CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE
Faculty of Economics and Management
System Engineering and Informatics

Diploma Thesis

Predictive Modelling in Selected Database

Professor:
Ing. Tomáš Hlavsa, PH.D.

Student:
Suada Myftari

CULS 2016, Prague

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Suada Myftari

Informatics

Thesis title

Predictive Modeling in selected database

Objectives of thesis

Diploma thesis deals with evaluation of customer (company) behavior. The main sense is to find out and analyze possible factors affecting the behavior.

Methodology

The analysis will be based on customer (company) database. There will be used predictive analytics that learns from data to predict the future behavior of individuals in order to drive better decisions. To reach the aim there will be employed statistical procedures, such as exploratory data analysis, regression analysis or multivariate statistical methods.

The proposed extent of the thesis

60 – 80 pages

Keywords

Predictive modeling, data, behavior, factor, statistical analysis

Recommended information sources

ABBOTT, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. USA, NJ, Somerset: Wiley, 2014. ISBN 978-1-118-72793-5.

AGRESTI, A. *Categorical data analysis.* Hoboken: John Wiley & Sons, 2013. ISBN 978-0-470-46363-5.

LAROSE, D T. *Discovering knowledge in data : an introduction to data mining.* Hoboken, N.J.: Wiley-Interscience, 2005. ISBN 0471666572.

SAS INSTITUTE., – CERRITO, P B. *Introduction to data mining using SAS Enterprise Miner.* Cary, N.C.: SAS Institute, 2006. ISBN 9781590478295.

SIEGEL, E. Predictive Analytics. Hoboken: John Wiley & Sons, 2013. ISBN 978-1-118-35685-2.

SMYTH, P. – MANNILA, H. – HAND, D J. *Principles of data mining.* Cambridge, Mass.: MIT Press, 2001. ISBN 026208290.

SOCIETY FOR MINING, METALLURGY, AND EXPLORATION (U.S.), – EARY, L E. – CASTENDYK, D N. *Mine pit lakes : characteristics, predictive modeling, and sustainability.* Littleton, Colo.: Society for Mining, Metallurgy & Exploration, 2009. ISBN 9780873353052.

Expected date of thesis defence

2016/17 WS – FEM

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 21. 10. 2015

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 11. 11. 2015

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 28. 11. 2016

Declaration

I declare that I have worked on my diploma thesis titled "Predictive Modelling in selected database" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the diploma thesis, I declare that the thesis does not break copyrights of any their person.

In Prague, ……. ……..……

Suada Myftari

Acknowledgement

I would like to thank professor Ing. Tomas Hlavsa Ph.D for the constructive advices and for being very collaborative throughout the process of preparing this diploma thesis. I dedicate all my work on this thesis to my parents and my brother for always supporting and believing in me. Many thanks to all the friends I made throughout these master studies who never stopped giving me support.

Predictive Modelling in Selected Database

Summary

The area of technology that we are undergoing made Big Data and all the issues related to it, the topic of the day. Today companies are exposed to a huge flow of information but the key of success is how smart they use the huge information they own to improve their product and services based on the feedback of customers, how they improve the working condition of their employees to have satisfied and motivated people to want to work on that company and do their job well, how fast are they to make the changes and how efficiently they are in delivering their own product and services.

This paper shows how Big Data creates value for the company, if used properly, through transparency, enabling analysis to discover what the needs are and to improve performance, segmentation to customize and personalize the strategies for every segment and what are the issues coming with Big Data as well as what are some of the techniques used for analyzing Big Data like cluster analysis, data mining, pattern recognition. Every process of business inside and outside the company is rich in information and therefore is a source of information. If companies pay attention to the data coming in and out and translating them in proper information and knowledge and to make smart use of them by making the right changes that they indicate, hence smarter and better decision and strategic business moves, they are already a step ahead and closer to satisfied customer which is key of their success and ongoing activity.

The issues arising is what is considered Big Data and the Vs related to it and how they can be analyzed and processed and the case study presented in this research treats data mining and predictive modelling techniques to analyze categorical data in order to understand customer behavior and factors affecting this behavior. Decision trees(random forest, boosted trees and C&RT) are used to analyze the behavior and to see which among them has the higher accuracy in predicting factors affecting behavior of customers.

Keywords: Big data, V-s of big data, data mining, random forest, boosted trees, C&RT, data preparation, data cleaning, predictive modelling.

Predictive Modelling in Selected Database

Souhrn

Význam Big Data v současné době narůstá. V současné době je nutné řešit neustálý tok informací, které je vhodné analyzovat. Klíčové informace umožňují zlepšit produkt či služby; obtížnější je ovšem se v informacích, které jsou obsaženy v rozsáhlých databázích, vyznat.
Tato diplomová práce se zabývá Big Data jakožto nástrojem pro tvorbu hodnot dané firmy. Pokud jsou analytické techniky Big Data využívány správně, umožní prozkoumat datový soubor a následně pak i vylepšit výkon, definovat strategie pro různé segmenty. Cílem práce je identifikace faktorů ovlivňujících chování zákazníka. Analýzy byly založeny na rozsáhlé zákaznické databázi, byly využity různé postupy z oblasti regresní analýzy a rozhodovacích stromů. Za nejvýznamnější faktory byl identifikován počet zaměstnanců a doba trvání posledního kontaktu.

Klíčová slova: Big data, V-s of big data, data mining, random forest, boosted trees, C&RT, příprava dat, čištění dat, prediktivní modelování.

Table of Contents

Table of Figures

Table of Bar Charts

Table of Tables

1. Introduction

1.1 Research Context

Today companies are exposed to a huge flow of information. Companies trading the same products and offering similar services have roughly the same cost of production and services and too similar ways of producing their products but the key of success is how smart they use the huge information they own to improve their product and services based on the feedback of customers, how they improve the working condition of their employees to have satisfied and motivated people to want to work on that company and do their job well, how fast are they to make the changes and how efficiently they are in delivering their own product and services. The society lives in the area of internet and therefore big data. Tons of data are generated every day especially with the phenomenon of emerging markets and the high usage of internet and social media.

Statistics of 2013 from United Nation Economic Commission for Europe (UNECE) show a trend of growth 40 bigger is expected to be in 2019 than the one in 2005. On SAS website a year earlier prediction gives a rate of 50 larger amounts of data will be created and stored than it was in 2012 and yet nowadays only a very small percentage of this data is analyzed and used. There is no worth of this immense amount of data if it is not being analyzed and taken information and knowledge out of them to put them in use for the ongoing of the company to be able to survive the very challenging, tough and competitive market. With the emerging of market and with the creation of a world market companies face more and more challenges and need to be fast and effective in making changes and improvements and an essential step is analyzing the data they gather and store. Every process of business inside and outside the company is rich in information and therefore is a source of information. If companies pay attention to the data coming in and out and translating them in proper information and knowledge and to make smart use of them by making the right changes that they indicate, hence smarter and better decision and strategic business moves, they are already a step ahead and closer to satisfied customer which is key of their success and ongoing activity.

With the disappearance of the trading borders and being exposed to this huge world, is increasing even the need for analyzing data and for people being capable of doing so and capable of seeing the right information insight this unthinkable amount of data. This is what motivated the start of

this thesis in order to go deeper and to understand how the process of analyzing the big data is done, what good use can be of it for businesses and how can this lead to better decision making? Why the big data are important?

2. Objectives

The main purpose of this thesis is to identify the factors that affect the consumer behavior represented by variable sales in the dataset chosen for analyses. To reach these final purpose different techniques will be used to build models and to compare the models in order to find the model that gives more satisfactory results to better understand the customer's behavior. The focus will be in using predictive modelling tools to make a prediction on the future behavior of the customers and how accurate will this model be.

3. Methodology

A sample of phone sales company is used to carry out the experiment. What is required from the data is identifying some of the factors that influence the behavior of the customers and make as well a prediction whether customers are willing or not to buy the products by phone so to adjust the campaign based on their interests. The dataset consists of 8367 customers in total which will be clean and prepare in analysis and at the final stage ready for analysis the data set will consist of 5841 cases which then will be separated into training and testing set based on the algorithm used for modeling.

The thesis starts by doing a review of some of the important literature in order to study the big data issues, how to create a value out of them and the data mining techniques to analyze the big data and to solve the problem mentioned above. During the thesis several data mining techniques like classification prediction techniques, clustering will be review and then applied later on for the practical part. The practical part will focus on the most popular techniques found from the literature review. A literature review will be conducted as well for the first part of the topic of this thesis which is the big data to understand why analyzing them is important before going to the techniques to be used to do the analysis.

Then based on literature consultation, a process of cleaning the data in order to get a good dataset ready for modeling, will be taken. After the data have been cleaned out and have been analyzed for possible relation among them, the modeling process will follow. Three main algorithms will

be used to get predictive models, *Random Forest, Boosted Trees* and *C&RT.* As the dataset has only categorical variables, after a research in technics for classification data mining the three mentioned algorithms where chosen. At the end of each model *Rapid Deployment* tool will be used to check the accuracy of every model.

For better evaluation of the performance of the techniques, some well-defined criteria need to be established. After a comparative analysis and considering the restriction in availability of softwares, the software chosen for implementation of data mining technics is *Statistica*.

4. Theoretical Context

4.1 What is big data?

The Internet of Things this revolutionary concept that refers to physical objects becoming a type of information system, according Loffler,Chui and Roberts in their article published for McKinsey&Company , churns out enormous amount of data to be analyzed. In this article they emphasize that these physical information systems are starting to be used, some even without human intervention which makes a gigantic source of tremendous volumes of data. And indeed, companies in their day to day activity and as they interact with each other and individuals, are creating this huge amount of digital data. The wide usage of smart phones, social medias and many other customer related devices as laptops, smartwatches allow the increasing number of users to contribute more and more in this big amount of raw data available out there. Every one of us with our daily usage of internet including browsing, searching, buying online, sharing information through social media, communicating online, tweeting create a huge personal stream of data. (Manyika, et al., 2011)

Cloud computing is the technology that also made possible to store and analyze big amount of data with lower cost, making it possible even for small and medium size companies not only for big ones with huge income and revenues. The storage and analyzes of big data is not only beneficial for private sector but for public sector as well and national economies. The world is under a wave of huge changes in the technologic sector, under a wave of innovation and growth. People are consuming more and more good and services and therefore the economy is flowing towards unavoidable changes. Big data can create values across sector of this huge global economy and studies shows that already many companies are using big data to create the value

which take the other companies to the position of having to explore how to do the same if they want to stay in the competitive position. (Manyika, et al., 2011)

There are many definitions of big data out there and this legitimizes the confusion for question like how big should a data set be to be called big data? Is it decided by the number of columns or by the number of rows? So is it just about volume?

Among many definitions around two of them did a good job in explaining the big data. First comes from Merv Adrian in an online article he published in Teradata Magazine where he defines big data as: Big Data exceeds the reach of commonly used hardware environment and software tools to capture manage and process it within a tolerable elapsed time for its user population. (Adrian, 2011)

The second definition comes from McKinsey Global Institute in there paper related to big data and according to McKinsey: "Big Data" refers to dataset with that size that goes beyond the ability of typical database software tools to capture, store, manage and analyze. (Manyika, et al., 2011)

Even though both definitions emphasize that the definition varies depending on the sector as different sector have different need of usage of data and have different need of volume of dataset to analyze and also depending on the type of software used to work with the datasets, they say that big data usually varies from a few dozens of terabytes to multiple petabytes. What they want to bring to point is that big data definition is changed and will change with the technology development, which is making gigantic steps ahead. So what was qualified as big data a couple of years ago is far from being considered big data today and the same will happen in the future?

4.1.1 The Vs of Big Data

But big data is not only about volume but it includes the so called 3Vs- Volume, Velocity and Variety. And in a SAS publication they add as well two other dimensions: Variability and Complexity. (SAS Institute Inc. , 2012)

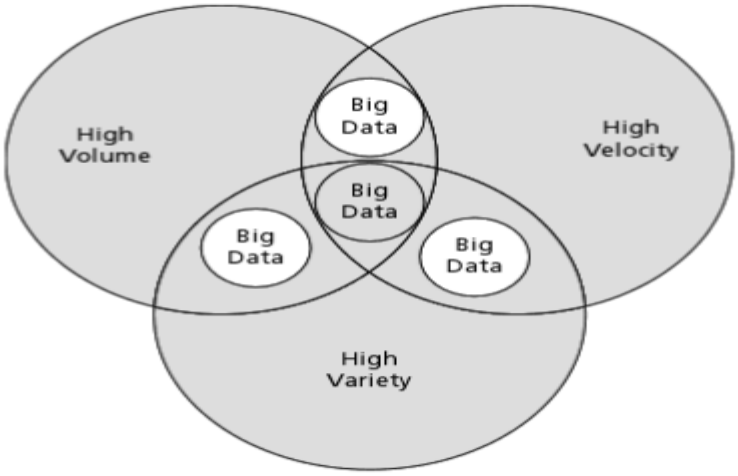


Figure 1 The Vs of Big Data (Gendron, 2014)

This means that not only we are getting more and more data but we are getting them faster and from different kind of sources and in more complex formats. And once more we emphasize that the important is not having big data. Having just big data does bring any values if you don't use it properly, if you don't analyze to create the value for the company. And this is the challenge to develop and get updated techniques and processes for analysis by using updated technologies and methodologies to have an effective and efficient final analysis of big data. And the actions that you as a business take afterward to make good use of the results is what really matter and what keeps you to a competitive level in this global economy that is breaking the borders.

Volume: refers to how much data the company generates. Companies and organization nowadays produce data every moment affected also by the high usage of internet. Their sources are business transactions, social medias, machine-to-machine data. There are different sources that make the organizations to create data. (Gendron, 2014) Sometimes this can be a legal requirement where some countries force the companies to do data collection by law which is mostly for security reason and to protect the privacy. A reason that drives the company to create data is the necessity to understand how customers rate her products and services. Another way of seeing how consumers rate your products and services is by analyzing the data generated by social networks which allow companies to have a two-way communication with its own customers. The big volumes of data are also result of mobile computing.

Velocity: refers to fast is the processing of the data available. And the organizations are having data streams into their data stores at a very fast rate. If critical decisions are depending on those

data than the company faces the so called high-speed problem of Big Data. A solution of this problem is the Event stream processing which analyzes data in real time to reduce the time for critical decision making. Example of these are: Fraud detection like banks checking every transaction through their systems to verify the validity of each of them, Mining the data from social media by doing a real time analysis of the all the data posting in social media to understand the rate that customers have for the product, Personalized marketing which is reachable by doing onclick analysis, search engine analysis so basically analysis every data we get from customers search on the internet. (Gendron, 2014) So based on the most often search results the companies build a personal marketing for every customer showing them advertisements mostly for the product they are interested more.

Variety: refers to different types of format that data are generated -structured data, email, video, audio, financial transactions, numeric data, unstructured data. (SAS Institute Inc. , 2012)The sources for these data are also an increasing number, just to mention few of them like social medias as Tweeter, Facebook, LinkedIn, You Tube; Internal data stores inside the company like CRM, ERP systems; the reviews for products on the organization website or on other websites connected with, Saas providers like Google Apps or Salesforce. (Gendron, 2014)

Variability: refers to the high speed and many varieties of data which makes it hard to manage the seasonal or day peaks as well as event-triggered peaks. Being the data mostly unstructured it makes even harder to work with them. (SAS Institute Inc. , 2012)

Complexity: refers to the many sources which generate the data that brings different formats which is hard to match and link with each other, also its hard to filter and clean them for a better analysis. Either way the organizations need to connect and correlate relationships among data not to let them get out of control. (SAS Institute Inc. , 2012) (Franks, 2012)

4.2 Issues with Big Data

Besides all the good values and advantages it gives to organizations and companies, Big Data don't come without risks or issues. And perhaps the biggest issue is privacy. The problem comes from both sides, not every individual is reliable as well as not every company and government is reliable which brings out the privacy issue. And this requires strong restraints in order to keep using the big potential of big data otherwise many resources of it might found themselves shut down to protect people privacy. Once the data are stored somewhere there is also the risk of them being stolen we have the examples of credit cards frauds or WikiLeaks the biggest scandal

of classified government documents that were stolen and posted online, being public to everyone to have access to them. Regulation from company itself and legal one is crucial in these cases. Company need to make clear enough and make people understand how they will store their data and how they will use them. This will make customer feel safer and confident to use the company services and products even though it will generate data about them. People are already being scared and worried about the huge amount of data being generated about them, how the web browsing history with be tracked and use by the others, how their locations and actions showed mobile phone and GPS systems will be tracked and used. Nowadays the telecommunication companies and mobile phone companies as well as search engine know everything about everyone. And it is understandable for people to fear this fact of being exposed all the time just by having a mobile phone and internet connection on it. Among the regulation the one that company imposes itself is better and less restrictive for the overall activity of companies rather than having the regulation being asked by government as it might be more restrictive but also bad for the image of the company because they come after the company is not doing well with policing itself. (Franks, 2012)

Another issue is the company getting more data that it can handle as it will make the company slow down and get steel without making any progress. The solution in this case is hiring the right people who know how to deal with big data and know how to address the right problems. Considering the many formats that data are generated, it does make it easy dealing with them as well. Traditional data sources were mostly structured data as it was too expensive already to collect data and yet even more to keep unnecessary data to which there was no use. You already know before what format will the data that you were collecting and storing going to have. This was making it very easy to work with this data. But nowadays with the huge flow of data, with the huge volume of data generated second by second we cannot expect the data to have a proper structure and to be ready to use. We should be ready as well to throw away some of the unnecessary data and to filter them to get to the proper form for analysis. Thus mostly the data generated are unstructured or semi-structured. Unstructured are the data that we don't have control over and have a non-predefined format like documents, texts, images, videos. The photos are made of pixels but the way the pixels will fit together to create the final photo that we see, it varies on each case. (Franks, 2012)

What can also be an issue with big data is the increase of costs very fast as more and more data are gathered in the company even before that companies and organization figure out what to do with it. What is important for companies is to capture a rhythm and create a pace that will allow them to keep up with the velocity of data gathered. This does mean that all 100% of every new data source should be analyzed but rather capturing some samples and analyzing them to understand what is important in this new sources and how they can be used. Starting from this the companies and organizations can effectively address and approach a data source on a larger scale. (Franks, 2012)

4.3 How does big data create value for the company?

As emphasized previously having big data is not a big deal but what you do with them is what create the values. There are several ways in which big data transforms the companies' potential into values and drive the way companies perform decision making, how they are designed, and managed. This how McKinsey in their publication think that Big Data bring value to the company:

Transparency: If all the data generated are accessible to all the relevant stakeholders depending on the area of interest they cover than this can lead to the considerable decrease of search and of processing time. This mean that instead of protecting the data inside the company the companies make them accessible to different level of employees and across separated departments this will reduce time to market and improve the quality of the product. It is all about knowledge taken out of the Big Data and the good use of it. (Manyika, et al., 2011)

Enabling analysis to discover what the needs are and to improve performance: As the data generated and stored can include very accurate and detailed data about performance on everything related to the company from inventories to personnel working from home days, sick days, PTOs the information technologies makes it possible to analyze the trends of performances and to understand what effects it so to work on managing the performance to the best.

Segmentation to customize and personalize the strategies for every segment: A benefit of big data is that allows companies to make very specific segmentations of customers and to focus on creating products and services exactly as required and showed by the segmentations needs. This is mostly a marketing strategy and risk management with the real time analysis of micro-segmentation to have a more accurate perception from the customer and to be more accurate in targeting them for promotions and advertising. (Manyika, et al., 2011)

The need to innovate business models, services and products: Big data are generated mostly because of the great developments of new technologies and the use of internet connection. This means that the need and demands of customers are changing and companies by analyzing the volumes of data generated and stored have very accurate information about those needs which pushes them towards creating new business models, new services and products, beside the improvement of the old ones. (Manyika, et al., 2011) For example this huge bloom of real-location data brought the need of creation of new location based products and services from insurances to navigations, depending on the places and the way that people drive. Apple as a well-known company for her technological products as well the marketing strategies her products always tends to analyze and satisfy customers need by making changes and improvements to the old products and by creating new products like apple watch based on the customer needs. The new change they are offering for their operation systems is the Siri features which allow users to control various functionalities of Mac operating system with their own voice as they are used to do on their iPhones.

4.4 Techniques for analyzing Big Data

More and more techniques are being researched and developed for having a better and more accurate data analysis also affected by the high range of industries. In the McKinsey report there is a list provided, which is mostly is applied in big datasets but of course some of them can be applied also on small datasets like in case of regression. We will shortly introduce some of them and on the following sections our focus will be on data mining. Techniques related to data mining will be treated on the last section.

Association rule learning: Includes a set of techniques used for data mining to find out relationships among variable in big databases. For example analyzing a market basket to see which products are usually bought together and focus the marketing campaigns on these products. (Manyika, et al., 2011)

Classification: is function for data mining for targeting certain categories. A classification model can be churn rate by identifying it high, low or medium depending on the tendencies of customer to leave. (Manyika, et al., 2011)

Cluster analysis: as we mentioned is an unsupervised classification used to group objects into smaller groups based on their similarities which are not known in advance but are noticed during the analysis. (Manyika, et al., 2011)

Data mining: a set of techniques to find important patterns from big datasets by using statistics and machine learning methods. This includes association rule learning, cluster analysis, classification and regression. For example we can mine customers to find segments that are most likely to stay, finding market baskets to identify and model the buying behavior of customers. (Manyika, et al., 2011)

Pattern recognition: deals with finding regularities in data through the use of some algorithms in order to classify the data into categories. (Bishop, 2006) And the models chosen for pattern recognition can be classified to different categories depending on the method used for data analysis and classification. They can be used independently or dependently to perform a pattern recognition task. The different models used are:

Statistical Model: where each pattern is described in terms of features. The features chosen don't allow different patterns to take place on non-overlapping features space.

Predictive modelling: a set of techniques to find the best mathematical model to have the best prediction of the changes of a wanted outcome. It is used for example in CRM to predict the likelihood that a customer with will stay or leave the company.

Regression: statistical techniques to check the dependency among variables. We build a model and see how the changes in independent variables will affect the dependent variable therefore we can make also forecasting for the future values of the variable.

Statistics: the science that deals with collecting data, organizing them and then interpreting the data. It is used to test the null hypothesis that test the relationship that might have happened by chance among variables and what kind of relationship is the casual one which mean whether it is statistically significant.

