USING DATA MINING IN CRM TO UNDERSTAND THE INSURANCE MARKET

دراسة سوق قطاع التأمين عن طريق تحليل بيانات نظام إدارة علاقات العملاء

By Wafa Qadadeh

Student ID number 2013128026

Dissertation submitted in partial fulfillment of the degree M.Sc. Informatics (Knowledge and Data Management)

Faculty of Engineering & IT

Dissertation Supervisor

Dr. Sherief Abdallah

November-2015

Abstract:

Understanding customers' interests is an important concept for designing marketing campaigns to improve businesses and increase revenue. The rapid growth of high dimensional databases and data warehouses, such as Customer Relationship Management (CRM), stressed the need for advanced data mining techniques. In this paper we investigate different data mining algorithms, specifically K-Means, SOM, and CHAID using the TIC CRM dataset. While K-Means has shown promising clustering results, SOM has outperformed in the sense of: speed, quality of clustering, and good visualization. Also we discuss how both techniques segmentation analysis can be useful in studying customer's interest. CHAID helps us to predict new target for customers' interest based on their demographic data.

الخلاصة:

تعتبر دراسة اهتمامات العملاء واحتياجاتهم من أهم الطرق المتبعة من قبل خبراء التسويق و التي تساعد في إيجاد سوق جديدة مما يؤدي إلى تطوير قطاع الأعمال وزيادة الربح.

وقد أدى النمو المتسارع للبيانات وتوفر مستودعات للبيانات مثل نظام إدارة علاقات المتعاملين، إلى ضرورة ملحة لاستخدام التكنولوجيا المتقدمة في تحليل هذه البيانات الضخمة.

من خلال هذا البحث سوف نقوم بدراسة تقنيات وطرق مختلفة لتحليل البيانات باستخدام بيانات نظام إدارة علاقات المتعاملين في قطاع التأمين. بالنسبة للتقنية SOM فقد تبين من خلال هذا البحث أنها من أكثر التقنيات كفاءة وسرعة في اكتشاف مجموعات العملاء المتشابهة بالإضافة لكافئتها العالية في تصوير أو عرض البيانات ذات الحجم والأبعاد الكبيرة مقارنة بتقنية K-means.

كما يتضمن البحث تحليل لخصائص مجموعات العملاء الناتجة عن استخدام التقنيتين السابقتين. هذا التحليل للخصائص أو الاهتمامات للعملاء قد يفيد في فهم السوق وبناء حملات التسويق للتأمين.

وأخيراً تستطيع تقنية CHAID أن تتنبأ بمستهدف جديد ومعرفة العملاء المقبلين على الاشتراك في وثيقة التأمين على الصحة التابعة للقطاع الخاص بناءاً على بياناتهم الديموغرافية.

Table of Contents

List of Figures

List Of Tables

Chapter 1:

1.1 Introduction.

The insurance sector has turned out to be a major element in certain economies. As per Cristea, M., 2014 research, there is an international correlation between insurance and the economic growth. Cristea, M argued that the mass of insurance to the Gross Domestic Product (GDP) of countries being over 10% in economically mature countries, and it is even greater as the economic growth is emerging (Cristea, M., 2014).

Insurance has been considered as Risk management tool for both society and businesses. For the society, insurance provides stability, and financial security against risks. It has also reduced both accident and loss from accident extraordinarily. It could be considered as a source of investment for society members' future. For example in social insurance people obtain benefits based on their contributions to an insurance program. This program could be survivor aids, unemployment insurance, or pensions (Keele, S, 2015).

For Businesses especially small ones, insurance can reduce the fears around any business. It also keeps managers attention on running their business while others are taking care of theft or natural disasters (Keele, S, 2015).

Insurance Companies nowadays are continuously working to increase their competitiveness. Companies should study the market before designing their marketing plans. Many marketing strategies could be implemented in insurance sector, such as: customers segmentation, direct marketing, and cross marketing (Thearling, K, 2015).

As per Thearling, K. the first task is to identity your customers segments using their buying behaviors. This buying behaviors could be found in Customer Relationship Management (CRM) systems. After identifying the segments marketing experts can use the resulted knowledge to feed their marketing software. Each segment would follow different marketing plan with different products and services instead of having one arbitrary plan.

The availability of big data in CRM has increased the need to use data mining technologies significantly. The usage of data mining algorithms can help businesses to find interesting knowledge in its customer's data both demographic and behavioral. Datamining is used in insurance in many tasks:

- Detect medical procedures that claimed together.
- Distinguish uncertain customers' behavior patterns and attrition.
- Know customers who will potentially purchase.
- Detect fraudulent behavior. (Padhy, N., 2012).

By mining this CRM dataset, marketing experts can use the extracted knowledge to develop marketing plans.

In our research we are interested in The Insurance Company (TIC) dataset. The high dimensions and the demographic data it includes could be considered as advantages. We had been looking for a real dataset with the same characteristics in the UAE insurance companies. After four months of research, unfortunately we could not convince any company to give us their data. As mentioned in Chapter 3 this dataset had been used in many papers participated in the COIL 2000 Challenge for the purpose of classification. Here we are trying to apply different techniques for different tasks. The main target of our study is to find the following:

- Can we discover meaningful clusters using different cluster analysis techniques applied on the TIC CRM data?
- Predict if a customer would buy a private health insurance policy based on demographic data from CRM.

We attempt to answer the first question by exploring cluster analysis. We identify some interesting patterns that can be used by marketing experts in insurance companies. In particular, we investigate two different data mining techniques. First, the famous K-means clustering algorithm. Then we combine K-means with the SOM technique in order to enhance the clustering, taking into consideration the previous work. Using this solution has been showing promising results in clustering CRM dataset and visualization. While to answer the second question the CHAID decision tree algorithm is used.

Clustering and visualization would be used to recognize the characteristics of a customer in CRM to design customer centric marketing plans. Also predicting new target considering understanding the customers demographic can create new marketing dimensions.

A few number of researches have been investigating this area. Starting by Namvar, M., 2010, the analysis had been done using K-means for profitability segmentation in a bank CRM. Namvar, M., 2010 had also used SOM as second level of clustering. Another segmentation was done by Khajvand, M., 2011, using K-means. Each segment lifetime value was calculated, its future value was estimated using ARIMA regression. Zadeh, R. B. K, 2011, had integrated three datasets to cluster customers using K-means. The target of the segmentation is to avoid attrition. More than one clustering algorithm, in addition to K-means, had been used by Dhandayudam, P., 2012, on a fertilizer company CRM. ANN was used by Emtiyaz, S., 2012 paper to visualize the data. While in Goonetilleke, T. O., 2013 the target was avoiding attrition by using classification algorithms. Mbarki, J., 2014 in his paper had used SOM to find interesting results in Automobile Insurance CRM data.

More details about the reviewed papers are covered in the literature review section.

1.2 Objectives, Key Questions and Scope.

The main target of this study is to answer the following research questions:

- Can we discover meaningful clusters using different clustering analysis applied on the high dimensional TIC data?
- Can machine learning be used to predict whether a customer would buy a private health insurance policy based on its demographic from CRM?

1.3 Contributions.

This paper proposes three contributions: two solutions for the first research question and one solution for the second question.

For the first problem, we apply K-means clustering algorithm to understand the characteristics of the customers who are interested in owning a caravan policy or not.

The second contribution is using the SOM technique (based on ANN) along with the K-means to enhance the clustering in TIC. As in Namvar, M., 2010, Back, A. H. H. T. E. B., 2011, and Branch, N., 2014 work, and in this paper, SOM shows promising results in clustering, and visualization. Clustering and visualization would be used to recognize the characteristics of a segment of customers. And therefore to design customer centric or cross marketing plans.

Our third contribution (which is a solution to the second problem) is predicting new target class using interpretable decision tree. The CHAID algorithm is used to decide who are interested in owning a private health insurance policy based on their demographics.

1.4 Dissertation Outline.

The remaining parts of this paper are organized as follows. In Chapter 2 we briefly define the science of data mining, CRM, and the technical background. The same chapter reviews other researches related to using data mining tasks and techniques in CRM data in different domains. It also shows points of interests in other high quality papers. Chapter 3 is about the dataset, tools, methodology, and the techniques used. Chapter 4 proposes our three solutions as an experimental evaluation of the used techniques showing the results of applying them on the TIC dataset. Finally Chapter 5 concludes this study with a general discussion about the proposed solutions results and the future work that could be done.

Chapter 2: Background.

2.1 Data Mining.

Data Mining or what is also known as knowledge discovery in data (KDD). Data mining is: the exercise of automatically studying large sets of data to discover previously unknown, non-trivial, interesting patterns to derive actionable information or make decisions. This automatic exploration is accomplished through building a model using a mathematical algorithm on a training dataset, then executing this model on a testing data to generalize it (Oracle.com, 2015). Data Mining is also “the process of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses” (Farooqi, M., 2012).

As per Kurt Thearling Data Mining is “the automated extraction of hidden predictive information from databases”. Using advanced technologies and tools of Data Mining, business owners can discover important information from their historical huge data warehouses. By this information they can build their proactive future decisions based on patterns and trends. As shown in Figure 1, data mining is a part of a large complicated process, where more than one algorithm or technique could be used. People thinking is not less important that machine learning algorithms, so experts should be involved to understand, analyze, and evaluate the results.

The major of machine learning algorithms is a blend of trial and error, and statistical approach. The main differences between traditional statistical and data mining approaches are: the first needs human interference to evaluate the accuracy of a model, so it is difficult to be fully automated, and hard to be used with big data sets. There are two scopes of capabilities data mining can provide to businesses: the automation of predicting trends and behaviors, and the automation of exploring previously unknown patterns (Thearling, K., 2015).

