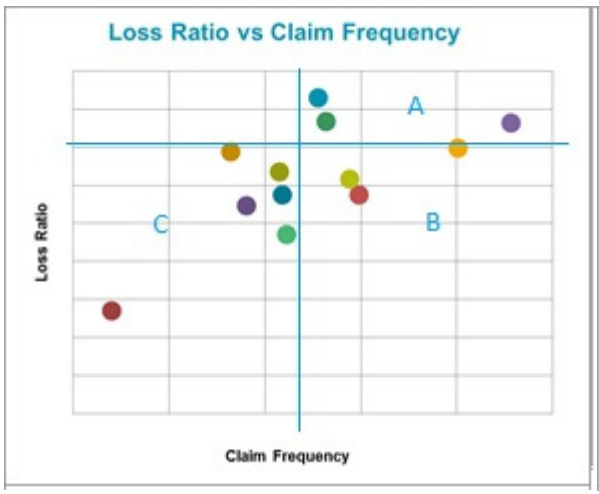


Figure 11: Loss Ratio vs Claim Frequency

Figure 12: Loss Ratio vs %Growth Policies

Equally important, combining analysis with new business growth, the insurer may monitor and control on which clusters they want to grow. Based on the analyses, it is found that the three highest growth of new business clusters are also the three highest loss ratio (Figure 12). Which means

that the insurer grows more on unprofitable clusters. Based on this fact, the insurer should consider carefully in which cluster they want to grow and take some actions accordingly.

Overall, based on the analyses above, it can be concluded that the KPI metrics give complement information to the technical characteristics of the clusters. They give different behavior of each cluster that lead the insurer to make the pricing strategy. For instance, the cluster model produces significantly different loss ratio and risk profile among clusters. As such, the more accurately pricing models can be built based on the identified groups and their characteristics. Furthermore, the insurer may not only take advantages for the pricing strategy but also for marketing campaign. The insurer may focus to certain clusters such as low risk clusters by targeting specific customers and portfolios according to technical characteristics of those clusters. As such, a desire for more focused marketing campaigns on the target group with high potential of profitability can be achieved.

10.3 Process continuity

The ongoing clustering analysis based on the most recent data should be performed. In addition, the cluster model should be updated regularly, particularly, because in dynamic industries new segments of the portfolios could emerge and existing segments can also evolve with new characteristics. The more accurately the groups and their characteristics are identified, the more accurately strategies can be taken. The knowledge discovered from the clustering analysis can be implemented in further insurance business applications. For instance, pricing is generally formulated based on risk profiles. As such, the more accurate pricing models can be built based on risk profile each clusters. In addition, pricing for renewal business can be adjusted according to the presented KPI metrics per cluster. Not only that, the insurer may target the most profitable clusters for marketing campaign.

9. SUMMARY

9.1. CONCLUSION

This project is started with business understanding and followed by data understanding. Several data treatments also have been done to prepare the data before building the model. First of all is to fix data error such as replacing old codes that are not used anymore with missing value, confirming inconsistency data and fixing incorrect values. Second treatment is handling the missing value with replacement and imputation. After that is handling outlier value, since k-means is very sensitive to the outlier. Finally, is doing dataset standardization, so that the values are comparable.

The first process to find useful portfolio segments is creating cluster model using k-means algorithm. The challenges in this process are to decide the input variables, the parameter values and to interpret the model. Common approach to this problem is running the algorithm multiple times with different input variables and parameter values including different number of k and different initialization seed. Each created model need to be validated. The validation is not straight forward, since it relies not only on the statistic outcome but also on the interpretation from business domain expert. After several trials of creating cluster model and discussion with business expert, finally thirteen input variables are defined. These input variables are important factors for pricing modelling which associated with main objective of this project. There are two cluster models that need to be created because the selected initial cluster model has one big cluster size, so that need to be split. The final cluster model has twelve clusters with different technical characteristics. The second process is extracting rule of the selected cluster model using decision tree. This contributes to an easy understanding of the cluster characteristics. The result of this process is a set of rules that describes how the clusters are formed which give final rule-based cluster model.

To increase knowledge of defined rule-based cluster model, four KPI dashboards have been created. With this dashboards, the company is able to monitor the characteristics and behavior of each clusters. Based on the analyses of technical characteristic and KPI metrics per clusters, it can be found that some insights can be revealed. The insurer may take benefits from discovered model to fine-tune further strategies such as to maintain relationships by decreasing the price of the cluster with low loss ratio and low risk profile for example, to attract new customers and to gain more profit. The insurer may build portfolio-oriented pricing which are more personalized pricing strategies based on the defined clusters. Furthermore, the company may targeting certain characteristics of clusters for marketing campaign. Lastly, the portfolio clustering may be integrated in a holistic pricing strategy for both new business growth and renewal book to achieve the ultimate optimized portfolio mix or cluster mix according to company strategy of growth and profitability.

Finally, the knowledge discovery from clustering model is a continuous process. The monitoring to the clusters and adjustment to the cluster model should be done regularly to get recent portfolio characteristics. The discovered characteristics can be beneficial to the insurer such as for pricing adjustment, risk management and marketing campaign. At last, the same method of rule-based cluster model can be implemented in other insurance industry.

