T.C.

ISTANBUL AREL UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

Industrial Engineering Sciences / Engineering Management Program

UNDERSTANDING CUSTOMER VALUE USING DATA MINING APPLICATIONS: A CASE STUDY OF AN INSURANCE BROKER

Thesis for the Degree of Master of Science

Author: **Fethi ATA**

Supervisor: Assist. Prof. Dr. Volkan ÇAKIR

T.C.

ISTANBUL AREL UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

Industrial Engineering Sciences / Engineering Management Program

UNDERSTANDING CUSTOMER VALUE USING DATA MINING APPLICATIONS: A CASE STUDY OF AN INSURANCE BROKER

Thesis for the Degree of Master of Science

Author: **Fethi ATA**

Supervisor: Assist. Prof. Dr. Volkan ÇAKIR

T.C.

İSTANBUL AREL ÜNİVERSİTESİ

FEN BİLİMLERİ ENSİTÜSÜ MÜDÜRLÜĞÜ

YÜKSEK LİSANS SINAV TUTANAĞI

…./…./2018

Enstitümüz Mühendislik Yönetimi Yüksek Lisans programı öğrencilerinde ***146401104*** numaralı ***Fethi ATA*** "*İstanbul Arel Üniversitesi Lisansüstü Eğitim – Öğretim Sınav Yönetmeliği*" nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği "**UNDERSTANDING CUSTOMER VALUE USING DATA MINING APPLICATIONS: A CASE STUDY OF AN INSURANCE BROKER**" konulu tezini, Yönetim Kurulumuzun ***…./…./2018*** tarih ve **……** sayılı toplantısında seçilen ve ……. Yerleşkesinde toplanan biz jüri üyeleri huzurunda, ilgili yönetmeliğin …… maddesi gereğince (…..) dakika süre ile aday tarafından savunulmuş ve sonuçta hakkında ***oyçokluğu/oybirliği*** ile ***Kabul/Red veya Düzeltme*** kararı verilmiştir.

İşbu tutanak 3 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

Danışman : Yrd. Doç. Dr. Volkan ÇAKIR

Üye : Yrd. Doç. Dr. .....

Üye : Yrd. Doç. Dr. .....

PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Fethi ATA

CONFIRMATION

I authorize the storage of the paper and electronic copies of my thesis in Archives of the Istanbul Arel University Institute of Natural and Applied Sciences on the following conditions:

- ☐ All of my thesis can be accessed from anywhere.
- ☐ My thesis can only be opened in Istanbul Arel University Campuses.
- ☐ I do not want access to my thesis to be opened for ..... years. If I do not apply for extension at the end of this period, access to my thesis can be allowed from anywhere.

Fethi ATA

ACKNOWLEDGMENTS

I would like to thank my supervisor Assist. Prof. Dr. Volkan Çakır, who encouraged, guided and motivated me during this study.

I would like to extend my thanks to my colleagues who helped me in every stage of thesis, guiding me ideas from business side and friends supporting me with sharing their tools and vision, and giving advices.

Special thanks are for my parents for their endless patience, love, support and encouragement from the beginning of this long story.

August 2018 **Fethi ATA**

TABLES OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIS : Artificial Immune System

AL : Artificial Intelligence

ANN : Artificial Neural Network

CRM : Customer Relation Management

CRVC : Commission Revenue Value per Customer

CRVP : Commission Revenue Value per Product

DBMS : Database Management System

DDL : Data Definition Language

DML : Data Manipulation Language

DW : Data Warehouse

EM : Expectation Maximization

ETL : Extraction, Transformation and Loading

JDBC : Java Database Connectivity

KD : K Dimensional

KNN : K Nearest Neighbors

LDA : Linear Discriminant Analysis

ML : Machine Learning

MLR : Multiple Linear Regression

ODBC : Open Database Connectivity

RDBMS : Relational Database Management System

SETM : SET-oriented Mining of association rules

SOM : Self Organizing Map

SVM : Support Vector Machine

SQL : Structured Query Language

XML : Extensible Markup Language

ÖZET

VERİ MADENCİLİĞİ TEKNİKLERİNİ KULLANARAK MÜŞTERİ DEĞERİNİ ANLAMA: BİR SİGORTA BROKERLİĞİ UYGULAMASI

Fethi ATA

Yüksek Lisans Tezi, Endüstri Mühendisliği Anabilim Dalı/Mühendislik Yönetimi Programı

Danışman: Yrd. Doç. Dr. Volkan ÇAKIR

Ağustos 2018, – 61 sayfa

Günümüz dünyasında, piyasadaki zor koşullar şirketleri daha yeni ve iyi rekabet etme yolları aramaya zorlamaktadır. Yoğun küresel rekabet ve hızla değişen teknolojik ortamlarda, müşterilerin çeşitli ihtiyaçlarını karşılamak ve müşteri memnuniyetini en üst düzeye çıkarmak, müşterilere değer vermek bir çok firmanın sahip olması gereken genel geçer kurallardan birisi haline dönüşmüştür. Teknolojik gelişmeler ile birlikte, şirketler ve kurumlar müşteri ve satış verilerini sürekli olarak depolamaktadırlar. Veri madenciliği, sürekli depolanan bu veriler içerisinde, daha önceden bilinmeyen, gizli, anlamlı, şirketlerin amaçlarına ulaşmaları doğrultusunda kullanışlı ve değerli bilgileri elde etmemize yarar. Kümeleme analizi bu veriler içerisinde benzerlikler olanları gruplamak için kullanılan, birliktelik kuralları ise benzer olayların gerçekleştiği kuralları tespit eden bir veri madenciği yöntemidir.

Kümeleme analizi, segmentasyon analizi veya taksonomi analizi olarak da adlandırılır. Daha spesifik olarak, gruplama önceden bilinmediyse, homojen vaka gruplarını belirlemeye çalışır. Birliktelik kuralı, ilişkisel veritabanları, işlem veritabanları ve diğer veri havuzu formları gibi çeşitli veri tabanlarında bulunan veri kümelerinden sık kalıpları, bağıntıları, ilişkileri veya nedensel yapıları bulmayı amaçlayan bir prosedürdür.

Bu çalışmanın amacı, sigorta aracılık sektöründe bir şirkete yönelik veri madenciliği araçlarını ve uygulamalarını kullanarak müşteri ilişkileri yönetimi faaliyetleri için bir temel oluşturmaktır. Müşteri ana verileri ve müşterilerin satış işlemleri, müşteri ilişkileri yönetimi faaliyetleri için kullanılabilecek anlamlı bilgilere dönüştürülür. Bu bağlamda, müşteriler arasında ve ürünler arasında bölümlendirme yapmak ve bunlar arasındaki ilişkileri bulmak için bir uygulama yapılmıştır.

Anahtar Kelimeler: Sigorta brokerliği, veri madenciliği, kümeleme analizi, birliktelik kuralları, müşteri ilişkileri yönetimi.

ABSTRACT

UNDERSTANDING CUSTOMER VALUE USING DATA MINING APPLICATIONS: A CASE STUDY OF AN INSURANCE BROKER

Fethi ATA

Thesis for the Degree of Master of Science, Department of Industrial Engineering/ Engineering Management Program

Supervisor: Assist. Prof. Dr. Volkan CAKIR

August 2018, – 61 pages

In today's world, difficult conditions in the market lead the companies to find new ways to compete better. With the intensive global competition and rapidly changing technological environments, meeting customers' various needs and maximizing the value of profitable customers are becoming the only viable option for many contemporary companies. Together with technological developments, companies and institutions constantly store customer and sales data. Data mining is a process to obtain useful, hidden, meaningful, unknown and valuable information in the way of reaching the aims of the companies in this continuously stored data. Clustering analysis is a method of data mining that is used to group similarities within these data, while Associative Rules identifies rules for similar events.

Clustering analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Association rule is a procedure, which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

The aim of this study is to propose a base for the Customer Relationship Management activities by using data mining tools and applications for a company in insurance brokerage sector. Customer master data and sales transactions of customers are converted to meaningful information that can be used for Customer Relationship Management activities. In this concern, an application was made use of to conduct a segmentation of customers and products, and to find relationships between them.

Keywords: Insurance brokerage, data mining, clustering analysis, associative rules, customer relationship management.

1 INTRODUCTON AND OBJECTIVES

1.1 Definition of the Problem

The rapid increase in technology and the rapid penetration of technology into the insurance sector led to the foreign capital companies operating domestically. This development brought about the globalization of the sector and the decrease in the market shares.

In the 21st century, it is expected that there will be a period in which the insurance sector in Turkey will move away from its traditional structure. In the course of this period, the insurance sector will also adapt itself to technological and commercial applications. The factors that are developing the conditions around the world such as increasing customer needs, customer diversity, customer expectations and competitive pressures force insurance companies to respond quickly to changing conditions and to have a more dynamic structure (Felix, 2015). Sekulovska has pointed out that the insurance industry is to understand the competitive factors that are affecting the emerging electronic commerce, and trade will create opportunities for the use of the internet (Sekulovska, 2012).

From day to day, development of technology has also accelerated people's access to information. In this case, companies are compelled to use technology better. Competitiveness is also directly proportional to how well companies use technology (Cappiello, 2018).

Increasing profitability and growth of insurance companies are driven by the use of data mining methods for analysis and presentation of the insurance product to the correct customer base. Offering right products to the right customers is very important. Because it measures customer satisfaction.

Data mining is a crucial facet for making association rules among the biggest range of itemsets (Prithiviraj & Porkodi, 2015). Data mining methods for predicting the future such as clustering analysis and associative rules can be used to identify customers and to offer the right products to the right customers in the insurance industry (Kumbhare & Chobe, 2014).

In the remainder of the thesis, we will describe the problems an insurance broker faces (Section 2), the methodologies used in the studies in insurance sector (Section 3) and the application of customer segmentation and associative rules for finding the relationships between insurance products (Section 4), ending with the results of the study (Section 5).

1.2 Purpose of the study

Sigortayeri A.Ş is an insurance broker, which has agreements with all insurance companies that operate in Turkey. Sigortayeri A.Ş communicates with these companies via web service integration and sells their insurance products online.

Sigortayeri A.Ş has the following channels to reach its customers.

- Web platforms
  - www.sigortayeri.com
  - www.pttsigorta.com
  - arac.sigortayeri.com (only for liability and full coverage car insurance)
  - IOS & Android Mobile Applications
- PTT branches (local offices)
- Telemarketing Team (direct telephone marketing)

If a customer wants to buy an insurance product online using one of the platforms about, he/she can list all the insurance companies' proposals on the screen by first filling out the form with required information. They can then compare these proposals and purchase the one that most suits to their needs.

Sigortayeri A.Ş provides all relevant service infrastructure for any types of insurance selling to PTT branches. PTT branches are able to sell the insurance policies of many insurance companies through online this service integration.

Telemarketing teams call the customers whose insurance policies are about to expire. Some customers are more valuable for the company than the others due to factors such as customer profitability, loyalty, etc. The company subjects the customers to a certain categorization based on their profile and attributes.

Compensation of revenue reductions resulting from reduced market share can be achieved by presenting the right product to the right customer and by creating products that are more suitable for new and targeted customers (Vijayarani & Sharmila, 2017).

Clustering analysis and associative rules will be used to identify customers and to offer the right products to the right customers (Kumbhare & Chobe, 2014). Data mining is a crucial facet for making association rules among the biggest range of itemsets (Prithiviraj & Porkodi, 2015).

The purpose of the study is to better understand the customer profile of Sigortayeri A.Ş. whereby the company can offer products that are more suitable for their customers.

1.3 Methods and Methodologies Used In This Research

In this research, Clustering Analysis and Association Rules of Predictive Data Mining algorithms are used.

1.3.1 Clustering Analysis

Clustering analysis is a method of data mining that tries to identify certain homogeneous constructs in a database or dataset. Clustering analysis is also known as taxonomy analysis or segmentation analysis. Once it is explained in more detail, it helps to find similar groups that are not known previously. Because clustering analysis is exploratory, it makes no difference between dependent and independent variables (StatisticsSolutions, 2018).

1.3.2 Association Rules

Association rule mining is a data mining method used to find correlations, causal structures, frequent patterns or relationship clusters in various databases and in data sets.

1.4 Research questions

In the future, as the technology progresses very fast, the data collected by intelligent chippers in vehicle or some devices which is used by the people. The data collected by the intelligent chips will be analyzed more accurately, and these estimations will be made to calculate the correct contribution share of the person or property. For example, with a chip placed on a vehicle or an application installed on a person's phone, the risk of accidental recording of that person's driving behavior can be accurately measured. Those who have a high risk of accidents will have their insurance policy costs relatively high compared to those who have a low accident risk.