The most commonly used techniques in data mining are:

- ANN: Artificial Neural Networks are Non-linear predictive models that look like biological neural networks in structure and work.
- Decision trees: look like trees and structure sets of decisions and rules.
- Genetic algorithms: use processes such as genetic combination, and mutation for optimization.
- Nearest neighbor method: classifies objects in a dataset based on a grouping of the classes of the k record(s) most similar to it in a dataset.
- Rules induction: the extraction of useful if-then rules from data based on statistical significance. (Thearling, K., 2015)

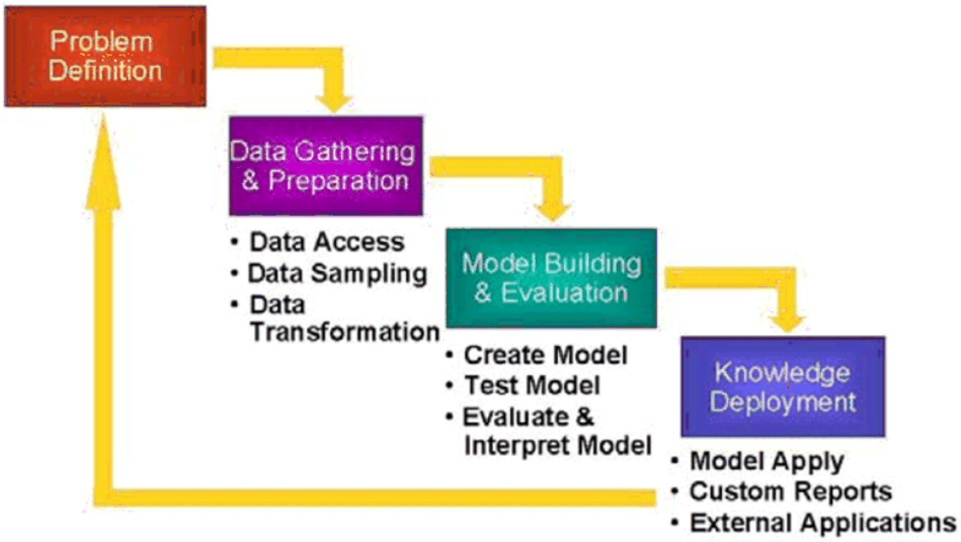


Figure 1: Data Mining Process. Adapted from www.docs.oracle.com

2.2 CRM as Big Data.

Customers Relation Management (CRM) is a knowledge management system used by companies to store and manage customers. CRM includes customers' contacts and data in order to sustain future relationship with them. As Farooqi, M., 2012 definition CRM is "the process of acquiring, retaining and growing profitable customer which requires a clear focus on service attributes that represent value to the customer and creates loyalty". Regardless of the tool or technology used, CRM is a business strategy that aims to study customers' needs and maintain relationships with potential ones. The data available in these systems could be collected from different channels. CRM channels could be: phone calls, direct meeting with customers, online transactions through e-commerce web sites, billing, online support, Surveys, or even social networks (Farooqi, M., 2012).

The large-volume, heterogeneous, independent sources with decentralized control of the available data, had made Big Data an emerging concept in the world of data mining. The different resources of gathering customers' data as well as the dimensions. In addition to the fast growth and to which depth this data is representing a customer taste or behavior, would categorize CRM as a Big Data approach (Wu, X., 2014).

Mining CRM to identify patterns, detect trends, study associations, segment market can help a company to understand their customer and build customer centric business. Marketing experts can use CRM demographic or behavioral to predict customers' needs in order to use it in direct promotions or cross-marketing. By analyzing customer retention and cost vs. revenue per service call, CRM mining would be a mean for continues benchmarking too. This drilling in big data won't be accomplished far away from data mining techniques.

In the remaining parts of this dissertation we will discuss how data mining techniques can be efficient in CRM analysis.

2.3 Literature Review.

2.3.1 Insurance CRM

In Goonetilleke, T. O., 2013 paper a study on life insurance company CRM data was conducted to analyze the customer's data and to avoid customer churn\attrition. The authors had argued that a big data with a multi-class problem was tackled to classify the customer's willingness to continue or not. The dataset used in this study was extracted from operational database. Unlike other insurance policies types, a life insurance policy is for an average of 18-20 years, so to build an efficient model the authors needed to mine data for a significant (big) period of time. The authors also are interested in using demographic data like gender, age, profession, etc. Term of policy, sum assured, premium, Agent, etc as the policy details was also used. The authors had started by visualizing the attributes to study dependency and correlation to select the relative attributes or the attributes to be combined. In addition to visualization Correlation-based Feature Selection (CFS), and Information Gain techniques were used. The ROC graph technique was used to evaluate the different classifiers accuracy. ROC indifferent to change in class distribution which is common in domains such as churn while the distribution of data is went towards one class label (most frequent).

The classifiers were used in Goonetilleke, T. O., 2013 paper to predict the classes are: J48 decision tree and ANN with a standard Multilayer Perceptron using BP. Many issues apart from the evaluation of the classifiers were addressed, such as the huge number of attributes, an efficient feature selection technique, maybe probabilistic one, could be used to solve this issue. Also human interference still required in different phases of the study.

Within Branch, N., 2014 paper a case study of applying SOM on Automobile Insurance CRM data, had been conducted. The target of this study is to analyze customers' clusters in order to improve the relationship with customers. The demographic data of customers had been used in order to develop: one-to-one marketing, personalized and satisfactory services. The proposed solutions could avoid ineffective strategies, increase revenue, and save money. The data used had been divided into four categories:

1. Annual salary, marital status, age, education, dwelling, gender, and profession as basic data.
2. Value of automobile, year, country of origin, type, and application as car characteristic.
3. Previous insurance company, no of years without accidents, as insurance information.
4. Insurance cost, discount, as insurance company selection factors.

The analysis of the above categories of data had been resulting into four customers segments. The clusters are named as service quality cluster, cost of service cluster, means of payment cluster, and acquaintances influence in choosing company cluster. These segments would help the organization in developing marketing strategies. The metrics, mean absolute error (MAE) and root mean square error (RMSE), had been used to evaluate the SOM, in addition to market experts' evaluation.

2.3.2 Other domains CRM.

Recently, a structured framework was developed to apply Recency, Frequency, and Monetary (RFM), customer's lifetime value (LTV) models (Namvar, M., 2010). The framework had used customers' demographic data to segment banking customers and design marketing strategies.

Namvar, M., 2010 analysis study was done using K-Means clustering technique and consisted of two main phases. First the CRM data were used to cluster the customers. Second the demographics variables (age, education, and occupation), which were chosen by SOM technique, was used to re-cluster the resulted segments from first phase. The customer value comparison used LTV instead of inter/intra cluster distances, in order to maximize the value of the customer.

More recent work proposed a Bank Customer segmentation framework, based on customer's LTV (Khajvand, M., 2011). The study had been handling different approach using the customer's lifetime value which could be more efficient and practical. The researcher had prepared a framework to segment the customers, calculate each segment lifetime value, and estimate the future value of each segment. Two levels of clustering had been implemented on a big dataset of customer's transactions. The transaction record includes deposit type, transaction date, balance before transaction, amount of transaction, etc. K-means and two step clustering algorithms were implemented. The customer's lifetime value had been calculated using a RFM model, the simplest and most powerful customer's LTV approximation model. Finally the study had been using a time series method (multiplicative seasonal ARIMA regression) to predict future value for each segment (Khajvand, M., 2011).

In Zadeh, R. B. K, 2011 another analysis was conducted using K-means to segment bank customers based on their behaviors. This study aims to help the bank to prepare retention strategies and gaining new customers. The dataset in Zadeh, R. B. K, 2011 work is an integration of three tables. First, the customers' demographic data table including age, gender, marital status, and etc. Second, the transaction table containing the customers' transactions. Third, the cards table includes the data for bank cards. Many important information was taken into consideration during this study and their attributes were combined with other customer's attributes such as: transaction type, frequency of a transaction, service type, bank type, and channel type (ATM, Web, and Terminal).

In Emtiyaz, S., 2012 investigated the problem of identifying potential customers in big datasets using the following methodology. First for accuracy, a semi-supervised techniques were used to build customer behavior modeling automatically. Second, the authors had used a neural network technique to visualize data. For the semi-supervised proposed technique, a multi-layered perceptron neural network with back propagation was used. The classifier is re-trained using subset of labeled data each time, then it is used to classify the testing data, the most confident unlabeled records along with their predicted classes are added to the training set to re-train the classifier (bootstrapping). The proposed technique had outperformed many other traditional techniques such as Neural Net, SVM, and Naïve Bayes in classifying customers in CRM. The proposed solution could be used to enhance CRM processes such as identifying valuable customers to retain or attract them.

In the research of Mbarki, J., 2014 the partitioning around Medoids clustering algorithm K- Medoids, was implemented on Telecom CRM post sales dataset which is stored in Teradata environment, to the purpose of segmenting customers behavior while selling new products. The K- Medoids uses the most centric object in a cluster to represent the cluster instead of the mean (K-Means) that may not belong to the cluster. This makes K-Medoids more robust than K-means which outperform with high number of Ks. Customer's preferences such as: age, contract type, quantity sold, used media, customer area importance, department, and billing history were used to define the segment. The results of the study had shown that K-Medoids clustering algorithm is very efficient in large datasets such as CRM.

In Back, A. H. H. T. E. B., 2011 research, the Self Organized Map (SOM) clustering technique was used to develop organization market strategy by analyzing profitable customers' data. The business of the organization used in Back, A. H. H. T. E. B., 2011 research is a consulting services to other businesses. The purchasing behavior data of their clients are stored in the company warehouse. The clients' demographic data such as Return on equity (REQ), turnover, change in turnover, Company age, balance sheet total, etc. were used in the analysis as well as the sales information. The results of the SOM had led to ten clusters, and after combining the resulted clusters with the products sales data, the organization could define three different types of clients: profitable, average, and non-profitable. The resulted clustering schema had helped the case organization to build its future marketing strategies and was evaluated by experts from the sales department.

In Dhandayudam, P., 2012 research a comparison between clustering techniques: K-Means, single link, complete link clustering, and an enhanced proposed clustering technique. The authors used customers RFM attributes from CRM data in a fertilizer manufacturing company. The proposed enhanced clustering technique had shown a significant improvement of the segmentation results, and it overcame the limitations of the other techniques. For evaluating the clustering results the authors had been using Mean Square Error (MSE), Intra cluster distance, Inter cluster distance, and the ratio of both distances. The proposed solution had shown the minimum MSE and intra distance, and increase the inter distance between clusters. The improved technique also had shown the minimum ratio of intra to inter distance.

In Rygielski, C., Wang, J. C., & Yen, D. C., 2002 paper, the author had been investigating more than one data mining technique to improve the business of two companies using their CRM data.

The first project was applied to Neo Vista Solutions Inc., by combining clustering and NN to improve the business. These two data mining techniques had helped the management to explore the minimum details that could improve its products' demand forecasting.