Time series analysis: techniques to analyze set of data representing values at successive times, to get meaningful characteristics out of data. It concerns with the fact that data points might have an internal structure (like autocorrelation, trend or seasonal factors) that should be addressed and analysis. Time series forecasting is a technique to predict future values based on known values or values of other series. (NIST/SEMATECH e-Handbook of Statistical Methods)

Visualization: techniques to present information in ways that everyone can easily see and understand their data. By creating diagrams, images, graphs, animation it makes it easier to understand, communicate and improve the results of the analysis of big data. Tableau is one of the latest well-known software that is specialized with visualization which has brought big

changes in this field. An example of visualization is Clustergram used for cluster analysis to display how individual representatives of a class are assigned to cluster as we increase the number of cluster which is a very important parameter in cluster analysis. From Clustergram we can see how the results of clustering differ with different numbers of clusters. (The Clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses, 2002)

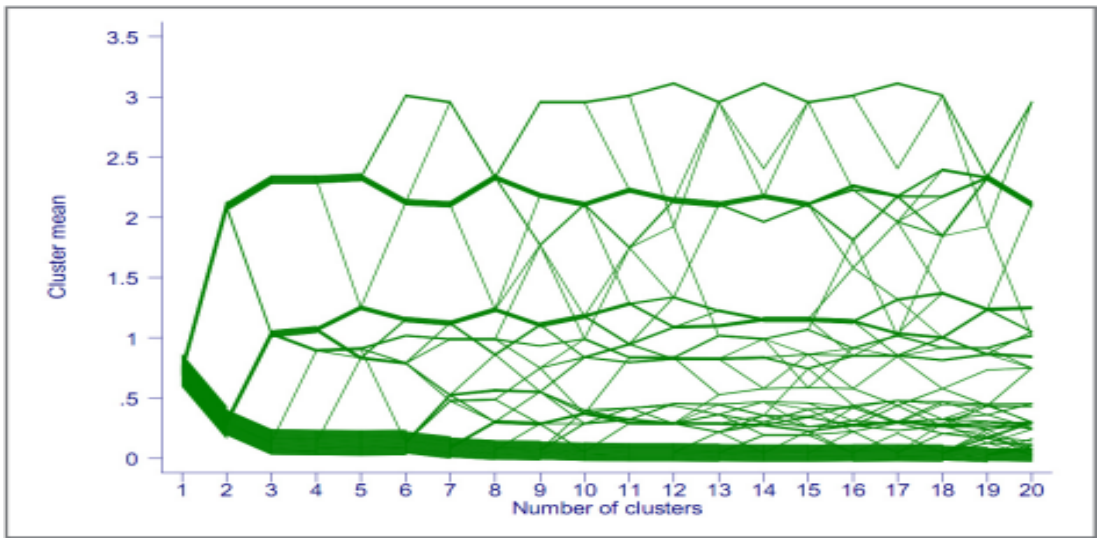


Figure 2 Clustergram (Manyika, et al., 2011)

4.5 What is data mining?

There are different definitions of data mining which emphasize that data mining is analysis of big set of data by specific software's to find out patterns and rules on how this data behave. As we said on the Big Data chapter, big data are worthless if we don't make good use of them. Data mining is a way of analyzing the Big Data and extracting the information needed out of them and thus creating value for the company and understanding the customer behavior.

According to Tuffery: Data mining is the set of methods and techniques for exploring and analyzing data sets, in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, associations or tendencies; special systems output the essentials of useful information while reducing the quantity of data. (Tufféry, 2011)

Based on his own definition of data mining Turffery also considers two categories of data mining: descriptive and predictive.

Descriptive data mining would be extracting the current existing information inside the data like clustering of individuals or searching for relationships between products or services. And predictive data mining would be when we try to extrapolate new information from the existing

one, which can be either quantitative in case of regression or qualitative in forms of scoring or classification

According to Linoff and Berry: Data mining is a business process for exploring large amounts of data to discover meaningful patterns and rules. The authors analyze every part of the definition and consider them important. As a business process data mining interacts with other business processing and it is an ongoing process from collection of data to analysis which will bring more data and will require more data mining. Data mining is an important process in the companies that want to keep a competitive position and understand more the markets and customers in order to improve their products and services. The large amount of data is all about big data which we deeply analyzed in the previous chapter. Never before companies had access to such big volumes of data and this is an advantage for the companies. It is a source for companies to better understand what goes right and what goes wrong and what needs to be improved. And the last part of the definition might be maybe the most important part. We can find a lot of patterns in data and we can extract a lot of information from it but it is important to focus only on the patterns that are meaningful and useful for businesses so to gain customer value in short and long term or to try to keep the customers which are most likely to leave. (Linoff, et al., 2011)

And the era of big data brought an explosion of more extensive techniques of data mining, as the size of the information is much larger and because the information is more diverse and immense in its nature and in its content. Simple and straightforward statistics are not enough anymore. For the companies is not enough just to know the location of its customers but it needs more detailed information about their age, average earnings, their buying behavior to target their needs better. And these business-driven needs as we mentioned before have changed the simple way if getting and analyzing data into a more complex data mining process. (Brown, 2012) From the below figure we can see the developments of technologies from ERP to CRM, eCommerce and nowadays to Big Data Analytics.

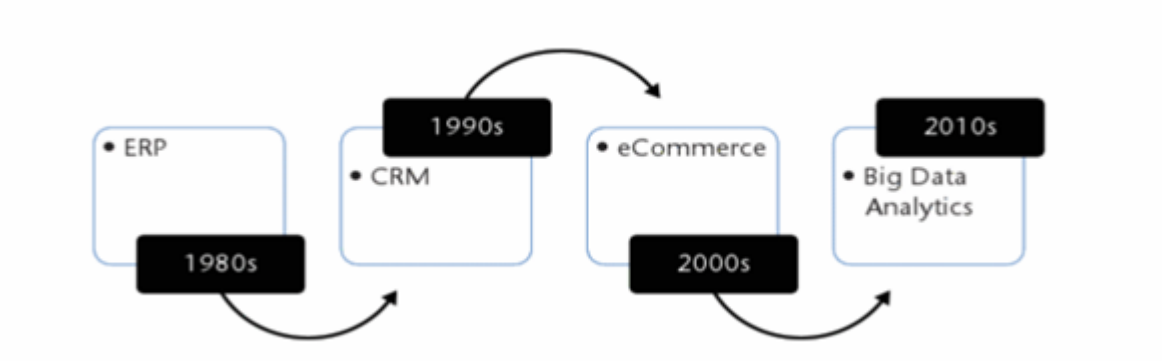
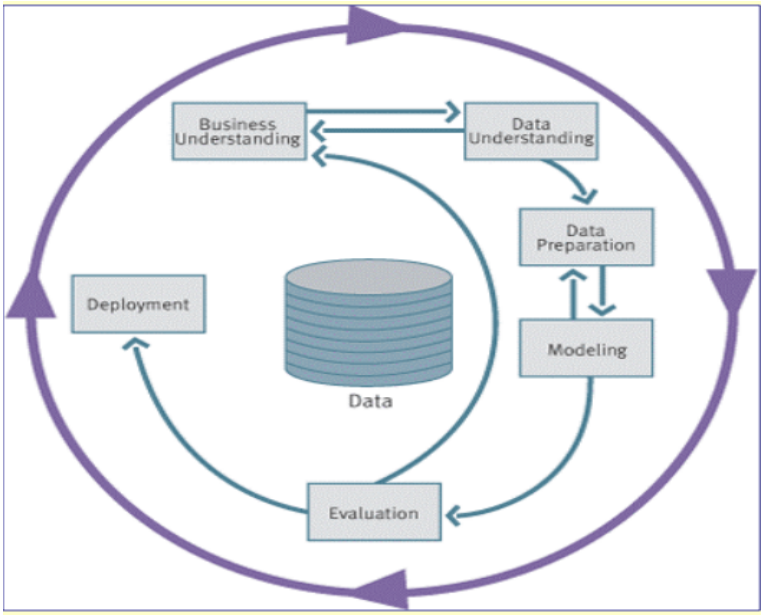


Figure 3 Timeline of recent technology development (Minelli , et al., 2012)

4.5.1 Data mining Process

CRISP-DM which is Cross Industry Standard Process for Data Mining a known and accepted methodology for data mining defines 6 steps through which the process goes as shown in the figure below. But the process of data mining does not finish once we figure out a solution because that can bring new and more complicated business questions so we will end up in new processes which will benefit from the previous ones.



Business understanding

First we need to understand what are the business requirement and what does it want to achieve with this data analysis. And then based on the business perspective we create a data mining problem and an initial plan to achieve the objectives settled. (Chapman, et al., 2000)

Figure 4 Data mining phases of CRISP-DM reference model (Chapman, et al., 2000)

Data understanding

This phase starts since the collection of data and it requires you to get more familiar with the data by having a first sight into the data, maybe discover subsets then can lead to some hypothesis related to information you want to find out. (Chapman, et al., 2000)

Data Preparation
This phase deals with constructing the final dataset from the initial raw data, even though as a phase it is most likely to be held many times as during your analysis you get a better understanding of the problem and realize which data are needed in the ongoing process and which one should be removed. This phase includes the selection of table, records, attributes considered important for the analysis, and also the changes and cleaning of the data for the modeling tools. So we can remove unnecessary data from the dataset and make changes on the variables like transforming qualitative variables to quantitate ones to make the analysis easier.

Modeling
After we have the dataset we need to build the model and to apply various modeling techniques for doing the modeling of the dataset. As some techniques might have specific requirements for the form a data we might find the need to go back the data preparation phase to make the necessary changes.

Evaluation
Once we built the model we need to evaluate it whether it achieves the business objectives. In this phase it is important to check whether important business issues are considered or left out.

Deployment
At the end we need to organize and present all the knowledge from this analysis in a way that the customers can use which involves including the knowledge and the model in the decision making processes of the company. Normally it is the customer who carries out the deployment steps and in the cases when the analysis has to do it that it is important for the customers to understand what should be done in order to make use of the created models. (Chapman, et al., 2000)

4.5.2 Tasks for each phase of the data mining process

The process goes deeper than the 6 phases discussed above. Each of the phases includes more detailed task which will give us a better understanding of each of the above phases.

1.1 Determine business objectives

On the business understanding phase first we need to determine the business objectives which mean to understand what a company wants to attain and often their objectives and constraints are

contestant therefore a proper balance is needed. A primary goal for the company would be how to keep the current customers from leaving by predicting when they are more willing to go to a competitor. For example how does the coverage and the fees of ATM effects the customers in leaving or staying? The success criteria might be the reduction of customer churn.

1.2 Assess situation

Second we need to evaluate the situation by getting more detailed facts to determine the goal of data analysis like resources, constraints, suppositions which should end up with a list of the resources (including personnel, business experts, data experts, data mining experts), data(access to live data, the ones stored on warehoused or the operation data), computing resources(the hardware platforms of the company) and the software resources(the tool, for data mining, software for visualization or other important software), list of requirements, assumptions and constraints of the project(making sure that we are allowed to use the data, considering the legal issues, security issues with the data, understandability and the quality of results, making assumption and hypothesis to verify them during data mining), list of constraints which might include constraints related to technology like the size of the dataset that practicable to use for modeling.

1.3 Determine data mining goals

Third from the goals in business terminology we need to move to the goals determined in technical terms. For example the business goal could be “increase sales” and the data mining goal would be” To predict how the sales will in increase by analyzing the given information about the purchases over the last years, about demography(salary, age, position, city etc.) and the price of the items”.

1.4 Produce project plan

At the end of the business understanding phase there should be the description of the plan for achieving the data mining goals and therefore the business goals, where there should be specified the steps to be performed during the rest part of the project, including techniques and tools.

The second phase Data understanding goes as well on different tasks, from collection of data to description, exploration and verification of the quality of the data.

2.1 Collection of initial data

Collection of initial data deals with accessing the data that are listed in the resources of the project. This includes loading of the dataset in the specific tools chosen for data understanding.

We end up with a report that lists the datasets obtained, the location of this dataset, the methods used to obtain the dataset and any possible problem faced.

2.2 Description of data

A description of the dataset is needed after the dataset is obtained which includes the format of the data, the quantity like number of records and fields in each table, the identities of the fields. This phase evaluates whether the data satisfies the important requirements.

2.3 Explore data

This tasks analysis the data to find initial hypothesis, relationship among attributes, properties of significant sub-populations, simple statistical analysis like frequency tables, two-way tables, and descriptive statistics. (Tufféry, 2011) Data exploration helps us to reduce the amount of information from the dataset so to focus on the most important aspects of the data.

2.4 Verify the quality of data

Verification of the quality of data concerns with verifying whether there are missing values, repeated values, is the data complete, are there error included and offering possible solution regarding the problem.

The third phase was Data preparation which includes tasks like select data, clean data, construct data, integrate data and format data.

3.1 Select the data

Based on the data mining goals, quality and possible technical limitation such a limits on the volume of data or the type of data we select the attribute so the columns and the records (rows) in the table that will be included in the analysis.

3.2 Clean the data

Data cleaning is mostly needed when we integrate heterogeneous data sources and it deals with detecting and removing possible errors and possible inconsistencies from data in order to raise the quality of data. Usually the quality problems with data are shown in single data collections, like files or databases and it might be due to misspellings during data entries, invalid data or missing information. And then in the cases that we need to integrate together multiple data sources for example in data warehouse or global web-based information systems, the need for data cleaning can result essential. This might come as a result to sources often having unnecessary data in different representations so different representation, overlapping or contradiction. And so to have more precise and consistent dataset we need to consolidate the

different data representations and to eliminate the duplicated information. If the data are used for decision making like in the case of data warehouses than the correctness of the data is vital as the duplicated or the missing information will lead to incorrect or misleading results. (Data Cleaning: Problems and Current Approaches, 2000)

3.3 Construction of data

In this task we deal with the needed changes in the data like entering new records or deriving new one from the existing values.

3.4 Integration of data

Integration of data refers to combing multiple data like tables and records sometimes from different sources into meaningful and valuable information. We can merge two or more tables which have different information about the same objects. (IBM)

3.5 Format Data

Formatting data refers to modifications made to the data which does not change the meaning of it but instead can be required from the modeling tool for example trimming all values to a maximum of 32 characters.

After we have a better understanding of the business goals and after having prepared the data for analysis now we can start finding the best model which makes the best prediction. Based on the model the company will then apply the suggested changes and improvements to better satisfy customer needs and to keep the competitive position in the market.

4.1 Select modeling techniques

Even though during the Business Understanding phase a tool for modeling might have been selected, this task is more specific for example decision tree building or generating a neural network. Many of the modeling techniques make specific assumption as well based on their focus of analysis like – no missing values allowed, class attributes must be symbolic.

4.2 Generate test design

Before building the model a procedure or e mechanism is needed to test the quality and the validity of the model. For example in the case classification we can use the error rates to measure the quality of the model. So typically the dataset is separated into training sets and test sets where the model is built on the training set and estimation of the quality is done on the separate test sets.

4.3 Build the model

Build model is focused on using the tools for modeling on the already prepared dataset to create one or more models. And very often based on the modeling tool there are a considerable number of parameter that can be adjusted. Usually the purpose of building the model is to have the best predictions in order to use them to business decisions. And the most important is the stability of the model which means that the model should be true when we apply it on the future data. No matter what data mining techniques we choose to use the basic steps for building predictive models are the same. The model set is split in three components: training set, test set and evaluation set. And out of the model set there is a fourth set called score set. The fourth sets should be completely separated thus not have any records which are in common since each of them performs a different purpose. As shown in the Figure below we create models using data from the past to make prediction for the future and this are the candidate models build during the training process. All this process is called training the model during which algorithms find patterns which are of predictive value. After having the build model we refine it by using the test set. In this step the model is used on the test sets data to ensure that it is stable and that it will perform well on unused data as the test set is different and separated from the training set. After we check the performance of the model using the evaluation set. The final model is further more applied on the score test to make prediction, used for business decisions. (Data mining: A conceptual overview, 2002)

4.4 Asses Model

After having the final model now we need to assess it to be sure that it meets data mining success criteria and that it passes the desired test criteria. This is a technical evaluation based on the outcome of the modeling. (Data mining: A conceptual overview, 2002)The data mining engineer discusses with the business analysts and the domain experts in order to check the results of data mining in the context of business. We go through the process building and the evaluation until we will find the best model or models and the most appropriate model will be the one that best meets the business objectives.

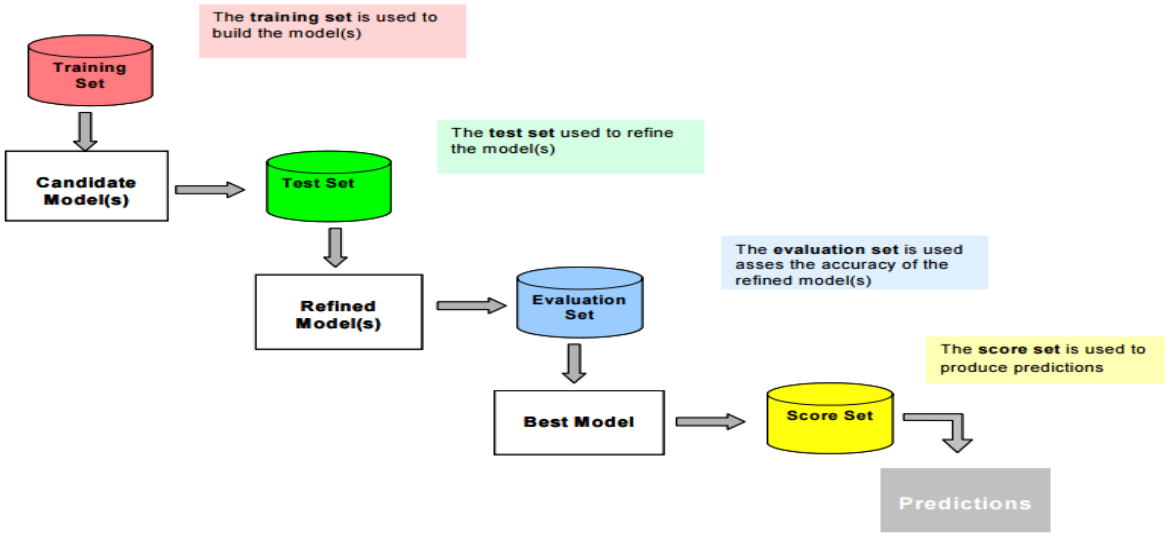


Figure 5 The process of building a predictive model (Data mining: A conceptual overview, 2002)

The assess model task is only related to the model itself, its precision and generality but the evaluation phase sees the model in the overall view of the project.

5.1 Evaluate results

This step evaluates the degree to which a model meets business objectives and tends to determine whether there is some business reason why the model might be deficient. Another way is to test the model or the models on test application in the real application, if the constraints related to time and budget permit it. (Chapman, et al., 2000)

5.2 Review of process

After assuring that the built model is satisfactory and satisfies the business needs, we now need to do a more deep review of the data mining engagement to see whether any factor or task is left out. The deep review also includes the quality assurance –like: Is the model built correctly?

5.3 Determine next steps

Depending on the evaluation task and on the process review the project team then decides how to proceed so whether to go to the next phase of deployment, to do an iteration there is place to improve the model or set up new data mining projects. (Data mining: A conceptual overview, 2002)

The last phase is the deployment and in many cases it is not the data analyst who undertake this phase but the client. The deployment can be either generating a report or a more complex process like a repeatable data mining process. This way even if the purpose of the model would

be the one to increase the knowledge of the data, to get more knowledge out of it, this knowledge has to be organized and presented in a way that the client can use.

6.1 Plan deployment

Now we need to deploy the data mining result or results into the business and in this task the evaluation results are taken and a strategy for deployment is developed.

6.2 Plan monitoring and maintenance

If the result of the data mining process gets part of the day-to day business and its environment that this tasks is very important in order to insure that the results of data mining are incorrectly used. In order to do so a monitoring plan is needed.

6.3 Produce final report

The overall process of data mining it needs to be documented as well by generated a final report which might be either a summary of the project or a presentation of the data mining results. (Chapman, et al., 2000)

6.4 Review Project

This task evaluates and summarizes the overall experience of the project. Documenting what went wrong and what went right, problems faced during the process, suggestions and hints related to the best suited data mining techniques to be used in similar situation. (Chapman, et al., 2000)

4.6 Data mining Techniques

Some the key techniques for datamining are:

- Classification
- Estimation
- Prediction
- Sequential patterns
- Association
- Clustering
- Description and visualization
- Decision Trees

Classification, estimation and prediction are examples of direct data mining or as it is called differently supervised learning which means that we use the available data to build a model that describes particular attributes in terms of the rest of the available attributes.

Classification can be of two types- supervised classification and unsupervised classification. In the first one we know in advance the set of classes and in the second one usually called clustering they are unknown but after classification we can give them a name. An example can be to detect the spam emails based on the message header and the content. (Jain, 2012)

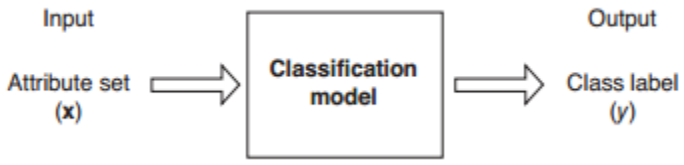


Figure 6 Classification as the task of mapping an input attribute set x into its class label y (Pang-Ning , et al.)

Estimation creates values for unknown continuous variables such as height or income or credit card balance when we have some input data. For example, we can estimate the number of children of a family based on the input data of mother's education. (Jain, 2012)

Prediction classifies the records according to future behaviors or estimated future values. An example of prediction would be to predict the customers that will leave within the next 3 months. In prediction the historical data can be used to build a model for the explanation of the current values and when it is applied to current input then the result is a prediction of the future values or future behaviors.

Sequential patterns mostly used for long-term data identifies trends or regular happenings of similar events. For example identify that certain products are bought together at different times of the year.

Association: The most used example is that of people who buy diapers also tend to buy beer. Association discovers the probability of the co-occurrence. The confidence by which we can tell that by buying a product the customer will buy another one as well. For example Cereal with 85% confidence implies milk. 85% of customers buying cereal buy milk as well, 20% of all customers buy both. 85% is called the confidence of the rule and the 20 % is called the support of the rule. We want to find all the association rules that satisfy the specified minimum support and minimum confidence constraints.

One of the most frequently used algorithm for association is called Apriori algorithm which generates candidate itemsets as shown in the below figure. (Pang-Ning , et al.)

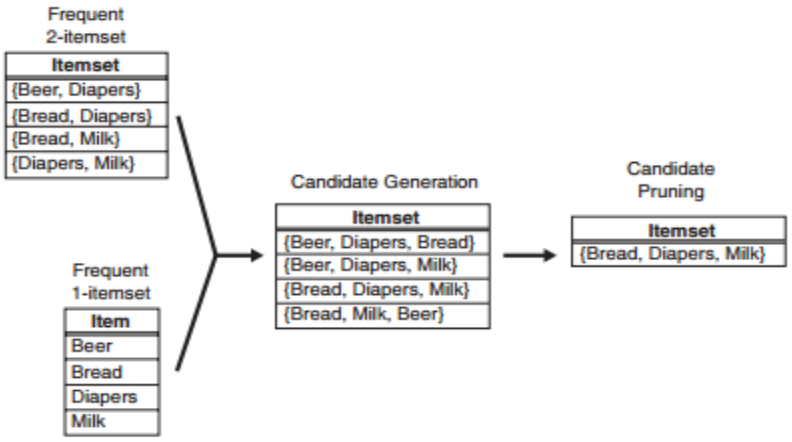


Figure 7 Generating and pruning candidate k-itemsets by merging a frequent(k-1) itemset with a frequent item (Pang-Ning , et al.)