1.5 Plan of Study

Choosing a thesis topic: 1 September 2017 - 15 October 2017

Choosing of Research Method: 1 October 2017 - 25 November 2017

Data Collection: 1 November 2017 - 15 December 2017

Scanning of Resources and Observations: 1 December 2017 - 25 January 2018

Writing of overview article for journal: 1 January 2018 - 30 January 2018

Writing of the project: 10 February 2018- 15 March 2018

Controlling of the project: 15 March 2018 - 20 April 2018

Making the required arrangements: 20 April 2018 – 5 June 2018

Final Checks: 5 June 2018 – 15 June 2018

Delivery of the thesis: 20 June 2018

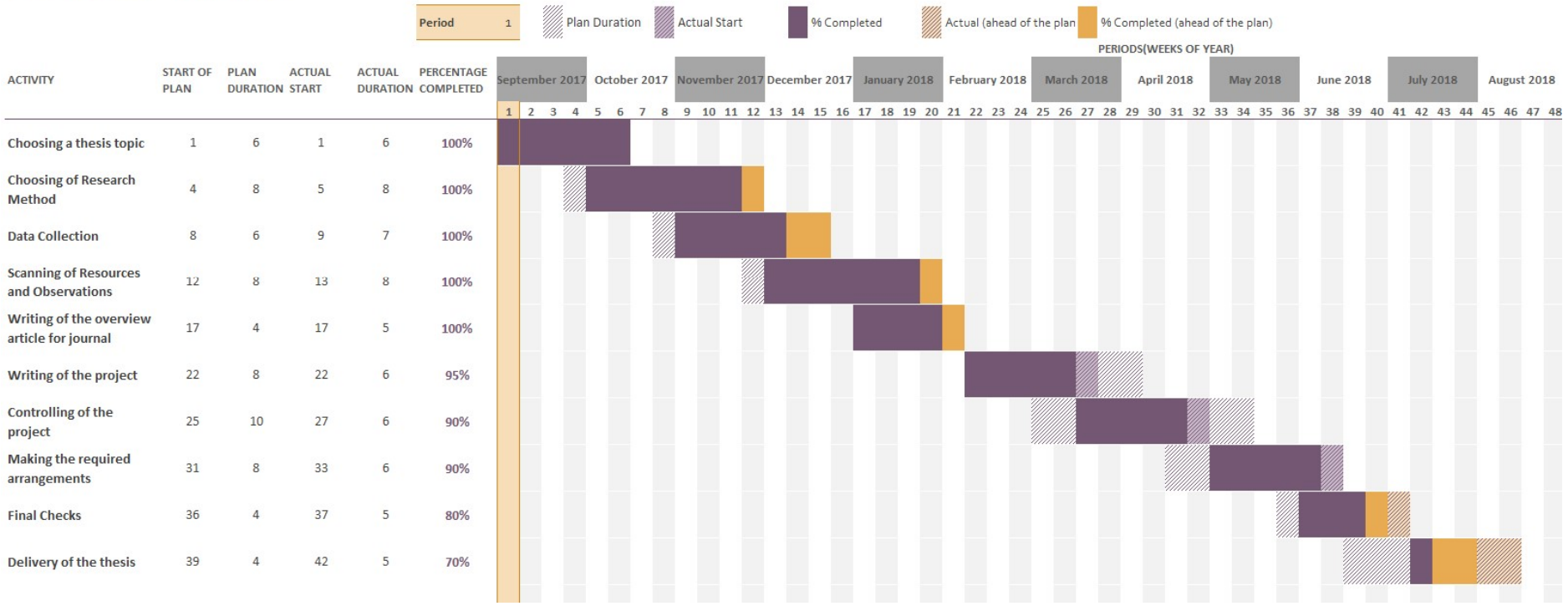


Figure 1-1 Gantt Chart Project Plan

2 LITERATURE OVERVIEW

2.1 History of Insurance

Insurance is a contract, defined by a policy, in which an individual or entity receives financial preservation or refund against losses from an insurance company (investopedia.com, 2018).

Insurance policies are used to preserve against the risk of financial losses, both big and small, which may damage the insured or their property, or that may be caused by damage or injury from a third party.

People can buy insurance policies at many stages of their life. For example, home, car, health, business or pension policies.

When someone buys a policy, he or she makes regular payments, known as premiums, to the insurer. If they make a claim, insurer pays the damage covered by the policy.

During the policy, the insured person cannot withdraw their money if they do not make any claims. With this money, which is not taken back and collected in the premium pool, the claims of other insurers who make claims can be covered.

The history of insurance began when Chinese merchants brought together the risk of losing the cargo carried by the Chinese rivers. In 3000 BC, traders and merchants in Sumer and Babylonia brought together the risks of protecting important cargo losses for pirates and thieves.

Up until the beginning of the 1900s, insurance was used in the fields of sea transportation and cargo against losses either from natural disasters or from thieves and pirates. After the commercial insurances, history of insurance continued with life insurances and fire insurances.

Before the 1900s, large financial services consisted of banking and insurance. Nevertheless, in the first years these two major areas of financial services were widely isolated because it was difficult to manage profitably without modern technology.

2.2 Historical Developments of Methodologies

Data mining is everywhere, but the story begins years ago with Edward Snowden and Moneyball. The following are important milestones in the history of data mining, as well as how data is assembled and merged with big data and data science.

Data mining is the process of discovering and exposing patterns in large data sets. Big Data is a computer science sub-unit that brings together many techniques from data science, statistics, database theory and machine learning.

After the death of Thomas Bayes, an article was published in 1763 about a theorem about the possibility of proving Bayes' theorem. Afterwards, Adrien-Marie Legendre and Carl Friedrich Gauss used setbacks in 1805 to find solar orbits. Later, in 1936, Alan Turing presented the idea of a Universal Machine that could do calculations like modern day computers. After Turing, Warren McCulloch and Walter Pitts were the first to create a conceptual model of neural networks in 1943. Lawrence J. Fogel founded a new company called Decision Science in 1965 for historical programming applications. Then it was possible to store and query terabytes and petabytes in complex database management systems in the 1970s. Later, in 1975, John Henry Holland wrote a groundbreaking book on genetic algorithms called Natural and Artificial Systems Adaptation. Then in the 1980s it became a general-purpose tool to create neural network models, meaning to protect a product called the DataBase Mining Workstation. Gregory Piatetsky-Shapiro has coined the term "Knowledge Discovery in Databases", called KDD, in 1989. Since then, Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik have proposed an improvement over the original support vector machine, which allows the creation of non-linear classifiers. This development is known as a supervised learning approach, which in 1992 identified and analyzed models used for regression and classification analysis. Later, Gregory Piatetsky-Shapiro has launched the news discipline Knowledge Discovery Nuggets (KDnuggets), originally used to associate researchers who participated in the KDD workshop. Later, William S. Cleveland presented data science as a liberal discipline in 2001. Later, Jeff Patil and Jeff Hammerbacher used it to define roles in LinkedIn and Facebook in 2001.

Finally, one of the most active techniques discovered is Deep Learning. Beyond other techniques, data reveal some of the greatest challenges in the world of artificial intelligence and data mining, depending on the acquisition of dependencies and sophisticated patterns (Li, 2018).

2.2.1 History of Clustering Analysis

Clustering analysis has a great history, dating back to the 1950s and before. It is conspicuous how different disciplines were the primary areas of work in clustering at miscellaneous times. Such applications were the operator of methodology development and practical deployment. In turn, disciplines like information science, library, computer science and statistics appreciate a great deal to the theory and practice of cluster analysis (Greenacre & Blasius, 2018).

2.2.2 History of Association Rules

One of the most used rule-based machine learning methods is association rules learning that was designed to find meaningful relationships between variables in large databases by using metrics and other measurable tools.

The researchers Agrawal and his friends have introduced relationship rules on the base of the concept of strong associations to define the consistency among products in large-scale transaction statements that were recorded by supermarket point-of-sale systems. The rule in the supermarket's sales data shows that there is some consistency between the customer’s purchases (for instance, the customer’s basket with buying such products as milk and cocoa together often contains chocolate-based food). This information may serve as a physical basis for decisions regarding marketing activities, such as product placements or promotional pricing (Dogan, 2014).

In addition to the examples mentioned above from the market basket analysis, association rules are used in many fileds such as bioinformatics, continuous production, intrusion detection and mining of the web. In contrast, rule-based learning generally does not take into account the order of items. It only considers the order of items between transactions or within a transaction.

Agrawal and his friends presented the mathematical model of the association rules. (Dogan, 2014)

The following algorithms are used for association rule analysis,

- Apriori
- DHP(Direct Hashing and Pruning)
- FP-growth algorithm

2.2.3 Summary of data mining techniques in insurance

Decision support system take a vital role with the changing and developing technologies in insurance field. So data mining techniques used to support the administrative and management tasks, efficient management of organization and financial data, the controlling of policies (Janakiraman & Umamaheswari, 2014).

When we summarize the role of data mining techniques how to help insurance sector, we can build the following table.

Table 2-1 Data mining & Insurance (Janakiraman & Umamaheswari, 2014)

Data Mining Techniques	Patterns
Clustering	• Segments associated to policy. • Finding out most likely used policy, most unlikely be used policy. • Analysis of customer wearing down or away in insurance sector. • Customer carry similar characteristics.
Classification& Prediction	• Classifying trend of behavior in the organization for successful or unsuccessful customer historical data. • In insurance sector estimating to discover what factors are going to attract new routes. • Estimating the accomplishment progress of segments during the whole of the accomplishment period. • Estimating insurance product action and approach. • Prediction of being retained most likely policy type, being left most likely policy type. • Classifying the historical customer records. • Estimation of consumer behavior. • Estimating the probability of achievement of policies.
Association	• Finding out of the association that promote business technique
Summarization	• Statistical concise information. • Supply summary information. • Miscellaneous multidimensional summary reports.

Based on these information and studies we will use Clustering Analysis for how customer how similar characteristics and how segments are related to the policy, we will also use Association Mining Rules for the relation between insurance products and discovery of such association that improve business technique.

2.3 Literature review in insurance domain

In this section, we provide the appropriate literature review regarding the researches that were made in the area of data mining for insurance domain including such techniques as clustering and associative rules fields.

Insurance domain is one of the most open and potentially beneficial fields that needs to be studied in terms of providing good and relevant services for customers.

The electronic trade market issues have been studied in the researches of Wu and Chou. They made an analysis of online customer segmentation by dividing them into multiple categories. They have suggested that this technique can provide better characterization and understanding of customer's buying behavior. The main idea of their study based on the fact that various data of customer demographic characteristics and buying activity including payment properties such as services satisfaction and internet usage are usually stored at the online shopping databases. Therefore the segmentation of appropriate customer data can help to company administrators establishing and managing good customer relations and improving their marketing strategies (Wu & Chou, 2011).

In contrast with the wide spread CRM value chain techniques that are focused on the impact of customer loyalty and satisfaction to the customer profitability, the role of self-construal (independent) and individual customer level as a part of CRM value creation chain for customer profitability increasing was deeply discussed in the paper of Qi and his friends (Qi, Qu , & Zhou, 2014).

Nakano and Kondo emphasized that customers can be clustered according to their purchase channels such as online stores, social media, mobile, media touchpoints of PC, as well as to their demographic and psychographic attributes (Nakano & Kondo, 2018).

2.4 Literature review of clustering analysis and associative rules usage in insurance business

Janakiraman and Umamaheswari claimed that the vital business decisions of insurance companies in many aspects depend on data mining technology. Efficient management of customer data is the crucial part of insurance industry. Data mining techniques may not only provide the optimal customer data operation but become the good instrument for the decision making process. The preventive regulations and processing of various requests serve an obstacle for the prompt extraction of appropriate information from the customer's databases. The data mining technology may help to avoid the wasting of time and provide the efficient data management (Janakiraman & Umamaheswari, 2014).

Some issues of the ways for selling of insurance products have been explored in the works of Desik and Behera. The reasons for that they have found in the recent regulatory changes and volatile market conditions in the insurance area. They have suggested a focus on customer–centric approach instead of product-centric conception. The authors considered customer data and knowledge management as a key technique for designing customized and well-suited offerings to attract and retain customers by any insurance company. The impact of advanced analytics on identifying of profitable cross-sell opportunities by the insurers is the one of the research points of this paper. By the suggestion of authors the improved analytics may help in designing strategies by targeting the right products for the right customers that can modify any cross-sell campaigns (Desik & Behera, 2015).

Fang and his friends have also focused on the idea of the predictable future cash flowing and historical buying attitude in insurance industry. They have recommended the method for investigation of the real insurance customer contribution (Fang, Jiang, & Song, 2016).

By following the author Fang and his friends the main two important factors for insurance company customers evaluating we can consider the premium profit and the claim risk since they characterize the main profit sources and major expenditures (Fang, Jiang, & Song, 2016). However, insurance broker may do not take into account the risk of claims because it deals with only different types of selling policies.

Schreiber have focused on the consequential arrangement of significant strategic decisions concerned to pricing, product innovations, and distribution channels to classify customer segments and activating incumbents in order to maintain profitability for a stable and continual market share (Schreiber, 2017).

Data mining in insurance sector, will be helpful for estimating medical coverage and fraud detection. Companies can predict or analyze the behavior of existing customers and they can recommend new insurance policies and offer campaigns to their customers and potential customers. Sale and profit rates of potential customers can be projected by examining the data of existing customers. In insurance industry, there are some fields of data mining below;

- Determine the risk factors that estimate losses, profits, and claims.
- Analysis of customer level
- Marketing and selling analysis
- Producing new types of insurance products
- Reinsurance
- Financial analysis
- Predicting the claim provision, which is outstanding
- Fraud detection

Decision support system take a vital role with the changing and developing technologies in insurance field. So data mining techniques used to support the administrative and management tasks, efficient management of organization and financial data, the controlling of policies (Janakiraman & Umamaheswari, 2014).

2.5 Conclusion

When we summarized the researches in the field of customer segmentation and in the field of insurance, we can create the following tab,

Table 2-2 Descriptive Sample Characteristics for customer segmentation

		(Schreiber, 2017)	(Fang, Jiang, & Song, 2016)	(Nakano & Kondo, 2018)	(Wu & Chou, 2011)	(Qi, Qu , & Zhou, 2014)	**Our Segmentation Application**
Descriptive Sample Characteristics (Attributes)	Age	X	X	X	X	X	**X**
	Gender	X	X	X	X	X	**X**
	Smoker	X					
	Marital Status	X			X		
	Claim Amount		X				
	Place of Residence - Region	X					**X**
	Level of Education	X		X			
	Net Household Income	X			X		
	Occupation		X			X	
	Insurance Premium	X					**X**
	Payment Type		X				
	Number of Family Members	X		X		X	
	Number of Children	X		X			
Additioanal Attributes	**Customer Lifetime Period**						**X**
	Repeat Sales				**X**		**X**
	Insurance Product Group		**X**				**X**
	***CRVC**						**X**

*CRVC is the Commission Revenue Value per Customer

3 METHODOLOGY

3.1 Data Science

We need to understand project requirements and objectives with a domain perspective. Then transform this knowledge into a data science problem designed to achieve the goals. Data science projects are usually designed and structured specifically around an industry sector's specific requirements (as shown below) or for a single organization. A successful data science project begins with question or need, which is well-defined.

In the table below 2014 - 2016, it is seen in which areas data mining is used at what rates (Gregory, 2018).