The second solution was applied for Applied Matrix Ltd. CHAID was used for market segmentation to initially identify significant interaction within predictors toward a certain promotion. For instance customer age could be used to create clusters of customers while used to predict the customers' response to certain promotions along with other variables like income. In this project CHAID had succeeded to reveal 16 segments, each segment can be used to predict response-to-promotion. After that CHAID segmentation was used by a regression model to measure the influence of each predictor on a certain condition. The solution had been showing interesting results. It was expected that implementing this techniques will increase the response rate by 10 percent, but surprisingly the response rate was increased by 30 percent.

Chapter 3: Methodology

3.1 The Dataset.

In this research we have been using THE INSURANCE COMPANY (TIC) 2000 dataset. This dataset was published by Peter van der Putten, and owned by a Dutch data mining company Sentient Machine Research, Amsterdam.

This data constituted the CoIL Challenge 2000 data mining competition. TIC dataset was collected from real world Customer Relationship Management (CRM) data. TIC consists of 9,822 customer records, 5,822 record for training and the remaining for testing. Each record has 86 attributes. The first 43 attributes are representing customers' demographics, and the remaining 43 are representing customers' behavior or products ownership. All of the features have nominal values, with the last one (the target attribute for COIL 2000) being binominal (P. van der Putten and M. van Someren, 2000). TIC 2000 dataset is available on: TIC 2000 homepage: http://www.wi.leidenuniv.nl/~putten/library/cc2000/ , and Edinburgh University http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html.

3.2 Tools.

To run the experiments for this research we had been using RapidMiner 5.3 as data mining tool. RapidMiner 5.3 is an open source free tool with an integrated environment for mining data, business analytics, and machine learning (Guido Deutsch, 2010).

3.3 Methodology.

The research that we have reported in this paper had as objective to explore the insurance company dataset using several techniques of data mining. In particular we want to investigate the clustering of the data, which was not covered by previous research. Also we aim to predict new insurance target.

We will start first by studying the characteristic of the dataset and its tendency toward clustering, we will implement two different clustering experiments using different conditions. At last we will use one of the classification methods with customers' demographics and behavior to interpret a decision tree.

3.3.1 Determining the clustering tendency of data.

Before starting the experiment and applying any clustering technique on this dataset, we have to study the tendency of the data to have clusters or similarity between objects. In big datasets, such as the one we have with high dimensions, the curse of dimensionality could have a critical effect on the similarity measure. To reveal the tendency of the data to be clustered we had plotted the histogram of the pairwise distances of all objects in our dataset. If the resulted graph contains two peaks, this means that the dataset contains clusters. One of the peak to represent the distance between objects in clusters, and the second peak to represent the average distance between objects as shown in Figure 2 (Steinbach, M., Ertöz, L., & Kumar, V., 2004). The histogram of our data set as shown in Figure 3, is showing that the TIC CRM has the tendency to be clustered.

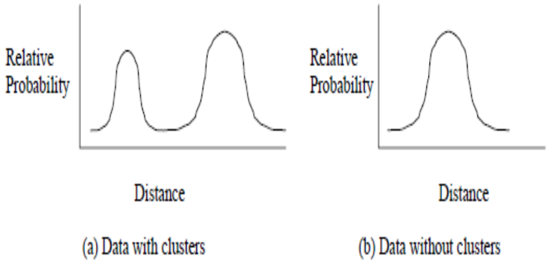


Figure 2: Plot of interpoint distances for data with and without clusters. Adapted from (Steinbach, M., Ertöz, L., & Kumar, V., 2004).

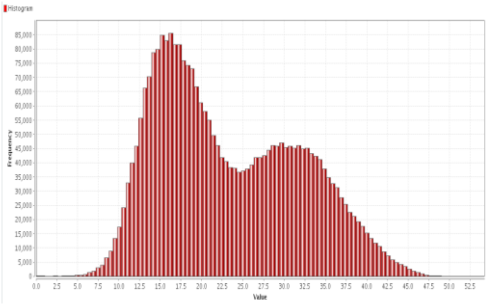
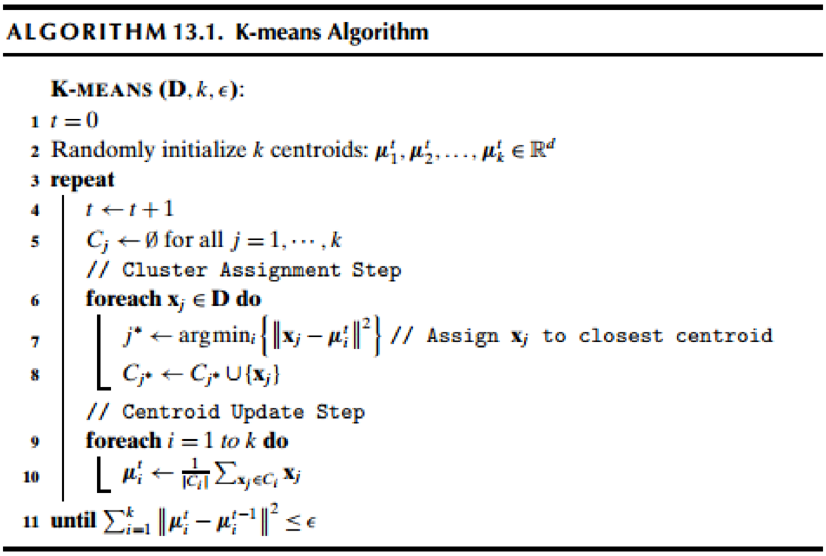


Figure 3: Analyzing Clustering Tendency-Distances Histogram for the TIC dataset.

3.3.2 K-means Algorithm.

K-means is a partitioning clustering algorithm, where each cluster is connected with a centroid or central point (mean of points). During training each object is assigned to a cluster with the closest centroid, usually Euclidean distance is used. Number of clusters K should be defined at the beginning and initial centroids are defined randomly. K-means Algorithm is shown in Figure 4 (Zaki, Mohammed J., 2014).



ALGORITHM 13.1. K-means Algorithm

```
K-MEANS ($\mathbf{D}, k, \epsilon$):
$t = 0$
Randomly initialize $k$ centroids: $\mu_1^t, \mu_2^t, \ldots, \mu_k^t \in \mathbb{R}^d$
repeat
    $t \leftarrow t + 1$
    $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$
    // Cluster Assignment Step
    foreach $\mathbf{x}_j \in \mathbf{D}$ do
        $j^* \leftarrow \text{argmin}_i \left\{ \left\| \mathbf{x}_j - \mu_i^t \right\|^2 \right\}$ // Assign $\mathbf{x}_j$ to closest centroid
        $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$
    // Centroid Update Step
    foreach $i = 1$ to $k$ do
        $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
until $\sum_{i=1}^k \left\| \mu_i^t - \mu_i^{t-1} \right\|^2 \le \epsilon$
```

Figure 4: K-means Clustering Algorithm, adapted from (Zaki, Mohammed J., 2014).

To define the best number of Ks we had been following the Elbow technique as in the next section.

3.3.3 Elbow method to Choose Number of Ks.

Elbow criterion is a way to define the best K in K-means clustering technique. In Elbow technique to define the best K we had been repeating the experiment 10 times with different number of clusters. In each time we had been plotting the Average within Centroids Distance for each K. Then we had picked the K where the graph has an angle that is followed by a drop and then no variance. At this angle the best K or number of clusters for this dataset is located. The best K is 5 for the first clustering experiment, while it is 6 for the second one as shown in Figure 5 and 6 (The Data Science Lab, 2013).

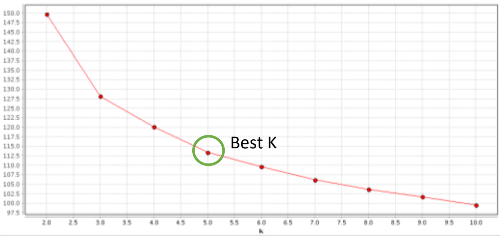


Figure 5: Elbow Method to determine the no of Ks- first Experiment

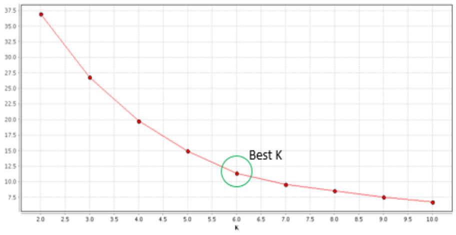


Figure 6: Elbow Method to determine Ks- Second Experiment

3.3.4 Self-Organized Maps (SOM).

For a high dimensional dataset, visualization is challenging. To simplify the presentation and explore meaningful relationships we had been using the Self Organized Maps (SOM) or Kohonen Maps. SOM are neural networks that converts multidimensional data into two dimensional data representing the relationships between data objects. The location of the nodes on the map represents the similarity (order) to its neighbor in the feature space. So by reducing the high dimensions to a map, visualization becomes easy and attractive, at the same time grouping similar data together is a mean of clustering (Kohonen, T., 1990).

SOM is a special type of Neural Network that uses competitive learning to respond to the samples. For a sample vector, weights from the same size of the output network (no. of nodes) are randomly defined, then the Euclidean distance (commonly used) between the sample and weights should be calculated. The node with the minimum distance is the winner, this winner is considered to represent a cluster of similar objects or neighborhood. Next is the adaptation phase where weights of all neighborhoods nodes should be adjusted using the following formula:

$$w_{ij,NEW} = w_{ij,CURRENT} + \eta(x_{ni} - w_{ij,CURRENT})$$

“Learning rate should be decreasing function of training epochs” (Kohonen, T., 1990). The adaptation will lead the weights to move toward the input attributes values, so it becomes more adapted to cluster similar records.