9.2. LIMITATION AND FURTHER WORK

This study focuses only at personal policy holders as the majority customer type in this company. It can be expanded by analyzing the cluster characteristics of commercial policy holders, which will be useful to understand whether commercial customers are profitable or not. In order to do that, further study can be taken; since commercial customers are profitable then company can expand the business to the commercial. Back to this study, it focuses only for motor policies, but similar study can be applied into different line of business such as health insurance, life insurance and others.

10.BIBLIOGRAPHY

Advanced Business Analytics (2012), SAS Institute, Inc.

Ahlemeyer A, Coleman, S.S (2014), A Practical Guide to Data Mining for Business and Industry, Willey.

Azevedo, A., Santos, M.F. (2008), KDD, SEMMA and CRISP-DM: A Parallel Overview, IADIS European Conference Data Mining.

Berry, M.J.A., Linoff, G.S. (2004), Data Mining Techniques: for Marketing, Sales and Customer Relationship Managemet, Wiley Publishing.

Brito, P.Q., Soares, C., Almeida, S., Monte, A. Byvoet, M. (2015), Customer Segmentation in a Large Database of an Online Customized Fashion Business, Robotics and Computer-Integrated Manufacturing, 36, 93-100.

Chapman, P. et al, (2000). CRISP-DM 1.0 – Step by step data mining guide.

Collica, R.S. (2006), CRM Segmentation and Clustering Using SAS Enterprise Miner, SAS Institute Inc.

Dolgui, A., Proth, J.M. (2010). Pricing strategies and Models, Annual Reviews in Control, 34, 101-110.

Ghoreyshi, S., Hosseinkhani, J., (2015), Developing a Clustering Model based on K-means Algorithm in order to Creating Different Policies for Policyholders in Insurance Industry, International Journal of Advanced Computer Science and Information Technology (IJACSIT), 4(2), 46-53.

Ghoson, A.M.A (2010), Decision Tree Induction and Clustering Techniques In SAS Enterprise Miner, SPSS Clementine, and IBM Intelligent Miner – a Comparative Analysis, Journal of Management & Information System 14(3), 57 – 70.

Gibert, K., Aluja, T., Cortes, U. (1998), Knowledge Discovery with Clustering Based on Rules. Interpreting Results, Principles of Data Mining and Knowledge Discovery, 1510, 83-92.

Guelman, L. Guillen, M. (2014a). A Causal Inference Approach to measure price elasticity in Automobile Insurance, Expert Systems with Applications, 41, 387-396.

Guelman, L., Guillén, M., Pérez-Marín, A.M. (2014b). A survey of personalized treatment models for pricing strategies in insurance, Insurance: Mathematics and Economics, 58, 68-76.

Guo, L. (2003). Applying Data Mining Techniques in Property/Casualty Insurance, Forums of the Casualty Actuarial Science.

Han, J., Kamber, M. (2006), Data Mining Concepts and Techniques. Morgan Kaufmann Publishers.

Hasan, M.S., Duan, Z.H., (2015), Hierarchical k-means: A Hybrid Clustering Algorithm and Its Application to Study Gene Expression in Lung Adenocarcinoma, Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology, 51-67. doi:10.1016/B978-0-12-802508-6.00004-1.

He, Z., Xu, X., Huang, J. Z., & Deng, S. (2004). Mining class outliers: Concepts, algorithms and applications in CRM. Expert Systems with Applications, 27, 681–697.

IBM SPSS Modeler CRISP-DM Guide.

Jurek, A., Zakrzewska, D., (2008), Improving Naïve Bayes Models of Insurance Risk by Unsupervised Classification, Proceedings of the International Multiconference on Computer Science and Information Technology, 137–144.

Kodinariya, T.M., Makwana, P.R. (2013), Review on determining number of Cluster in K-means Clustering, International Journal of Advance Research in Computer Science and Management Studies 1(6), 90.

Liao, S.h., Chen, Y.j, Lin, Y.t. (2011), Mining customer knowledge to implement online shopping and home delivery for hypermarkets, Expert Systems with Applications, 38, 3982-3991.

Madhulatha, T.S. (2012), An Overview on Clustering Methods, IOSR Journal of Engineering, 2(4), 719-725.

Matignon, R. (2007), Data Mining Using SAS Enterprise Miner, John Wiley & Sons, Inc.

Mirkin, B. (2005), Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall/CRC.

Pelessoni, R., Picech, L. (1998), An Application of Unsupervised Neural Network in General Insurance: The Determination of Tariff Classes, General Insurance Convention and Astin Colloquium.

Peng, Y., Kou, G., Shi, Y., Chen, Z. (2005), Improving Clustering Analysis for Credit Card Accounts Classification, Computational Science – ICCS 2005, 548–553. Doi: 10.1007/11428862_75

Refaat, M. (2007), Data Preparation for Data Mining Using SAS, Morgan Kaufmann Publishers.

SAS Reference Help, SAS Enterprise Miner.

SAS Technical Report A-108 (1983), Cubic Clustering Criterion, SAS Institute, Inc.

Stomer, T. (2013). Optimizing Insurance Pricing by Incorporating Consumers' Perceptions of Risk Classification, Working Papaers on Risk Management and Insurance, 140.

Vaughn, E. J., Vaughn T., Fundamentals of Risk and Insurance, 2008, 10th Edition, John Wiley & Sons.

Wang, X., Keogh, E. (2008), A Clustering Analysis for Target Groups Identification by Locality in Motor Insurance Industry, Soft Computing Applications in Business, 113-127.

Werner, G., Modlin, C. (2010). Basic Ratemaking, Casualty Actuarial Society.