Table 3.1-Data Mining Usage Areas

Industries and Fields where you applied Data Science, Data Mining and Analytics in 2016?			
Industries and Field	**2014 % of voters**	**2015 % of voters**	**2016 % of voters**
CRM/Consumer analytics	22.2%	18.6%	16.3%
Finance	10.9%	15.4%	15.0%
Banking	16.7%	14.3%	13.4%
Advertising	10.4%	8.9%	12.0%
Science	13.6%	11.7%	12.0%
Health care	16.3%	13.4%	12.0%
Fraud Detection	13.6%	10.0%	11.1%
Retail	13.6%	9.1%	10.3%
Insurance	8.6%	7.4%	9.2%
E-commerce	9.5%	10.3%	8.9%
Telecom / Cable	9.0%	7.7%	8.3%
Social Media / Social Networks	8.6%	10.3%	8.3%
Software	7.2%	6.0%	7.2%
Network Infrastructure / IT	--	6.6%	7.2%
Energy / Oil / Gas	9.5%	8.9%	7.1%
Education	7.7%	10.0%	7.1%
Credit Scoring	8.1%	7.1%	6.9%
Supply Chain	--	--	6.5%
Pharma / Medical	7.2%	6.0%	6.5%
Other	13.6%	8.9%	6.3%
Investment / Stocks	5.0%	4.3%	6.2%
Biotech/Genomics	6.8%	4.9%	5.8%
Manufacturing	9.0%	6.9%	5.6%
Military / Government	6.3%	7.1%	5.6%
Search / Web content mining	6.3%	6.0%	5.4%
Automotive/Self-Driving Cars	5.9%	4.3%	4.5%
Direct Marketing/ Fundraising	7.2%	5.1%	4.3%
Mining	--	3.7%	4.2%
Travel / Hospitality	3.2%	2.6%	4.0%
Entertainment/ Music/ TV/Movies	1.8%	3.1%	4.0%
HR/workforce analytics	5.9%	6.3%	3.6%
Mobile apps	2.3%	1.4%	3.3%
Agriculture	--	2.9%	3.3%
Games	1.8%	4.0%	2.9%
Security / Anti-terrorism	2.3%	2.3%	2.7%
Social Good/Non-profit	1.4%	2.3%	2.0%
Social Policy/Survey analysis	1.8%	1.7%	1.8%
Anti-spam / Junk email	1.8%	0.3%	1.1%

3.2 Data Preparation

Data set creation from one or more data sources is also used for data preparation, research and modeling. It is a good practice to recognize the data, to investigate the data in the initial study and to start with a data set in the firstly in order to better understand possible data quality problems.

Data preparation is time-consuming and fault risk-appearing process. We could use the term "garbage-in-garbage-out" for data analyzing projects where data is gathered with many invalid, missing, out-of-range values. However, careful analysis of poorly scanned data for such types of problems can lead to misleading results. In addition, the quality of prepared data is the crucial aspect for the success of data science projects.

3.2.1 Data

In generally, data is information that typically can be measured and written by categorical (counting) or numerical. Variables render service as placeholders for data. There are two kinds of variables, categorical and numerical.

We can consider a continuous or numerical variable that may accept any value within an infinite or finite range (for example, blood pressure, temperature, height, weight etc.). By studying two kinds of numerical data, ratio and interval, we may find out that by interval scale data can be subtracted and added but cannot be meaningfully divided or multiplied for the reason that there is no true zero. For instance, it is not possible to accept that one week is twice as hot as another week. Otherwise, data on a ratio scale has true zero and can be divided, multiplied, added or subtracted (for example, weight).

When we say about variable that can accept two or more values (categories), we mean a discrete or categorical variable. Categorical data is dived into 2 types, ordinal and nominal. Ordinal data has a real ordering in the categories. For example, "level of energy" with three orderly categories (high, medium and low). In contrast, nominal data does not have a real ordering in the categories. For instance, "gender" with two categories, female and male.

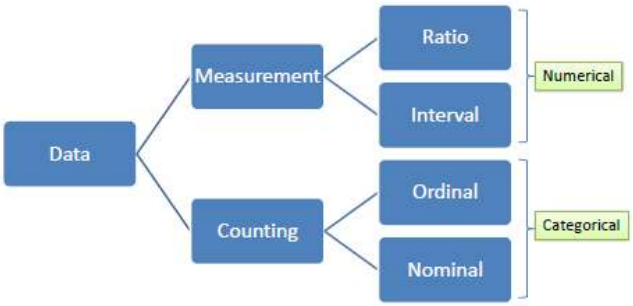


Figure 3-1 Categorization of Data (Sayad, 2018)

3.2.2 Dataset

Dataset, which contains a set of collected data, usually can be shown in a table form. Every column of the table identifies a particular variable and each row indicates the given member of the data.

Columns

ID	Outlook	Temp	Humidity	Windy	Play Golf
1	Rainy	85	92	False	No
2	Rainy	80	88	True	No
3	Overcast	83	86	False	Yes
4	Sunny	70	80	False	Yes
5	Sunny	68	?	False	Yes
6	Sunny	65	58	True	No
7	Overcast	64	62	True	Yes
8	Rainy	72	95	?	No
9	Rainy	?	70	False	Yes
10	Sunny	75	72	False	Yes
11	Rainy	75	74	True	Yes
12	?	72	78	True	Yes
13	Overcast	81	66	False	Yes
14	Sunny	71	79	True	No

Rows

Values

Figure 3-2 Sample Dataset

There are many available possibilities for rows, columns and values.

- Variables, Attributes, Fields, Columns
- Vectors, Instances, Cases, Objects, Examples, Records, Rows
- Data, Values

In the predictor model, attributes or predictors are input variables, and the output variable is the target or class attribute, the values of the estimators, and the function of the prediction model.

3.2.3 Database

The technology and science of managing and saving data so users can fetch, update, insert or delete such data.

A Database manages the persisted information that the users can easily access and modify. The relational databases, which are mostly used database system, are consist from tables with columns and rows. A table keeps same type of information whose attributes are columns. There are relations between tables according to concepts based on common keys (columns), thus relevant data can be retrieved from related tables in a convenient way, which express the term “relational database”. A Database Management System (DBMS) operates data for providing accessibility by multiple users and maintains information to keep integrity and efficiency. The software applications or clients are connected to databases through ODBC (Open Database Connectivity) or JDBC (Java Database Connectivity).

Figure 3-3 Database Schema

3.2.4 SQL

SQL is stands for Structured Query Language. It is a database computer language for administrating data in RDBMS (Relational Database Management Systems).

SQL Data Definition Language (DDL) allows database tables to be created, deleted or altered. In a database, it can be defined that Indexes (keys), impose restrictions and specifying links between database tables.

Indexes (keys) impose restrictions between database tables and specifying links between tables can be defined.

SQL allows users to manipulate and access data with the commands below,

Table 3-1 Database SQL Commands

Database SQL Commands	**Description**
CREATE TABLE	creating a new table
ALTER TABLE	altering a table
DROP TABLE	deleting a table
CREATE INDEX	creating an index
DROP INDEX	deleting an index
SELECT	fetching of data from a database
INSERT INTO	inserting new data into a database
UPDATE	updating data with the new ones in a database
DELETE	deleting data in a database

3.2.5 ETL (Extraction, Transformation and Loading)

When we use ETL process, we envisage the process of taking out data from data sources and loading it into data destinations by using a set of transformation functions. Therefore, data extraction means the ability to take out data from a diversity of data sources, such as relational databases, flat and XML files, streaming data or ODBC/JDBC data sources. Data transformation means the ability to merge, split, convert, clean and collect data. Data loading uses to supply the ability to put data into destination databases with delete, insert, update statements, or in bulk.

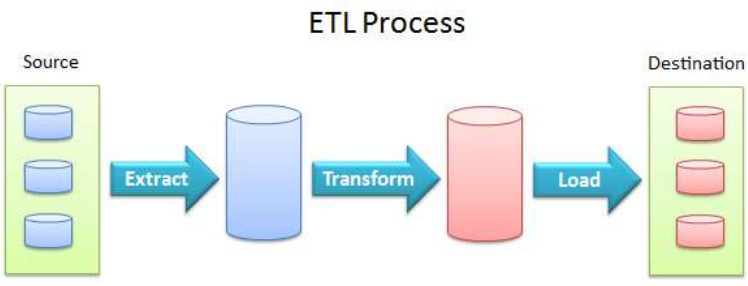


Figure 3-4 ETL (Extract-Transform-Load)

3.3 Data Science

Data Science or Data Mining is the analysis of data to predict the past and predict the future. It is a multidisciplinary field that combines data science, statistics, artificial intelligence, machine learning and database technology. The value of data intelligence applications is known to be very high. Many businesses record very large amounts of data in their operations over the years, and data mining can provide valuable information from these data. Later, businesses can turn out more information, more sales, more profits and more customers. This also applies to engineering and insurance.

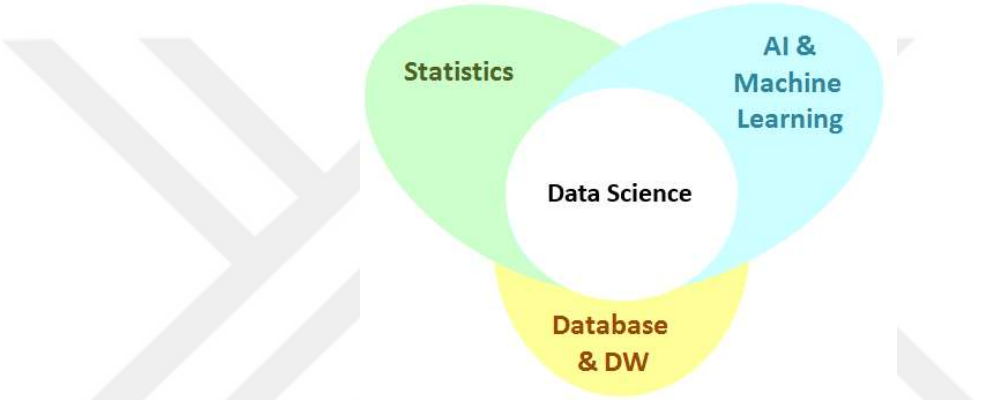


Figure 3-5 Data Science (Sayad, 2018)

3.4 Statistics

It is the science of recording, analyzing, summarizing, organizing, classifying and disclosing data.

3.5 Artificial Intelligence

The operation of computer algorithms that govern intelligent simulation is used to demonstrate actions that are normally thought of as requiring intelligence.

3.6 Machine Learning

Machine learning (ML) is a category of algorithm that serves software applications to be more accurate in predicting outputs without being specifically programmed. The study of computer algorithms are being automatically improved through experience.

3.7 Data warehousing

Data warehouse (DW) is a technology and science of gathering, managing and storing data with advanced multi-dimensional reporting services with the help of the decision-making processes.

Data understanding is separated into two parts in the most general structure. (Sayad, 2018)

These are,

- Explaining the past,
- Predicting the future,

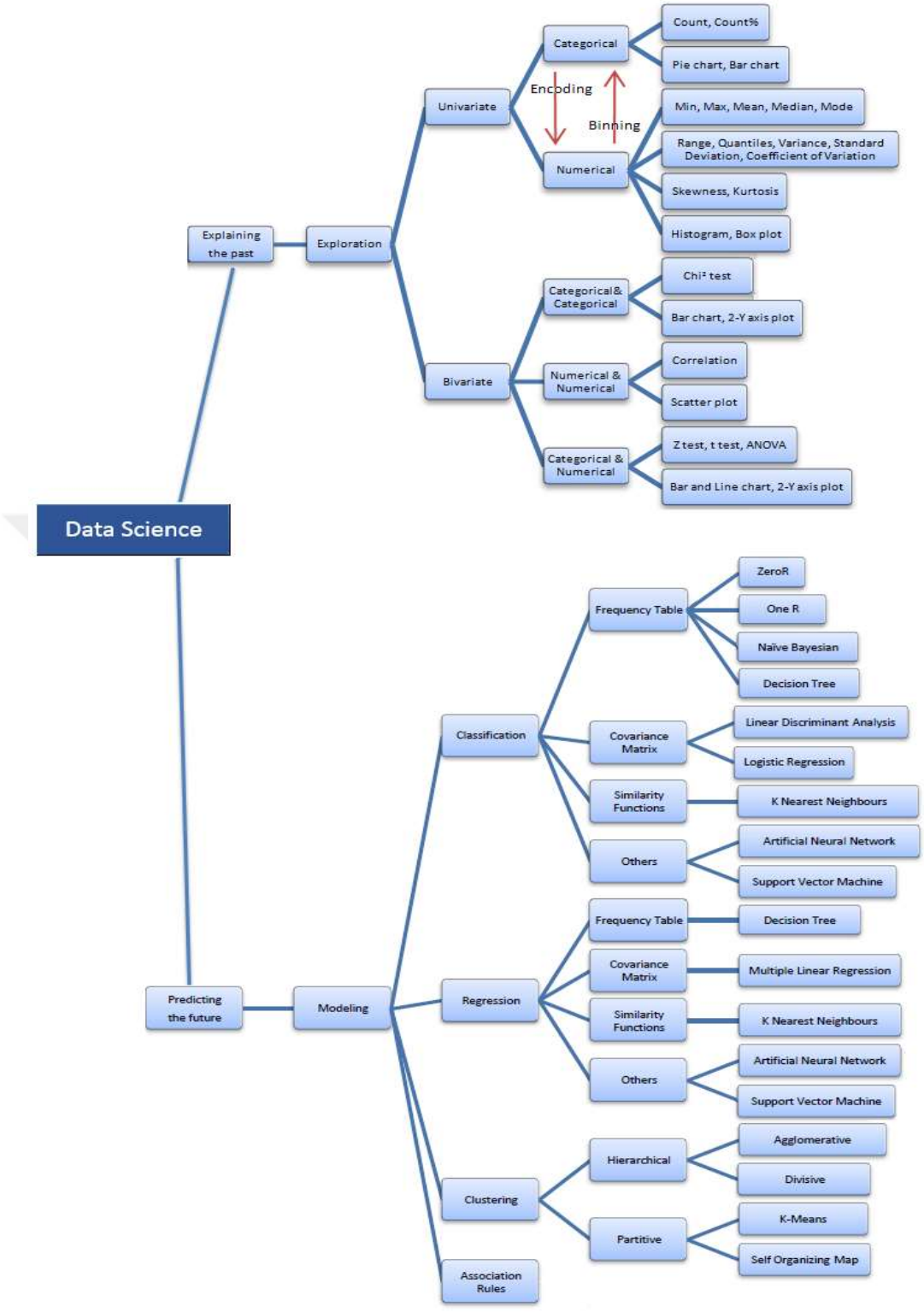


Figure 3-6 Data Science Schema (Sayad, 2018)

3.8 Explaining the Past

Data science clarifies the past through data exploration.

Data Exploration is about explaining the data by means of visualization and statistical techniques. We explore data in order to get important aspects of that data into the center of attraction for further analysis.

In general, we can separate Data Exploration in two main topics,

- Univariate Analysis
- Bivariate Analysis

3.8.1 Univariate Analysis

Univariate analysis discovers factors (attributes) one by one. Factors can be either numerical or categorical There are different visualization and statistical techniques of investigation for each kind of variable. Numerical variables can be converted into categorical similar part by a process called grouping or separation. It is also possible to convert a categorical variable into its numerical similar part by a process named encoding. Finally, suitable handling of missing values is a vital issue in mining data. Numerical and categorical variables are separated into two groups.