The variation in the color scheme, shown in Figure 7, from red to dark blue is showing the average Euclidean distance between adjacent nodes. The net shown in Figure 7 called the SOM grid or map, this grid can be used to understand or read the dataset distribution. In this graphical representation red areas are representing dissimilarity (large Euclidean distances between objects). While the dark blue areas are representing similarity (small Euclidean distances between objects). Then the color degrade between orange to blue showing the reduction in distance between nodes. We also can use this color scheme to visualize each feature or attribute, so these colors will represent the value of each individual attribute (dark blue for small values, red for large values). The representation of individual attributes using this coloring schema will generate a grid called feature plane, as shown in Figures 31 to 40 in the Appendix section. (http://www.viscovery.net/self-organizing-maps)

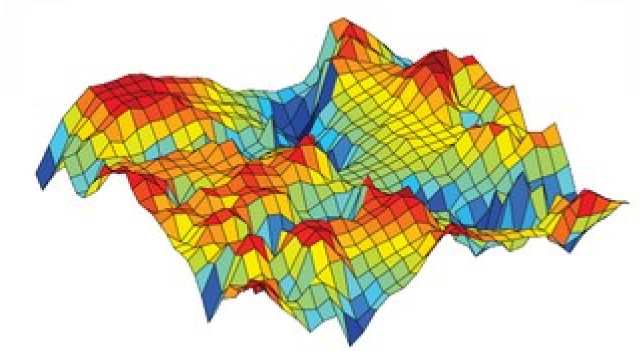


Figure 7: Adapted from http://www.viscovery.net

3.3.5 CHAID Decision Tree.

Based on Elkan, C. 2001 evaluation of COIL challenge 2000 competition, and as the author had argued the randomness nature of the dataset. And because of the randomness, he had claimed that to conduct a successful data mining predictive task, an analyst should take care of the following. First to use an effective feature selection technique. Second look for interesting unknown patterns. In our experiment we will look for interesting patterns that could inspire a marketing consultant in any insurance company.

We will use the CHAID decision tree classifier where a feature selection based on chi-squared attribute relevance test is imbedded within the model. The reason of using this method is because it generates interpretable trees. These readable trees can reveal the demographics (features) as well as the behavior of customers that can lead to purchase a certain insurance policy (target).

CHAID stands for Chi-square Automatic Interaction Detector. The usage of chi-square test to study whether the class and the assessed predictor are independent, makes this technique one of the most powerful and efficient techniques along with IG. This test is used by the model at each level to evaluate the best features to be used to split the tree. The feature with the smallest p-value will be selected for splitting. (Tsiptsis, K., 2011). So CHIAD works exactly same as any decision tree with only one difference, Chi-square is used to determine the best feature to split the tree instead of IG in decision trees.

Chi-square statistic is a statistical test used to judge if a distribution of observed frequencies varies from the hypothetical expected frequencies. Nominal data is the appropriate data type for Chi-square, so this test uses frequencies instead of means and variances. While IG tests the uncertainty (entropy) in predicting the class having the feature, Chi-square measures the independency between the target and the features. (Geng, X., 2007)

Although CHAID is limited to categorical attributes, therefore cannot be used for quantitative or mixed categorical-quantitative classifiers. At the same time it is an effective way for heuristic search. It searches in large tables quickly so it saves the computing time. CHAID generates smaller trees because of the usage of pre-pruning techniques. This makes the resulting tree best fit for the purpose of interpretation or studying what features led to a certain target (Wilkinson, L., 1992).

Chapter 4: Analysis.

4.1 Experiment 1: Clustering by K-means

4.1.1 Execution:

To run this experiment we had been using the most 21 informative attributes to the target attribute (caravan policy). Then we had applied the Clustering (K-means) operator on the reduced dataset as shown in Figure 8. As recognized from the table of centroids in Figure 9, cluster 0 and cluster 2 are very near to each other. So they could share many characteristics as what we will see in the next sections.

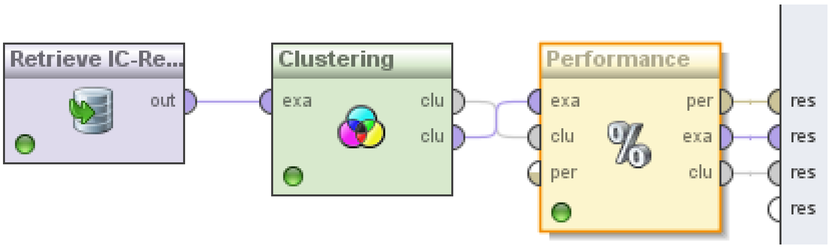
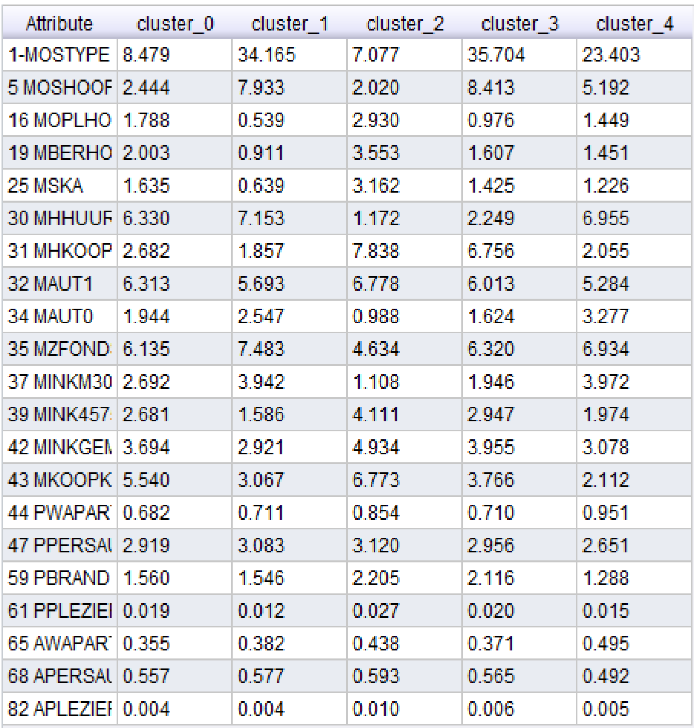


Figure 8: Rapidminer process

Attribute Description
Customer Subtype
Customer main type
High level education
High status
Social class A
Rented house
Home owners
1 car
No car
National Health Service
Income < 30.000
Income 45-75.000
Average income
Purchasing power class
Contribution private third party
Contribution car policies
Contribution fire policies
Contribution boat policies
Number of private third party
Number of car policies
Number of boat policies

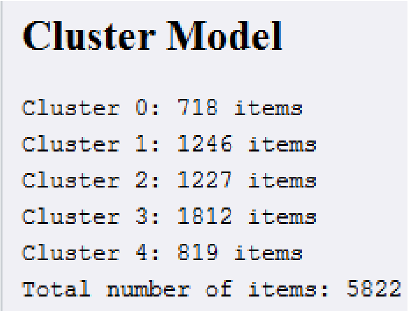


Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
1-MOSTYPE	8.479	34.165	7.077	35.704	23.403
5 MOSHOOF	2.444	7.933	2.020	8.413	5.192
16 MOPLHO	1.788	0.539	2.930	0.976	1.449
19 MBERHO	2.003	0.911	3.553	1.607	1.451
25 MSKA	1.635	0.639	3.162	1.425	1.226
30 MHHUUF	6.330	7.153	1.172	2.249	6.955
31 MHKOOP	2.682	1.857	7.838	6.756	2.055
32 MAUT1	6.313	5.693	6.778	6.013	5.284
34 MAUT0	1.944	2.547	0.988	1.624	3.277
35 MZFOND	6.135	7.483	4.634	6.320	6.934
37 MINKM30	2.692	3.942	1.108	1.946	3.972
39 MINK457	2.681	1.586	4.111	2.947	1.974
42 MINKGEI	3.694	2.921	4.934	3.955	3.078
43 MKOOPK	5.540	3.067	6.773	3.766	2.112
44 PWAPAR	0.682	0.711	0.854	0.710	0.951
47 PPERSAI	2.919	3.083	3.120	2.956	2.651
59 PBRAND	1.560	1.546	2.205	2.116	1.288
61 PPLEZIEI	0.019	0.012	0.027	0.020	0.015
65 AWAPAR	0.355	0.382	0.438	0.371	0.495
68 APERSAI	0.557	0.577	0.593	0.565	0.492
82 APLEZIEI	0.004	0.004	0.010	0.006	0.005

Figure 9: Centroid Table

4.1.2 Evaluation:

The results of the first experiment, K-means on a reduced dataset, had shown the following results. The value of Davies Bouldin is very small, means the intra-distance (between points in the same cluster) is very small and the inter-distance (between clusters) is very big. Small Davies Bouldin means we have a good clustering. In our experiment the value of Davies Bouldin is 1.632 the smallest for K=5.



Cluster Model

```
Cluster 0: 718 items
Cluster 1: 1246 items
Cluster 2: 1227 items
Cluster 3: 1812 items
Cluster 4: 819 items
Total number of items: 5822
```

Figure 10: K-Means Results.

4.1.3 Analysis:

To read the clustering we had visualized the clusters, attributes, and the target class as a scatter plot for the most informative 10 attributes. All these scatter plots are shown in the appendix section Figure 21 to Figure 30. We had analyzed the resulted plot and had recognized the following interesting results:

<u>Cluster A (C0):</u>

Singles, families with Adults, Seniors, Retired, and Religious Farmers in this cluster are more likely to subscribe in a caravan Policy. Because a caravan is made of wood, customers in this cluster usually have 3-4 contribution in fire policy.

Cluster B (C1):

Business Men, Retired, and Cruising Seniors customers tend to do not own a caravan policy. This could be explained by the nature of these families of customers. For example cruising seniors would prefer a sea trip instead of camp out in a caravan. A few number of them have boat insurance policies. At the same time their low purchasing power and average income could be a reason.

Cluster C (C2):

Customers in this segment are Living well singles, average families, or Business Men. People in this segment are more likely to own a Caravan Policy according to their Intermediate to High Purchasing Power and Average Income.

Cluster D (C3):

This cluster includes Seniors, Retired, and Religious Farmers. These people with 4-7 car policy contributions are more likely to subscribe in a caravan policy. This is logically because any caravan needs to be pulled by a car.

Cluster E (C4):

While Seniors, Retired, and Religious Farmers with good income have higher tendency to own a caravan and boat insurance as well. Also customers with 2 contributions of private third party insurance are more likely to own a caravan policy.

4.2 Experiment 2: SOM

4.2.1 Execution:

We had been using Rapidminer 5.3 to conduct the experiment as shown in Figure 11. We had started the experiment by avoiding duplicate in behavior. To do this we had studied the correlation between attributes and ignored the associated (High Weights) ones. After that we had been using the SOM technique. SOM is not only efficient in reducing dimensions, but also for visualizing. SOM visualizes the clustering in an attractive, readable, easy, and fast way. Then we had re-clustered the 2D features resulted from SOM using K-means algorithm.