Williams, G.J., Huang, Z., (1997), Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases, Advanced Topics in Artificial Intelligence, 340-348. Doi: 10.1007/3-540-63797-4_87

Wu, J. (2012), Advances in K-means Clustering: A Data Mining Thinking, Springer.

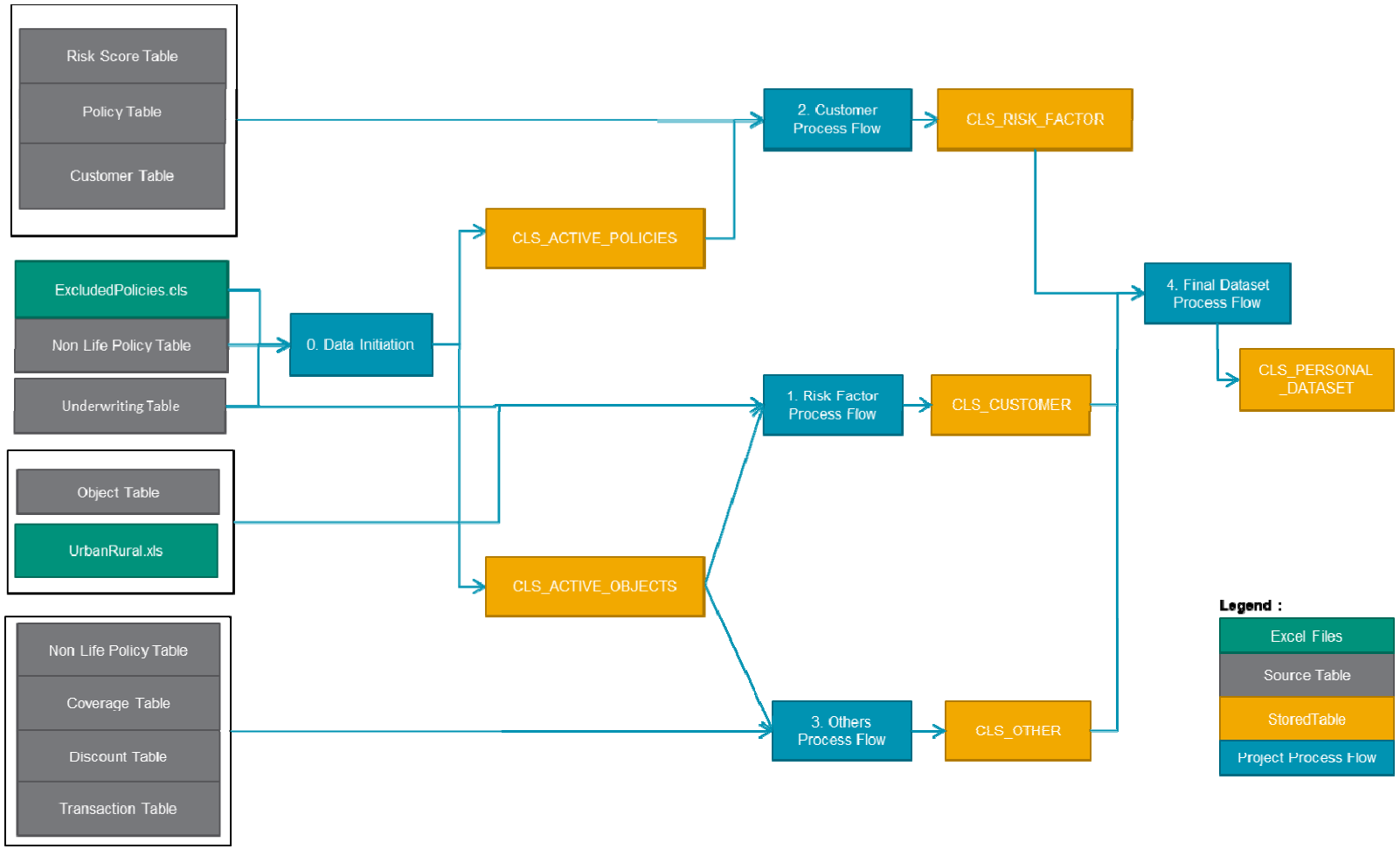
Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M. (2001a). Clustering Technique for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry, International Journal of Intelligent Systems in Accounting, Finance & Management, 10, 39-50.

Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M. (2001b). Modelling the Effect of Premium Changes on Motor Insurance Customer Retention Rates Using Neural Networks, Proceedings of the International Conference on Computational Science-Part II, 390-399.

APPENDIX A. DATASET

A.1 Dataset Process Flow Diagram

Below is the dataset building process flow that uses 10 table sources and 2 excel files.