3.8.2 Bivariate Analysis

Bivariate analysis may be a synchronous analysis of 2 factors or variables. It explores the construct of relationship between 2 factors or variable, a relationship and also the similarity of this relationship, or whether or not there's a distinction between the 2 variables, and whether or not these variations area unit vital. There are 3 kinds of quantity analysis. These are,

- Numerical & Numerical
- Categorical & Categorical
- Numerical & Categorical

3.9 Predicting the Future

Data science tries to estimate the future by means of modeling.

Modeling

Predictive modeling means the process by which a model is created to estimate a result. We can distinguish the classification as a categorical result, and the regression as process that uses numeric results. Descriptive modeling or clustering we can use as a process for separation of observations into clusters which can be considered as similar for the same clusters. The association rules can help to find meaningful relationships between observations.

3.9.1 Classification

Classification could be a knowledge science issue of attempt to understand the worth of a categorical variable (class or target) by cunstructing a model supported one or additional categorical and/or numerical variables (predictors or attributes).

Tehere are four main groups in classification algorithms:

1. Frequency Table
    - ZeroR
    - OneR
    - Naive Bayesian
    - Decision Tree
2. Covariance Matrix
    - LDA (Linear Discriminant Analysis)
    - Logistic Regression
3. Similarity Functions
    - KNN (K Nearest Neighbors)
4. Others
    - ANN (Artificial Neural Network)
    - SVM (Support Vector Machine)

3.9.2 Regression

Regression could be an information science issue of attempting to estimate the worth of target (numerical variable) by constructing a model supported one or a lot of computer (categorical and numerical variables).

There are four main topics about regression (Sayad, 2018),

1. Frequency Table
    - Decision Tree
2. Covariance Matrix
    - (MLR) Multiple Linear Regression
3. Similarity Function
    - (KNN) K Nearest Neighbors
4. Others
    - (ANN) Artificial Neural Networks
    - (SVM) Support Vector Machine

3.9.3 Clustering

A cluster could be an area of knowledge that are similar. Clustering (also named unsupervised learning) is that the process to separate a dataset into teams specified the members of every cluster have similarity (close) as doable to at least one another, and completely different teams have dissimilarity (far) as doable from each other.

Clustering cannot cowl antecedently unknown relationships in a very dataset. There are a unit tons of applications for cluster analysis. let's say, in business, cluster analysis are often wont to explore and describe client segments for biology and selling functions, it are often used for classification of animals and plants such their options.

There are 2 main types in cluster algorithms:

1. Hierarchical
    - Agglomerative
    - Divisive
2. Partitive
    - K Means
    - (SOM) Self-Organizing Map

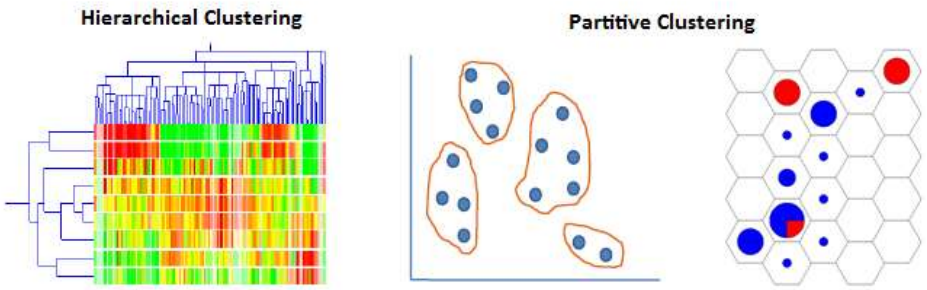


Figure 3-7 Clustering Analysis Groups

Requirements for being a good clustering method are,

- The ability to find some or all of the hidden clusters.
- Ability dealing with various types of attributes.
- Within - cluster similarity and between - cluster dissimilarity.
- Can be dealing with outliers and noise.
- Can be handling high dimensionality.
- Interpretable, usable and scalable.

The essential part of clustering is a way to identify the similarity between two objects. High or low similarity occurs in clusters or between clusters respectively. Generally, the measure for distance such as Minkowski, Manhattan and Euclidean is used to scale similarity or dissimilarity between objects. A function of distance indicates a lower value for pairs of objects that are dissimilar.

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Hierarchical Clustering

"Hierarchical" is a type of clustering that makes a predetermined sorting from high to bottom for clusters. For instance, all folders and files on the hard disk are hierarchically organized. Hierarchical clustering has two types named agglomerative and divisive.

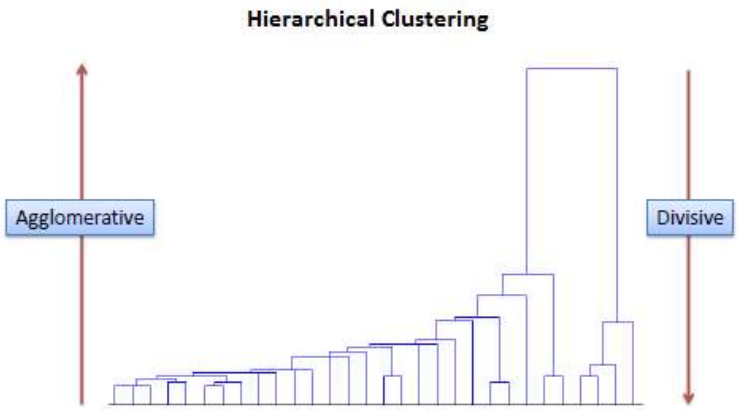


Figure 3-8 Hierarchical Clustering

Divisive method

We assign all of the observations to a single cluster in top-down or divisive clustering method and then make partition the cluster to two least similar clusters. Lastly, until there is one cluster for each observation, we proceed recursively on each cluster. There is proof that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some conditions but is conceptually more complex.

Agglomerative method

We assign each observation to its own cluster in bottom-up or agglomerative clustering method. Afterwards, calculate the similarity (i.e., distance) between all of the clusters and merge the 2 most similar clusters. Lastly, until there is only one cluster left, repeat steps 2 and 3. The related algorithm is shown below.

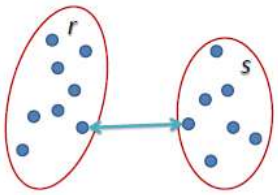
Given:
A set X of objects $\{x_1,\ldots,x_n\}$
A distance function $dist(c_1,c_2)$
for $i = 1$ to n
  $c_i = \{x_i\}$
end for
$C = \{c_1,\ldots,c_n\}$
$l = n+1$
while C.size > 1 **do**
  - (c_{min1},c_{min2}) = minimum $dist(c_i,c_j)$ for all c_i,c_j in C
  - remove c_{min1} and c_{min2} from C
  - add $\{c_{min1},c_{min2}\}$ to C
  - $l = l + 1$

end while

Determination of the the proximity matrix including the distance between each point using a distance function before any clustering is needed. The matrix is updated to show the differences of clusters. The following three methods vary in how the distance between each cluster is measured.

Single Linkage

The distance between two clusters, which in single linkage hierarchical clustering, is defined as the closest interval between two points in each cluster. For instance, the shortest distance between the two points of each cluster ("s" and "r") can be considered as a distance between clusters "s" and "r".

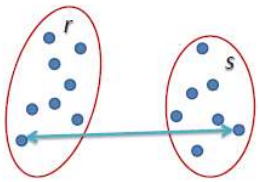


$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Figure 3-9 Single Linkage

Complete Linkage

The distance between two clusters, which in complete linkage hierarchical clustering, is defined as the great interval between two points in each cluster. For example, the distance between two farthest points of the clusters "s" and "r" can be considered as a distance between two mentioned clusters.

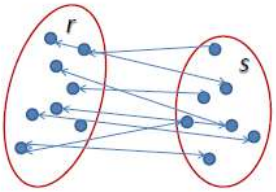


$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Figure 3-10 Complete Linkage

Average Linkage

The interval between two clusters, which in average linkage hierarchical clustering, is described as the average linkage between each point in one cluster to every point of the other cluster. For instance, the distance between clusters “s” and “r” to the left is equal to the medium length each arrow between binding the points of one cluster to the other.

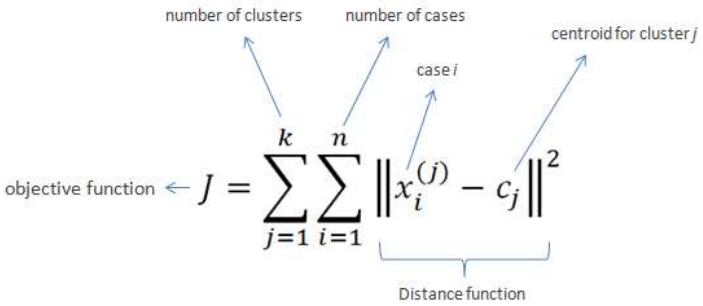


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Figure 3-11 Average Linkage

3.9.3.1 K-Means Clustering

K-Means clustering algorithm aims to split n objects into k clusters in which each object belongs to the cluster with the most closed mean. This method produces accurately k different clusters of biggest possible difference. The best number of clusters k leading to the biggest separation (distance) is not known as a priori and must be calculated from the data. The objective of K-Means clustering is to reduce total intra-cluster variance or the squared error function, which is written below,



Algorithm

1. Clustering the data into k groups where k is defined firstly.
2. Selecting k points at random as cluster middle point.
3. Assigning objects to their most closed cluster middle point according to the function named Euclidean distance.
4. Calculating the mean or centroid of all objects in every cluster.
5. Until the same points are assigned to each cluster in consecutive rounds, repeat steps 2, 3 and 4.

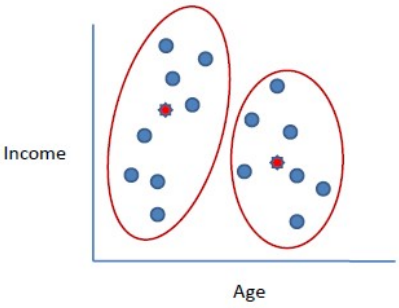


Figure 3-12 K-Means Clustering

K-Means is relatively a genius method but we have to define the number of clusters previously. The results of clustering are responsive to initialization and often completes optimally at a local. Unfortunately, there is no global theoretical method to compute the optimal number of clusters. A practical approach is to compare the results of multiple runs with different k and choose the best one based on a firstly defined criterion. Generally, a large k value probably reduces the errors but enhances the risk (Wang & Bai, 2016).

3.9.3.2 X-Means Clustering

X-Means (Extending K-means) algorithm is a kind of K-Means algorithm, which try to estimate count of cluster by the reputation of updating cluster centers.

- Determines the cluster number itself
- Determine the min/max number of clusters itself. For example, you may be able to create up to 8 minimum 3 clusters from this dataset.
- Distance metrics are self-explanatory. There is a concept called KD-tree.
- Never receives nominal data.

Clustering methods are used to group the data/observations into a few segments so that data within any segment are alike while data across segments are different. Cluster centroids are chosen randomly through a fixed number of K-clusters. The algorithm partitions the given data into K-clusters, each one having its own cluster membership and assigns each data point to the closest centroid. It then computes again the centroid using current cluster association and if the clustering does not converge, the process will be repeated until a specified number of times. X-means clustering is a variation of K-means clustering that treats cluster allocations by repetitively attempting partition and keeping the optimal resultant splits, until some criterion is reached.

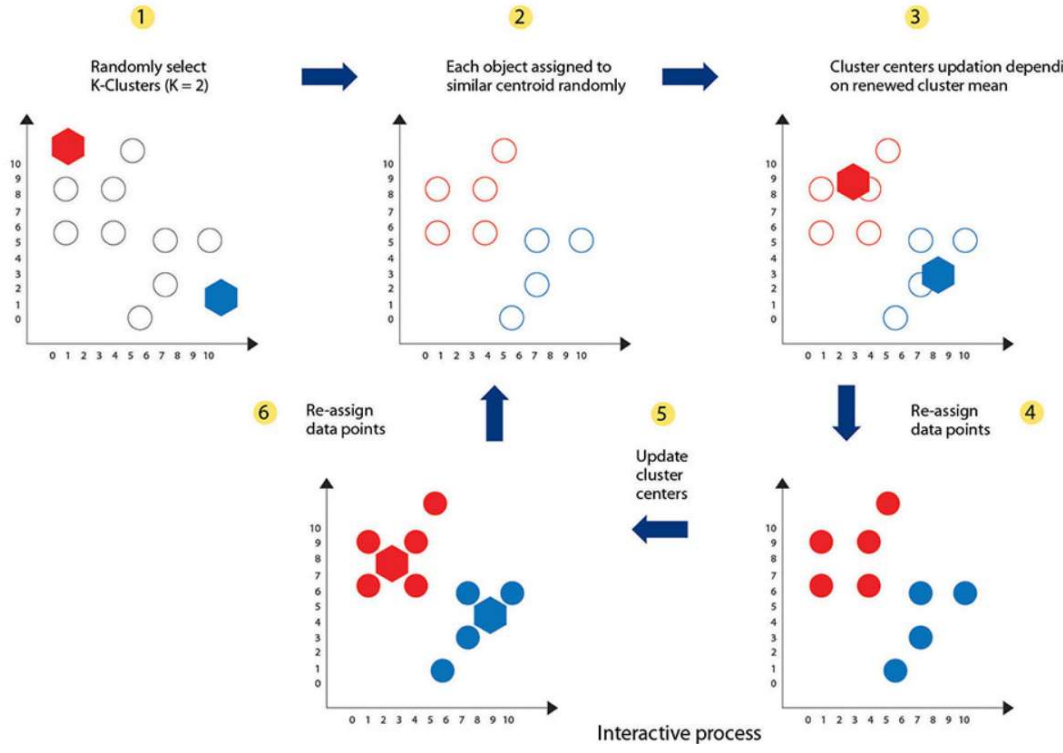
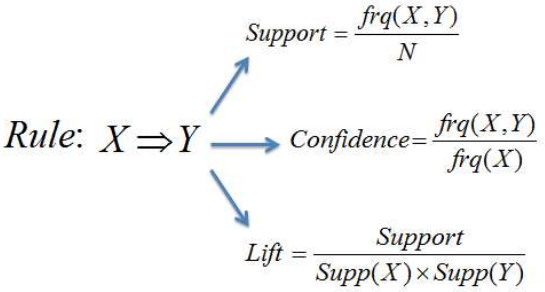


Figure 3-13 X-Means Clustering Algorithm

3.9.4 Association Rules

The Association Rules find all the item sets that use larger sets of items to produce the desired rules with the minimum support and greater trust with the lowest possible confidence. The lift of a rule is the ratio of the monitored support to that expected if Y and X were not depending. Widely used and typical example of association rules application is market basket analysis.