There are many approaches for cluster analysis and even though there is no best solution for determining the exact number of clusters to extract three of many approaches will be given below. Partional and Hierarchical clustering are basic clustering techniques and Model-Based Clustering is a data mining clustering technique which can be either partitional or hierarchical.

Partitional Clustering

There are two types of partional clustering the distance–based and the density-based clustering. In distance based the most popular is the K-means in which n objects are partition into k clusters, number that the analyst has to specify before. In this clusters each object will belong to the cluster that has the nearest mean. As there is no best solution of how to choose the optimal number of clusters a way can be to have multiple runs of the methods with different k and choose the best among them based on a criterion specified before. One might think to increase the number of k but this even might probably decrease the error will bring a higher risk of overfitting. The objective of the K-means clustering as shown from the figure is to minimize the squared error function. (Sayad)

Hierarchical clustering

Hierarchical clustering consists on creating clusters that have a pre-decided ordering like the folders and files in hard disk are. It has two types Divisive and Agglomerative.

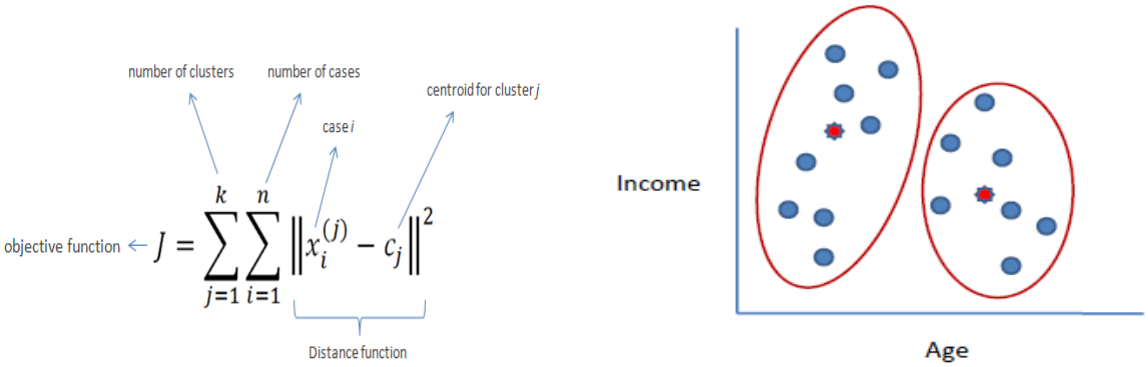
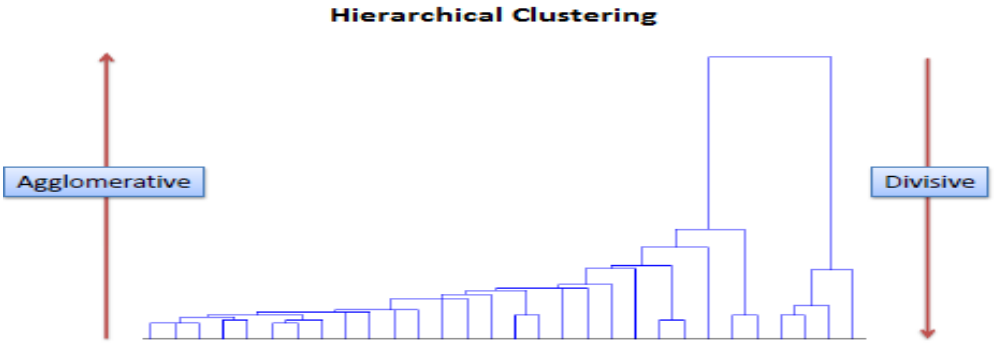


Figure 8 k-means clustering (Sayad)

In the divisive we start with a single cluster and partition it into two least similar cluster and we continue recursively till there is one cluster for each observation. Agglomerative method goes on the opposite way which means it assigns each observation to its own cluster. Then checks the similarities between each cluster and join the ones that are most similar. The method finishes when only one cluster is left. (Sayad)



(Sayad)

Figure 9 Hierarchical clustering (Sayad)

Model based clustering

In this type of clustering several data models are assumed and then maximum likelihood estimation and Baeys criteria is applied to identify the most probable model and number of clusters. Algorithms used for model-based clustering find good approximations of model parameter to best fit the data. This type of clustering based structure or the model and the way improve models to identify partitioning can be either partitional or hierarchical. (Andritsos, 2002)

Figure 11 shows an example of a plotting cluster solution to see the cluster results and in this case there are 5 group cluster solution.

Description and Visualization

Visualization is a very important for of descriptive data mining. Meaningful visualization makes analysis easy to understand as humans are more used to extract meaning from what is visual. There are two types of knowledge discovery goals: Verification and Discovery. So either we use visualization to verify the hypothesis or the results of a method or to discover patterns for predicting the future behavior of some entities. (Jain, 2012)

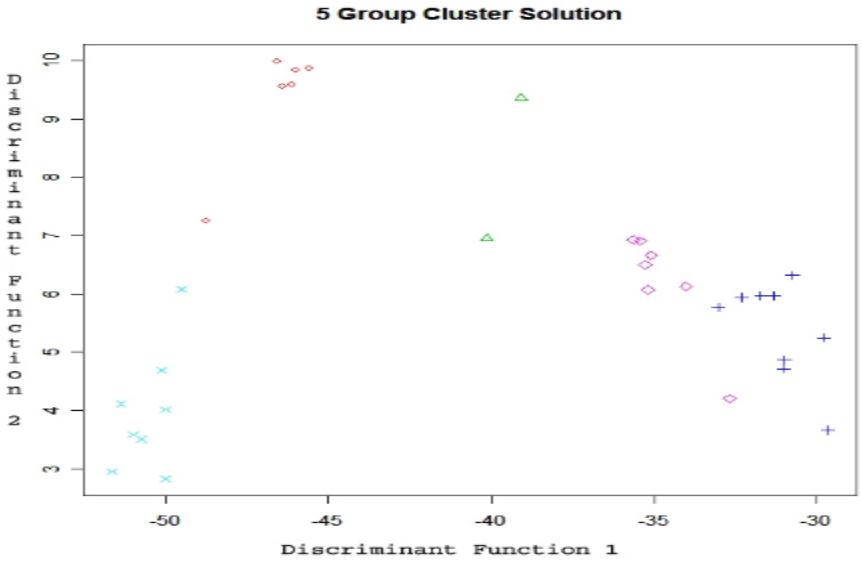


Figure 10 Group cluster solution (Kabacoff, 2014)

Decision Trees

Decision tree presents the classification or regression models in a tree structure. It separates a dataset into smaller subsets and at the same time associated decision tree develops incrementally. The process starts with a question that has two or more answers. Each answer then will direct it to more questions to help in classifying or identifying the data, in this way the data can be categorized and based on each answer a predication can be made. At the end it will end up with a tree with decision nodes and leaf nodes. In the figure above it is shown that the decision node for example outlook has two or more branches in our case Sunny, Overcast, and Rainy. (Sayad)

Figure 11 Decision tree (Sayad)

The leaf node shown in the Figure is Play represents a classification or decision. Decision tree works both with numerical and categorical data. (Sayad)

The focus of the thesis will now move to the practical analysis of the dataset, mostly it will be on understanding company customers in order to improve marketing strategies, sales and the operation for customer support.

5. Practical Study

In this practical part will be analyzed a sample of phone sales company. What is required from the data is identifying the factors that influence customer behavior. The dataset consists of 8367 customers in total which will be clean and prepare in analysis and at the final stage ready for analysis the data set will consist of 5841 cases which then will be separated into training and testing set based on the algorithm used for modeling.

The dataset found is focused on sales so the analyses will also be focused in finding out which of the variables effect the sales, what leads to shorter sales cycle and have some sales predictions.

The idea is to check the importance of the effect of some of the variables to build the right model at the end.

Sales are already a binary variable with values yes-in case of sales and no-in case of not sales. Based on the data given the analysis with focus to check whether the sales are effected by the job type which mean people having a certain job type for instance management tend to have positive sales than other job types, if the education level effects the positivity of sales, whether having loan effects or not the sale, which is the age interval that is more willing to buy. Also if the economical indexes like CPI, CCI, Euribor month rate and employment rate effect the sales or not. *CPI*-consumer price index which is a measure of the average change over time in the price

paid by urban households for a set of consumer goods and service, sometimes is referred to as a cost-of-living index (Institute for Research on Poverty, 2014). *CCI*-consumer confidence index is the outlook that consumers have towards the economy and their own personal finance situation, this is an important factor that determines the willingness of consumers to spend, borrow and save. *Euribor month rate* where *Euribor* stand for (Euro Interbank Offered Rate) which is a benchmark giving an indication of the average rate at which banks lend unsecured funding in the euro interbank market for a given period, it's the rate at which euro interbank term deposits are offered by one prime bank to another within the EMU(The Economic and Monetary Union) zone (European Money Market Institute, 2014)And the final economic indicator is indicator is *employment rate* which is the measure of the extent to which available labor resources(people available to work) are being used. (Organization for Economic Co-operation and developmnet, 2016)

The data shows that sales are made by phone and there are two types of contact type telephone and cellular which can also be part of analyses to see if it effects having sales or not and also the day of the week and the time when the call are made if they make any difference on the sales.

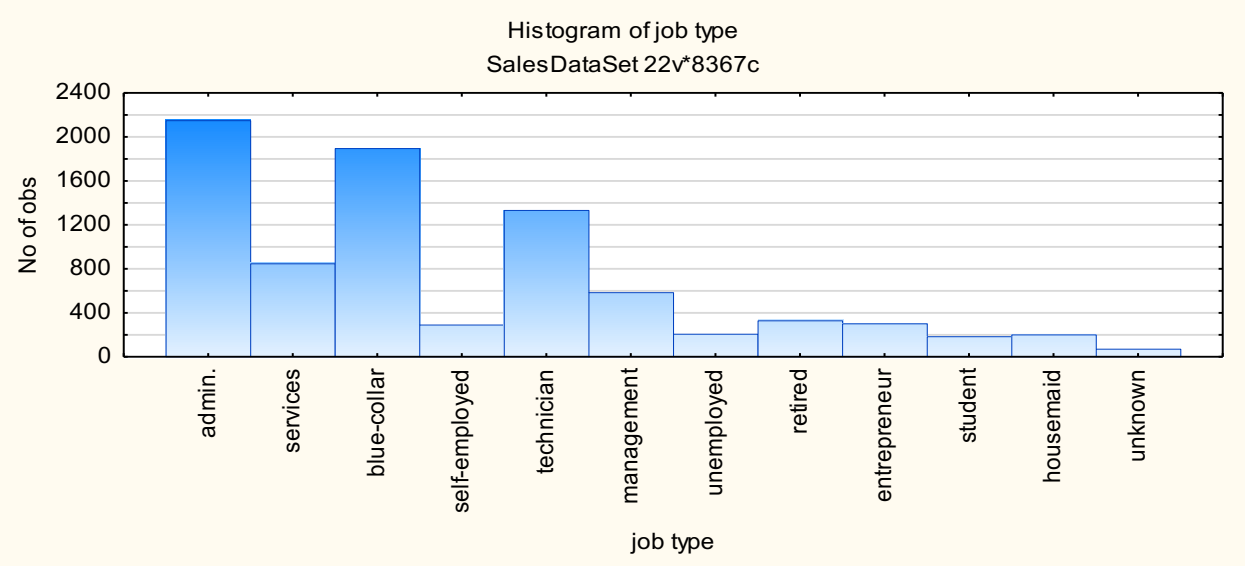
Based on the above analyses result it can then be built a proper model and also be done some sales prediction for the future. Statistica will be the software used throughout the analysis for applying several statistical functions on the dataset.

5.1 Data preparation

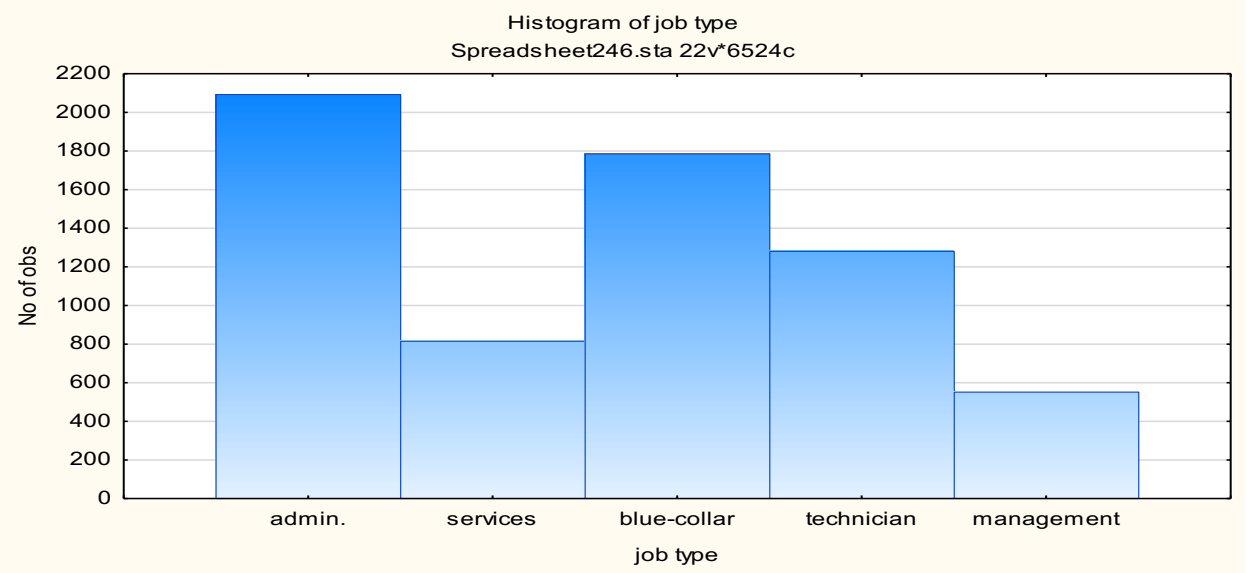
Data preparation consists on constructing a good dataset to be used for exploration and modelling. Cleaning of missing values, invalid values, duplicate cases or outliers and extreme values that don't contribute to the analysis and that may lead to wrong conclusions, are some of the techniques to construct a good dataset ready for further analysis. As the data are categorical and because outlier implies distance or position and categories don't have distance or positon what we can have is unusual values which means that a category has less than 5% of the total sample. When a given ***categorical variable has a level with less than 5%*** of the data, there are two ways to treat it we can either set the category to a new value by recoding or combining with other categories or mark the cases with that level as missing data and delete them from the dataset. In order for them not to affect the overall analysis of the dataset some of the extreme values are considered as missing value and are dismissed from the dataset, some of them will be recode when it will make sense the combination of several categories as we will see from the

respective Bar Charts. Statistica offers convenient commands for processing missing data and recode the extreme as they don't contribute to the analysis. After treating most of the extreme values as missing values we end up having sparse data which in this case need to be removed to have a better analysis of the dataset at the end.

Bar Charts 1 and 2 shows the variable Job Type before and after the extreme values are examined. We can notice that several categories like "student", "household", "entrepreneur"," retired", "un-employed" and "self-employed" are dismissed from the dataset as they had a level of less than 5 % of the cases, as well as the "unknown" category which is seen as missing values.

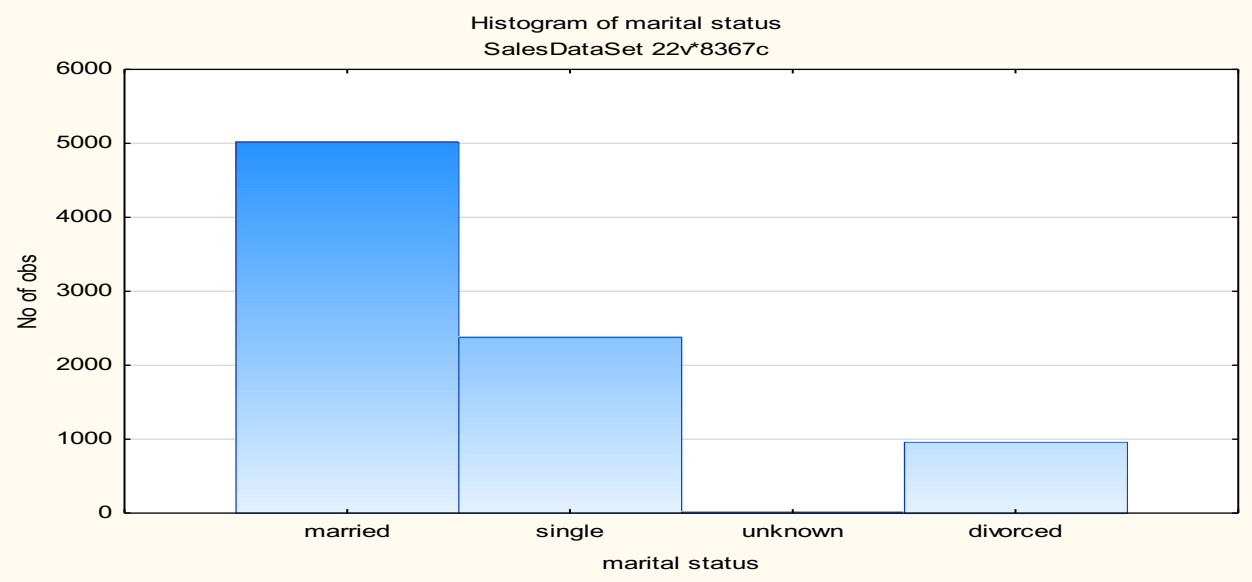


Bar Chart 1. Job type before analysis [source: own]

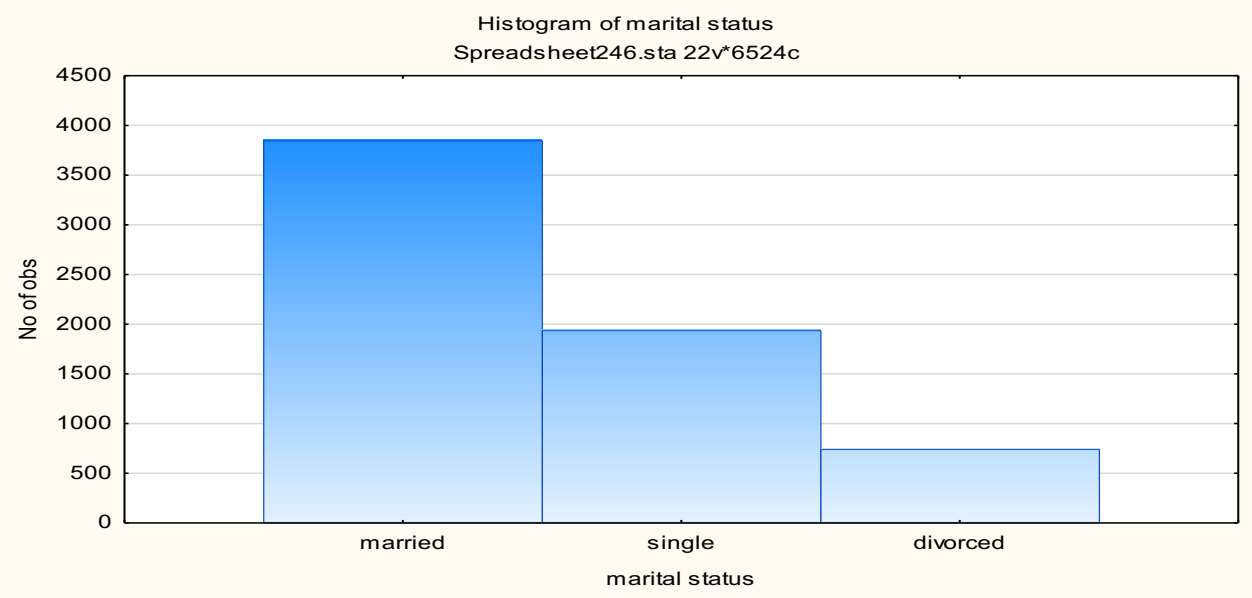


Bar Chart 2. Job type after removing extreme values [source: own]

In case of marital status as shown from the Bar Chart 3 the "unknown" category will be treated as missing values. As there are enough data and the missing values seem to be random therefore the unknown values for most of the variables were considered as missing values and were dismissed. As they do not add anything to the analysis, it gives us freedom to remove them from the dataset and we will have a variable "marital status" with only three categories as shown in Bar Chart 4.