A.2 Correlation Analysis

Input Variables	Driver	YearsDrivi	Capital	Vehicle	Vehicle	Vehicle	WeightPo	Vehicle	VehicleNum	VehicleNum	Customer	Policy	Payment	Claim	ClaimFreq	AvgClai
DriverAge	1.00	0.83	-0.05	0.16	0.01	-0.03	0.02	0.03	0.02	-0.02	0.06	0.06	-0.20	-0.03	-0.32	-0.01
YearsDrivingLicense	0.83	1.00	0.00	0.10	0.05	0.03	0.01	0.08	0.00	-0.02	0.09	0.09	-0.19	-0.02	-0.22	0.00
CapitalCar	-0.05	0.00	1.00	-0.47	0.22	0.53	-0.07	0.30	0.09	0.11	-0.03	-0.06	-0.03	0.06	0.41	0.02
VehicleAge	0.16	0.10	-0.47	1.00	-0.11	-0.29	0.10	0.00	-0.06	-0.23	0.04	0.06	0.00	-0.06	-0.54	-0.02
VehicleWeight	0.01	0.05	0.22	-0.11	1.00	0.42	0.07	0.67	0.10	0.13	-0.04	-0.05	-0.04	0.03	0.19	0.01
VehiclePower	-0.03	0.03	0.53	-0.29	0.42	1.00	-0.16	0.66	0.18	0.13	-0.02	-0.03	-0.04	0.04	0.37	0.02
WeightPowerRatio	0.02	0.01	-0.07	0.10	0.07	-0.16	1.00	-0.01	-0.07	-0.06	0.01	0.01	0.00	-0.01	-0.09	0.00
VehicleCapacity	0.03	0.08	0.30	0.00	0.67	0.66	-0.01	1.00	0.07	0.06	-0.02	-0.03	-0.05	0.03	0.21	0.01
VehicleNumOfSeats	0.02	0.00	0.09	-0.06	0.10	0.18	-0.07	0.07	1.00	0.52	0.00	0.01	0.01	0.02	0.18	0.01
VehicleNumOfDoors	-0.02	-0.02	0.11	-0.23	0.13	0.13	-0.06	0.06	0.52	1.00	-0.22	-0.22	0.01	0.03	0.24	0.00
CustomerTenure	0.06	0.09	-0.03	0.04	-0.04	-0.02	0.01	-0.02	0.00	-0.22	1.00	0.92	-0.01	-0.01	0.06	0.05
PolicyTenure	0.06	0.09	-0.06	0.06	-0.05	-0.03	0.01	-0.03	0.01	-0.22	0.92	1.00	0.00	-0.01	0.05	0.05
PaymentType	-0.20	-0.19	-0.03	0.00	-0.04	-0.04	0.00	-0.05	0.01	0.01	-0.01	0.00	1.00	0.03	0.06	0.01
ClaimFreq	-0.03	-0.02	0.06	-0.06	0.03	0.04	-0.01	0.03	0.02	0.03	-0.01	-0.01	0.03	1.00	0.11	0.11
ClaimFreqByProfile	-0.32	-0.22	0.41	-0.54	0.19	0.37	-0.09	0.21	0.18	0.24	0.06	0.05	0.06	0.11	1.00	0.05
AvgClaimCost	-0.01	0.00	0.02	-0.02	0.01	0.02	0.00	0.01	0.01	0.00	0.05	0.05	0.01	0.11	0.05	1.00
AvgPremiumChg	0.03	0.04	-0.03	0.02	-0.02	-0.02	0.01	-0.03	0.01	-0.01	0.06	0.06	-0.01	0.08	-0.03	0.04
AvgChgFreq	-0.04	-0.02	0.03	-0.05	0.01	0.02	0.00	0.00	-0.04	-0.02	-0.03	-0.03	-0.02	0.01	0.02	0.03
NumOfLOB	-0.02	0.03	0.09	-0.09	0.05	0.09	-0.01	0.07	-0.01	-0.01	0.11	0.09	0.00	0.02	0.07	0.01
NumOfNonLifeLOB	-0.07	-0.03	0.08	-0.06	0.04	0.07	-0.01	0.06	-0.01	-0.01	0.10	0.07	0.07	0.02	0.06	0.01
NumOfMotorPolicies	0.03	0.06	0.01	0.05	0.07	0.04	0.01	0.10	-0.13	-0.12	0.11	-0.01	-0.02	-0.01	-0.06	-0.01
NumOfLifePolicies	0.14	0.15	0.09	-0.04	0.05	0.08	0.00	0.07	-0.01	-0.02	0.04	0.02	-0.10	0.00	0.01	0.00
NumOfHealthPolicies	-0.13	-0.11	0.01	-0.02	0.01	0.03	0.00	0.02	0.00	0.01	-0.03	-0.03	0.15	0.03	0.04	0.00
NumOfAccidentPolicies	0.03	0.05	0.07	-0.05	0.02	0.06	-0.01	0.03	-0.01	-0.02	0.12	0.10	0.02	0.02	0.04	0.01
NumOfMultiRiskPolicies	0.03	0.06	0.06	-0.04	0.03	0.06	-0.01	0.04	-0.01	0.00	0.09	0.06	-0.01	0.01	0.03	0.00
NumOfcoverage	-0.08	-0.03	0.64	-0.54	0.17	0.37	-0.08	0.16	0.21	0.24	-0.02	-0.03	0.02	0.09	0.47	0.04
Cov_Collision	-0.06	-0.01	0.64	-0.50	0.15	0.32	-0.06	0.13	0.11	0.12	0.07	0.06	0.02	0.08	0.42	0.04
TotalDiscount	0.00	-0.01	-0.02	-0.01	0.02	-0.01	0.00	-0.02	0.02	0.14	-0.48	-0.51	0.06	-0.01	-0.06	-0.02
BankScore	-0.27	-0.23	-0.02	0.01	0.01	0.00	0.00	0.01	-0.01	-0.01	0.02	0.02	0.31	0.04	0.07	0.01
CameoEducationScore	-0.01	0.04	0.17	-0.18	0.00	0.13	-0.04	0.02	0.04	0.01	0.11	0.11	0.00	0.03	0.20	0.01
CameoFamilyScore	-0.12	-0.10	0.02	-0.07	-0.02	0.02	-0.02	-0.03	0.03	0.03	0.02	0.02	0.04	0.01	0.10	0.00
CameoIncomeFocus	0.02	0.04	0.15	-0.18	-0.01	0.10	-0.03	-0.01	0.06	0.03	0.08	0.08	0.04	0.02	0.25	0.02