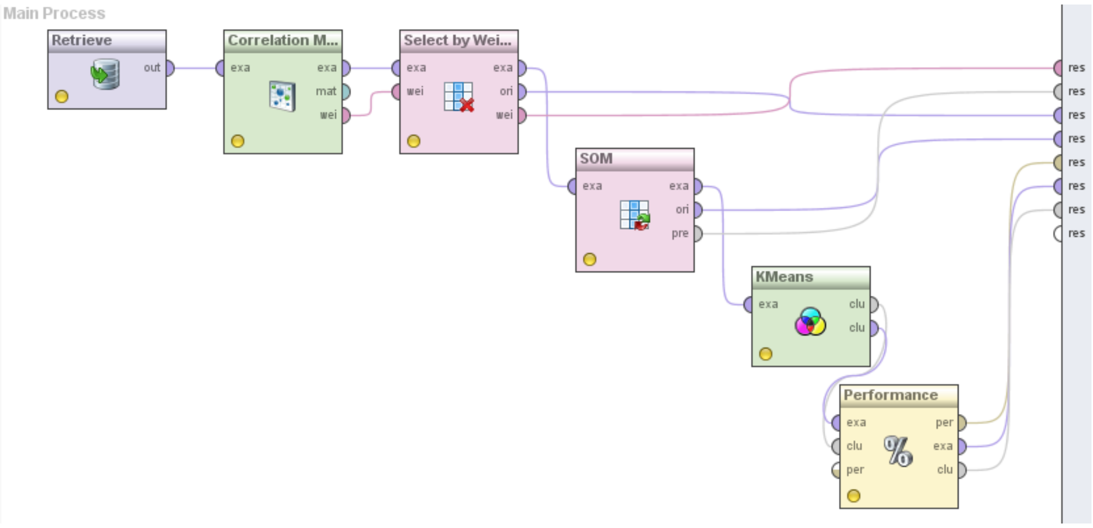
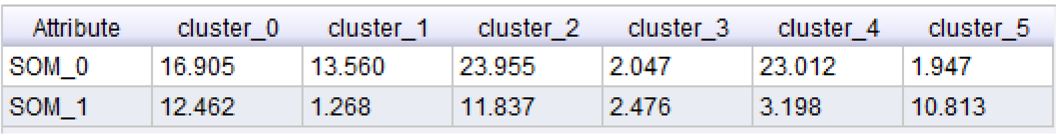


Figure 11: Rapidminer Process



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
SOM_0	16.905	13.560	23.955	2.047	23.012	1.947
SOM_1	12.462	1.268	11.837	2.476	3.198	10.813

Figure 12: SOM Centroid Table

The SOM model generated by the experiment is shown in Figure 13. This model is showing how SOM can visualize high dimensional dataset and the relation between samples in 2D map.

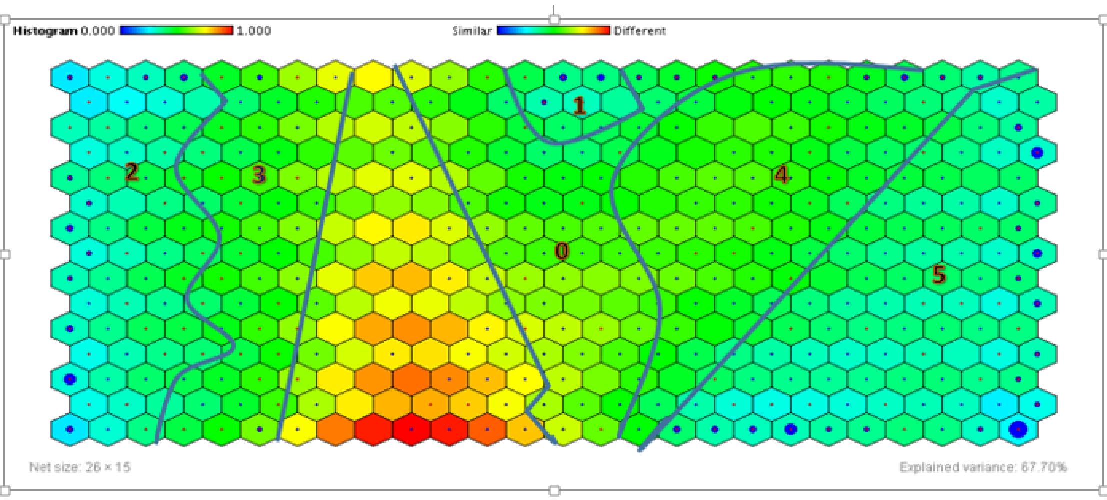


Figure 13: SOM

The following figure is showing the scatter plot of the new 2D feature space resulted from SOM within the 6 resulted clusters.

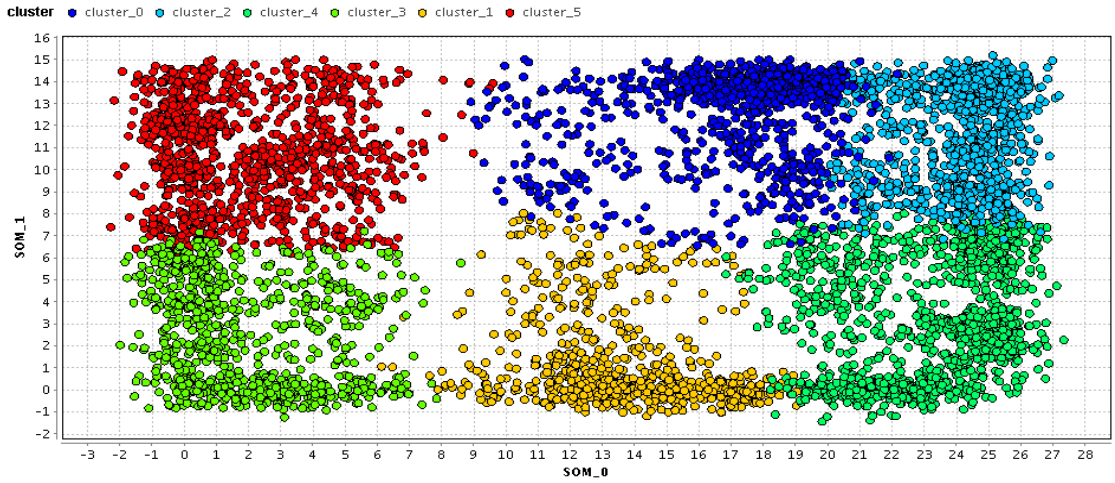
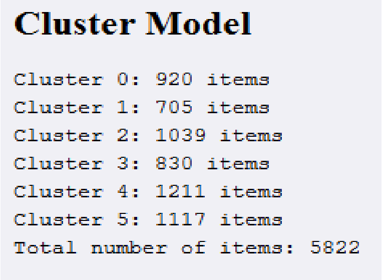


Figure 14: SOM dimensions scatter plot

4.2.2 Evaluation:

We have used Davies Bouldin metric to assess how good the clustering without reference to external information. Low Davies Bouldin Index means that intra-distance is very small and the inter-distance is very big leading to an ideal clustering. For our experiment the value of Davies Bouldin measure is 0.699.



```
Cluster Model

Cluster 0: 920 items
Cluster 1: 705 items
Cluster 2: 1039 items
Cluster 3: 830 items
Cluster 4: 1211 items
Cluster 5: 1117 items
Total number of items: 5822
```

Figure 15: SOM results

4.2.3 Analysis:

To study the behavior of customers within each cluster we have been using the feature plane for each attribute as shown in Appendix. The variation of color within the same feature is representing different values, dark blue for low values, to red for high values. We conclude our experiment with the following analysis of customers' demographics and behavior. By comparing the clustering map with the feature planes of the most 10 interesting attributes we can read the following:

<u>Cluster A (C0):</u>

This is the cluster of Singles Religious Seniors. A customer belongs to this cluster has Higher average income than Cluster B, high purchasing power, high contribution for car policies , high contribution of private third party insurance.

<u>Cluster B (C1):</u>

Is the cluster of Living well Seniors. Customers in this cluster have a high average income and purchasing power, but lower than Cluster A, a high contribution for car policies, a high contribution of private third party insurance.

Cluster C (C2):

The customers in this segment are Middle Class Provincials and Home Owners families. This leads to higher contribution in fire policy. They have high income and purchasing power. They prefer private third party insurance than Cluster D. They also select private health service more than national.

Cluster D (C3):

This is the cluster of Affluent starting young families, with fair number of home owners. This cluster also includes intermediate average income people with high purchasing power. They have low contribution of private third party insurance.

Cluster E (C4):

These customers spend less with higher income than Cluster F, and have high contribution private third party insurance.

Cluster F (C5):

In this segment customers spend more with lower income than Cluster E. It is the segment of Large Families, Employed child, Village families, Couples with teens, mixed small town dwellers, or Traditional families.

By using the SOM in parallel with the feature planes reading, any insurance company would be able to understand its customer's demographics and behaviors. Using this analysis in a cross-marketing campaign to offer the customers new policies could be done as follows:

1- Customers, who prefer National Health Service, insurance packages with national hospitals or clinics.

2- Customer, who prefers private health service than national, an insurance packages with private hospitals or clinics.
3- Customer, who spends less with higher income than other clusters, a life insurance policy.
4- Customer, who spends less with higher income than other clusters, and has no caravan policy, but has car and fire policies, a caravan policy.
5- Customers, with complete family and high to intermediate income and have car insurance, accidents insurance policy.
6- Customers, who are home owners, property insurance policies.

4.3 Results Comparison:

After implementing the above proposes solutions on TIC dataset, the following results are obtained:

Parameters	K-Means	SOM With K-Means
No. Of Clusters	5	6
Execution Time	16 seconds	4 seconds
Davies Bouldin	1.632	0.699
Visualization	Difficult to visualize clusters	Clusters can be recognized easily

Table 1 : Experiments Results

As recognized from Table 1, using SOM along with the K-Means algorithm to cluster the dataset had outperformed using K-Means alone in many dimensions.

4.4 Experiment 3: CHAID

4.4.1 Execution:

In our third proposed solution, we had changed the task from exploration to predicting who could buy a private health insurance. We had used the CHAID decision tree to conduct our experiment, where interpretation is an added value.

We had started the experiment by sampling the two classes equally. National Health policy attribute has a high correlation with the target attributes as shown in Figure 16. So we had removed the National Health policy attribute from the dataset.