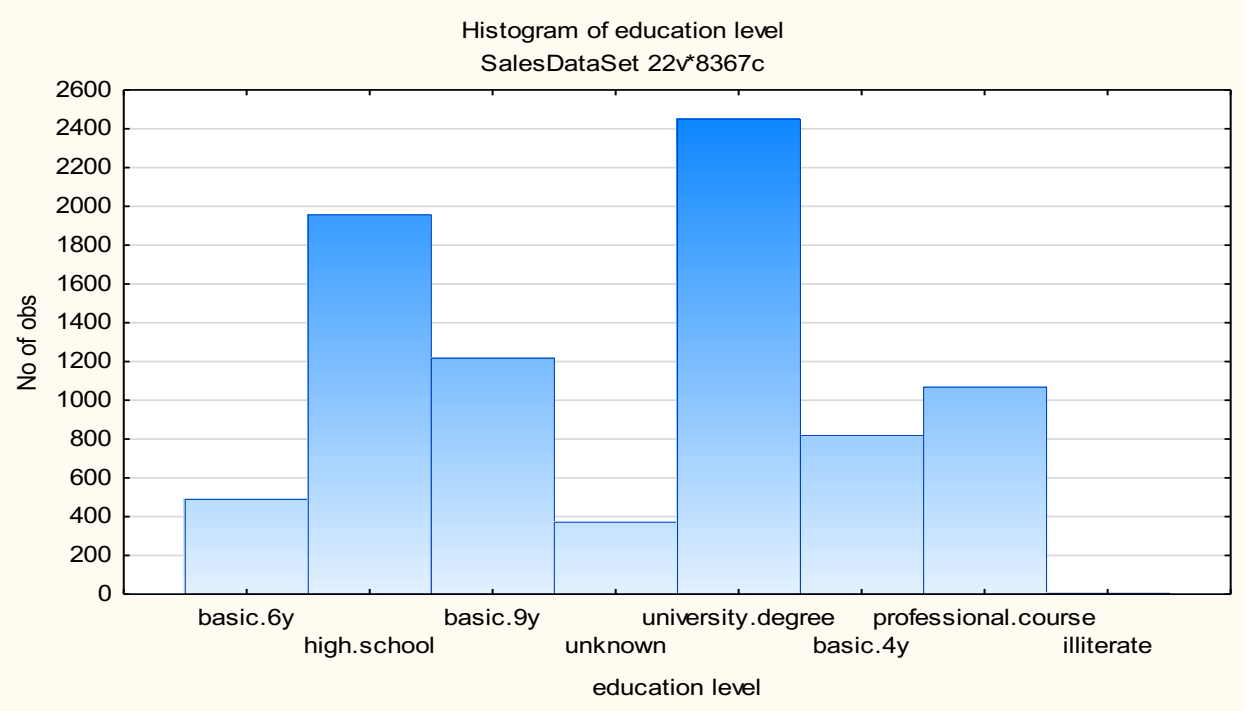


Bar Chart 3.Marital status before analysis[source: own]

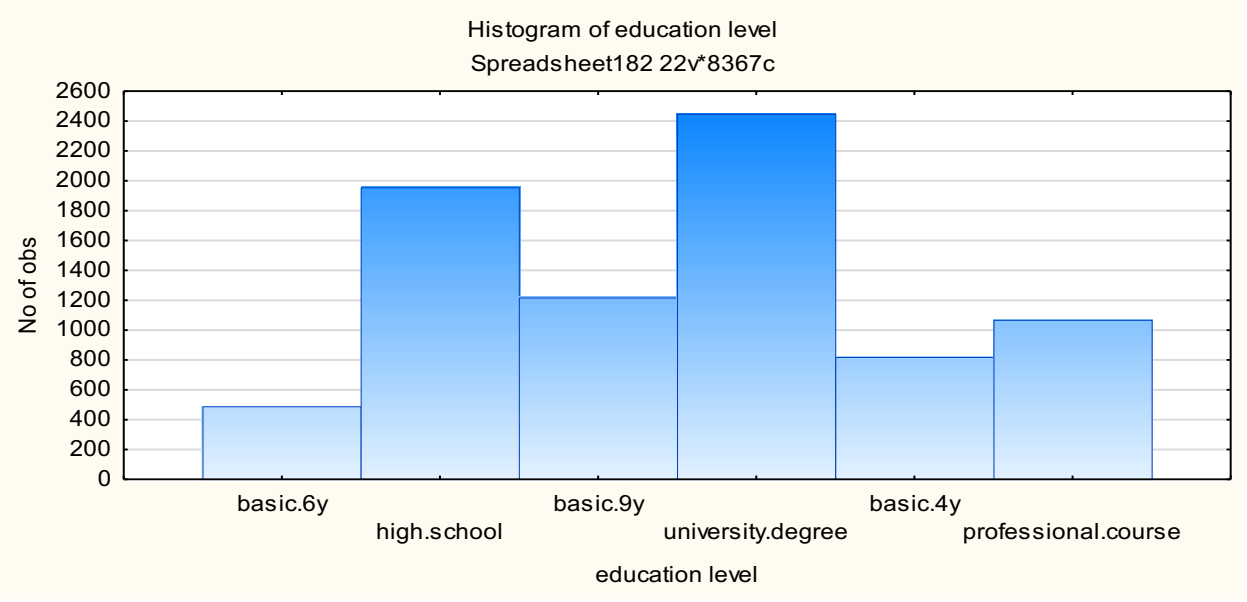


Bar Chart 4.Marital status after removing extreme values [source: own]

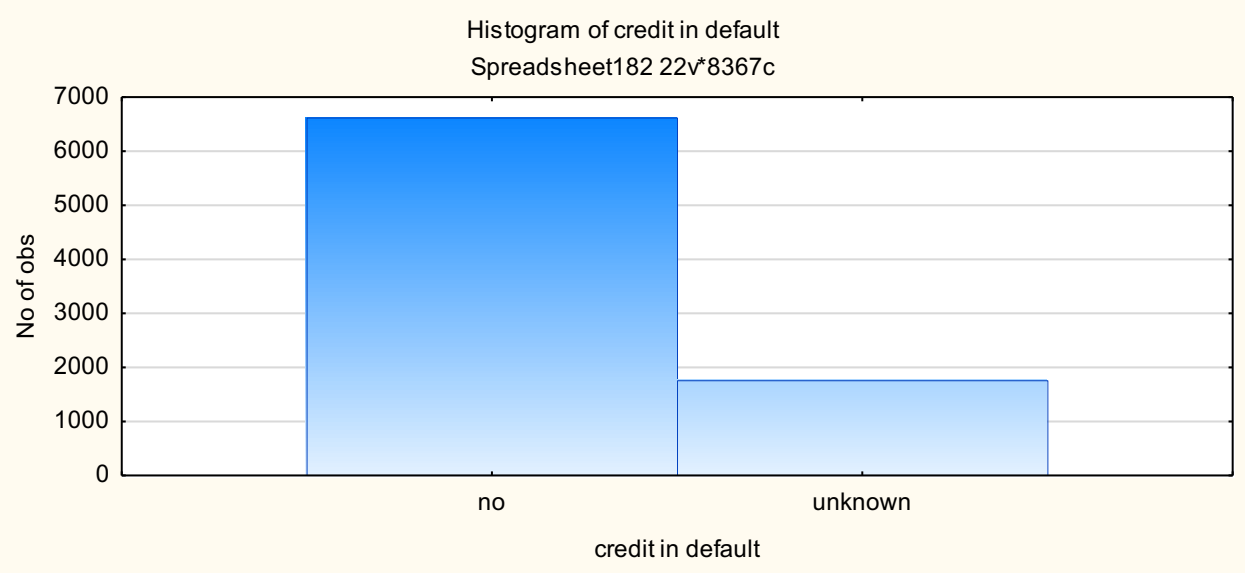
The education level Bar Chart 5 show that the illiterate category and the unknown category for mentioned before will be removed from the dataset. Illiterate category makes less than 5 % of the cases and the unknown category is considered as missing values. After removing the mentioned categories education level will be represented by categories that are shown in Bar Chart 6.



Bar Chart 5.Educational level before analysis [source: own]

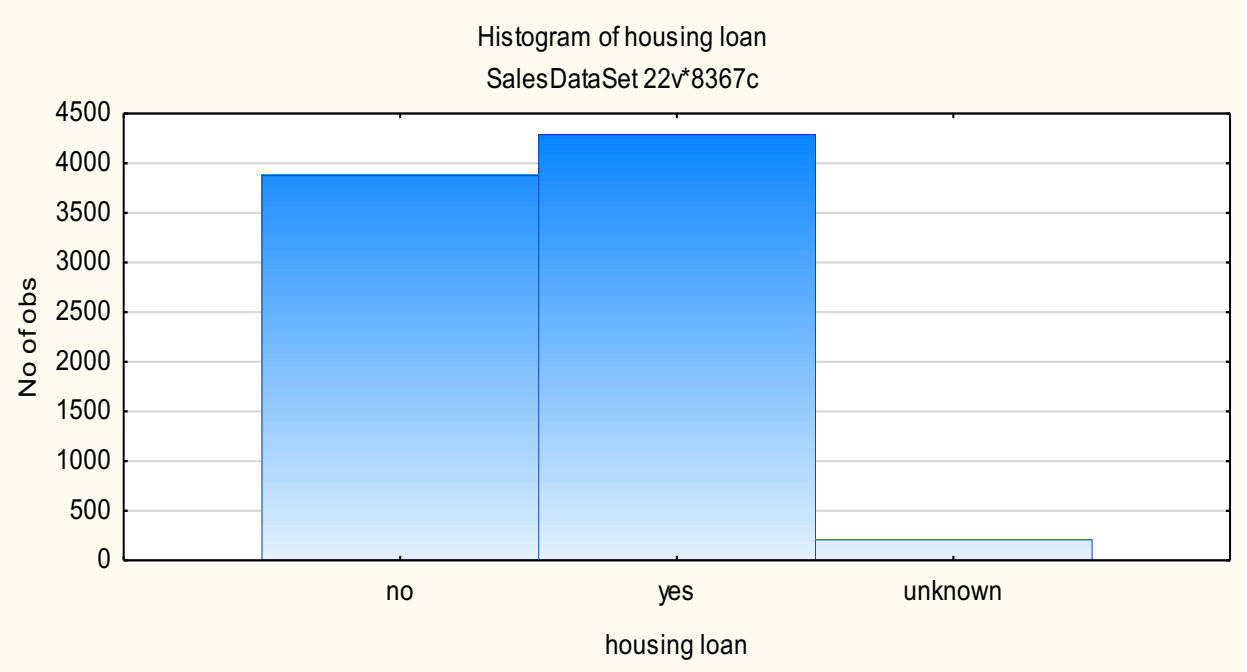


Bar Chart 6.Educational level after removing extreme values [source: own]

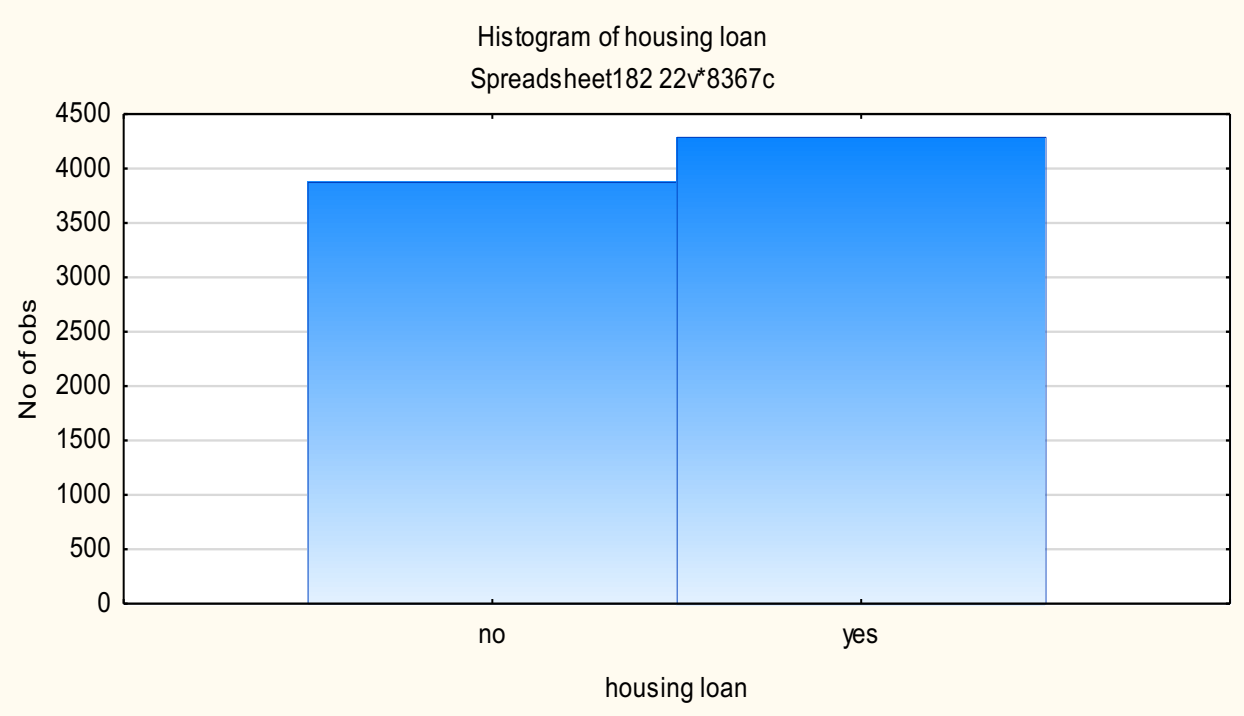


Bar Chart 7.Credit in default [source: own]

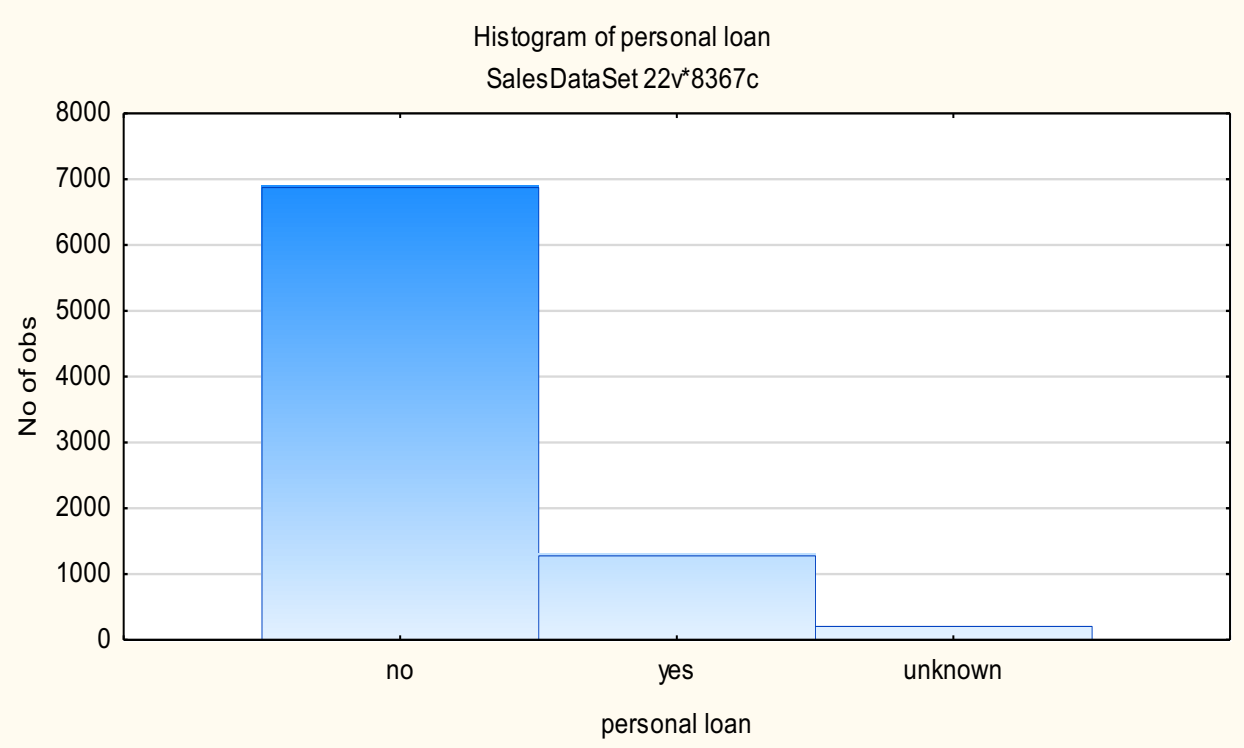
From Bar Chart 7 the category 'yes' was an extreme value and the category unknown will be considered as missing values as the variable is categorical which will lead us to a variable with only one category so we will remove the variable from the analysis. The unknown category is considered again as missing values for variables Housing Loan and Personal Loan as shown from Bar Chart 8 and 9 for housing loan and Bar Chart 10 and 11 for personal loan and therefore it will be removed.



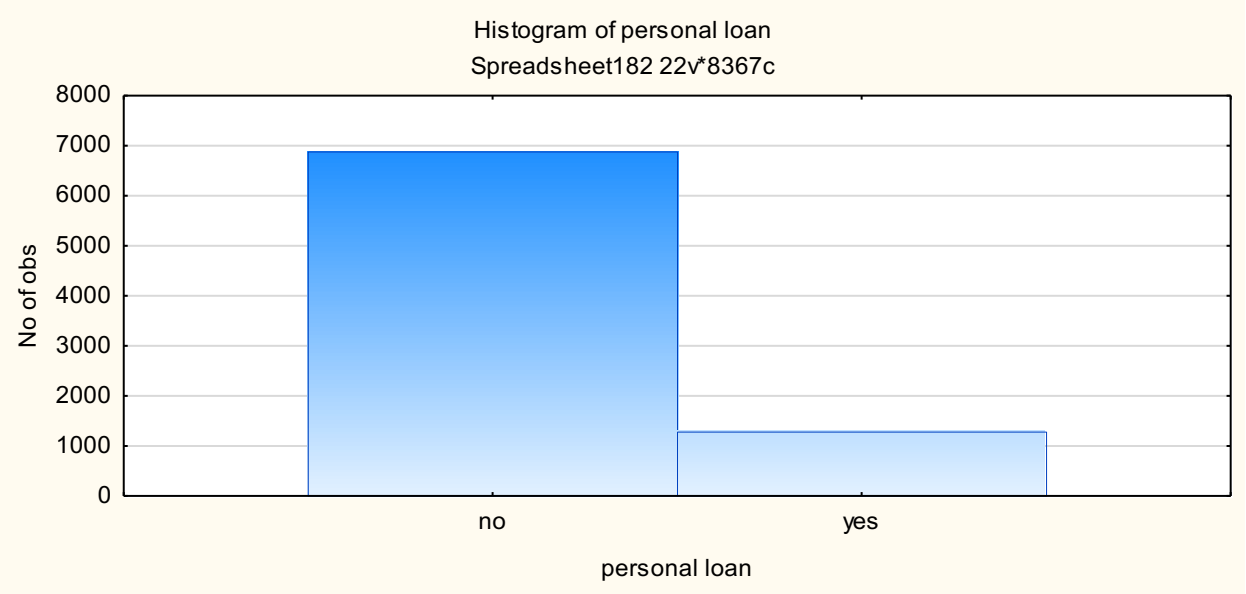
Bar Chart 8.Housing loan before analysis [source: own]



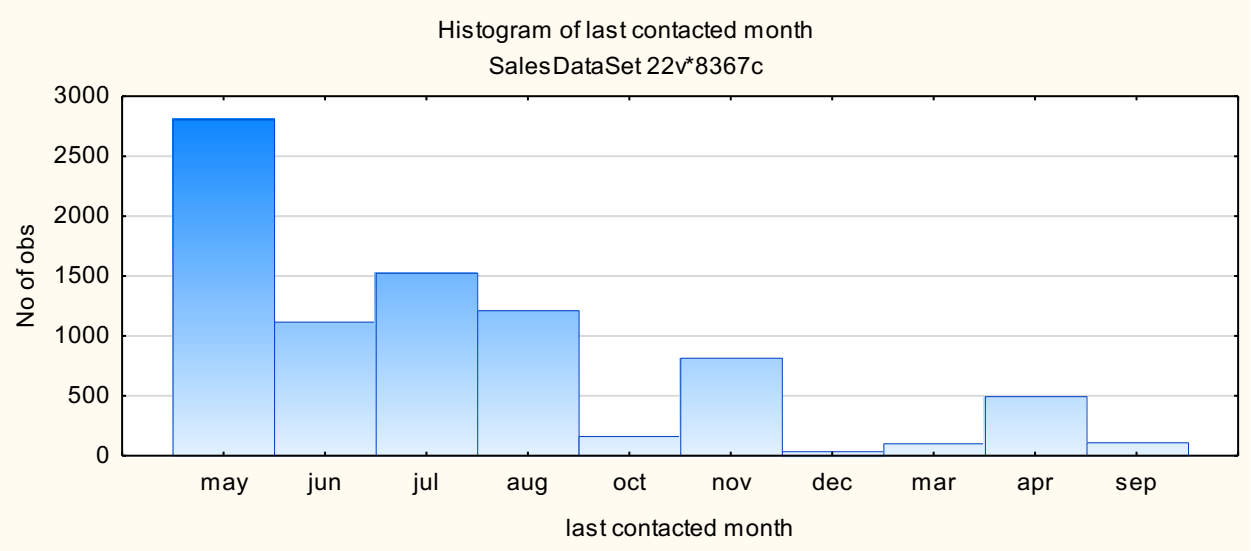
Bar Chart 9.Housing loan after removing extreme values [source: own]



Bar Chart 10.Personal loan before analysis [source: own]

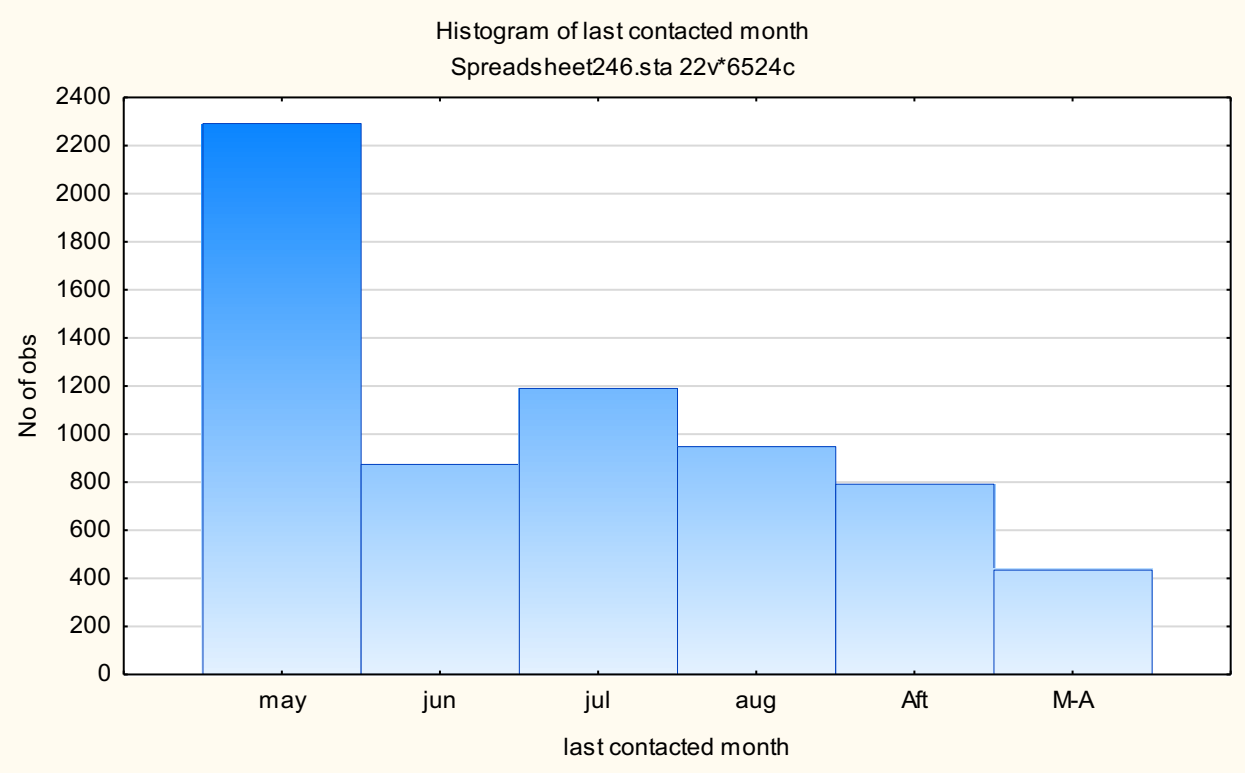


Bar Chart 11.Personal loan after removing extreme values [source: own]

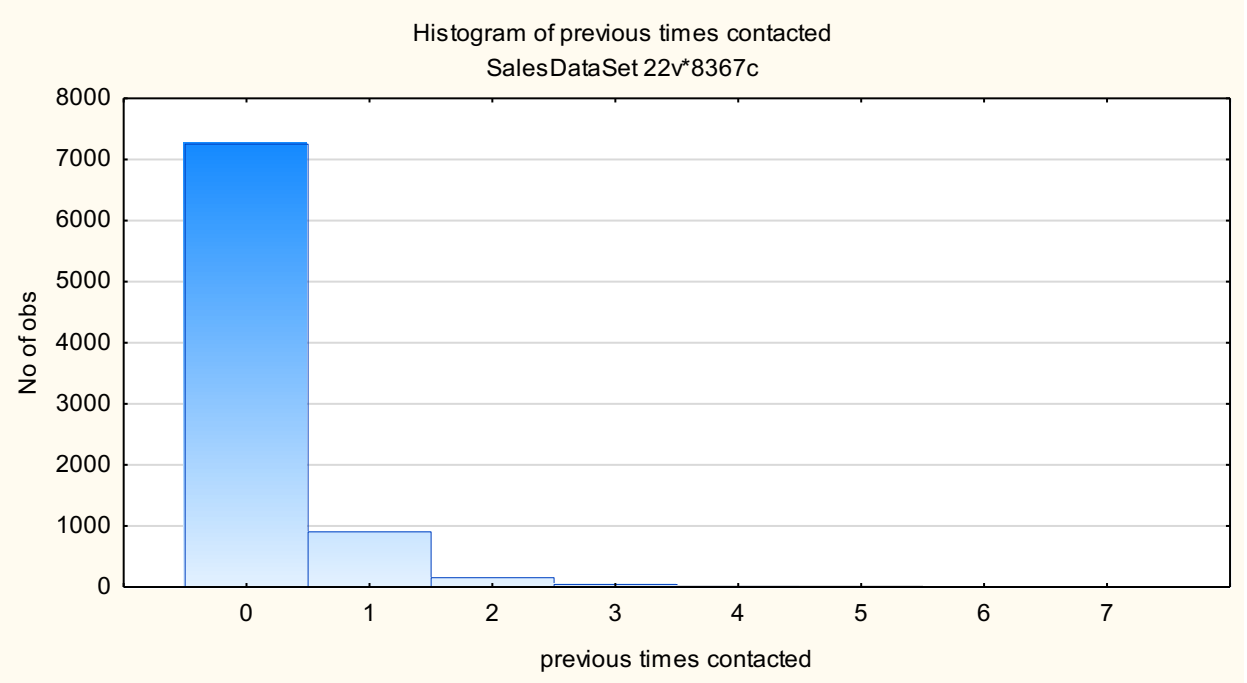


Bar Chart 12.Last contacted month before analysis [source: own]

In the Bar Chart 12 of 'last contacted month' we will combine the month March(mar) and April (apr) in one category M-A and September(sep), October (oct), November(nov) and December(dec) in one category Aft(after September). And Bar Chart 13 this new categories added after recode.