Input Variables	AvgPr	AvgChgFr	NumOf	NumOf	NumOf	NumOf	NumOfHe	NumOfA	NumOfMulti	NumOfcove	Cov_Collis	Total	Bank	Educa	CameoFa	CameoI
DriverAge	0.03	-0.04	-0.02	-0.07	0.03	0.14	-0.13	0.03	0.03	-0.08	-0.06	0.00	-0.27	-0.01	-0.12	0.02
YearsDrivingLicense	0.04	-0.02	0.03	-0.03	0.06	0.15	-0.11	0.05	0.06	-0.03	-0.01	-0.01	-0.23	0.04	-0.10	0.04
CapitalCar	-0.03	0.03	0.09	0.08	0.01	0.09	0.01	0.07	0.06	0.64	0.64	-0.02	-0.02	0.17	0.02	0.15
VehicleAge	0.02	-0.05	-0.09	-0.06	0.05	-0.04	-0.02	-0.05	-0.04	-0.54	-0.50	-0.01	0.01	-0.18	-0.07	-0.18
VehicleWeight	-0.02	0.01	0.05	0.04	0.07	0.05	0.01	0.02	0.03	0.17	0.15	0.02	0.01	0.00	-0.02	-0.01
VehiclePower	-0.02	0.02	0.09	0.07	0.04	0.08	0.03	0.06	0.06	0.37	0.32	-0.01	0.00	0.13	0.02	0.10
WeightPowerRatio	0.01	0.00	-0.01	-0.01	0.01	0.00	0.00	-0.01	-0.01	-0.08	-0.06	0.00	0.00	-0.04	-0.02	-0.03
VehicleCapacity	-0.03	0.00	0.07	0.06	0.10	0.07	0.02	0.03	0.04	0.16	0.13	-0.02	0.01	0.02	-0.03	-0.01
VehicleNumOfSeats	0.01	-0.04	-0.01	-0.01	-0.13	-0.01	0.00	-0.01	-0.01	0.21	0.11	0.02	-0.01	0.04	0.03	0.06
VehicleNumOfDoors	-0.01	-0.02	-0.01	-0.01	-0.12	-0.02	0.01	-0.02	0.00	0.24	0.12	0.14	-0.01	0.01	0.03	0.03
CustomerTenure	0.06	-0.03	0.11	0.10	0.11	0.04	-0.03	0.12	0.09	-0.02	0.07	-0.48	0.02	0.11	0.02	0.08
PolicyTenure	0.06	-0.03	0.09	0.07	-0.01	0.02	-0.03	0.10	0.06	-0.03	0.06	-0.51	0.02	0.11	0.02	0.08
PaymentType	-0.01	-0.02	0.00	0.07	-0.02	-0.10	0.15	0.02	-0.01	0.02	0.02	0.06	0.31	0.00	0.04	0.04
ClaimFreq	0.08	0.01	0.02	0.02	-0.01	0.00	0.03	0.02	0.01	0.09	0.08	-0.01	0.04	0.03	0.01	0.02
ClaimFreqByProfile	-0.03	0.02	0.07	0.06	-0.06	0.01	0.04	0.04	0.03	0.47	0.42	-0.06	0.07	0.20	0.10	0.25
AvgClaimCost	0.04	0.03	0.01	0.01	-0.01	0.00	0.00	0.01	0.00	0.04	0.04	-0.02	0.01	0.01	0.00	0.02
AvgPremiumChg	1.00	-0.06	0.02	0.03	0.00	0.01	0.01	0.00	0.02	-0.02	-0.02	0.03	0.01	0.00	0.01	0.00
AvgChgFreq	-0.06	1.00	0.02	0.02	0.01	0.01	0.01	0.00	0.02	0.02	0.03	0.01	0.01	0.01	0.00	0.02
NumOfLOB	0.02	0.02	1.00	0.71	0.18	0.43	0.39	0.35	0.51	0.10	0.10	-0.04	0.07	0.09	0.02	0.04
NumOfNonLifeLOB	0.03	0.02	0.71	1.00	0.18	0.16	0.41	0.31	0.59	0.08	0.08	-0.04	0.15	0.08	0.02	0.04
NumOfMotorPolicies	0.00	0.01	0.18	0.18	1.00	0.10	0.04	0.09	0.18	-0.06	-0.02	-0.04	-0.03	0.00	-0.03	-0.03
NumOfLifePolicies	0.01	0.01	0.43	0.16	0.10	1.00	0.02	0.10	0.18	0.05	0.06	0.00	-0.09	0.05	-0.02	0.02
NumOfHealthPolicies	0.01	0.01	0.39	0.41	0.04	0.02	1.00	0.13	0.09	0.01	0.00	0.05	0.23	0.00	0.02	0.02
NumOfAccidentPolicies	0.00	0.00	0.35	0.31	0.09	0.10	0.13	1.00	0.15	0.08	0.09	-0.08	0.04	0.09	0.00	0.05
NumOfMultiRiskPolicies	0.02	0.02	0.51	0.59	0.18	0.18	0.09	0.15	1.00	0.06	0.06	-0.03	0.03	0.07	0.03	0.04
NumOfcoverage	-0.02	0.02	0.10	0.08	-0.06	0.05	0.01	0.08	0.06	1.00	0.91	-0.04	-0.02	0.18	0.03	0.18
Cov_Collision	-0.02	0.03	0.10	0.08	-0.02	0.06	0.00	0.09	0.06	0.91	1.00	-0.10	-0.01	0.20	0.03	0.21
TotalDiscount	0.03	0.01	-0.04	-0.04	-0.04	0.00	0.05	-0.08	-0.03	-0.04	-0.10	1.00	0.01	-0.07	-0.02	-0.03
BankScore	0.01	0.01	0.07	0.15	-0.03	-0.09	0.23	0.04	0.03	-0.02	-0.01	0.01	1.00	0.01	0.04	0.02
CameoEducationScore	0.00	0.01	0.09	0.08	0.00	0.05	0.00	0.09	0.07	0.18	0.20	-0.07	0.01	1.00	0.31	0.39
CameoFamilyScore	0.01	0.00	0.02	0.02	-0.03	-0.02	0.02	0.00	0.03	0.03	0.03	-0.02	0.04	0.31	1.00	0.04
CameoIncomeFocus	0.00	0.02	0.04	0.04	-0.03	0.02	0.02	0.05	0.04	0.18	0.21	-0.03	0.02	0.39	0.04	1.00