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Figure 3-14 Associative Rules Sample

3.9.4.1 Apriori Algorithm

With the help of scanning the database Several times, there can be found items that associate with the merge, pruning, and minimum support criteria.

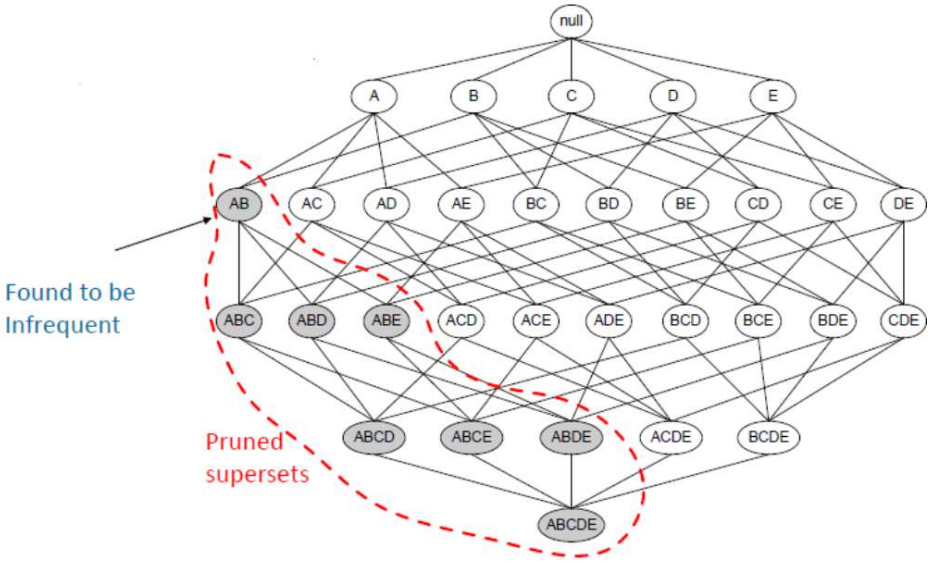


Figure 3-15 Apriori Algorithm (Dogan, 2014)

1. Only the large item sets of the preceding pass without considering the database transactions created candidate item sets.

2. All item sets whose size is higher than one attend the large item set of the preceding pass.

3. The item set that is created and has a subset, which is not large, is erased. Other item sets, which are staying, are the candidate ones.

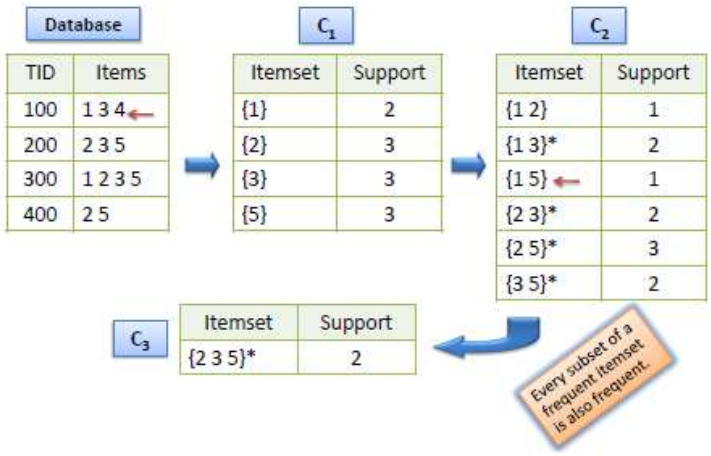


Figure 3-16 Apriori Algorithm

The Apriori algorithm takes advantage of the fact that any sub item set of a frequent item set is also a frequent item set. The algorithm can accordingly, decrees the number of candidates being observed by only exploring the item sets whose support count is bigger than the minimum support count. All sparse item sets can be trim if it has an infrequent subset.

Interests Measurements Used in Association Rules

- **Coverage**

In some resources, it is also referred to as miraculous support or previous support. It measures how often a rule X->Y can be applied on the database.

It is simply equal to the support of the given X value or the probability value.

Coverage (X->Y) = supp (X) = P (X)

It can be calculated as. It is the base for other calculation methods.

- **All-Confidence**

All the rules that can be generated from the item set Z are to have at least one dependency on Z. Confidence

The following formula can be used for this measurement method.

$$Confidence(Z) = \frac{Supp(Z)}{\max(Supp(z \in Z))} = \frac{P(Z)}{\max(P(z \in Z))}$$

The value of max (support (z) of the above form points to the element with the highest support in Z. The value of supp (Z) is the previous unit of measure, which is the scope calculation, is described above.

- **Collective Strength**

The following formula can be used for collective strength measurement.

$$C(Z) = \frac{1 - v(Z)}{1 - E[v(Z)]} * \frac{E[v(Z)]}{v(Z)}$$

The notation v (Z) in the above form specifies the violation value and E [] indicates the expected value.

The Collective Strength value can carry a positive value between zero and infinity. A value of 0 indicates that there is a negative correlation between the elements and a positive relationship if it is infinite. Where the Collective Strength value is used as a measure, the joint force value approaches 1 if the values in the data set are higher in the case that the observed values of the medium and low likelihoods are not included in the Z value and the expected values violate the expected values.

- **Conviction**

It can be calculated as follows,

$$Conviction(X \rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \rightarrow Y)} = \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)}$$

Conviction has been developed as an alternative to measuring value. When the value of Conviction is calculated, the probabilities of occurrence of X elements without Y element are calculated; if the frequencies X and Y are connected, then the likelihood of X being independent of Y is calculated.

In this sense, the value of Conviction is similar to the value of interest (lift will be explained below). However, the value of interest is a directional measure and can not be used to evaluate sequential information.

- **Leverage**

The leverage method measures the difference between the co-occurrences of Y and X, and calculates the statistical relation between Y and X values.

$$\text{Leverage}(X \rightarrow Y) = P(X \text{ ve } Y) - (P(X)P(Y))$$

It can be calculated as.

In order to make it easier to understand, we can think to sell Y and X products together on a sales data to understand how much Y and X are sold separately.

- **Interest**

It is also used in some sources (lift) to understand how much they go together if Y and X are independent statistically.

$$\text{lift}(X \to Y) = \text{lift}(Y \to X) = \frac{conf(X \to Y)}{supp(Y)} = \frac{cons(Y \to X)}{supp(X)} = \frac{P(Y \wedge X)}{P(X)P(Y)}$$

Note that the above formula is similar to the previous leverage value. The only difference is that the splitting operation is used here when the leverage is measured. That is, the value of interest produces a proportional result.

4 APPLICATION

Once we can see the entity relationship diagram for our database schema below.

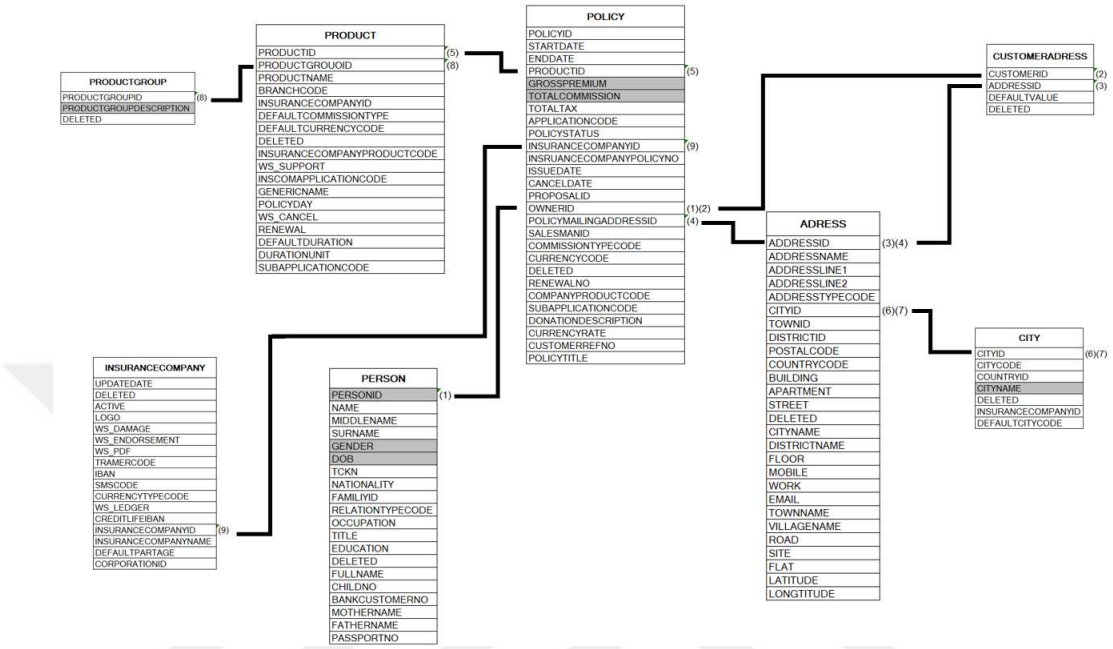


Figure 4-1 Entity Relationship Diagram in the insurance broker

For an insurance broker we also add some attributes to the sample characteristics. We are going to use the following characteristics shown the table 2.2 in chapter-2 as bold in our study. They are,

- Customer Lifetime Period
- Repeat Sales
- Insurance Product
- CRVC (Commission Revenue Value per Customer)

4.1 Data Preparation

In order to obtain the data to be used in this study, an SQL query was written and the following data have been extracted. There are mainly seven headings,

- Age
- Gender
- City
- Commission Revenue Value per Customer (CRVC)

- Repeat Sales
- Customer Lifetime Period
- Insurance Product Group

	B	C	D	E	F	G	H	I	J	K	L	M	Q	S	V	W	X	Y	Z	AA
1	PERSON ID	CUSTOMER NAME	BUILDING INSURANCE	CONTENTS INSURANCE	PERSONAL ACCIDENT INSURANCE	UNEMPLOYMENT INSURANCE	FULL COVERAGE CAR INSURANCE	HOME INSURANCE	HOME AND CONTENTS INSURANCE	HEALTH INSURANCE	TRAVEL INSURANCE	LIABILITY CAR INSURANCE	REPEAT SALES	AGE	*CRVC($)	CITY (NUMERICAL)	GENDER (NUMERICAL)	GROSS PREMIUM($)	TOTAL COMMIS SION($)	CUSTOMER LIFETIME PERIOD
2	1797370	ABABAIKERI XIAOKAITI	1	0	0	0	0	0	0	0	0	0	1		5,48	1	0	28,91	3,47	0,33
3	636795	ABAMÜSLİM AKBULUT	0	1	0	0	0	0	0	0	0	0	1	33	10,95	0	-1	11,11	7,04	0,25
4	600954	ABAMÜSLİM ERGİN	0	1	0	0	0	0	0	0	0	0	1	59	10,95	1	-1	11,11	7,04	0,25
5	2412237	ABAYDİN YÜKSEL	1	0	0	0	0	0	0	0	0	0	1	50	5,48	-1	-1	11,38	1,93	1,00
6	112290	ABBAS ATABAY	0	2	2	0	0	0	0	0	0	0	4	60	54,17	-1	-1	69,29	16,53	0,67
7	359450	ABBAS AYDIN	0	1	0	0	0	0	0	0	0	0	1	62	10,95	-1	-1	13,64	8,64	0,20
8	2547	ABBAS BEKTAŞ	0	1	0	0	0	0	0	0	0	0	1	37	10,95	-1	-1	62,53	30,48	0,17
9	544516	ABBAS CEVİZ	0	1	0	0	0	0	0	0	0	0	1	61	10,95	0	-1	13,64	8,64	0,20
10	2210190	ABBAS CEYLAN	0	0	0	0	2	0	0	0	0	2	4	70	219,40	-1	-1	847,99	105,4	3,00
11	721805	ABBAS ÇAĞIŞLAR	0	0	0	0	2	0	0	0	0	2	4	37	219,40	1	-1	1479,8	210,9	1,17
12	943966	ABBAS ÇARBAKA	1	0	0	0	0	0	0	0	0	0	1	82	5,48	1	-1	45,66	4,57	0,33

Figure 4-2 Data in the insurance broker

*CRVC is the Commission Revenue Value per Customer.

We have done some trimming on the data fetched by the SQL.

Firstly, there are some 0 (zero) values within the total commission column. We have trimmed these values from the data. This is because an insurance policy sold to a customer on a given day was cancelled on the same day and the customer took refund of his/her money.

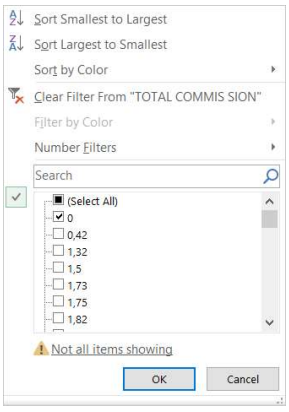


Figure 4-3 Commission data trimming

Secondly, we have some data showing values that are less than the gross premium of 3.00 Dollars. We have deleted the data since they are not significant for the company.

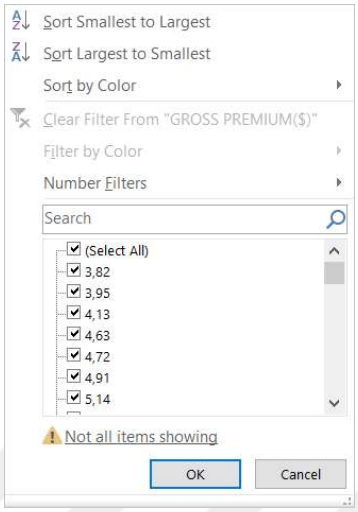


Figure 4-4 Gross Premium data trimming

Finally, once the dataset was summarized statistically, we were able to come up with the following table. In this table, it is seen that the median of total commission is 7.04 Dollars and the standard deviation is calculated to be 51.08 Dollars. In addition, the average of the customers is approximately 47.82. The average gross premium is 95.66 Dollars and the average total commission is 16.61 Dollars.

Table 4-1 Statistical Details

Statistical Descriptions	Value
Median for Total Commission	7.04 Dollars
Standard Deviation for Total Commission	51.08 Dollars
Average Age	47.82 Years Old
Average Gross Premium	95.66 Dollars
Average Total Commission	16.61 Dollars

Since it is not possible to use categorical data directly in our application, some of the columns are used directly and some of them after converting from categorical data to numeric data. There are 3 columns which need to be converted namely city, gender, and total commission (CRVC) columns.