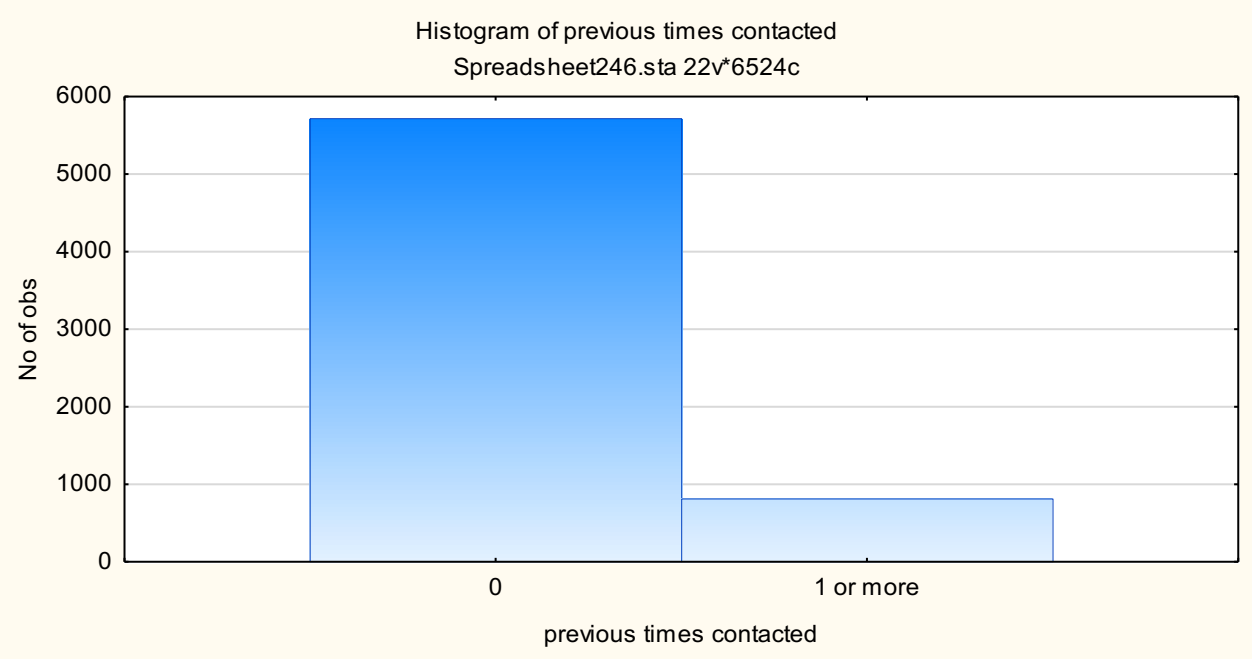


Bar Chart 13.Last contacted month after the recode [source: own]

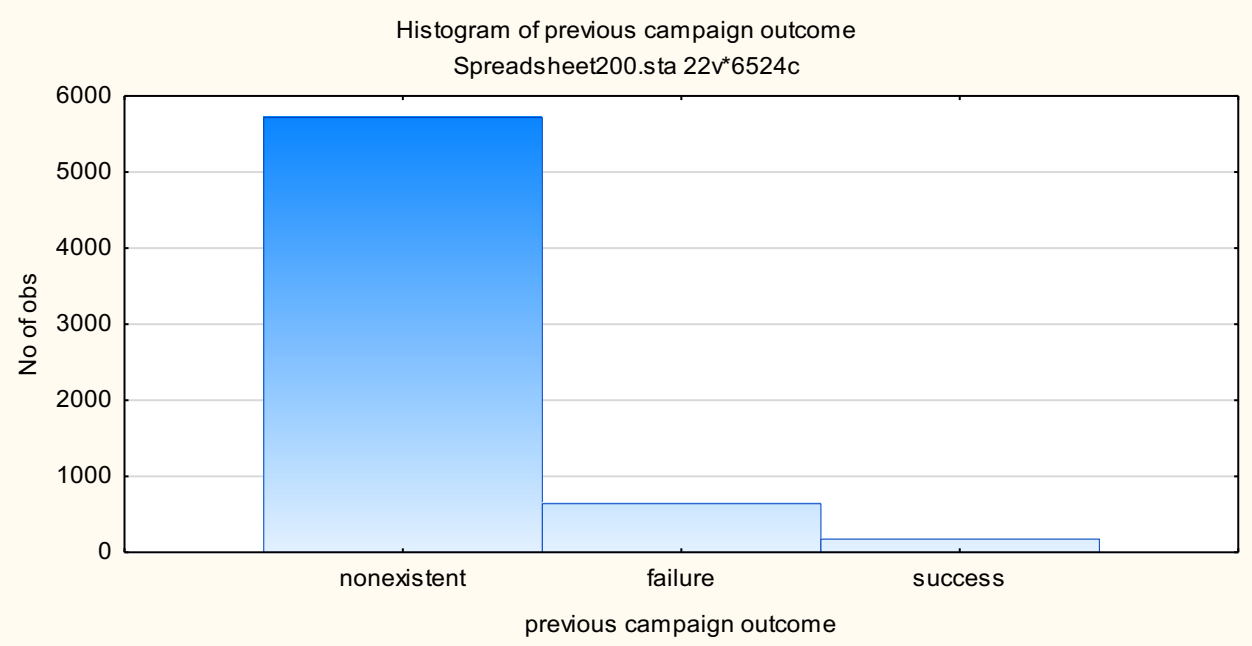


Bar Chart 14.Previous times contacted before analysis [source: own]

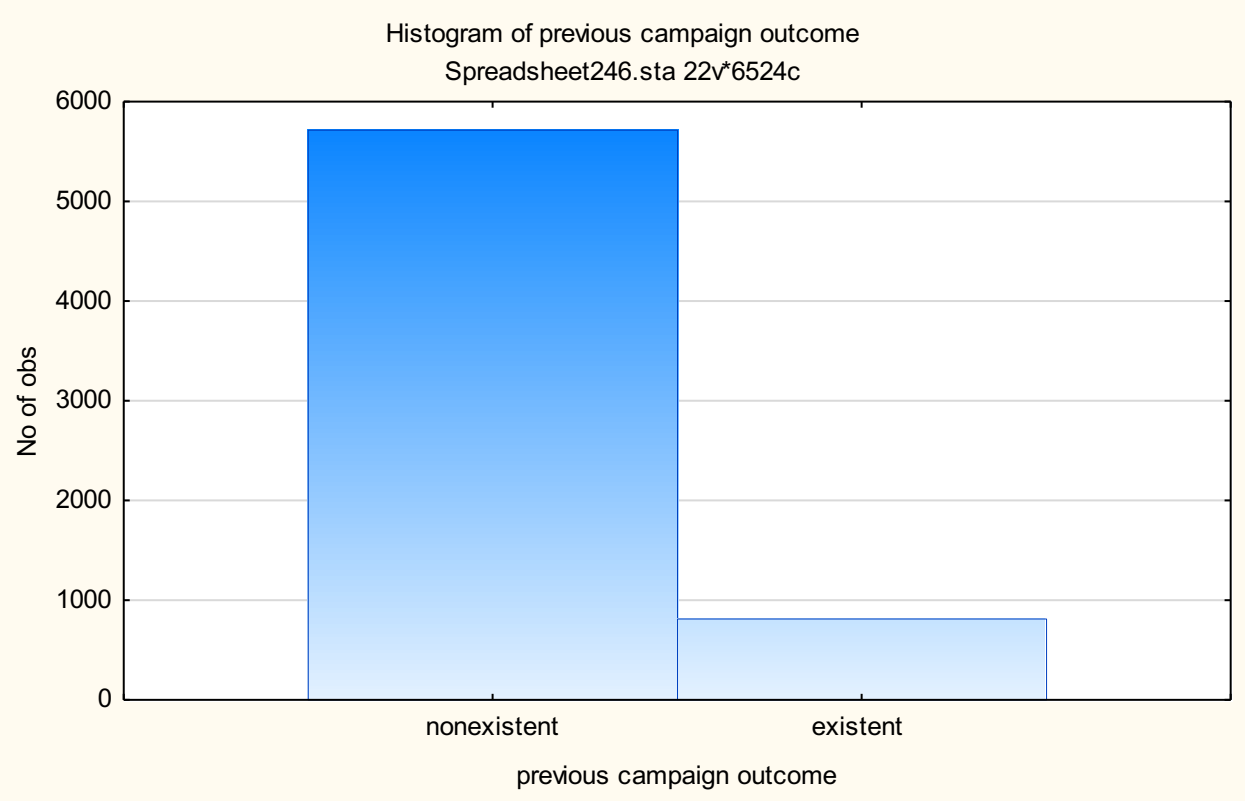
From Bar Chart 14 we can notice from the bar chart of previous times contacted that there are very few cases in the case of categories 1, 2, 3, 4, 5 ,6 and 7 so we are going to combine all these categories in one category called '1 or more' as shown in Bar chart 15.



Bar Chart 15.Previous times contacted after the recode [source: own]

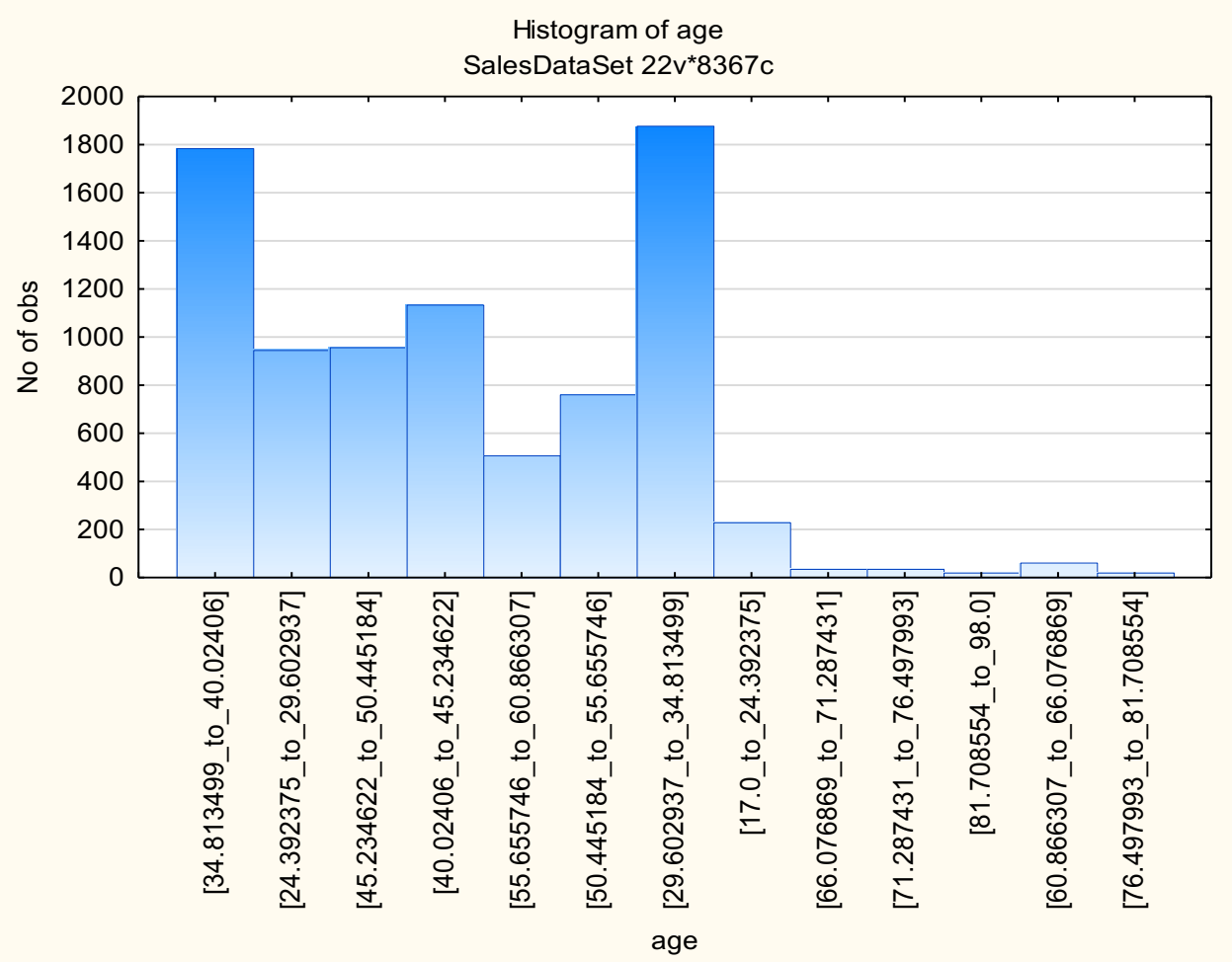


Bar Chart 16.Histogram of previous campaign outcome before analysis [source: own]

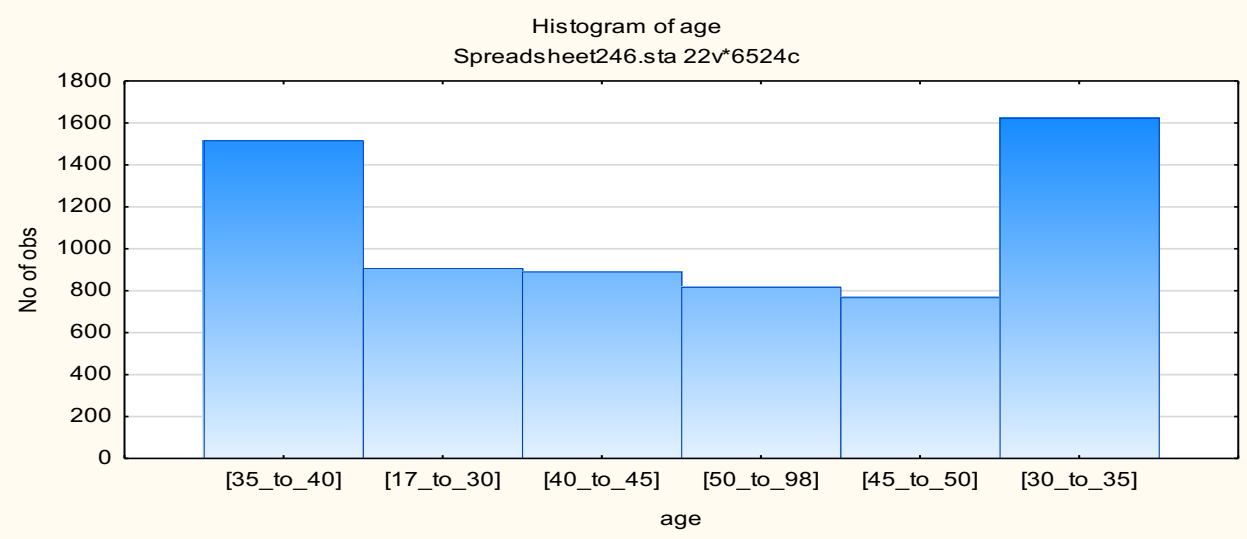


Bar Chart 17.Histogram of previous campaign outcome [source: own]

From Bar Chart 16 we can see that categories "failure" and "success" have less then 5% of the case and therefore as seen in Bar Chart 17 these two categories will be combined in one category named" existent". Age chart, in Bar Chart 18 we can see that there are very few cases in the last categories starting from the category [50.445184 _to_ 55.655746] and onwards so the variable will be recoded. We are going to combine all of them in one category [50_to_98.0]. As the way the ages are written doesn't make much sense we are going to rename the categories by rounding the ages as seen in Bar Chart 19 where can be seen the last changes made to variable "Age" after recode and rounding of categories.

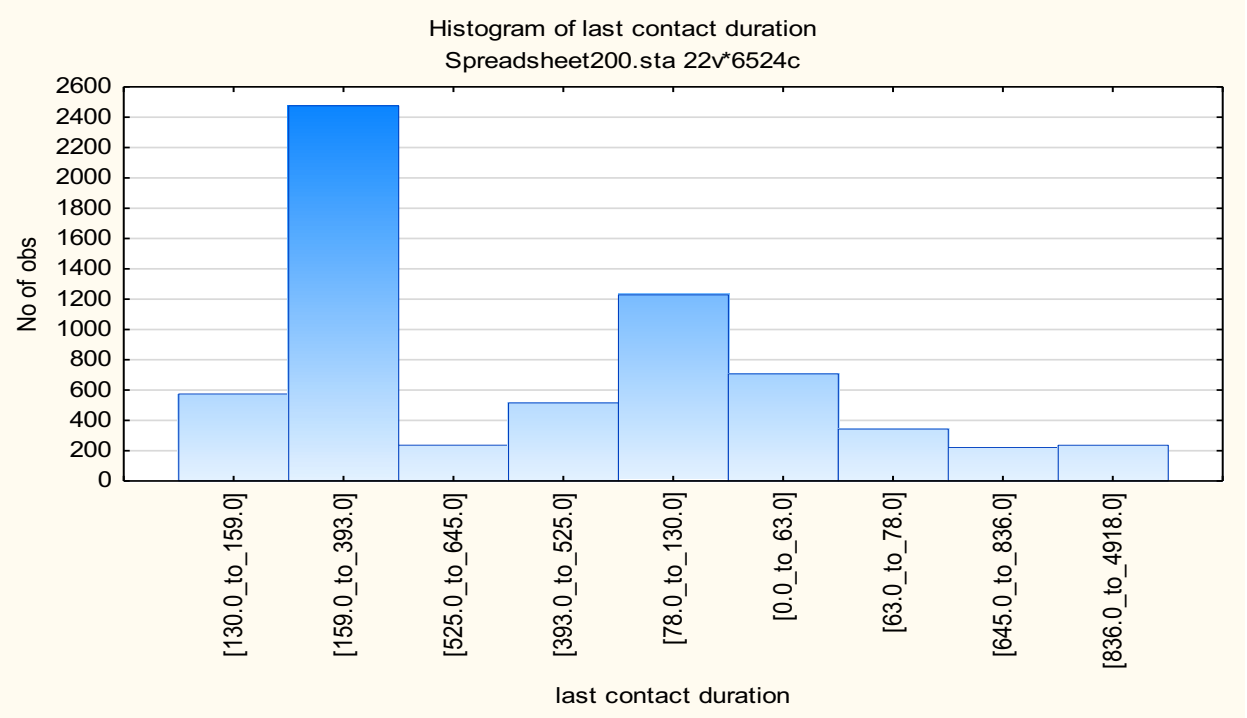


Bar Chart 18.Age before analysis [source: own]

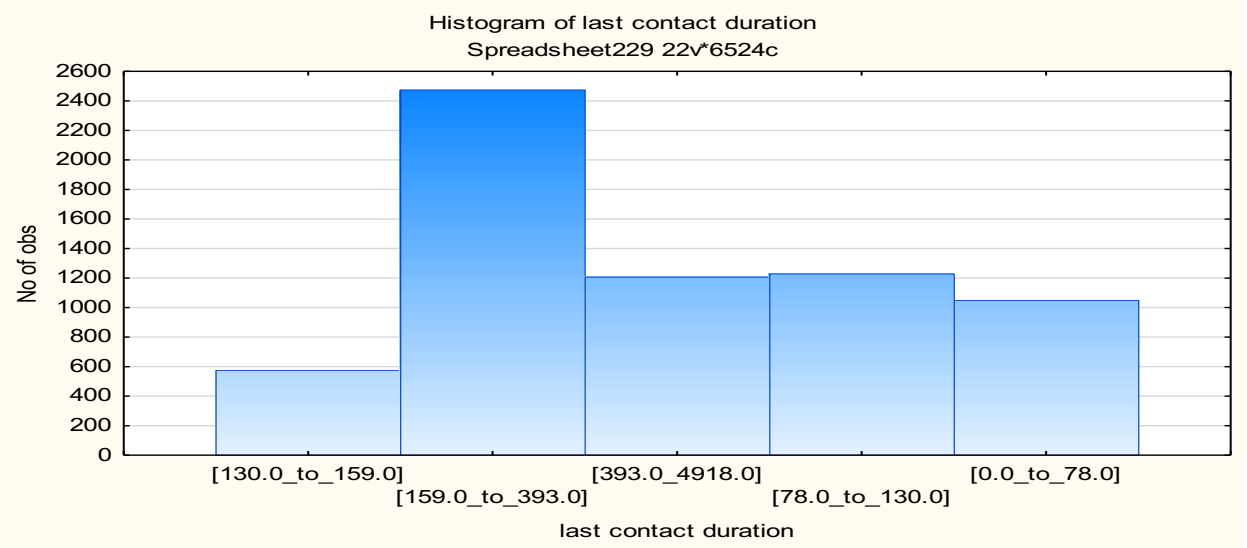


Bar Chart 19.Age after combining the last categories [source: own]

Bar Chart 20 shows that many categories in “last contact duration” have less than 5 % of the cases and therefore we will combine some of the categories in one category by ending up in only 5 categories as shown in Bar Chart 21.