APPENDIX B. CLUSTER MODEL

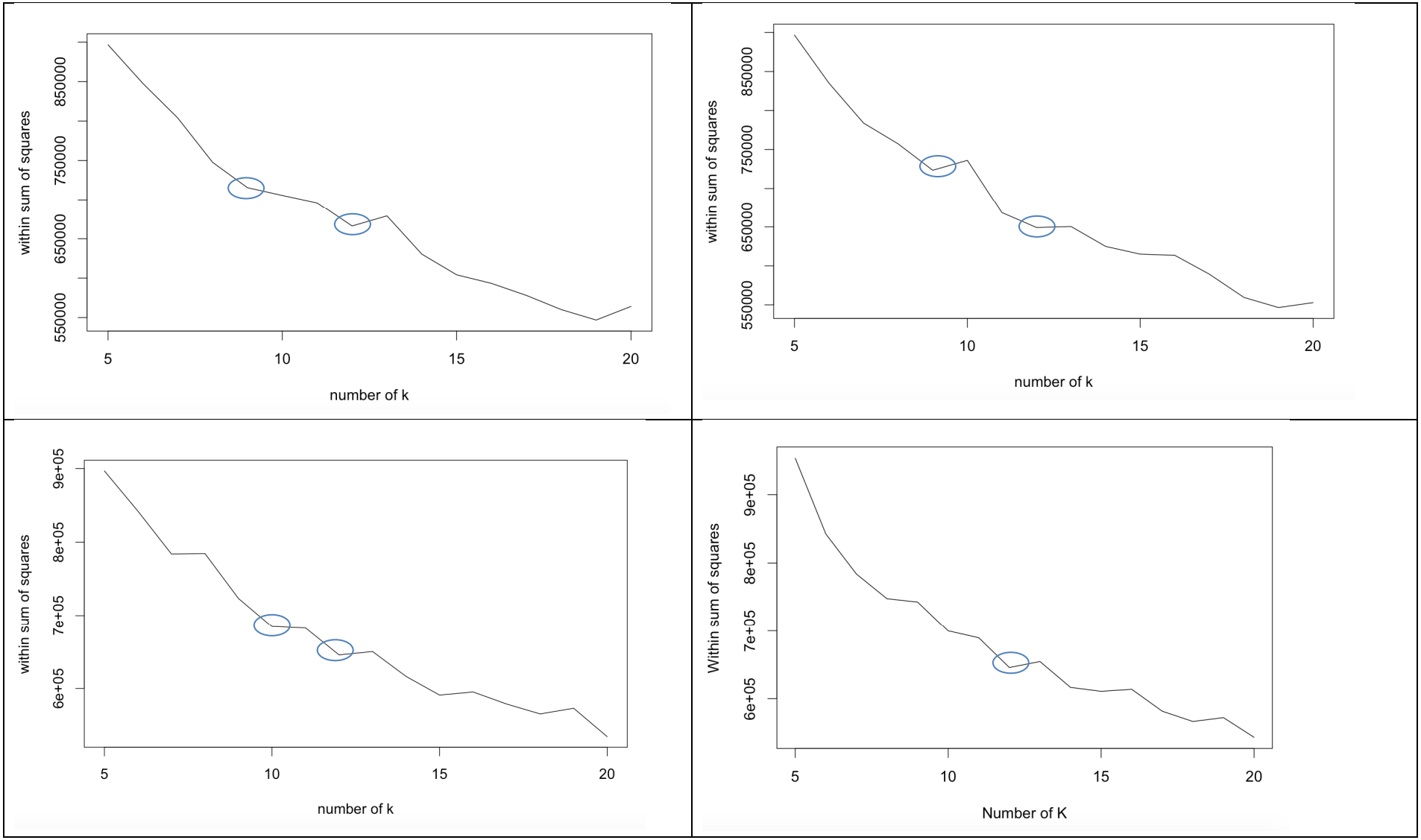
B.1 Variable Combinations

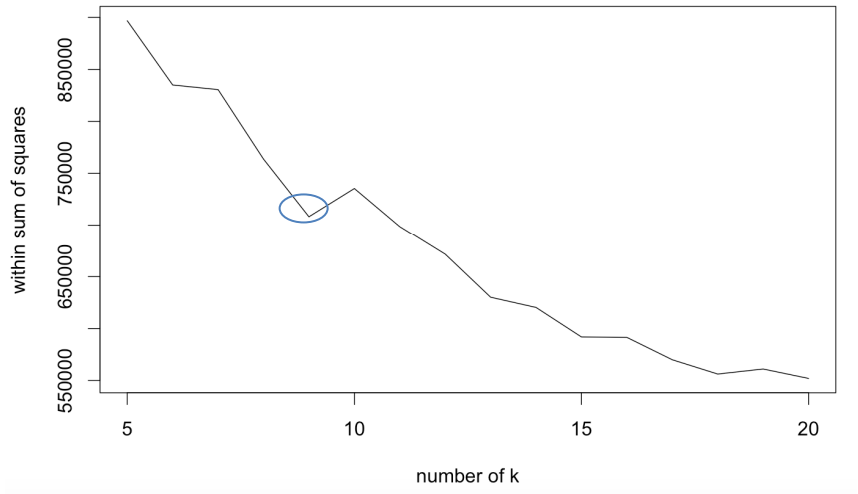
Below is the list of variable combination that have been tried.

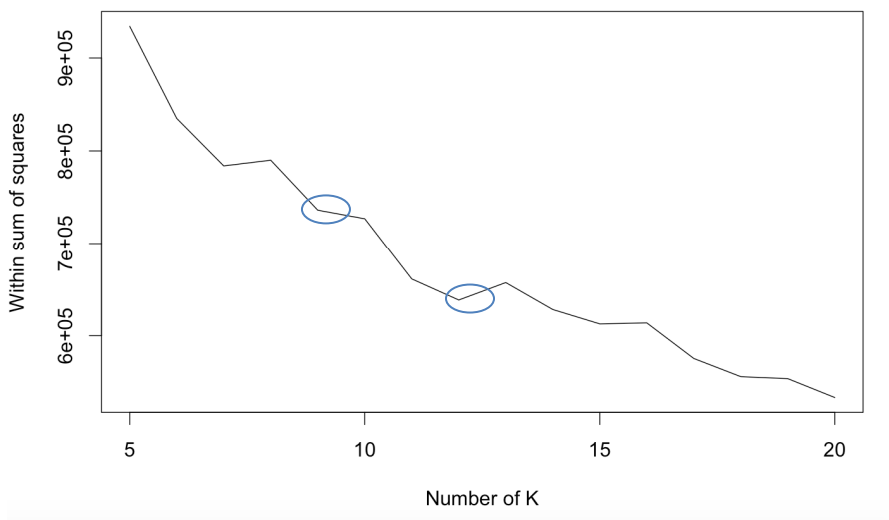
Variable Name	VC1	VC2	VC3	VC4	VC5	VC6	VC7	VC8	VC9	VC10	VC11	VC12	VC13	VC14	VC15	VC16
Vehicle Age	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Vehicle Weight	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Vehicle Power	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Weight Power Ratio	V	V	V			V	V	V	V	V	V	V	V	V	V	V
Vehicle Capacity	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Vehicle Num. of Doors						V	V	V	V							
Vehicle Num.of Seats						V	V	V	V	V						
Private Vehicle Ind					V	V	V			V	V	V	V	V	V	V
Vehicle Gasoline Ind						V	V			V	V	V	V	V	V	V
Driver Age	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Driver Married Ind						V	V			V	V	V	V	V	V	V
Driver Gender Ind						V	V			V	V	V	V	V	V	V
Capital Car	V	V	V			V	V	V	V	V	V	V	V	V	V	V
Avg Claim Cost	V	V	V	V	V	V		V		V	V	V	V	V		
Claim Freq	V	V	V	V	V	V		V		V	V	V	V	V		
Claim Freq By Profile	V	V	V	V	V	V		V		V	V	V	V	V		
Risk Score	V		V	V	V	V	V	V	V	V	V	V	V	V	V	V
Policy Tenure	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Number of Coverage	V	V	V	V	V					V	V		V	V		V
CovCollisionInd												V			V	
AvgPremiumChg	V	V														
NumOfLOB	V	V	V	V	V											
NumOfMotorPolicies	V	V	V	V	V											
Total Discount	V	V														
EducationScore													V	V		
FamilyScore													V			
IncomeFocus													V			
UrbanRural													V	V		