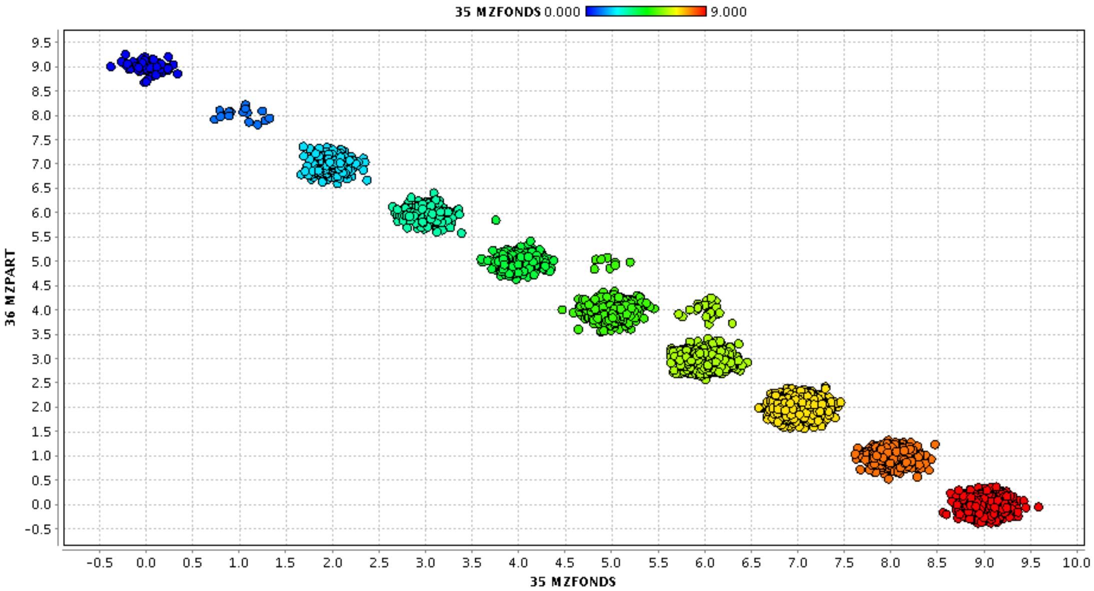


Figure 16: Scatter Plot for (X-National - Y- Private Health Insurance Policy)

Then we conducted the experiment by using cross-validation as a second level of sampling. The x-validation operator guarantees the accuracy of a model by simulating a real experiment, and avoid over-fitting. The CHAID prediction model full process is shown in Figure 17.

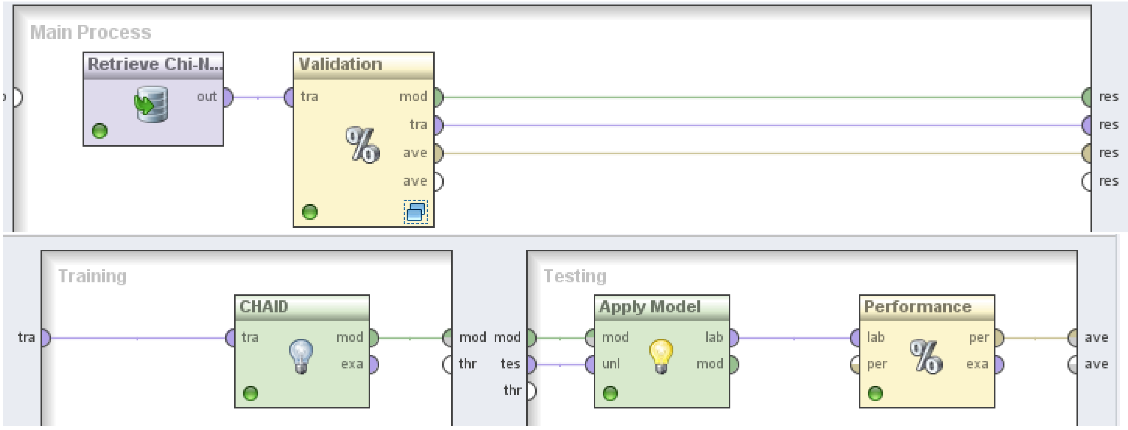
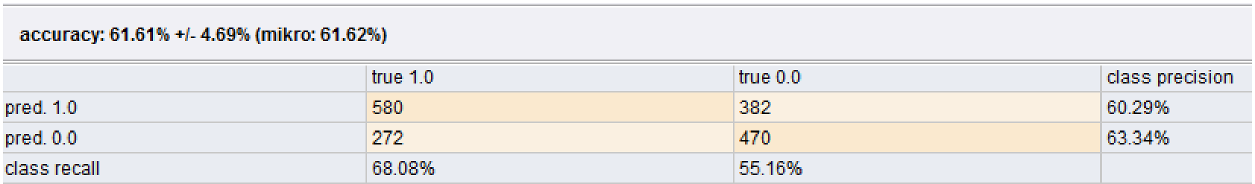


Figure 17: RapidMiner Experiment- CHAID.

4.4.2 Evaluation:

The resulted decision tree, as shown in Figure 19, to predict who can own a private health insurance policy had been showing interesting interpretation to visualize the model.



accuracy: 61.61% +/- 4.69% (mikro: 61.62%)

	true 1.0	true 0.0	class precision
pred. 1.0	580	382	60.29%
pred. 0.0	272	470	63.34%
class recall	68.08%	55.16%	

Figure 18: CHAID sampling classes equally

The interpretation of CHAID tree as recognized in Figure 19, had been showing the importance of three attributes in predicting the target. In our solution the CHAID algorithm had used the Average of House Holds, the Average of Age, and the No of Houses owned by each household to split the tree. In the other hand using DT for the same dataset cannot find any relative attribute to predict the target class and cannot split or generate any tree.

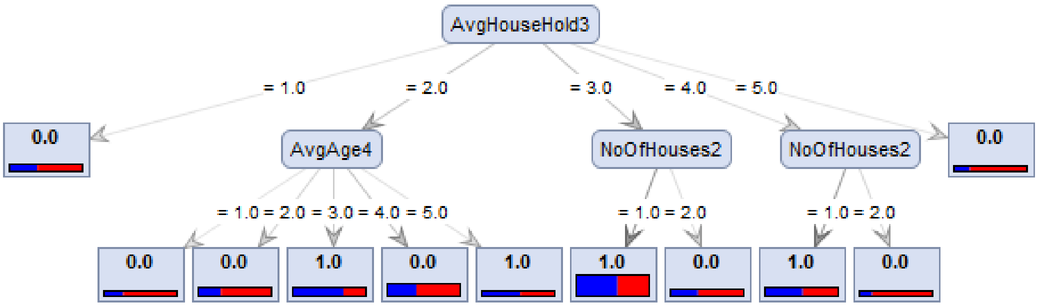
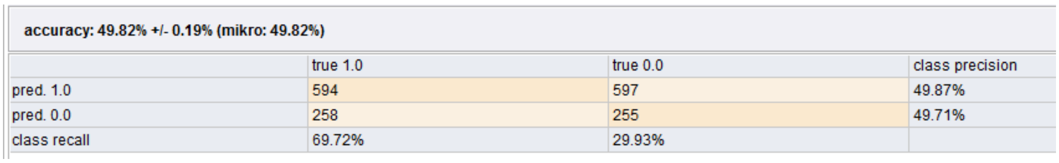


Figure 19: CHAID Sampling Classes Equally



accuracy: 49.82% +/- 0.19% (mikro: 49.82%)

	true 1.0	true 0.0	class precision
pred. 1.0	594	597	49.87%
pred. 0.0	258	255	49.71%
class recall	69.72%	29.93%	

Figure 20: DT Sampling classes equally

As shown in Figures 18 and 20, from the previous two experiments we can recognize the difference between CHAID and DT feature selection techniques. CHAID had selected the relevant variables based on independence/dependence, while DT had selected the relevant variable based on information gain or the most informative feature to the class.

4.4.3 Analysis:

With CHAID technique the generated tree is shown in Figure 19. As recognized Households with average of 2, age from category 3 (40- 50 years) or Seniors have higher interest toward owning a private health insurance. Also from age category 5 (60- 70 years) have a good tendency of owning private health insurance policy. At the same time this interest is not bad when the average of households is 4 and with 1 house only.

This solution can answer two important questions: First, who will buy a private insurance policy? Second, what are the characteristics of this customer recognized from the generated tree? By investigating this information carefully, a company would use this knowledge to build its direct marketing campaign. The campaign would focus on the company potential customers instead of wasting the time with unpromising customers.

Chapter 5: Conclusion & Future work.

The research that we have reported in this paper had two objectives:

To explore several clusters in TIC dataset.

While trying to achieve the above goal we've recorded two contributions. The first considered result was recorded by using K-means algorithm. We've clustered the characteristics of customers who are likely to buy a Caravan policy (belong to the same class). Studying these segments (its demographic and behavioral features) could be useful in designing marketing Campaigns. Since the characteristics and behavior (i.e. Car Policy) in a cluster are known, experts would recommend other policies. Policies owned by customers from the same segment, but not by all its members, could attract others who are sharing the same demographic (same cluster).

Our second contribution was achieved by using SOM along with K-means. This solution has shown interesting results in high dimensional data visualization. Using SOM has reduced the Davies Bouldin by 60% and the computational time by 75% compared with K-means alone. A good visualization of the dataset was accomplished too, so customer's behavior and demographics are readable. A number of marketing plan could be built on top of this segmentation as mentioned in section 4.2.3. One of the ideas is offering Customers, who are home owners, property insurance policies. For customers, who spends less with high income, we can offer a life insurance policy, etc.

To predict who may be a private health insurance policy owner based on demographic attributes.

Predicting different target is our third contribution. Using CHAID technique in constructing the decision tree has been showing remarkable results. Based on dependency between the attributes and the class, the resulted tree has shown interesting knowledge. This knowledge could be used in direct marketing. So sales people would increase the revenue by spending extra time with potential customers who would join this program.

To conclude, trying different data mining methods to study the demographic and behavior of a customer in CRM is really interesting. The intuition of the analyst in finding interesting patterns or insights is the spirit of data mining. So whatever the technique used, people would analyze the results/knowledge in different ways.

The usage of Elbow technique, as we've seen in the experiment, would give us a confident clustering. Also using SOM could improve the clustering quality and enhance the visualization of the data.