4.1.1 Age

When examine the data by age group of the customers, It can be seen the following results. Our customer age group, which we sells the best percentage, are 26-60 age groups. We will use customers' age info directly in our application. Because it is a numerical field.

36-40 age group is the first 14.83 percentage. 31-35 age group is the second with 13.64 percentage. 41-45 age group is the third with 11.99 percentage.

Table 4-2 Age Details

Age Group	Total Commission ($)	Percent
0 - 20	1756.50	0.54
21 - 25	57136.40	4.41
26 - 30	125603.09	9.70
31 - 35	176621.89	13.64
36 - 40	191954.69	14.83
41 - 45	155206.68	11.99
46 - 50	133261.78	10.29
51 - 55	129388.49	9.99
56 - 60	122900.12	9.49
61 - 65	83173.90	6.42
66 - 70	54614.40	4.22
71 - 75	31420.67	2.43
76 - 80	17471.23	1.35
81 - 105	8906.45	0.69

When we draw a chart for age and total commission, we can see the figure below,

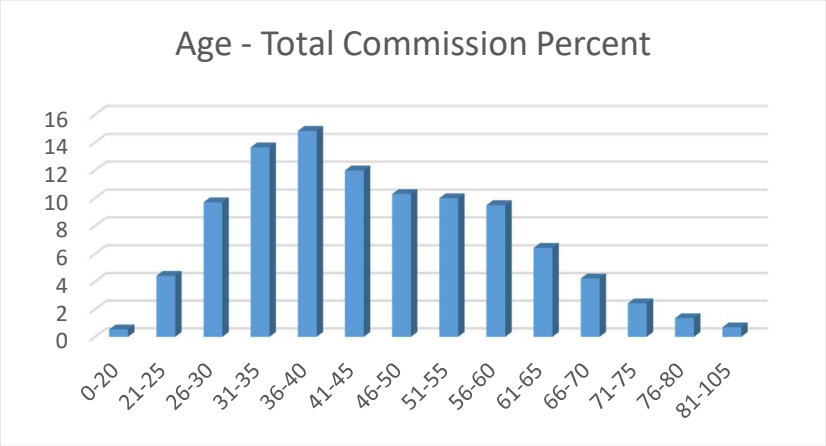


Figure 4-5 Age Details Chart

4.1.2 Gender

We cannot use gender information directly in our application. Because it is a categorical value. We converted gender data the table below,

Table 4-3 Gender Converting to Numerical

Gender	Gender (Numerical)
Male	-1
Female	1

Once we summarized the dataset by the personal characteristics, i.e. gender column, we were able to come up with the following table. It is seen that 72.5 % of the customers are male, and the rest are female.

Table 4-4 Gender characteristics

	Male	Female
Count	56391	21386
Percentage	72,503	27,496

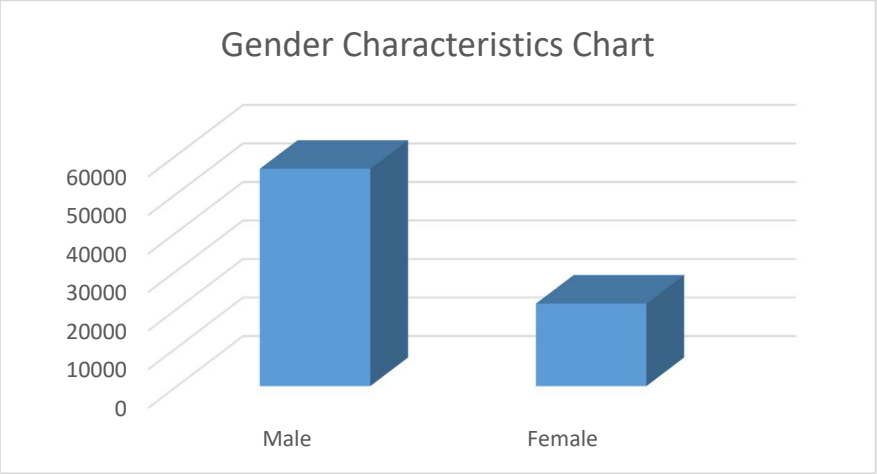


Figure 4-6 Gender Characteristics Chart

4.1.3 City

When we have summarized the dataset with the city column, we can see the following table. We can see the 49.87 percentage of our sales is in the Istanbul city. Other Big Cities (Ankara, Izmir, Bursa, and Kocaeli) follow Istanbul with the percentage of 16.79.

City data is converted to Numerical City Values column like below. Because, city data fetched in SQL as a categorical data. Therefore, we converted city data to a numerical data.

Table 4-5 Cities Details

Cities	Total Commission ($)	Percent	Numerical City Values
Istanbul	645667	49.87	1
Other Big Cities	217404	16.79	0
The Others	431615	33.34	-1

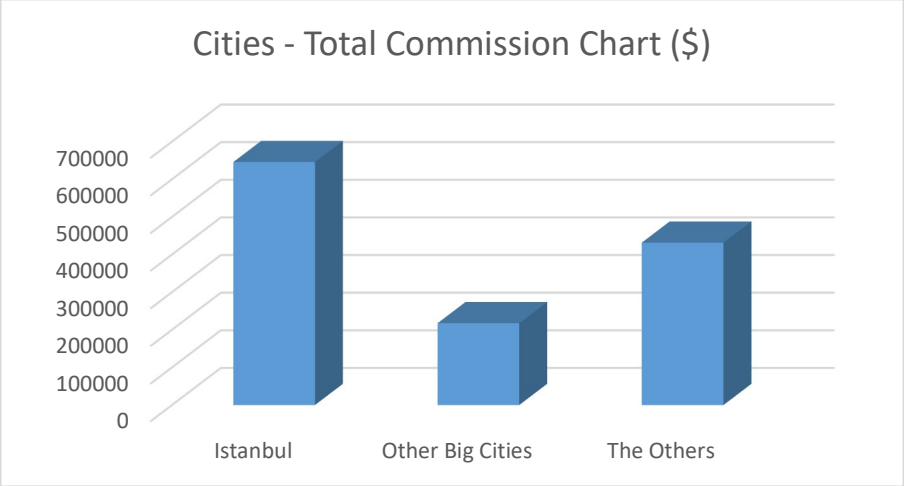


Figure 4-7 Cities Details Chart

4.1.4 Repeat Sales

Repeat sales is the sales amount that the customer buys a product more than once. We used repeat sales directly in our application because it is a numerical value.

Once we summarized the dataset by the Repeat Sales, we were able to come up with the following table. Average Repeat Sales a product is about 1.286.

Table 4-6 Personal characteristics summary (repeat sales)

Characteristics	Average
Repeat Sales for a product	1,286

4.1.5 Customer Lifetime Period

Customer Lifetime Period is the period extending from the record date of the customer up to the present in the insurance broker.

Once we summarized the dataset by the Customer Lifetime Period, we were able to come up with the following table. Average Customer Lifetime Period is about 0.58.

Table 4-7 Personal characteristics summary (customer)

Characteristics	Average
Customer Lifetime Period	0,58

Table 4-8 Customer Lifetime Value Calculation Table

Years	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018(now)
Multiplier	1/10	1/9	1/8	1/7	1/6	1/5	1/4	1/3	1/2	1/1
Customer X	1 product	2 product	No sales	No sales	1 product	No sales	No sales	3 product	No sales	No sales
Customer Lifetime Value (Sample)	1x(1/10)+	2x(1/9)+	0x(1/8)+	0x(1/7)+	1x(1/6)+	0x(1/5)+	0x(1/4)+	3x(1/3)+	0x(1/2)+	0x(1/1)

***Customer Lifetime Value = 1x(1/10)+ 2x(1/9)+ 0x(1/8)+ 0x(1/7)+ 1x(1/6)+ 0x(1/5)+ 0x(1/4)+ 3x(1/3)+ 0x(1/2)+ 0x(1/1)**

There is a sample Customer Lifetime Value calculation for a customer above.

We summarized the dataset by the customer lifetime period; we can see the following table. We can understand from the table that we are expanding our customer base. However, we can lose some customers after 4 years. Therefore, we should take some precautions not to lose the customers.

Table 4-9 Customer Lifetime Period Detail

Customer Lifetime Period	Total Commission ($)	Percent
0 – 0,1	562	0,04
0,11 – 0,2	233093	18,00
0,21 – 0,3	121564	9,39
0,31 – 0,4	114767	8,86
0,41 – 0,5	197573	15,26
0,51 – 0,9	98241	7,59
0,91 – 1	199509	15,41
1,01 – 1,5	68556	5,30
1,51 – 3	146893	11,35
3,01 – 5	61215	4,73
5,01 – 7,5	21351	1,65
7,51 – 10	15803	1,22
10 – 32	15560	1,20

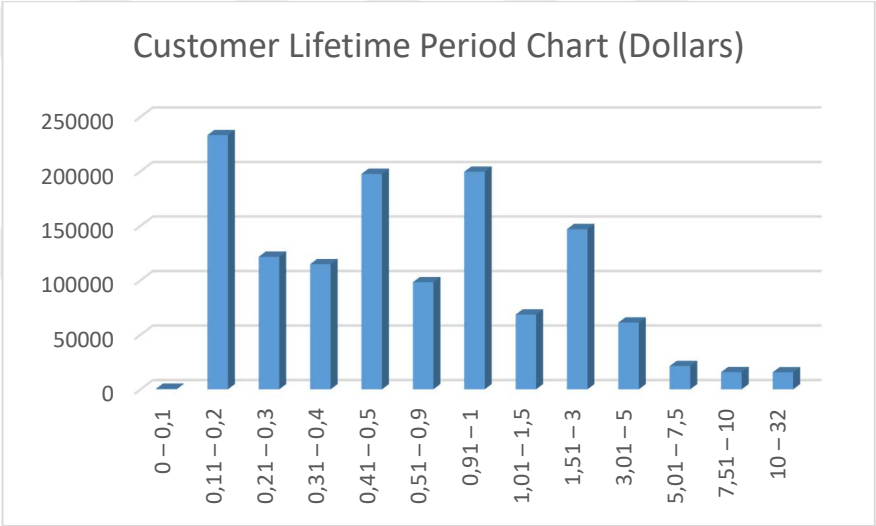


Figure 4-8 Customer Lifetime Period Detail Chart

4.1.6 Insurance Product Group

After then, we have summarized the dataset by the product group we can see the following results. A big percent of our company's revenue comes from Full Coverage Car Insurance (Kasko), Liability Car Insurance (Trafik), Contents Insurance (Eşya), Building Insurance (Dask) policies with the percent of 27.93, 25.22, 15.48 and 12.36. Personal Accident Insurance (Ferdi Kaza) and Home and Contents Insurance (Ev ve Eşya) are following them with the percentage of 4.94 and 4.75.

Then, when we summarized the dataset by the policy count, we can see the following table. We can see the best-selling insurance product as a Building Insurance with the percentage of 41.52. Then Contents Insurance with the percentage of 26.07. Liability Car Insurance and Full Coverage Car Insurance follow the Building Insurance and Contents Insurance with the percentage of 11.70 and 7.37.

Table 4-10 The insurance broker's revenue summary

Product Group	Gross Premium ($)	Total Comission ($)	Policy Count	Gross Premium Percent	Total Comission Percent	Policy Count Percent	Gross Premium Per Policy($)	*CRVP ($)
Building Insurance	1123159,82	159963,18	41453	15,05	12,36	41,52	40,60	**5,49**
Contents Insurance	460519,97	200416,71	26025	6,17	15,48	26,07	26,51	**10,95**
Personal Accident Insurance	345189,33	63972,65	5647	4,63	4,94	5,66	91,59	**16,11**
Unemployment Insurance	33072,29	19344,01	1306	0,44	1,49	1,31	37,94	**21,07**
Full Coverage Car Insurance	2425343,72	361607,87	7355	32,51	27,93	7,37	494,07	**69,92**
Home Insurance	186575,11	31487,78	809	2,50	2,43	0,81	345,54	**55,36**
Home and Contents Insurance	218533,47	61494,60	2658	2,93	4,75	2,66	123,18	**32,90**
Health Insurance	190082,82	44904,40	1273	2,55	3,47	1,27	223,72	**50,17**
Travel Insurance	153409,34	24927,82	1638	2,06	1,93	1,64	140,32	**21,64**
Liability Car Insurance	2324564,71	326566,78	11682	31,16	25,22	11,70	298,14	**39,76**

*CRVP is the Commission Revenue Value per Product.

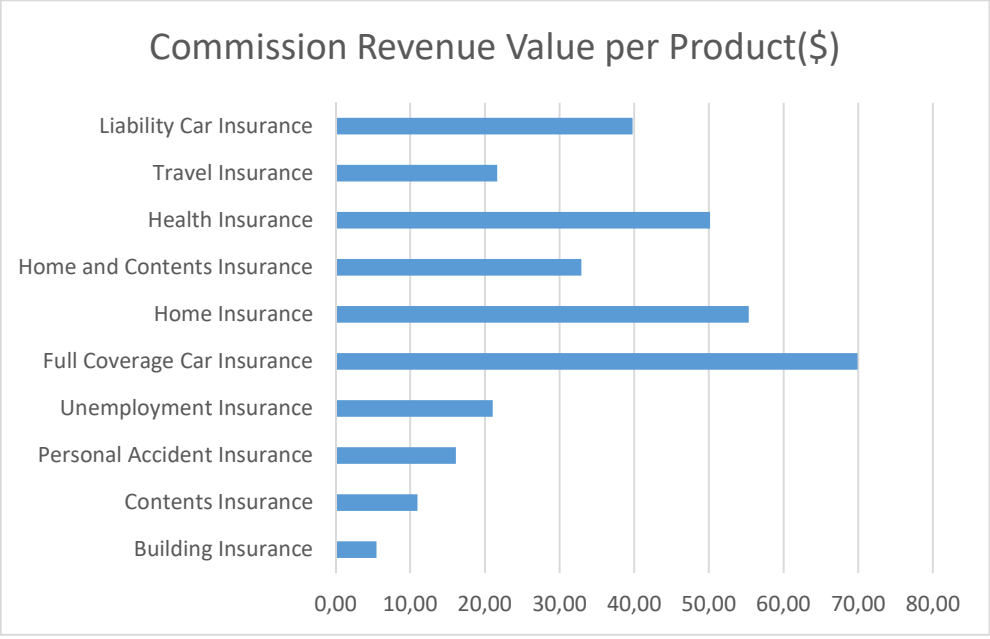


Figure 4-9 The insurance broker's revenue summary Chart

4.1.7 Commission Revenue Value per Customer (CRVC)

CUSTOMER ID	BUILDING INSURANCE	CONTENTS INSURANCE	PERSONAL ACCIDENT INSURANCE	UNEMPLOYMENT INSURANCE	FULL COVERAGE CAR INSURANCE	HOME INSURANCE	HOME AND CONTENTS INSURANCE	HEALTH INSURANCE	TRAVEL INSURANCE	LIABILITY CAR INSURANCE	CRVC	FORMULA FOR CRVC
90	1	2	0	0	0	0	0	0	0	0	27,39	5,49x1 + 10,95x2
91	1	0	0	0	0	0	0	0	0	0	5,49	5,49x1
92	0	1	0	0	0	0	0	0	0	0	10,95	10,95x1

Figure 4-10 The Commission Revenue Value per Customer (CRVC)

$$(CRVC) \quad U_j = \sum_{i=1}^{\forall i} \sum_{j=1}^{\forall j} c_i x_{ij} \qquad [4.1]$$

$$(Product\ Group)\ \ i\ = 1\ to\ 10$$

$$(\text{Customer})\ \ j = 1\ to\ {\sim}75000$$

$$(\text{Commission Revenue Value per Product})\ \ c_i = CRVP$$

There is a revenue on the table 4-9 above for per a product.