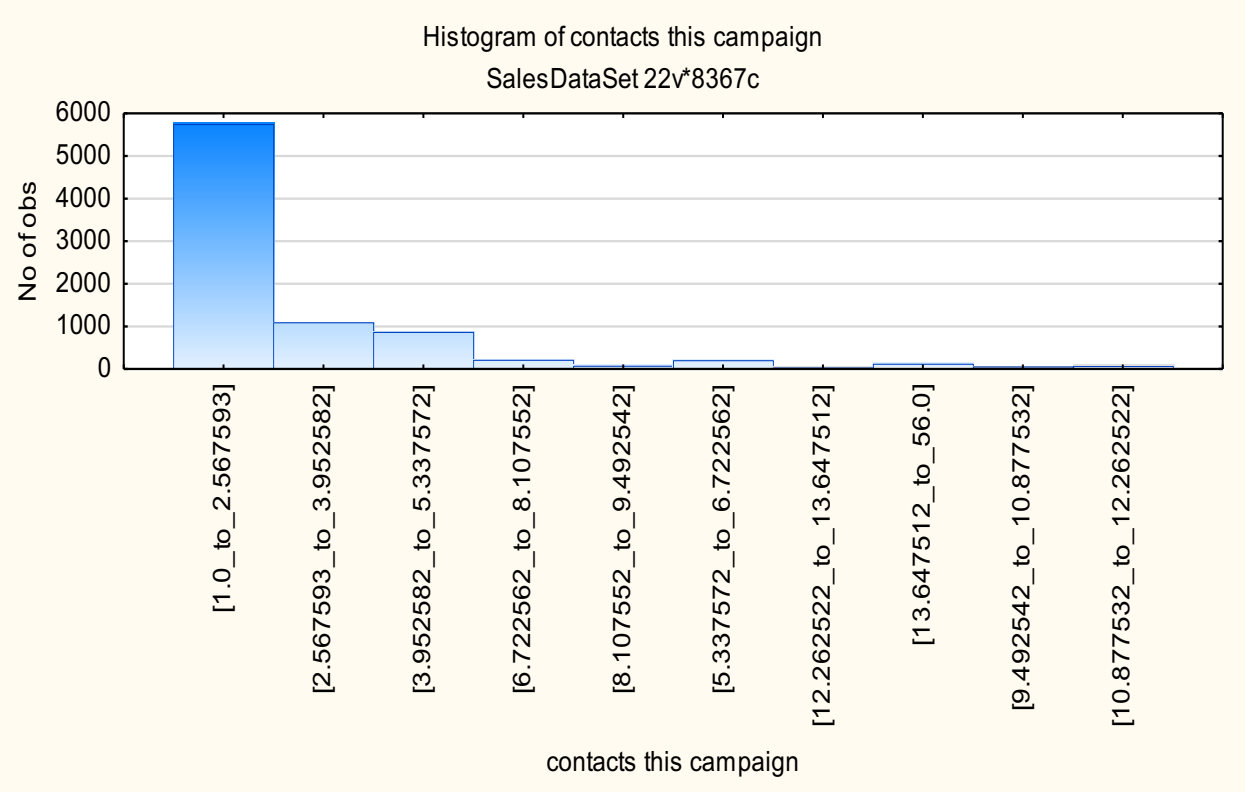
Bar Chart 20.Last Contact duration before analysis [source: own]



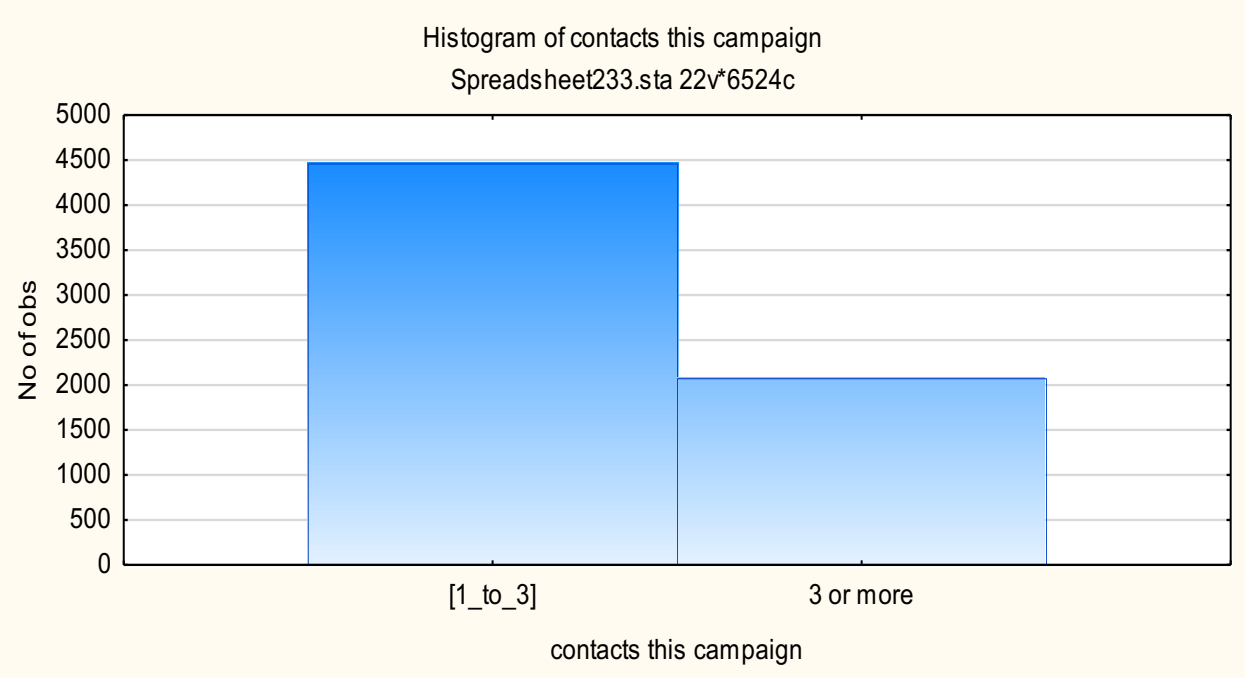
Bar Chart 21.Last contact duration after recode [source: own]

Bar Chart 22 for variable “contacts this campaign” shows the same problem as with variable “last contact duration”, many categories having less then 5% of the total cases. Bar Chart 23

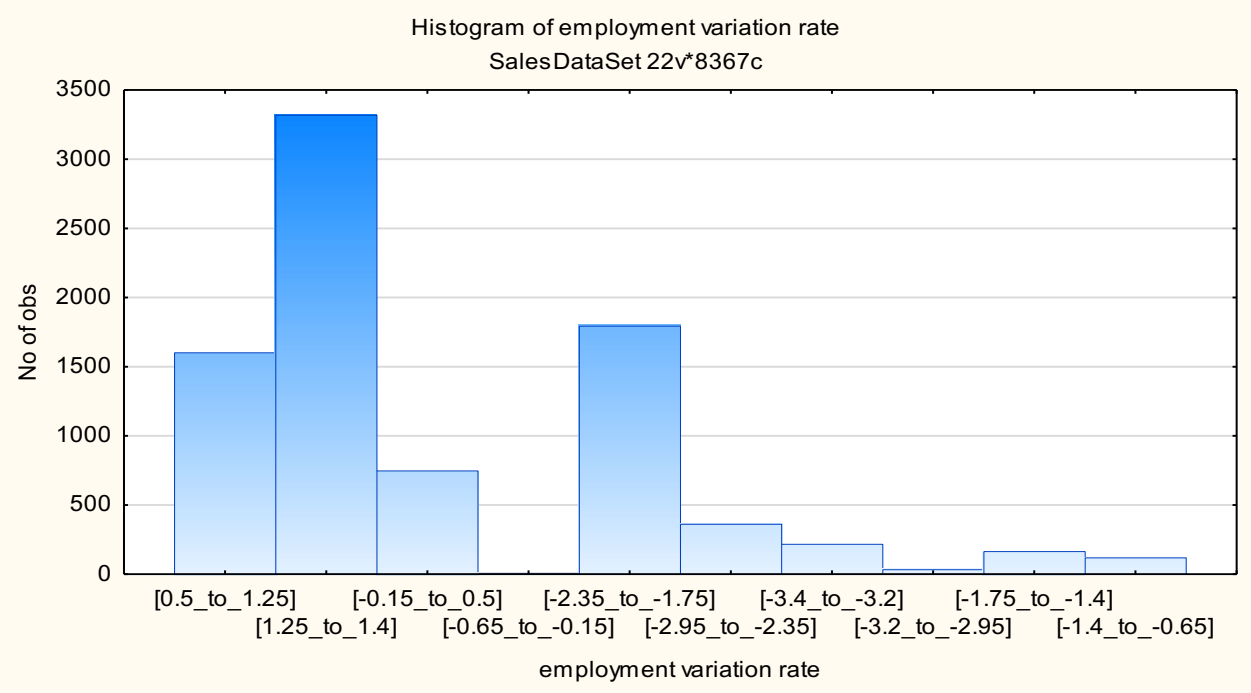
shows how some categories are combined in the category [1_to_3] and the rest in the category “3 or more” , as well the values of the intervals are rounded for easier understanding.



Bar Chart 22.Contacts this campaign before analysis [source: own]

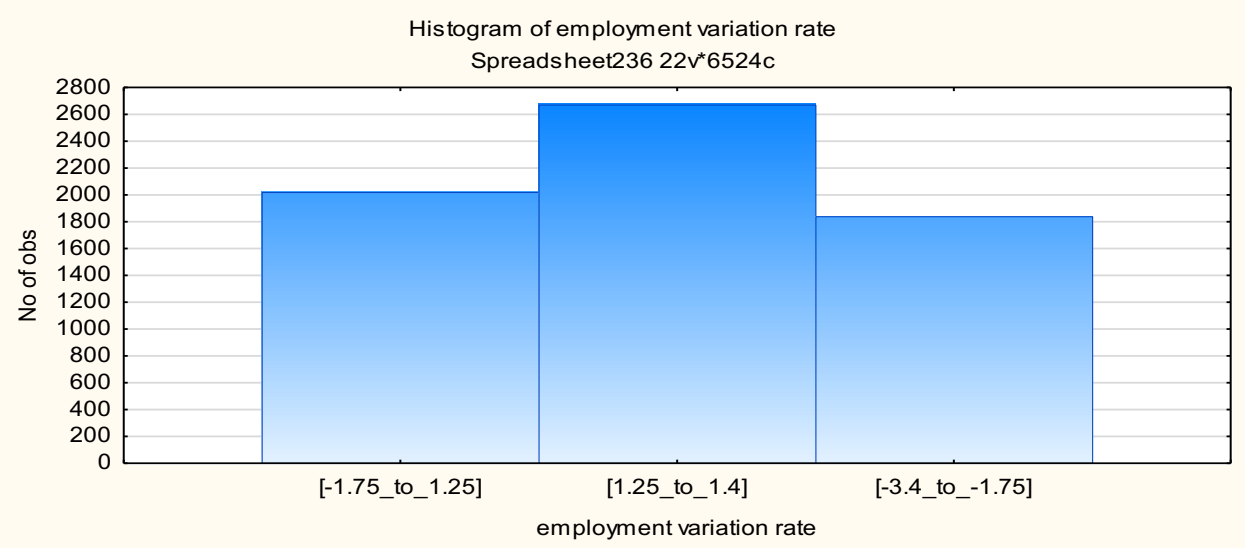


Bar Chart 23.Contacts this campaign after recode [source: own]



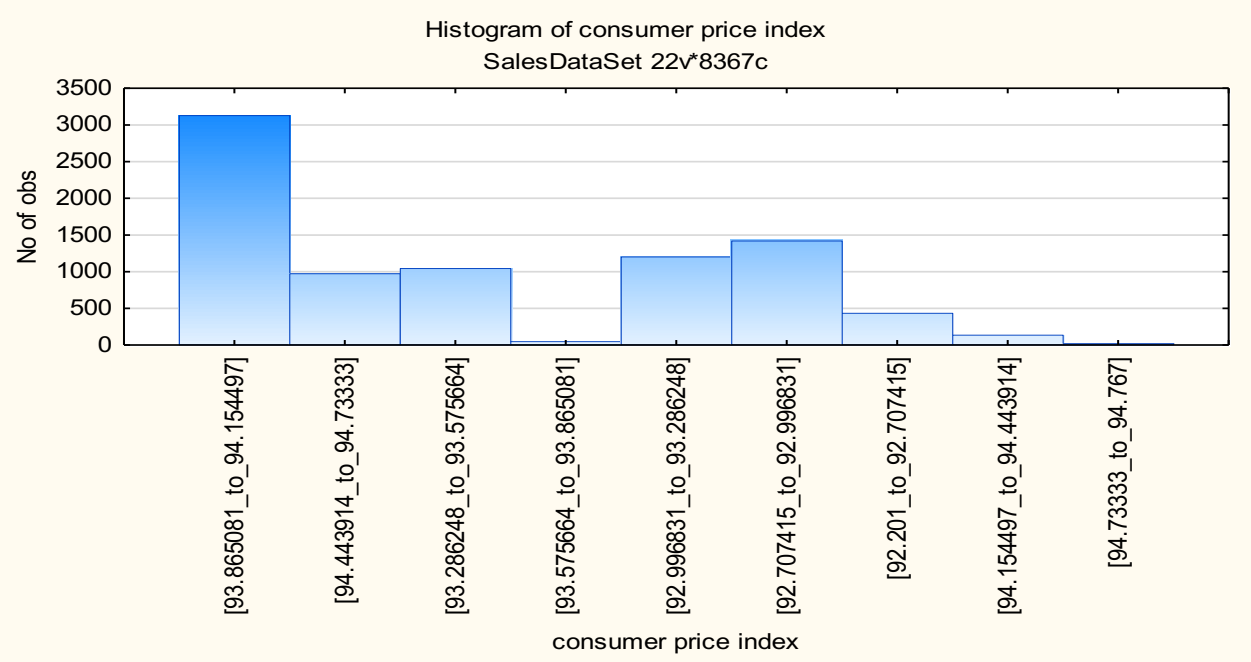
Bar Chart 24.Employment variation rate before analysis [source: own]

Bar Chart 24 and 25 show that categories [0.5_to_1.25], [-0.15_to_0.5], [-0.65_to_-0.15], [-1.4_to_-0.65] and [-1.75_to_-1.4] for variable employment variation rate will be combined in one category [-1.75_to_1.25] and the categories [-2.95_to_-2.35],[-3.2_to_-2.95] and [-3.4_to_-3.2] ,[-2.35_to_ -1.75] will be combine in another category [-3.4_to_-1.75].

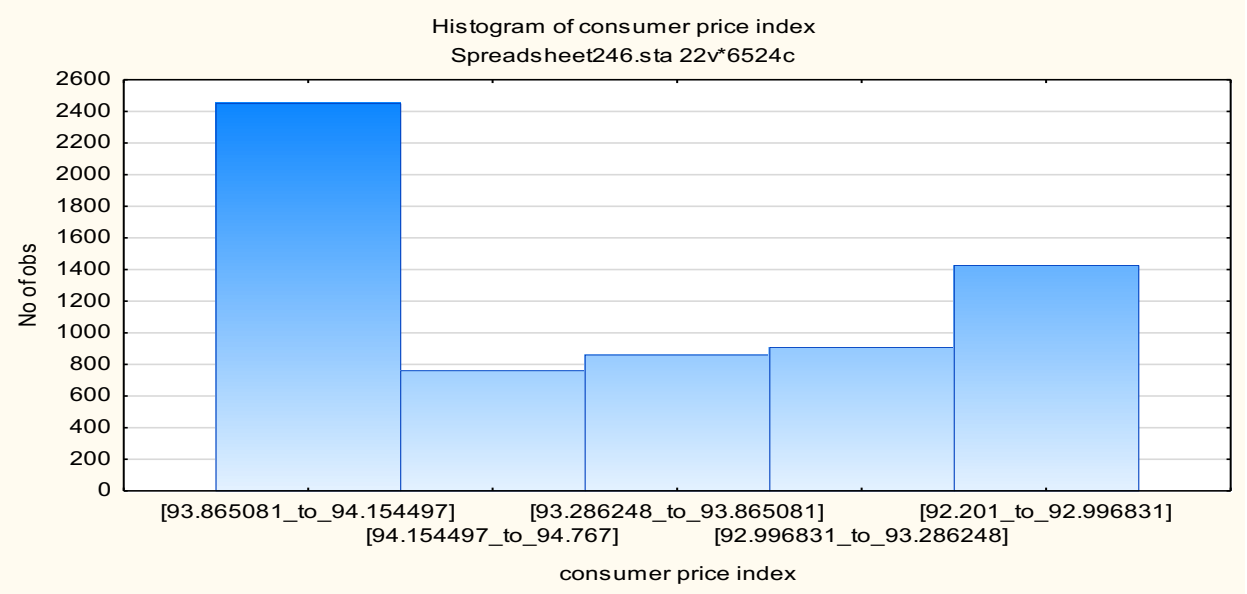


Bar Chart 25.Employment variation rate after recode [source: own]

Bar chart 26 shows variable ‘consumer price index” before analysis and we can clearly notice that some of the values have less than 5 % of the cases. All the categories having less than 5 % with be combined between them or with other categories having more than 5 % in order to keep the continuity of the intervals and at the end we will get 5 categories as seen in bar chart 27.

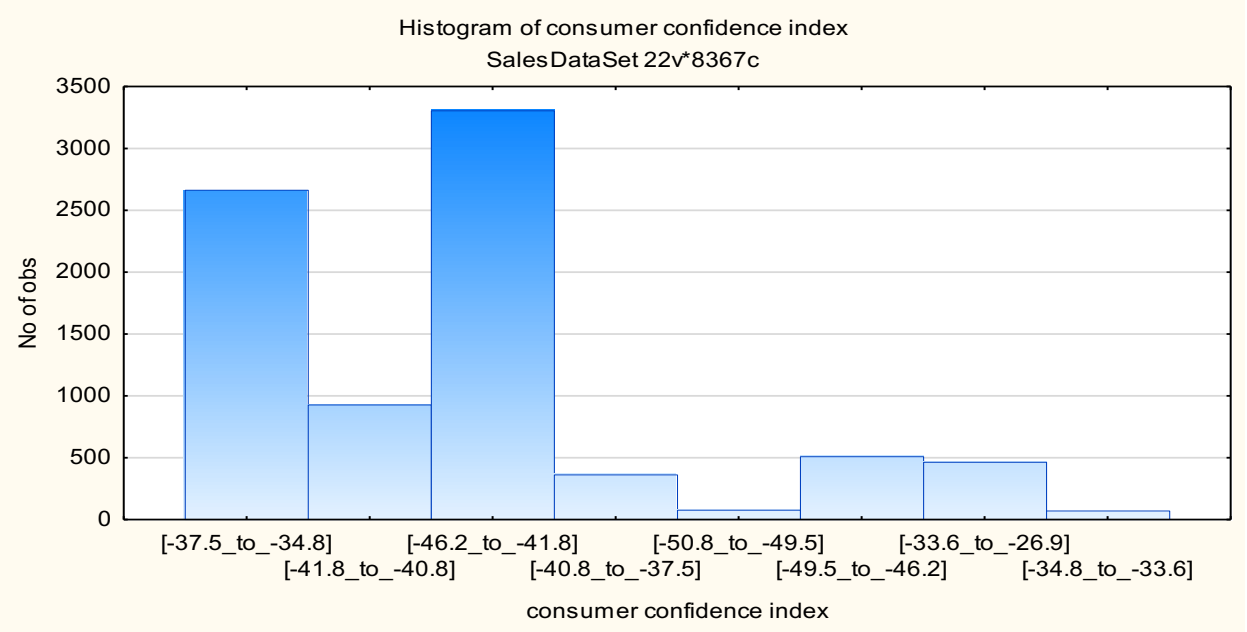


Bar Chart 26.Consumer price index before analysis [source: own]

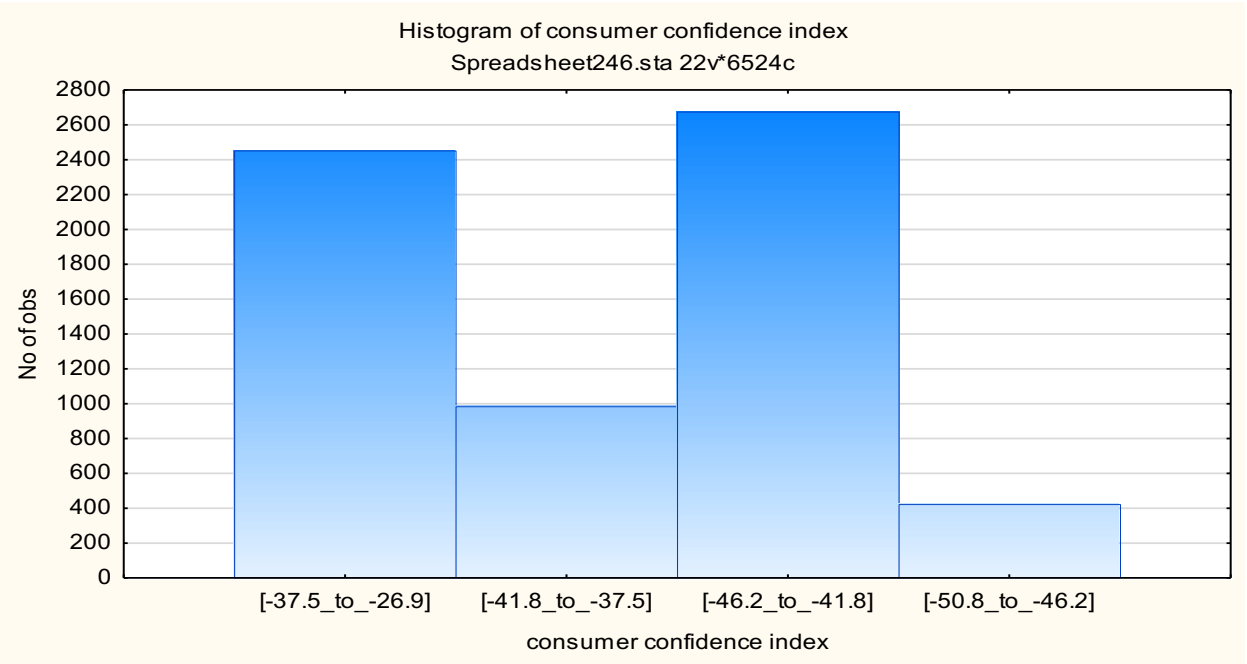


Bar Chart 27.Consumer price index after recode [source: own]

The same analogy as with the consumer price index will be applied to "consumer confidence index". Bar chart 28 shows variable "consumer confidence index" before analysis and we can see that 5 categories have less than 5 %. As we did previously with consumer price index, in order to keep the continuity of the intervals the categories with be combined in the end in four categories as shown in Bar chart 29.