As a future plan, extra efforts could be spent toward applying our proposed solutions on data from local market, and in different domains. Also studying the behavior of a customer for a longer time could be effective, since the insurance service is a long term service. Finally integrating different datasets such as customers' car accidents or health records could reveal interesting knowledge and new dimensions.

References.

Ahuja, V., & Medury, Y. (2011). Corporate blogs as tools for consumer segmentation-using cluster analysis for consumer profiling. *Journal of Targeting, Measurement and Analysis for Marketing*, *19*(3), 173-182.

Back, A. H. H. T. E. B. (2011). Customer portfolio analysis using the SOM. *International Journal of Business Information Systems*, *8*(4), 396-412.

Bay, S. D., Kibler, D., Pazzani, M. J., & Smyth, P. (2000). The UCI KDD archive of large data sets for data mining research and experimentation. *ACM SIGKDD Explorations Newsletter*, *2*(2), 81-85.

Branch, N. (2014). A Case Study of Applying SOM in Market Segmentation of Automobile Insurance Customers.

Chopra, B., Bhambri, V., & Krishan, B. (2011). Implementation of Data Mining Techniques for Strategic CRM Issues. *Department of Computer Sciences, babs_niit@ yahoo. com, vivek_bhambri@ rediff. com, alram_krishan@ yahoo. co. in*.

Cristea, M., Marcu, N., & Cârstina, S. (2014). The relationship between insurance and economic growth in Romania compared to the main results in Europe–A theoretical and empirical analysis. *Procedia Economics and Finance*,*8*, 226-235.

Dhandayudam, P., & Krishnamurthi, I. (2012). An improved Clustering Algorithm for customer segmentation. *International Journal of Engineering Science and Technology*, *4*(2), 99-102.

Emtiyaz, S., & Keyvanpour, M. (2012). Customers behavior modeling by semi-supervised learning in customer relationship management. *arXiv preprint arXiv:1201.1670*.

Elkan, C. (2001, August). Magical thinking in data mining: lessons from CoIL challenge 2000. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 426-431). ACM.

Farooqi, M., & Raza, K. (2012). A Comprehensive Study of CRM through Data Mining Techniques. *arXiv preprint arXiv:1205.1126*.

Gupta, A., & Dubey, G. (2012). Identifying Buying Preferences of Customers in Real Estate Industry Using Data Mining Techniques.

Goonetilleke, T. O., & Caldera, H. A. (2013). Mining Life Insurance Data for Customer Attrition Analysis. *Journal of Industrial and Intelligent Information Vol*, *1*(1).

Geng, X., Liu, T. Y., Qin, T., & Li, H. (2007, July). Feature selection for ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 407-414). ACM.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464-472.

Harman, M., Jia, Y., & Zhang, Y. (2012, June). App store mining and analysis: MSR for app stores. In *Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on* (pp. 108-111). IEEE.

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science*, *3*, 1327-1332.

Kaymak, U., & Setnes, M. Target selection based on fuzzy clustering: a volume prototype approach to CoIL Challenge, 2000.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Keele, S., & Alpert, P. T. (2015). Preparing for Retirement in Uncertain Times. *AJN The American Journal of Nursing*, *115*(1), 50-55.

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(5), 431-443.

Mbarki, J., & Jaara, E. M. (2014). Deployment of Partitioning Around Medoids Clustering Algorithm on a Set of Objects Derived from Analytical CRM Data.

Nachev, A. (2008). Solving a Direct Marketing Problem by Three Types of ARTMAP Neural Networks.

Namvar, M., Gholamian, M. R., & KhakAbi, S. (2010, January). A two phase clustering method for intelligent customer segmentation. In *Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on* (pp. 215-219). IEEE.

Nanni, M., Trasarti, R., Monreale, A., Grossi, V., & Pedreschi, D. (2013).*Distributed monitoring of cluster quality for car insurance customer segmentation*. Technical Report, University di Pisa.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Pulakkazhy, S., & Balan, R. V. S. (2013). Data mining in banking and its applications-a review. *Journal of Computer Science*, *9*(10), 1252.

P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company Case.
Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51-59.

Rüping, S. (2004). A simple method for estimating conditional probabilities in svms. In *LWA 2004-Lernen-Wissensentdeckung-Adaptivität. Humboldt-Universität*.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in society*, *24*(4), 483-502.

Ranjit Jeba Thangaiah, P. (2012). Neural Networks in ERP and CRM.*International Journal of Computer Applications*, *39*(1), 1-3.

Soeini, R. A., & Rodpysh, K. V. (2012). Applying Data Mining to Insurance Customer Churn Management. *International Proceedings of Computer Science and Information Technology*, *30*, 82-92.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In *New Directions in Statistical Physics* (pp. 273-309). Springer Berlin Heidelberg.

Thangaiah, P. R. J. (2012). Neural Networks in ERP and CRM. *Neural Networks*, *39*(1).

Tsiptsis, K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

Tudor, A. D. E. L. A., Bara, A. D. E. L. A., & Botha, I. U. L. I. A. N. A. (2011). Solutions for analyzing CRM systems-data mining algorithms. *International Journal of Computers*, *5*(4), 485-493.

Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *Knowledge and Data Engineering, IEEE Transactions on*, *25*(5), 961-973.

Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART. In*Proceedings of Sawtooth Software Conference*.

Yadav, S., Desai, A., & Yadev, V. Knowledge Management in CRM using Data mining Techniques. *International Journal of Scientific & Engineering Research*.

Zadeh, R. B. K., Faraahi, A., & Mastali, A. (2011, January). Profiling bank customers behavior using cluster analysis for profitability. In International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia.

Zadrozny, B., & Elkan, C. (2002, July). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699). ACM.

Zaki, Mohammed J., and Wagner Meira Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge UP, 2014.

Padhy, N., Mishra, D., & Panigrahi, R. (2012). The survey of data mining applications and feature scope. *arXiv preprint arXiv:1211.5723*.

Appendix 1:

Since reading the clustering in a high dimensional dataset is challenging. In the following figures we have plotted the attribute inside the clusters with respect to the class. The most informative 10 attributes values were used to represent the dataset. From Figure 21 to Figure 30 we can read the following:

In Figure 21: Customers with 0-2 car policy contributions has less chance to own a Caravan. Customers with 4-7 car policy contributions, especially in C3, more likely to own a Caravan Policy.

In Figure 22: Customers with 3-4 contribution in fire policies, in C0, C3, are more likely to own a Caravan Policy. Few has more than 6.

In Figure 23: Most of the customers have <= 4 car policies. C0, C2, C3, C4 More likely to own a Caravan Policy.

In Figure 24: Customers in C0, C2 are from type < = 5. Customers in C3, C4 are from type > = 6. C1 has customers from 4-7 type, Career Loners, Living well, cruising seniors, and retired and religious families tend to do not own a Caravan policy. Low Purchasing Power & Average Income). C0, C2, C3, C4 are more likely to own a Caravan Policy (Intermediate to High Purchasing Power & Average Income).

In Figure 25: C0 > 5. C2 2-8 has more ones, purchasing power is >=6. C3, C4 <=6. C1 <=4 Customers tend to do not own a Caravan policy.

In Figure 26: Max contribution in C4 with higher potential to own a Caravan Policy (compare with main type L2). Most of Customers has 0 boat policy contribution.

In Figure 27: C0, C4 have max income and more likely to own a Caravan Policy. C0, C4 has >=7. C1, C2, C3 <=6.

In Figure 28: C0, C1, C2, C4 more likely to own a Caravan Policy when they own 1, or 2 boat policies. Max no in C4 with higher potential to own a

Caravan Policy. All Clusters could have Caravan Policy contribution when no of boat policies=0, min in C1.

In Figure 29: C0, C2 < 13, tend to have more Caravan policy owners (High Purchasing Power & Average Income). C1 14 – 30 tend to have less Caravan policy owners. C3, C4 25-40 more likely to own a Caravan policy (Intermediate to High Purchasing Power & Average Income).

In Figure 30: Most of the records 0, then 2 contributions. The customer with 2 contributions more likely to own a caravan policy. At the same time customer with 0 contribution showed a tendency toward owning a Caravan Policy. C4 has the most tendency.

Scatter Plot attribute-target-clustering (Top 10 informative):