The factor value is calculated according the table 4-9 and figure 4-10. In order to calculate factor value, first we are calculating the revenue for selling one policy for the insurance type. For example, for selling a building insurance policy our company can earn 5.49 Dollars. In addition, revenue per contents insurance is 10.95 Dollars (Shown on Figure 4-10). So when a customer takes 1 building insurance policy and 2 contents insurance policy the factor value for that customer is 5,49x1 + 10,95x2 = 27,39.

4.2 Xmeans Analysis

For clustering analysis, we will use numerical columns data, which is prepared. We will use;

- Repeat Sales
- Age
- Customer Lifetime Period
- CRVC (Commission Revenue Value per Product)
- City (Numerical)
- Gender (Numerical)

CUSTOMER ID	REPEAT SALES	AGE	*CRVC($)	CITY (NUMERICAL)	GENDER (NUMERICAL)	CUSTOMER LIFETIME PERIOD
2	1	33	10,95	0	-1	0,25
3	1	59	10,95	1	-1	0,25
4	1	50	5,49	-1	-1	1,00
5	4	60	54,13	-1	-1	0,67
6	1	62	10,95	-1	-1	0,20
7	1	37	10,95	-1	-1	0,17
8	1	61	10,95	0	-1	0,20

Figure 4-11 Data format is used in clustering analysis

Our data pattern is like below.

```
@RELATION insurance

@ATTRIBUTE REPEAT_SALES                 REAL
@ATTRIBUTE AGE                          REAL
@ATTRIBUTE CRVC                         REAL
@ATTRIBUTE CITY_NUMERICAL               REAL
@ATTRIBUTE GENDER_NUMERICAL             REAL
@ATTRIBUTE CUSTOMER_LIFETIME_PERIOD     REAL

@DATA
1,33,10.95,0,-1,0.25
1,59,10.95,1,-1,0.25
1,50,5.49,-1,-1,1.00
4,60,54.13,-1,-1,0.67
1,62,10.95,-1,-1,0.20
1,37,10.95,-1,-1,0.17
1,61,10.95,0,-1,0.20
```

Figure 4-12 Our data pattern is used in clustering analysis

Table 4-11 Weka Xmeans percentage results

	Data Count	Percentage
Cluster 1	10707	14%
Cluster 2	32878	42%
Cluster 3	10679	14%
Cluster 4	23513	30%

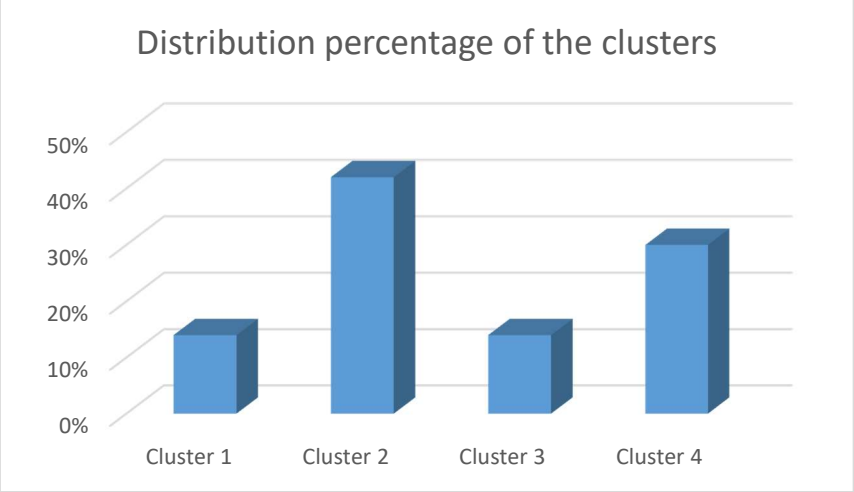


Figure 4-13 Distribution Percentage of the clusters

Table 4-12 Weka Xmeans results

	Repeat Sales	Age	CRVC	City	Gender	Customer Lifetime Period
Cluster 1	1.28	49.16	20.86	1.00	1.00	0.57
Cluster 2	1.35	47.18	26.07	0.75	-1.00	0.59
Cluster 3	1.18	48.35	21.20	-0.71	1.00	0.53
Cluster 4	1.24	47.87	22.73	-1.00	-1.00	0.58

Table 4-13 Weka Xmeans percentage table

	CRVC Percent	Cluster Percent
Cluster 1	22,95%	14%
Cluster 2	28,69%	42%
Cluster 3	23,33%	14%
Cluster 4	25,01%	30%

When we create a chart, we can see the following,

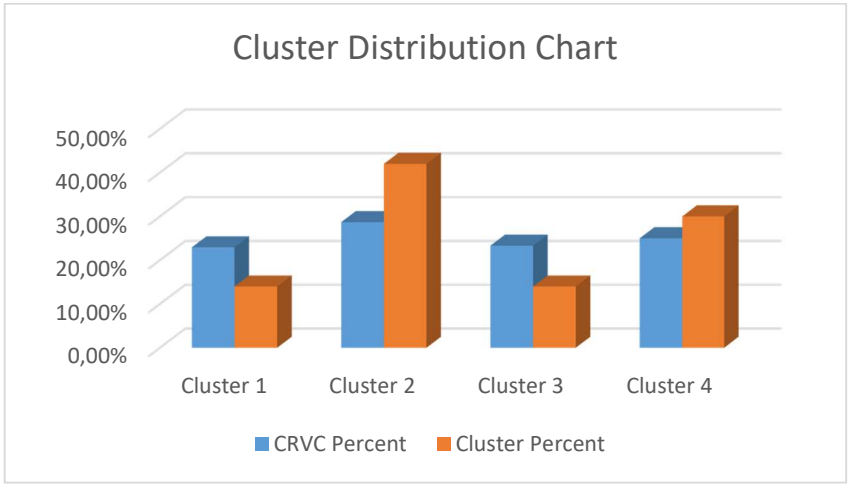


Figure 4-14 Clusters Distribution Chart

After we create a trends chart we can see the following,

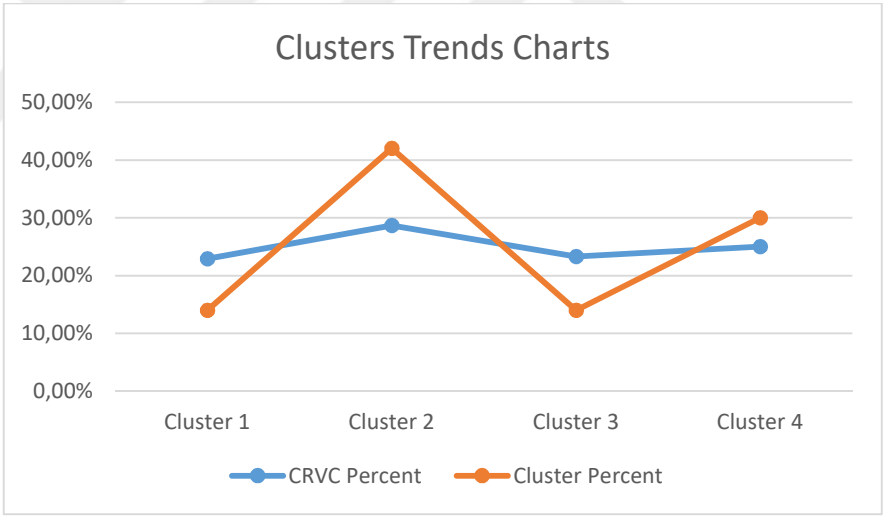


Figure 4-15 Clusters Trends Charts

As we can see on the table 4-10 and table 4-11, there are three cluster.

First one is named Cluster 1 has 10707 customer data and it includes 14 percent of the whole data. Average Repeat Sales of this cluster is 1.28 and average age of this cluster is 49.16. On the other hand, the Customer Lifetime Period of this cluster is 0.57 and gender of the cluster is female. This cluster is located in Istanbul with the value of 1.00. The CRVC of this cluster is 20.86 Dollars.

The second is named Cluster 2 has 32878 customers data. It includes 42 percent of whole data implemented. The average Repeat Sales for this cluster is 1.35 and the average age of this cluster is 47.18. The average Customer Lifetime Period of this cluster for our country is 0.59 and city is 0.75 so the customer mostly lives in Istanbul. Completely the customer in this cluster is male. The CRVC of this cluster is high than the others with 26.07 Dollars.

The third cluster is named Cluster 3 has 10679 customers data and it includes 14 percent of whole data implemented. The average repeat sales and age of this cluster are 1.18 and 48.35. The average Customer Lifetime Period is lower than the others with 0.53. Customers in this cluster are also living far away from Istanbul and the other big cities with the numbering -0.71. Gender of the cluster is 1.00 means all of this cluster have female customers. The CRVC of this cluster is 21.20 Dollars.

The fourth cluster is named Cluster 4 has 23513 customers data and it includes 30 percent of whole data implemented. The average repeat sales and age of this cluster are 1.24 and 47.87. The average Customer Lifetime Period is 0.58. Customers in this cluster are also living the other cities with the numbering -1.00. Gender of the cluster is -1.00 means all of this cluster have male customers. The CRVC of this cluster is 22.73 Dollars.

Consequently, when we consider these 4 cluster, we can see that the most important cluster for our company is Cluster 2, which has 32878 customers' data with the 42 percent of whole data and male. Because its CRVC value is maximum with 26.07 Dollars. The other important clusters are Cluster 4, Cluster 3 and Cluster 1 orderly. They are include 23513 customers, 10679 customers and 10707 customers with the 30 percentage, 14 percentage and 15 percentage of the whole data. The CRVC value of these clusters are 22.73, 21.20 and 20.86.

In this title, Xmeans algorithm shows us Cluster 2 is the most important instance. In addition, the customers in this instance more valuable from the other ones. When we sort the other instances in this algorithm, we can see cluster 4, cluster 3 and cluster 1.

4.3 Association Study

Analyzing the patterns of past transactions in the dataset, making predictions about the future by making use of this information, and extracting the associations of each of the obscenities in the dataset with each other.

For association rules analysis we will use categorical columns data, which is prepared. We will use only the insurance type's columns that is because we want to know hidden patterns in insurance types. Therefore, our data will be as shown below.

- Building Insurance
- Contents Insurance
- Personal Accident Insurance
- Unemployment Insurance
- Full Coverage Car Insurance
- Home Insurance
- Home And Contents Insurance
- Health Insurance
- Travel Insurance
- Liability Car Insurance

CUSTOMER ID	BUILDING INSURANCE	CONTENTS INSURANCE	PERSONAL ACCIDENT INSURANCE	UNEMPLOYMENT INSURANCE	FULL COVERAGE CAR INSURANCE	HOME INSURANCE	HOME AND CONTENTS INSURANCE	HEALTH INSURANCE	TRAVEL INSURANCE	LIABILITY CAR INSURANCE
1	F	T	T	F	F	F	F	F	F	T
2	T	F	F	F	T	F	T	F	F	T
3	F	F	T	F	T	F	T	F	F	T
4	F	F	T	F	T	F	F	F	F	T
5	F	T	T	F	F	F	T	F	F	F
6	F	T	F	F	T	F	F	F	F	T
7	T	F	T	F	T	F	F	F	F	F
8	F	T	T	F	F	F	T	F	F	F

Figure 4-16 Weka association rules data

Our data pattern is like below,

```
@RELATION insurance

@attribute 'BUILDING INSURANCE' { 'T', 'F'}
@attribute 'CONTENTS INSURANCE' { 'T', 'F'}
@attribute 'PERSONAL ACCIDENT INSURANCE' { 'T', 'F'}
@attribute 'UNEMPLOYMENT INSURANCE' { 'T', 'F'}
@attribute 'FULL COVERAGE CAR INSURANCE' { 'T', 'F'}
@attribute 'HOME INSURANCE' { 'T', 'F'}
@attribute 'HOME AND CONTENTS INSURANCE' { 'T', 'F'}
@attribute 'HEALTH INSURANCE' { 'T', 'F'}
@attribute 'TRAVEL INSURANCE' { 'T', 'F'}
@attribute 'LIABILITY CAR INSURANCE' { 'T', 'F'}

@data
F,T,T,F,F,F,F,F,F,T
T,F,F,F,T,F,T,F,F,T
F,F,T,F,T,F,T,F,F,T
F,F,T,F,T,F,F,F,F,T
F,T,T,F,F,F,T,F,F,F
F,T,F,F,T,F,F,F,F,T
T,F,T,F,T,F,F,F,F,F
F,T,T,F,F,F,T,F,F,F
```

Figure 4-17 Data pattern for association rules

In this study, Apriori algorithm, which is widely used as an association rules to detect hidden relational patterns in data, is used.

After uploading the data file to the Weka, we can see some visual charts like below. On the Figure 4-15 are seeing true or false count of Building Insurance. On the Figure 4-16, we can see true or false counts of all insurance types.