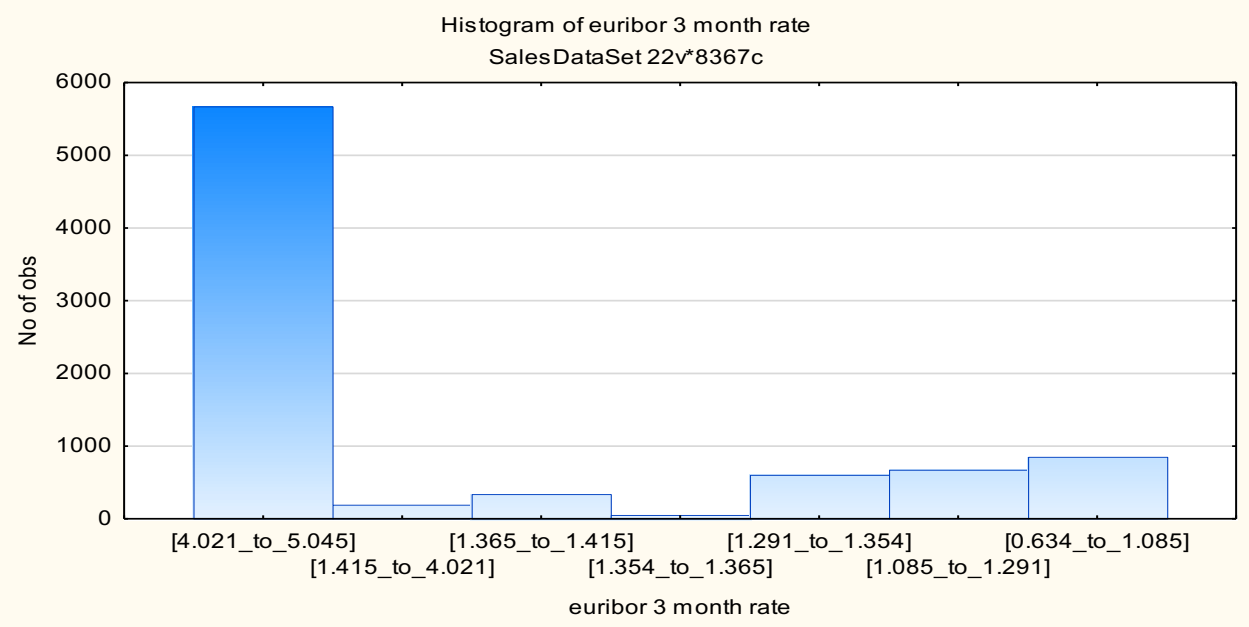


Bar Chart 28.Consumer confidence index before analysis [source: own]



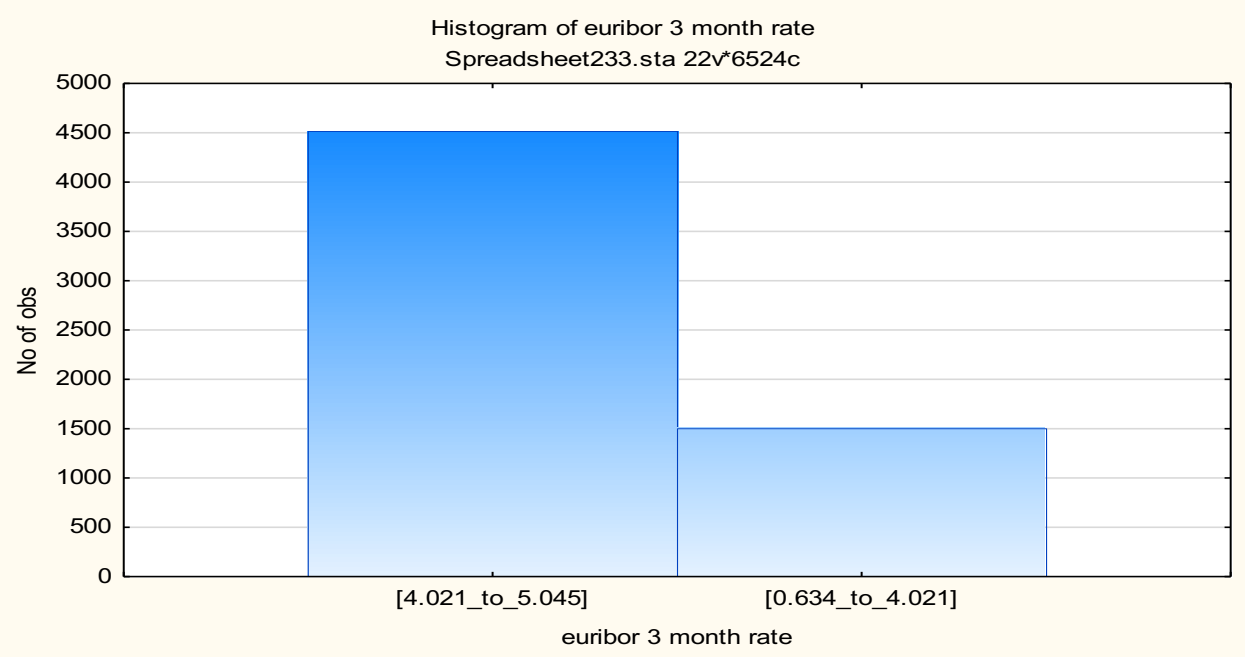
Bar Chart 29.Consumer confidence index after recode [source: own]

Bar chart 30 shows the euribor 3 month rate and we can see a lot of extreme values which will be recoded in order to have categories which do not contain extreme values.



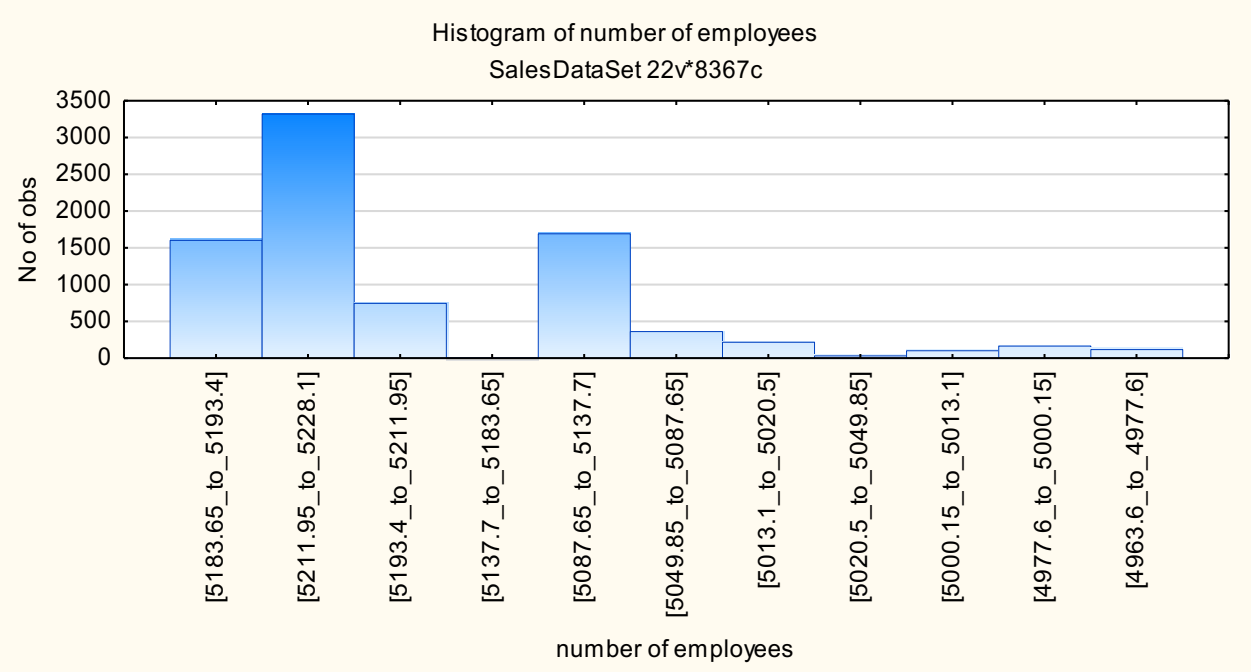
Bar Chart 30.Euribor 3 month rate before analysis [source: own]

Bar chart 31 shows how 6 categories of the euribor month rate are combined in one category [0.634_to_4.021] and the variable ends having now only two categories that do not have any more extreme values.

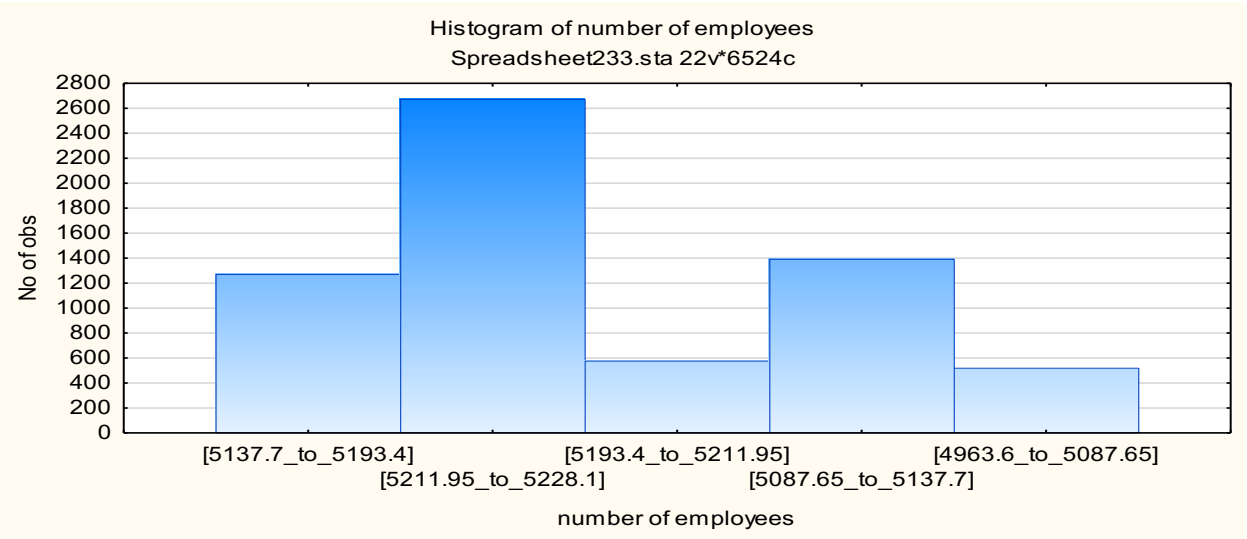


Bar Chart 31.Euribor 3 month rate recode [source: own]

Bar Chart 32 shows us variable number of employees where we can notice that different categories representing extreme values, meaning less then 5% of all cases. As we did previously with most of the variables, in order for this extreme values not to effect the overall analysis we will combine them together to have categories that have more than 5 % of the case. At the end after recode there will be only 5 categories for variable "number of employees" as shown in Bar chart 33 instead of 11 categories shown in bar chart 32.



Bar Chart 32.Number of employees [source: own]



Bar Chart 33.Number of employees after recode and removing the remaining extreme values [source: own]

As from all the changes made there with be missing values showed in the dataset we with filter the sparse data. Sparse data are variables or cases with too many missing values.

5.2 Data Exploration

After cleaning the dataset now we can go on with the exploration of the dataset to understand our data better. We will describe the data by the means of statistical and visualization techniques in order to bring important aspects of the data into the focus for further analysis. Both univariate and bivariate analysis with be applied to understand the behavior of a single variable and also the relationship among variables in the bivariate analysis. As our dataset is mostly categorical we will be limited to use the frequencies tables, combination charts and **Pearson Chi-square** test to analyze the relationship among variables. In order to use Pearson there are two assumptions: the *first one* is the random selection of the sample and *the second one* is the expected frequencies are not very small. The reason for this second assumption is that Chi-square tests the underlying probabilities in each cell and when the expected cell frequencies fall, for example below 5, the probabilities cannot be estimated precisely enough.

Our focus with be to check the relationship that all the variables will have with the variable sales as we want to know what leads to shorter sales cycle and so on. Following we will show for each variable 2-way summary table and the statistics table following that one. Based on the Chi-square and the p-value we will check the follow hypothesis for every variable for α=0.05:

H_0: Variable x and variable sales are not associated.

H_a: Variable x and sales are associated.

As Table1 shows that the expected frequencies are all >5 we can use the Chi-Square to check the relationship among job type and sales. With a Chi-squared test statistic equal 9.7 and an associated p=0.046< 0.05 which mean that we will reject the null hypothesis H_0 so we can conclude that the Job type and Sales are associated which means that Sales are affected by the job type of the customers.

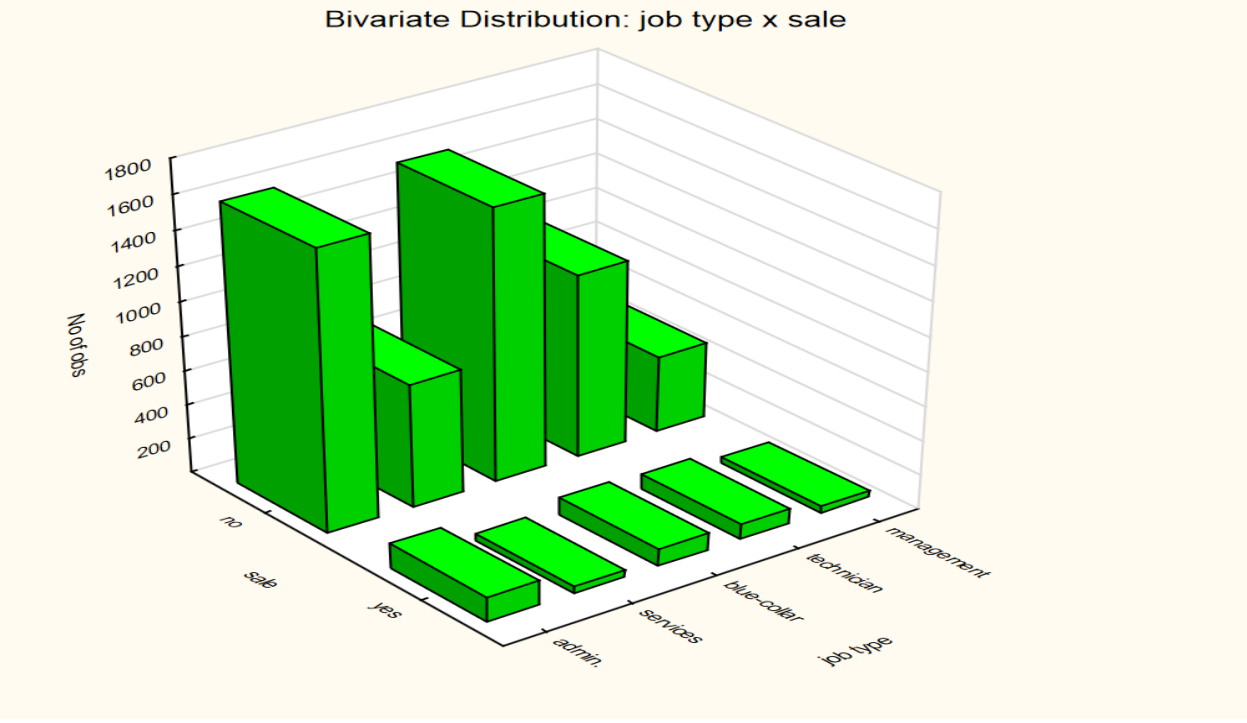
job type	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
	sale no	sale yes	Row Totals
admin.	1641.109	123.8915	1765.000
services	708.513	53.4874	762.000
blue-collar	1559.286	117.7144	1677.000
technician	1072.067	80.9331	1153.000
management	450.026	33.9736	484.000
Totals	5431.000	410.0000	5841.000

Table 1.2-Way Summary Table: Expected Frequencies Job Type x Sales [source: own]

Statistic	Statistics: job type(5) x sale(2) (Spreadsheet246.sta)		
	Chi-square	df	p
Pearson Chi-square	9.700619	df=4	p=.04578
M-L Chi-square	9.838855	df=4	p=.04323
Uncertainty coefficient	X=.0005569	Y=.0033140	X|Y=.00095

Table 2.Statistics Person Chi -square Job type x Sales [source: own]

And from the Bar Chart 34 and Table 1 representing 2-way summary table we can see that customers who work as administrator are the one that tend to do higher sales follow by blue-collar and followed by the third category technician. Services and management category are the ones that have the lowest sales for the company.



Bar Chart 34.Bivariate Distribution: job type x sale [source: own]

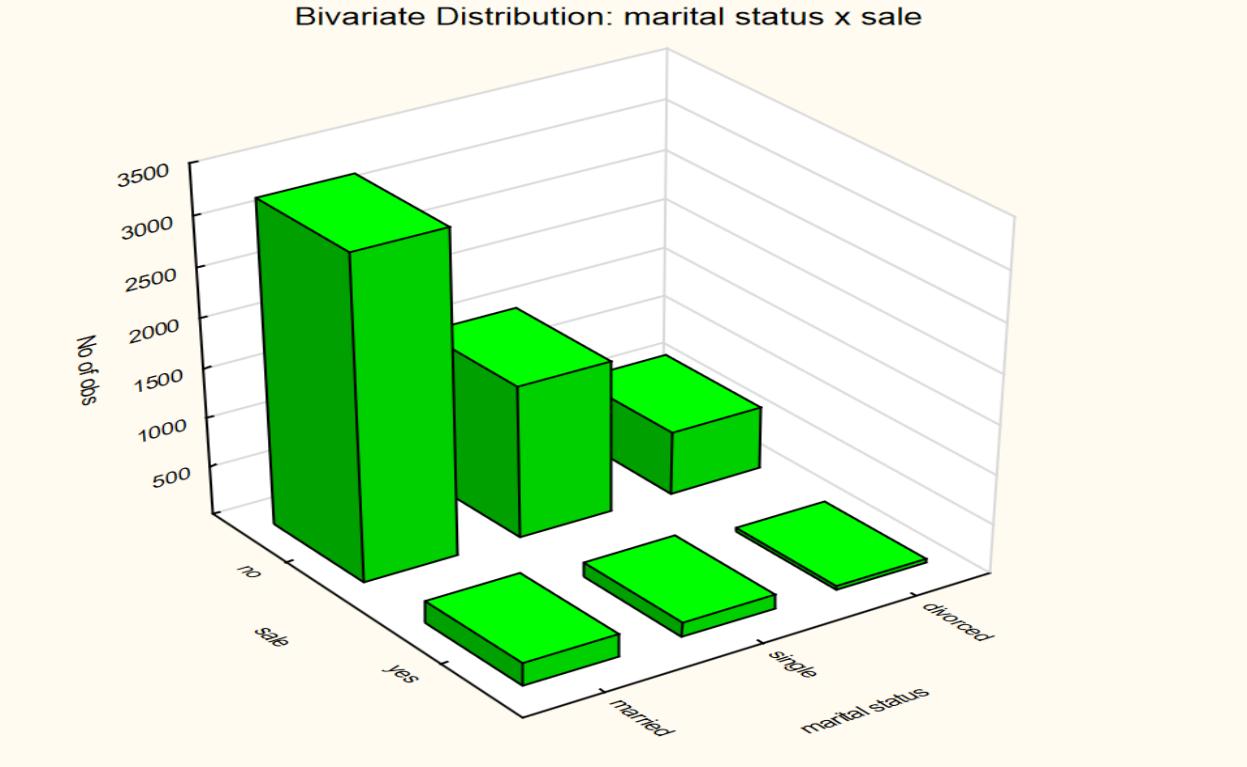
marital status	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta) sale no	sale yes	Row Totals
married	3236.657	244.3434	3481.000
single	1567.654	118.3462	1686.000
divorced	626.690	47.3104	674.000
Totals	5431.000	410.0000	5841.000

Table 3.2-Way Summary Table Marital Status x Sales [source: own]

Statistic	Statistics: marital status(3) x sale(2) (Spreadsheet246.sta) Chi-square	df	p
Pearson Chi-square	10.16657	df=2	p=.00620
M-L Chi-square	9.853064	df=2	p=.00725
Uncertainty coefficient	X=.0009205	Y=.0033188	X|Y=.00144

Table 4.Statistics Pearson Chi -square Marital Status x Sales [source: own]

The Chi-squared test statistic is 10.16657 with an associated p=0.00620 < 0.05 so we can conclude that the Marital Status and Sales are associated and therefore the sales will be determined as well from what marital status the customer will have. From Bar Chart 35 we can see that married customer have higher purchases almost as high as double of the single customer which on the other hand make purchases almost more than half of the divorced customers.