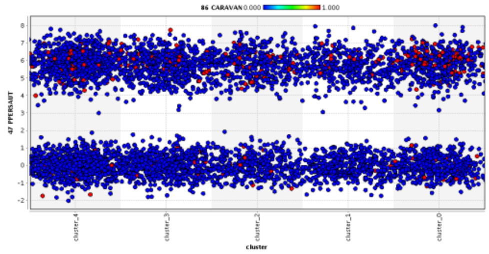


Figure 21: Contribution Car Policies

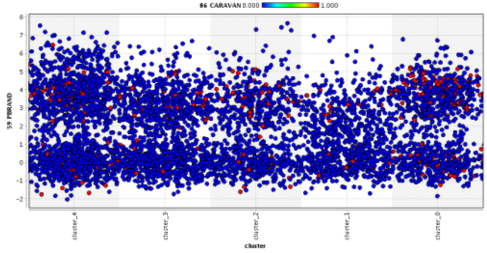


Figure 22: Contribution Fire Policies

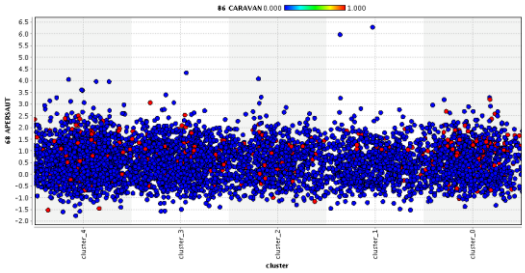


Figure 23: Number of Care policies

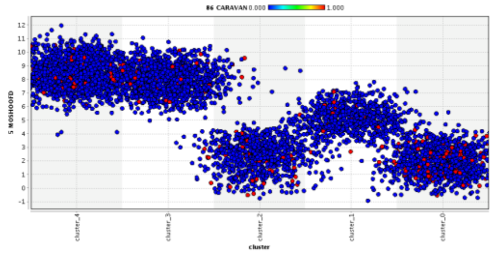


Figure 24: Customer main type

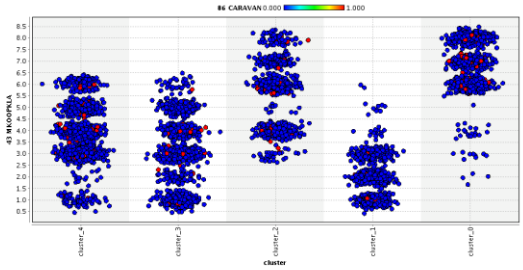


Figure 25: Purchasing power class

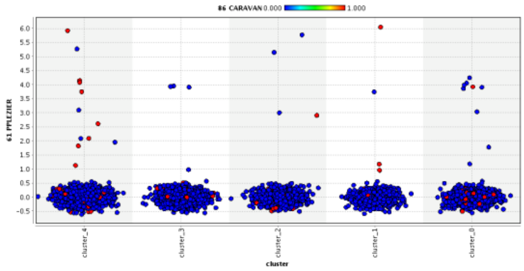


Figure 26: Contribution boat policies

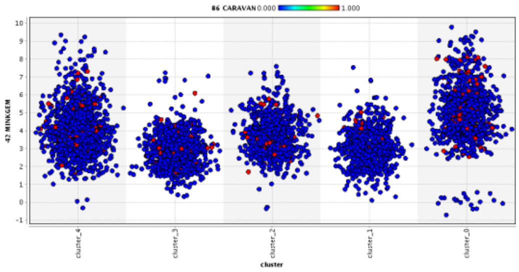


Figure 27: Average income

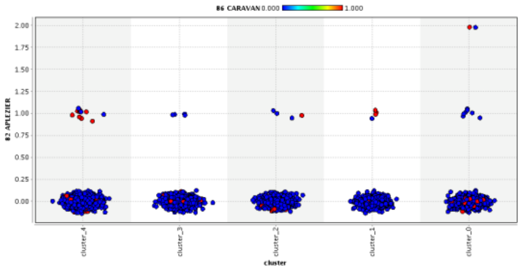


Figure 28: Number of Boat policies

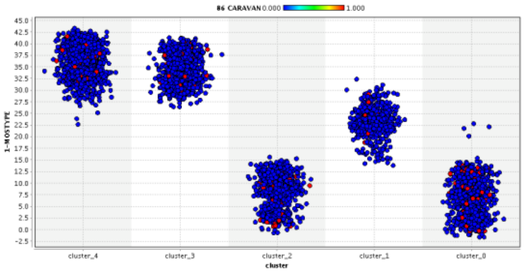


Figure 29: Customer Subtype

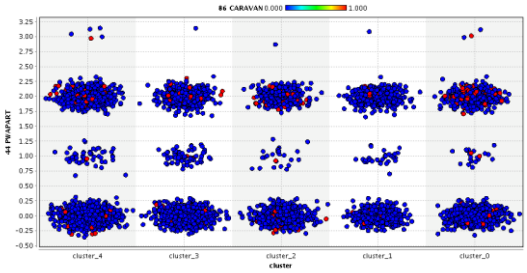


Figure 30: Contribution private third party insurance

Appendix 2:

Each of the following grids is showing the feature plan of an attribute. We are using the most informative 10 attributes in the TIC dataset (informative to the caravan policy class). The feature plan of each attribute is a visualization technique of its value. The variation in color from red (big value) to dark blue (small value) builds this map. This technique called Self Organized Map (SOM).

In Figure 31: a variation in color from light green to dark green representing the most common values of Contribution car policies (0, 5, and 6).

In Figure 32: the attribute Contribution fire policies shows a variation in colors from light blue to medium blue representing the most common values of this attributes (0, 3, and 4).

In Figure 33: the feature plan of the number of car policies attribute reflects its values. The change in colors from dark blue to medium blue visualizing the most common values of this attributes (0, 1, and 2).

In Figure 34: The transformation in the colors from dark blue to orange representing the values of customer main type attribute (1- 10).

In Figure 35: the values of purchasing power class (2- 8) are visualized using the SOM feature plan. A variation in the color from medium blue to yellow is created.

In Figure 36: Most of the records has 0 value for contribution boat policies attribute. The value is represented as dark blue color feature plan.

In Figure 37: the feature plan of the attribute average income. Most of the records upgraded from 1-4 forming a grid from light blue to medium green.

In Figure 38: Most of the records has 0 value for Number of boat policies attribute, this explains the dark blue color feature plan.

In Figure 39: the feature plan of the attribute Customer Subtype. The value of this attribute varies from 1 to 41 and that explains the development in color from dark blue to red.

In Figure 40: the feature plan of the attribute Private third party insurance. Most of the records have the values 0 or 2 forming a grid from dark blue to light green.

Feature Planes (only Top 10 informative):

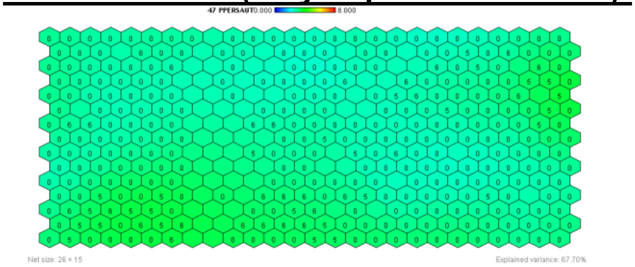


Figure 31: Contribution car policies 0-8

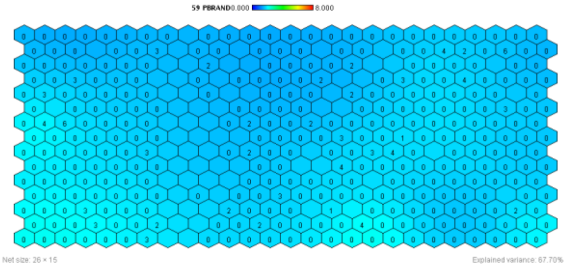


Figure 32: Contribution fire policies 0-8

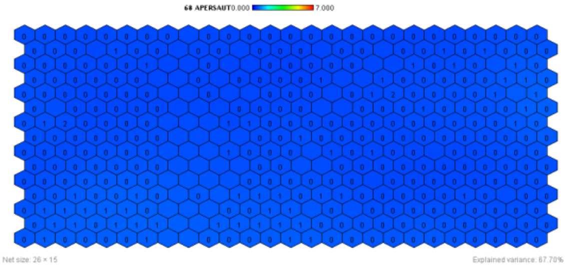


Figure 33: Number of car policies 0-7

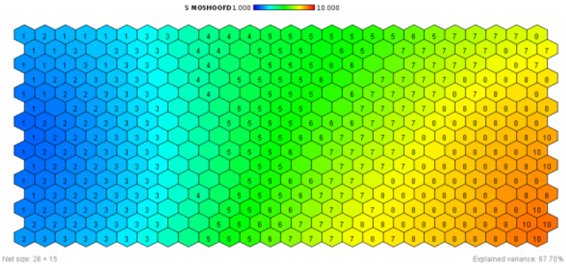


Figure 34: Customer main type.

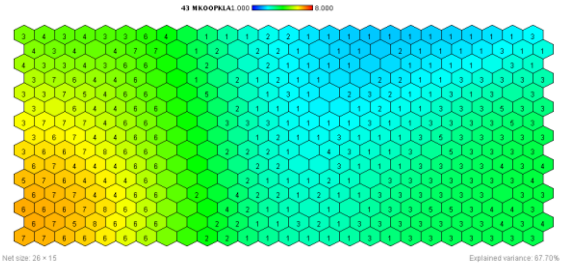


Figure 35: Purchasing power class

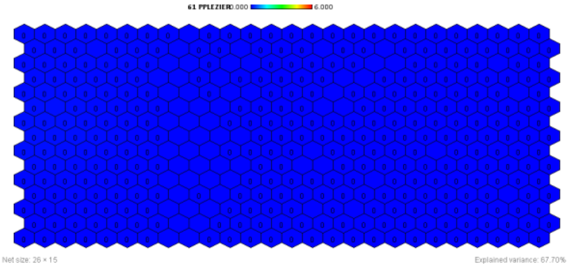


Figure 36: Contribution boat policies

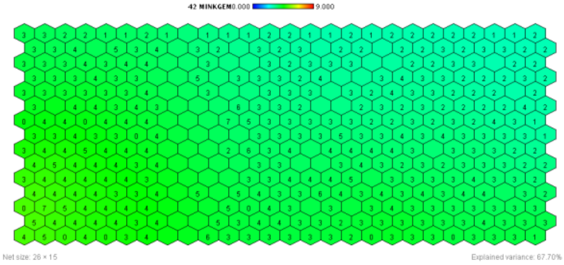


Figure 37: Average income

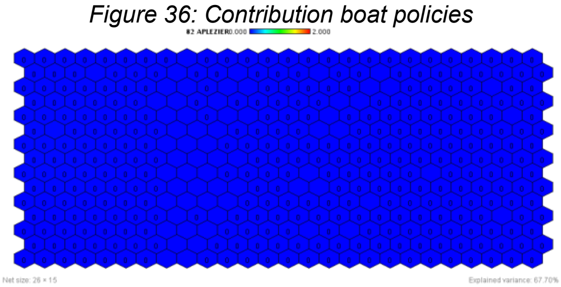


Figure 38: Number of boat policies

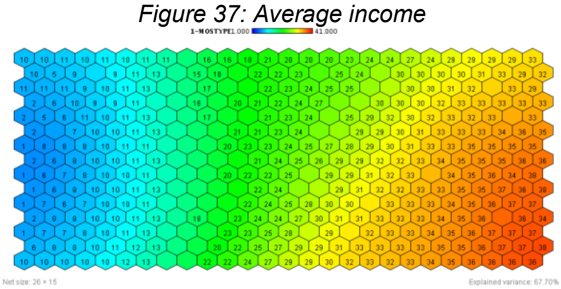


Figure 39: Customer Subtype.

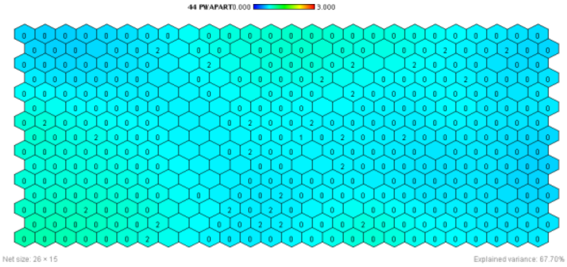


Figure 40: Private third party insurance.