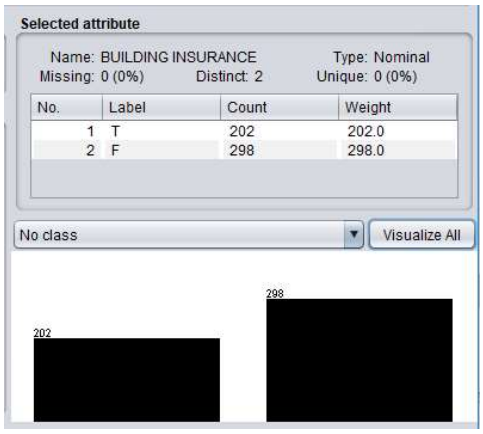


Figure 4-18 Building Insurance count in the Weka visualization

In the Figure 4-16, we can see the count of sales for each insurance product like below,

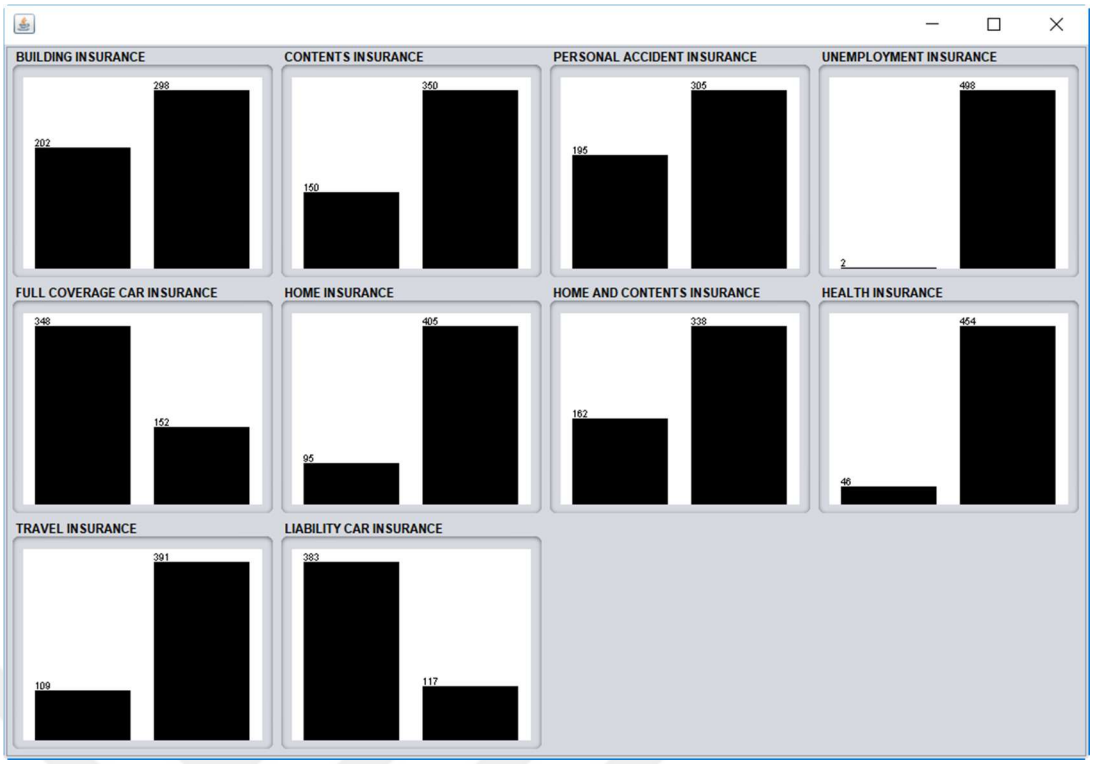


Figure 4-19 All insurance types in the Weka visualization

After uploading data on the Weka, we can choose number of rules. By default there is 10 written as you see below. If you want more algorithm, you can enter 20 or more rules. We will change it as 20 because we want to see best 20 rules in this study.

After choosing number of rules and clicking the start button, we can obtain the following results on the table 4-12.

Table 4-14 Apriori application results

RULE	PRODUCT GROUP	PRODUCT SOLD	PRODUCT GROUP	PRODUCT SOLD	COUNT		PRODUCT GROUP	PRODUCT SOLD	PRODUCT GROUP	PRODUCT SOLD	COUNT	conf.	lift	lev	conv
1	HEALTH INSURANCE	F			454	=>	UNEMPLOYMENT INSURANCE	F			453	1	1	0	0.91
2	HOME INSURANCE	F			405	=>	UNEMPLOYMENT INSURANCE	F			404	1	1	0	0.81
3	TRAVEL ISURANCE	F			391	=>	UNEMPLOYMENT INSURANCE	F			390	1	1	0	0.78
4	HOME INSURANCE	F	HEALTH INSURANCE	F	368	=>	UNEMPLOYMENT INSURANCE	F			367	1	1	0	0.74
5	HEALTH INSURANCE	F	TRAVEL INSURANCE	F	362	=>	UNEMPLOYMENT INSURANCE	F			361	1	1	0	0.72
6	HEALTH INSURANCE	F	LIABILITY CAR INSURANCE	F	354	=>	UNEMPLOYMENT INSURANCE	F			353	1	1	0	0.71
7	CONTENTS INSURANCE	F			350	=>	UNEMPLOYMENT INSURANCE	F			349	1	1	0	0.70
8	FULL COVERAGE CAR INSURANCE	T			348	=>	UNEMPLOYMENT INSURANCE	F			347	1	1	0	0.70
9	HOME AND CONTENTS INSURANCE	F			338	=>	UNEMPLOYMENT INSURANCE	F			337	1	1	0	0.68
10	HOME INSURANCE	F	LIABILITY CAR INSURANCE	T	331	=>	UNEMPLOYMENT INSURANCE	F			330	1	1	0	0.66
11	FULL COVERAGE CAR INSURANCE	T	LIABILITY CAR INSURANCE	T	330	=>	UNEMPLOYMENT INSURANCE	F			329	1	1	0	0.66
12	LIABILITY CAR INSURANCE	T			383	=>	UNEMPLOYMENT INSURANCE	F			381	0.99	1	-0	0.51
13	FULL COVERAGE CAR INSURANCE	T			348	=>	LIABILITY CAR INSURANCE	T			330	0.95	1.24	0.13	4.29
14	UNEMPLOYMENT INSURANCE	F	FULL COVERAGE CAR INSURANCE	T	347	=>	LIABILITY CAR INSURANCE	T			329	0.95	1.24	0.13	4.27
15	FULL COVERAGE CAR INSURANCE	T	UNEMPLOYMENT INSURANCE	F	348	=>	LIABILITY CAR INSURANCE	T			329	0.95	1.24	0.13	4.14
16	TRAVEL ISURANCE	F			381	=>	HEALTH INSURANCE	F			353	0.93	1.02	0.01	1.21
17	TRAVEL ISURANCE	F			391	=>	HEALTH INSURANCE	F			362	0.93	1.02	0.01	1.20
18	UNEMPLOYMENT INSURANCE	F	TRAVEL INSURANCE	F	390	=>	HEALTH INSURANCE	F			361	0.93	1.02	0.01	1.20
19	LIABILITY CAR INSURANCE	T			383	=>	HEALTH INSURANCE	F			354	0.92	1.02	0.01	1.17
20	TRAVEL ISURANCE	F			391	=>	UNEMPLOYMENT INSURANCE	F	HEALTH INSURANCE	F	361	0.92	1.02	0.01	1.19

When we examine the output of this calculation,

On the 1st, 3rd, 5th, 6th, 16th, and 18th rules we can see that when health insurance false means did not sold unemployment insurance is not sold also. It means you cannot sell the unemployment insurance product to the people whom do not purchase the health product. In the same way, vice versa of this situation is true. You cannot sell a health product to someone who do not buy any unemployment insurance.

From the 2nd, 4th, and 10th rules, we can recognize that if the home insurance product is false, the unemployment insurance product also is false. It means you cannot sell an unemployment insurance product to the customer who do not buy the home insurance product. You cannot make a cross selling campaigns between these two product.

On the 3rd, 5th, and 20th rules we can see that you cannot sell unemployment insurance to the customer who travel insurance is not sold. You cannot make a cross selling campaigns between these two product.

On the 6th, 10th, 11th, 12th, and 14th rules, we can see that when the liability car insurance is false, unemployment insurance is false. It means you cannot sell unemployment insurance product to the customers who have not liability car insurance product. There are no relation between the two products.

On the 7th rule, we can see that when the contents insurance product is false, the unemployment insurance product also is false. It means they are following together on the negative direction. Therefore, you cannot make a cross sell campaigns between them.

On the 8th, 11th, and 15th, we can see also that the full coverage car insurance product and unemployment insurance product are following together in the negative direction. Therefore, we cannot make a cross sell campaigns between them.

On the 9th rule, we can see that home and contents insurance product and unemployment insurance product are following together in negative direction. If someone do not take home and contents insurance product, you cannot sell to the person an unemployment insurance product.

On the 14th and 15th rules, we can see a relationship between full coverage car insurance product and liability car insurance product. Therefore, between then there is a positive correlation. We can make some cross sell campaigns between them.

In summary, there is only positive correlation between the two car insurance products. There is no any relation between the other insurance products. In this application, we can understand that there are only two products to make a cross sell campaigns, full coverage car insurance and liability car insurance.

5 CONCLUSION

5.1 Results

In conclusion, we made 3 analysis in this study. They are,

- numerical summary analysis,
- clustering analysis,
- and association analysis

In the numerical summary analysis, we have overviewed the data and we can read it briefly. With this analysis, we can see the customer base in the distribution of region, in the distribution of ages, in the distribution of products and commission values.

In the clustering analysis, we have seen how should be divide the customer base, how many instance we have and the most important one is that we can see how our customer base is distributed by the factor value.

In the association rules analysis, we have seen the relationships between the products. We have analyzed the products movement on the customer base.

On this study, we examined a data in an insurance broker. After our study, when we look at the administrative point of view, there are many decisions, which we can take. Such as who is our best customer base and who is the worst, also how campaigns should we make in which products.

5.2 Limits and Difficulties

A more detailed examination could be made for address information of the customers by examining in deeply. However, I could not do this because the real address of the client could not be saved here in detailed such as district and town. We could also work on the actual city information that customers live in, for example, whether they live in the city center or in the rural area, and in which region they live and what type of insurance they prefer.

5.3 Suggestion

We used Xmeans clustering analysis technique for clustering analysis. It can be used in up-to-date techniques such as Self Organizing Map (SOM) or Expectation Maximization (EM).

We used the Apriori method in the association rule algorithm to identify products close to each other, to identify products containing a cooperative rule, and to campaign on these products. AIS and SETM algorithms can also be used in this study. In the same way, it would be possible to divide the customers into groups according to customer's revenue for the company and to apply an association analysis according to these groups.

It is also can be done to divide customer data in to the groups by the years, and xmeans algorithm can be applied each years' data separately.

Weka data mining program is used in this study because it is not need to have a license for using Weka program. The SPSS program could also be used if it is not need a license to use it.

6 REFERENCES

Cappiello, A. (2018). *Technology and the Insurance Industry.* Switzerland: Springer. doi:10.1007/978-3-319-74712-5

Desik, P., & Behera, S. (2015, January). An Analytics-Driven Method for Profitable Cross-Selling of Insurance Products. *The IUP Journal of Knowledge Management, XIII*, 59-70.

Dogan, B. (2014). Sigortacılık Sektöründe Müşteri İlişkileri Yöntemi İçin Birliktelik Kuralı Kullanılması. *Marmara Fen Bilimleri Dergisi*, 107.

Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. *Computers & Industrial Engineering*, 554–564. doi:http://dx.doi.org/10.1016/j.cie.2016.09.011

Felix, E. (2015, September 29). Marketing Challenges of Satisfying Consumers Changing Expectations. *International Journal of Marketing Studies, 7*(5). doi:10.5539/ijms.v7n5p41

Greenacre, M., & Blasius, J. (2018). History of Cluster Analysis. In M. Greenacre, & J. Blasius, *Visualization and Verbalization of Data* (p. 118). Boca Raton - New York: Taylor & Francis Group.

Gregory, P. (2018, 1 1). */2016/12/poll-analytics-data-mining-data-science-applied-2016.html*. Retrieved from www.kdnuggets.com: https://www.kdnuggets.com/2016/12/poll-analytics-data-mining-data-science-applied-2016.html

investopedia.com. (2018, 4 18). */terms/i/insurance.asp*. Retrieved from www.investopedia.com: https://www.investopedia.com/terms/i/insurance.asp

Janakiraman, D., & Umamaheswari, K. (2014, June). Role of Data mining in Insurance Industry. *An international journal of advanced computer technology, III*(VI), 961-966.

Kumbhare, T. A., & Chobe, P. V. (2014). An Overview of Association Rule Mining Algorithms. *International Journal of Computer Science and Information Technologies, 5*, 927-930.

Li, R. (2018, 04 18). */data/history-of-data-mining/*. Retrieved from https://hackerbits.com: https://hackerbits.com/data/history-of-data-mining/

Nakano, S., & Kondo, F. N. (2018). Customer segmentation with purchase channels and media touchpoints using single source panel data. *Journal of Retailing and Consumer Services*, 142-152. doi:10.1016/j.jretconser.2017.11.012

Prithiviraj, P., & Porkodi, R. (2015). A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study. *American Journal Of Conputer Sciences and Engineering Survey*, 98-119.

Qi, J.-Y., Qu , Q.-X., & Zhou, Y.-P. (2014). How does customer self-construal moderate CRM value creation chain? *Electronic Commerce Research and Applications*, 295-304. doi:10.1016/j.elerap.2014.06.003

Sayad, D. (2018, 1 1). */data_mining_map.htm*. Retrieved from www.saedsayad.com: http://www.saedsayad.com/data_mining_map.htm

Schreiber, F. (2017, MArch 9). Identification of customer groups in the German term life market: a benefit segmentation. *Ann Oper Res*, 254:365–399. doi:10.1007/s10479-017-2446-y

Sekulovska, P. (2012). Internet Business Models For E-Insurance and Conditions In Republic of Macedonia. *Procedia - Social and Behavioral Sciences*, 163-168. doi:10.1016/j.sbspro.2012.05.016

StatisticsSolutions. (2018, 04 17). */cluster-analysis-2/*. Retrieved from www.statisticssolutions.com: http://www.statisticssolutions.com/cluster-analysis-2/

Vijayarani, S., & Sharmila, S. (2017). Comparative analysis of association rule mining algorithms. *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).* Shanghai, China. doi:10.1109/CISP-BMEI.2017.8302317

Wang, X., & Bai, Y. (2016, 9 27). The global Minmax k-means algorithm. *SpringerPlus*, 101-113. doi:10.1186/s40064-016-3329-4

Wu, R.-S., & Chou, P.-H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 331-341. doi:10.1016/j.elerap.2010.11.002