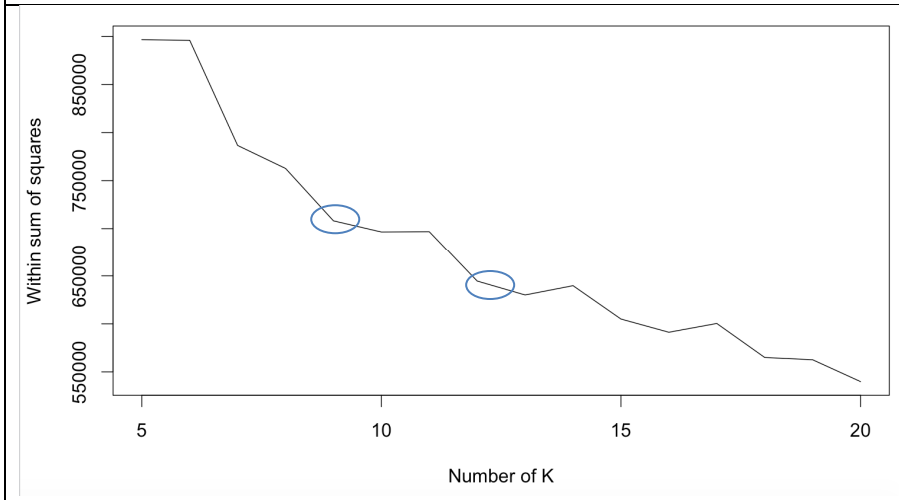
B.2 Elbow Chart

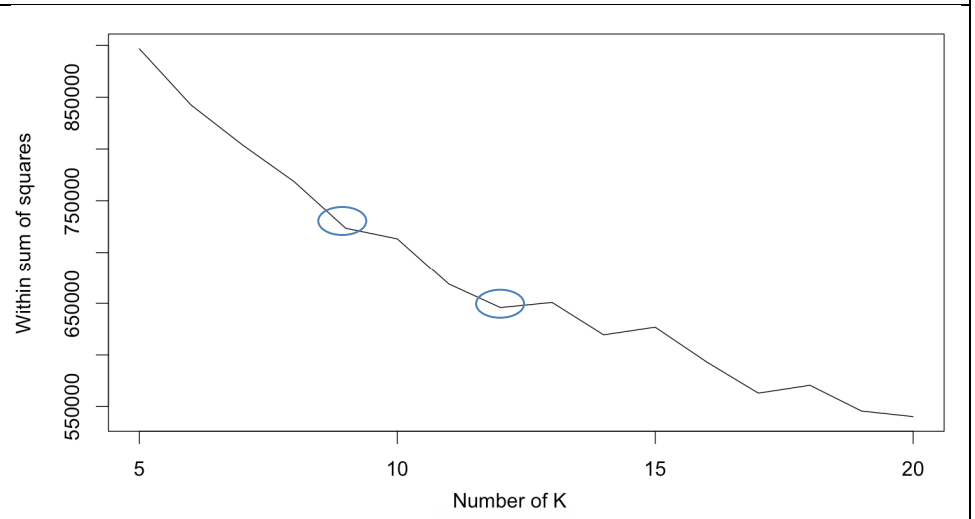
Below are the graphs of elbow method. Several k-means iterations have been run to get more elbow charts.




within sum of squares
550000
650000
750000
850000
5
10
15
20
number of k


Within sum of squares
6e+05
7e+05
8e+05
9e+05
5
10
15
20
Number of K


Within sum of squares
550000
650000
750000
850000
5
10
15
20
Number of K


Within sum of squares
550000
650000
750000
850000
5
10
15
20
Number of K

B.3 Rule Extraction

Below is the final Rule-based Cluster Model.

Cluster Label	Private Vehicles (Y/N)	Policy Tenure	Collision (TPL / Own Damage)	Married (Y/N)	Risk Score	Vehicle Sum Insured	Vehicle Capacity	Vehicle Weight	Vehicle Power	Vehicle Weight Power Ratio	Gasoline (Y/N)
Non Private Vehicles, Low Weight and Capacity (trailers, motorbikes, small tractors)	Non Private Vehicles						< 1150	< 2345			
High weight or high capacity vehicles. Example vehicles are vans car, Pick-ups, Truck, Heavy Machinery, powerful Motorbikes	Non Private Vehicles							>=2345			
	Non Private Vehicles						>=1150 or Missing	< 2345			
	Private Vehicles	<= 7	TPL		<=7 or Missing					>=28	N
Private Vehicles, Single/Divorced Drivers, Gasoline	Private Vehicles	<= 7	TPL	N	<=7 or Missing						Y or Missing
Private Vehicles, Married Drivers, Gasoline, Low Vehicle Power	Private Vehicles	<= 7	TPL	Y or Missing	<=7 or Missing				< 71		Y or Missing
Private Vehicles, Married Drivers, Gasoline, High Vehicle Power	Private Vehicles	<= 7	TPL	Y or Missing	<=7 or Missing				>=71 or Missing		Y or Missing
Private Vehicles, Single/Divorced Drivers, Diesel, Low/Medium Capacity	Private Vehicles	<= 7	TPL	N	<=7 or Missing		< 1750			<28 or Missing	N
Private Vehicles, Single/Divorced Drivers, Diesel, High Capacity,	Private Vehicles	<= 7	TPL	N	<=7 or Missing		>=1750 or Missing			<28 or Missing	N
Private Vehicles, Married Drivers, Diesel	Private Vehicles	<= 7	TPL	Y or Missing	<=7 or Missing					<28 or Missing	N
Private Vehicles, The worst risk score	Private Vehicles	<= 7	TPL		> 7						Y or Missing
Private Vehicles, Own Damage, Low Power and Low Capital Car	Private Vehicles	<= 7	Own Damage			<20570			< 126 or Missing		
Private Cars, Own Damage, High Power or High Capital Car with Low power (new vehicles)	Private Vehicles	<= 7	Own Damage			>=20570			< 126 or Missing		
	Private Vehicles	<= 7	Own Damage						>= 126		
Private Vehicles, Old policies	Private Vehicles	> 7									