Bar Chart 35.Bivariate Distribution: marital status x sale [source: own]

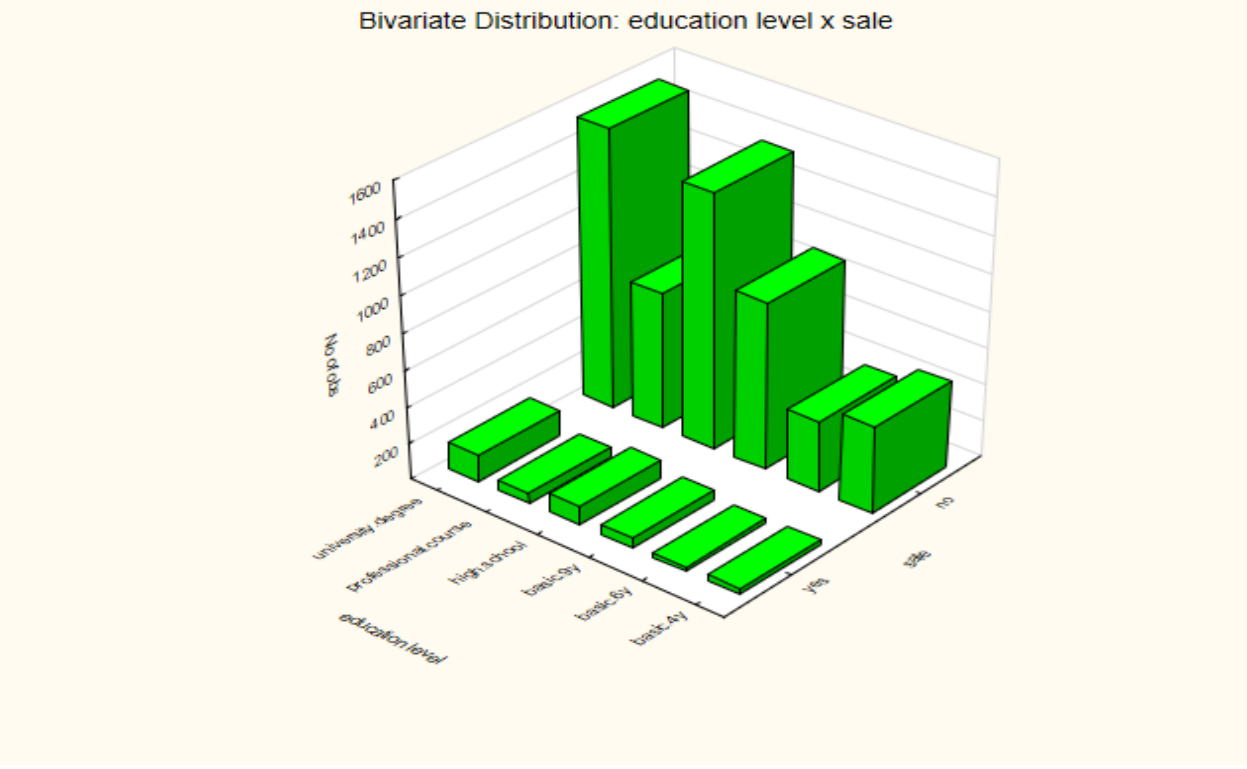
education level	2-Way Summary Table: Observed Frequencies (Spreadsheet8) Marked cells have counts > 10		
	sale no	sale yes	Row Totals
university.degree	1519	146	1665
professional.course	749	58	807
high.school	1387	100	1487
basic.9y	915	54	969
basic.6y	386	22	408
basic.4y	475	30	505
Totals	5431	410	5841

Table 5.2-Way Summary Table Education Level x Sales [source: own]

Statistic	Statistics: education level(6) x sale(2) (Spreadsheet246.sta)		
	Chi-square	df	p
Pearson Chi-square	13.70231	df=5	p=.01762
M-L Chi-square	13.57184	df=5	p=.01857
Uncertainty coefficient	X=.0006936	Y=.0045714	X|Y=.00120

Table 6.Statistics Pearson Chi -square Education Level x Sales [source: own]

The Chi-squared test statistic is 13.70231 with an associated p=0.01762 < 0.05 so we can conclude that the Education Level and Sales are associated which mean that the level of sales is affected by the level of education that customer have.



Bar Chart 36.Bivariate Distribution: education level x sale [source: own]

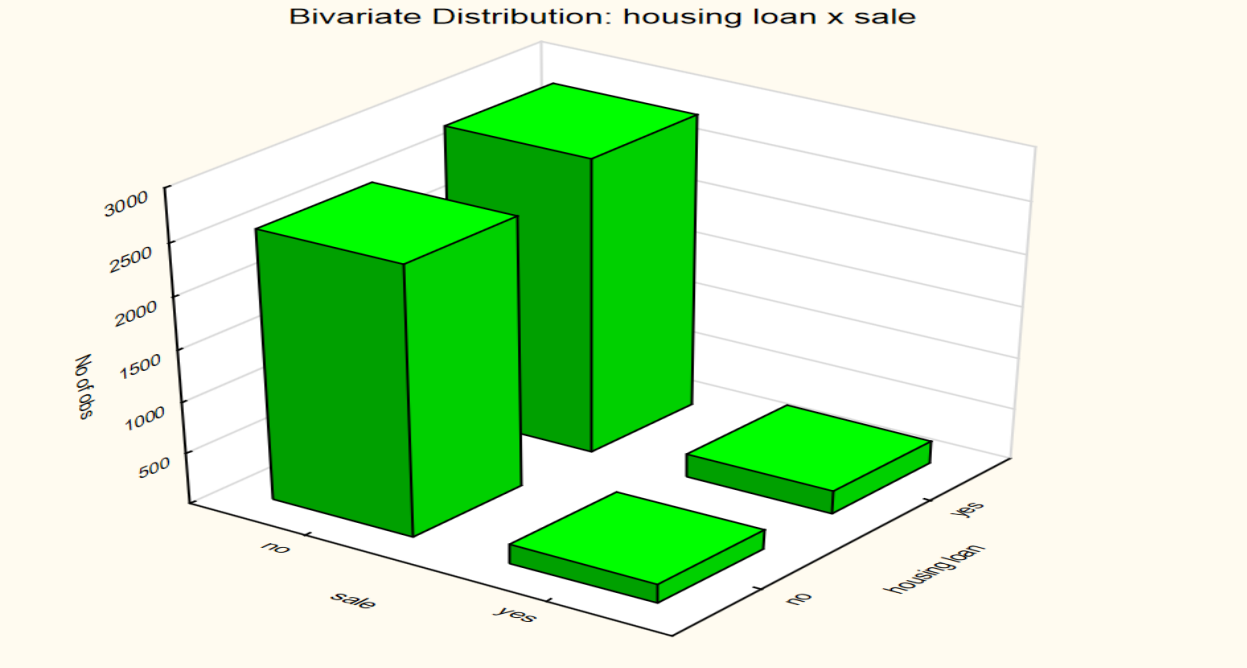
Bar Chart 36 shows that customers with a university degree have the highest level of purchases followed up by the categories high school and basic 9-year education. Customers with professional courses made almost half of purchases than customers with university degree. And the lowest level of sales is made by customer who have basic 6 year of education. We can see that it is not a clear correlation, meaning that the higher education has the higher level of sales for the company as basic 6 year has less than basic 4 year and professional courses have less than basic 9-years of education.

	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
housing loan	sale no	sale yes	Row Totals
no	2586.722	195.2782	2782.000
yes	2844.278	214.7218	3059.000
Totals	5431.000	410.0000	5841.000

Table 7.2-Way summary table: Expected Frequencies [source: own]

	Statistics: housing loan(2) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	.5570703	df=1	p=.45544
M-L Chi-square	.5577603	df=1	p=.45516
Uncertainty coefficient	X=.0000690	Y=.0001879	X|Y=.00010

Table 8.Statistics Pearson Chi –square Housing Loan x Sales [source: own]



Bar Chart 37.Bivariate Distribution: housing loan x sale [source: own]

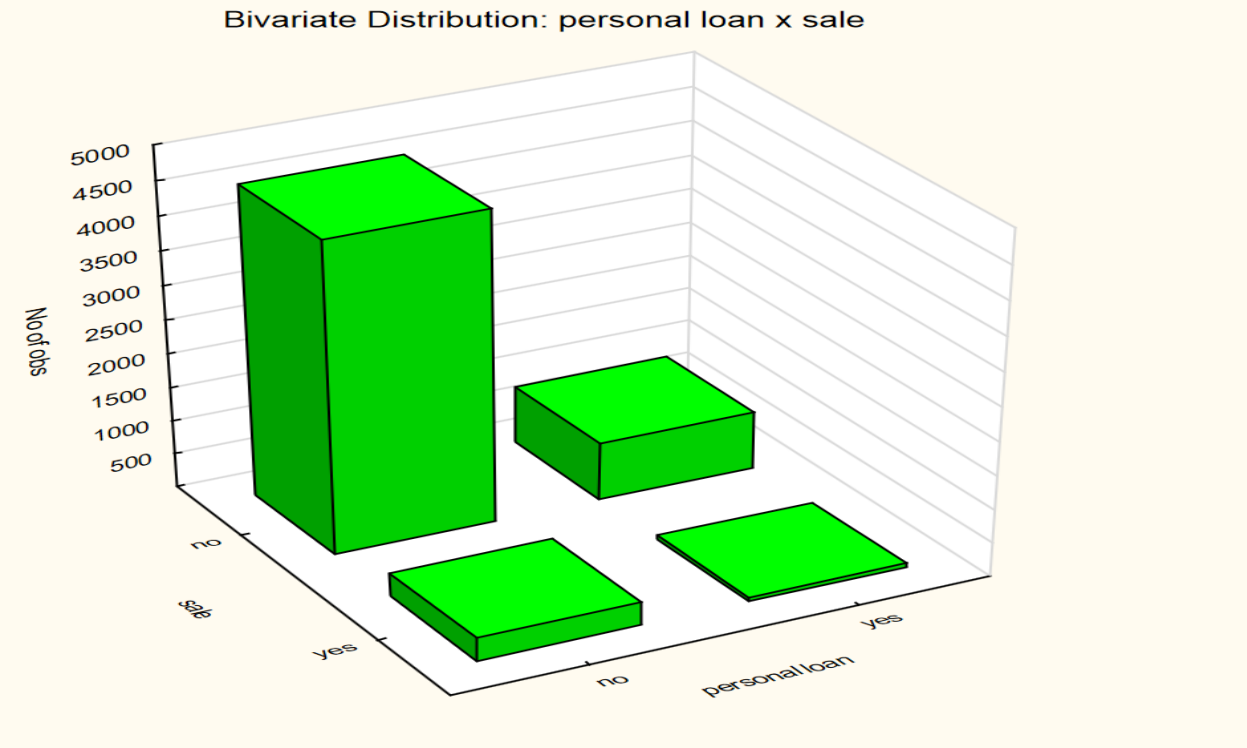
The Chi-squared test statistic shown in Table 8 is 0.557 with an associated p=0.455 > 0.05 so we can conclude that the Housing Loan and Sales are not associated. For this reason, we will remove "housing loan" variable from our model as whether people have housing loan or not, it does not affect the level of sales. Bar Chart 37 shows that people which have housing loan have done slightly more purchases that people who do not have housing loan but as mentioned above the overall conclusion was that variable "housing loan" doesn't not affect the sales.

	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
personal loan	sale no	sale yes	Row Totals
no	4572.789	345.2114	4918.000
yes	858.211	64.7886	923.000
Totals	5431.000	410.0000	5841.000

Table 9.2-Way summary table: Expected Frequencies [source: own]

	Statistics: personal loan(2) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	.1533094	df=1	p=.69539
M-L Chi-square	.1549996	df=1	p=.69380
Uncertainty coefficient	X=.0000304	Y=.0000522	X|Y=.00004

Table 10.Statistics Pearson Chi –square Personal Loan x Sales [source: own]



Bar Chart 38.Bivariate Distribution: personal loan x sale [source: own]

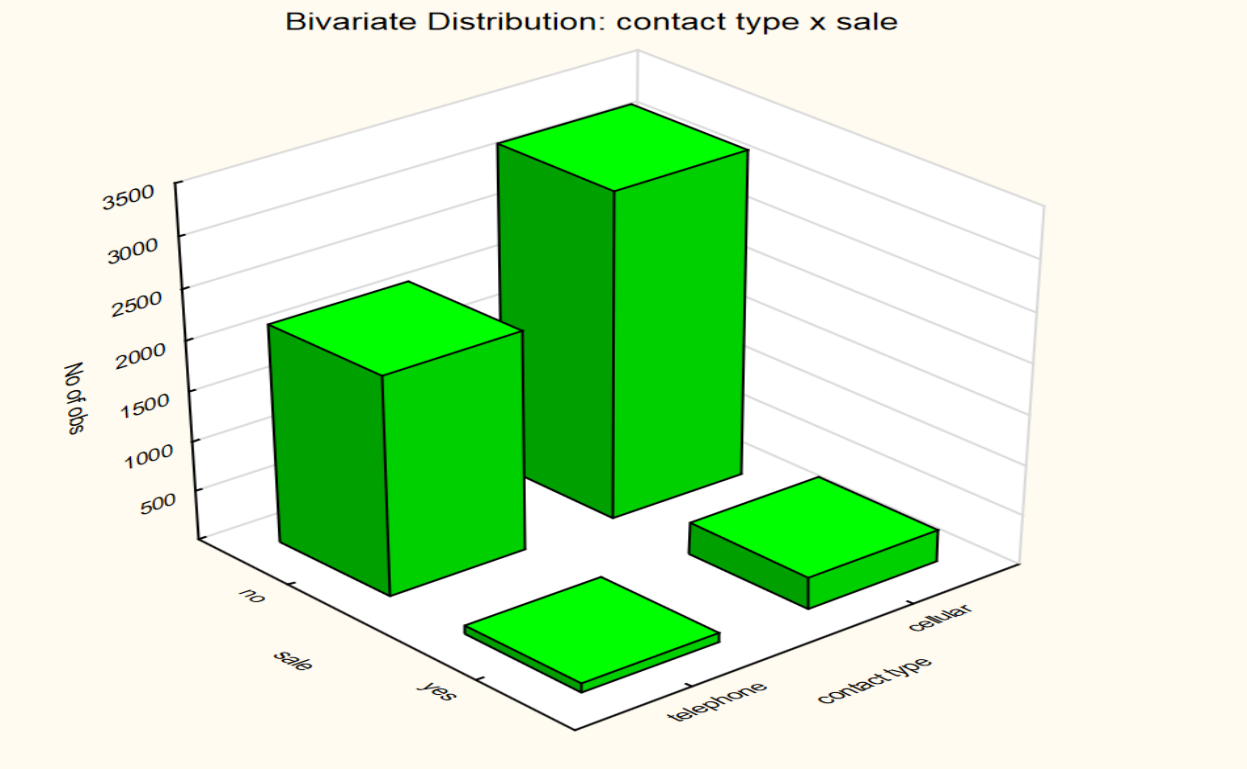
The Chi-squared test statistic as seen from Table 10 is 0.1533 with an associated p=0.695> 0.05 which leads us to accepting the null hypothesis and concluding that the Personal Loan and Sales are not associated. Therefore, sales are not affected whether the customers have personal loan or not. Bar chart 38 shows that people who do not have personal loan have made more purchase that people who do have personal loan.

	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
contact type	sale no	sale yes	Row Totals
telephone	2113.450	159.5497	2273.000
cellular	3317.550	250.4503	3568.000
Totals	5431.000	410.0000	5841.000

Table 11.2-Way Summary Table Contact Type x Sales [source: own]

	Statistics: contact type(2) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	56.49236	df=1	p=.00000
M-L Chi-square	61.05793	df=1	p=.00000
Uncertainty coefficient	X=.0078201	Y=.0205662	X|Y=.01133

Table 12.Statistics Pearson Chi –square Contact type x Sales [source: own]



Bar Chart 39.Bivariate Distribution: contact type x sale [source: own]

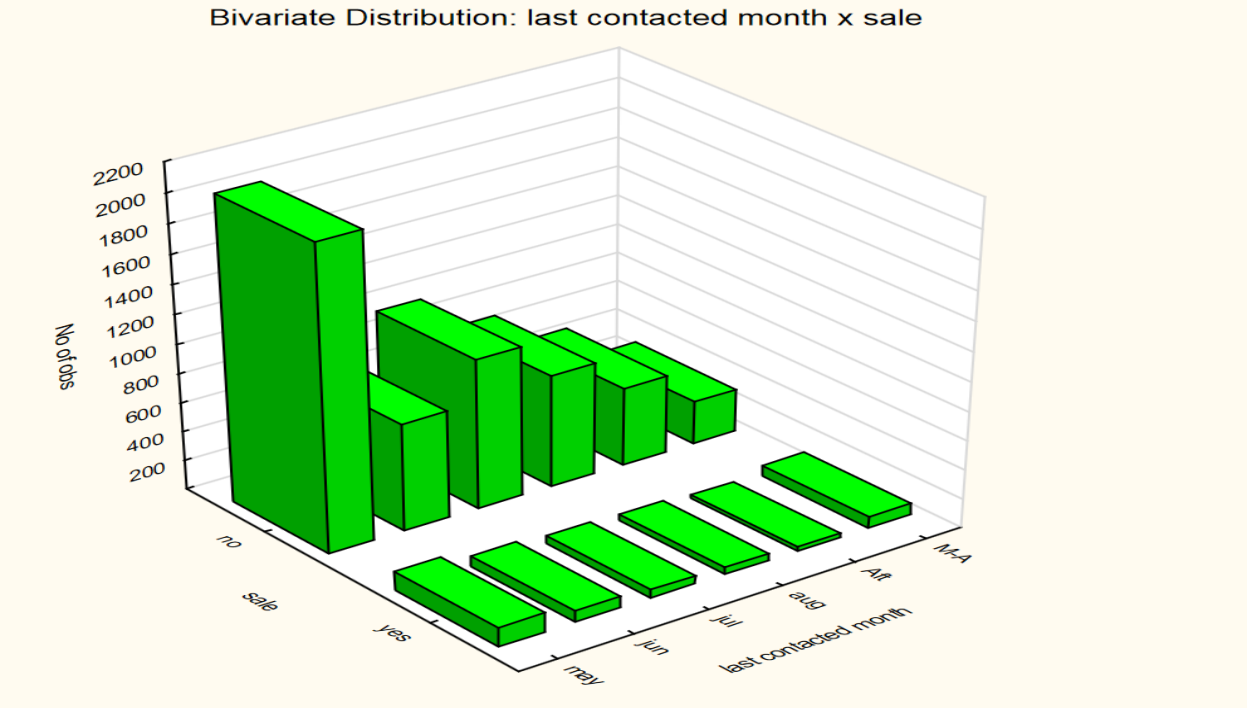
The Chi-squared test statistic shown in Table 12 is 56.49 with an associated $p < 0.05$ so we can conclude that the Contact Type and Sales are associated meaning that the level of sales had a relation with the type of contact that was used to contact the costumers. And Bar Chart 39 shows that people who were contacted through cellular (mobile phone) have made more purchases than people who were contacted through telephone.

	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
last contacted month	sale no	sale yes	Row Totals
may	2046.504	154.4958	2201.000
jun	755.003	56.9971	812.000
jul	1004.191	75.8089	1080.000
aug	755.933	57.0673	813.000
Aft	521.621	39.3785	561.000
M-A	347.748	26.2524	374.000
Totals	5431.000	410.0000	5841.000

Table 13.2-Way Summary Table Last contacted month x Sales [source: own]

	Statistics: last contacted month(6) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	138.2347	df=5	p=0.0000
M-L Chi-square	102.4735	df=5	p=0.0000
Uncertainty coefficient	X=.0053827	Y=.0345162	X|Y=.00931

Table 14.Statistics Pearson Chi –square Last contacted month x Sales [source: own]



Bar Chart 40.Bivariate Distribution: last contacted month x sale [source: own]

From Table 14 we can see that Chi-squared test statistic is 138.2347 with an associated $p < 0.05$ so we can conclude that the Last Contacted Month and Sales are associated. We can tell that level of sales is affect by the month the customers were contacted. Bar Chart 40 clearly shows that the level of sales was higher when customers where contacted on May, followed by July when people had almost half of purchases made on May, followed then by June and August with a very tight level of purchase. The end of the year from September to December represented by category "Aft" had very low level of sales followed by the lowest level of sales made on March and April represented by category "M-A".

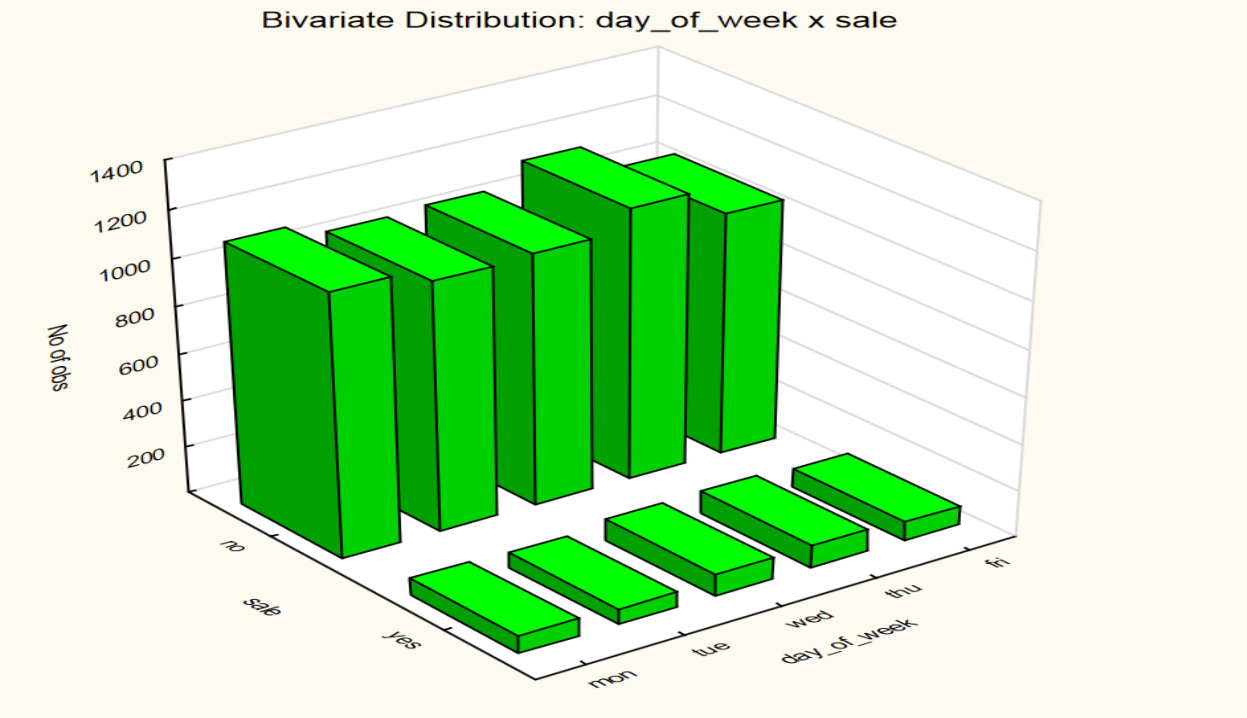
	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
day_of_week	sale no	sale yes	Row Totals
mon	1108.329	83.6706	1192.000
tue	1044.173	78.8273	1123.000
wed	1079.505	81.4946	1161.000
thu	1164.118	87.8822	1252.000
fri	1034.875	78.1253	1113.000
Totals	5431.000	410.0000	5841.000

Table 15.2-Way Summary Table Day of the week x Sales [source: own]

	Statistics: day_of_week(5) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	6.387640	df=4	p=.17201
M-L Chi-square	6.472847	df=4	p=.16651
Uncertainty coefficient	X=.0003445	Y=.0021803	X|Y=.00059

Table 16.Statistics Pearson Chi –square Day of the week x Sales [source: own]

Table 16 Chi-squared test statistic is 6.39 with an associated $p=0.17 > 0.05$ and this means that we have to accept the null hypothesis so Day of week and Sales do not have an association. Day of the week will be removed from the model as sales are not affected by the day of the week that customers are contacted. Bar chart 41 shows that the sales and non-sales are almost constant in all days. But we can notice that on Monday and Thursday there is slightly higher level of non-sales as well as of sales comparing to the other days of the week and yet the overall conclusion as mentioned is that the day of the week does not affect the sales.



Bar Chart 41.Bivariate Distribution: day of week x sale[source: own]

part of week	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta) sale no	sale yes	Row Totals
early	2156.221	162.7786	2319.000
mid	1082.295	81.7052	1164.000
late	2192.484	165.5162	2358.000
Totals	5431.000	410.0000	5841.000

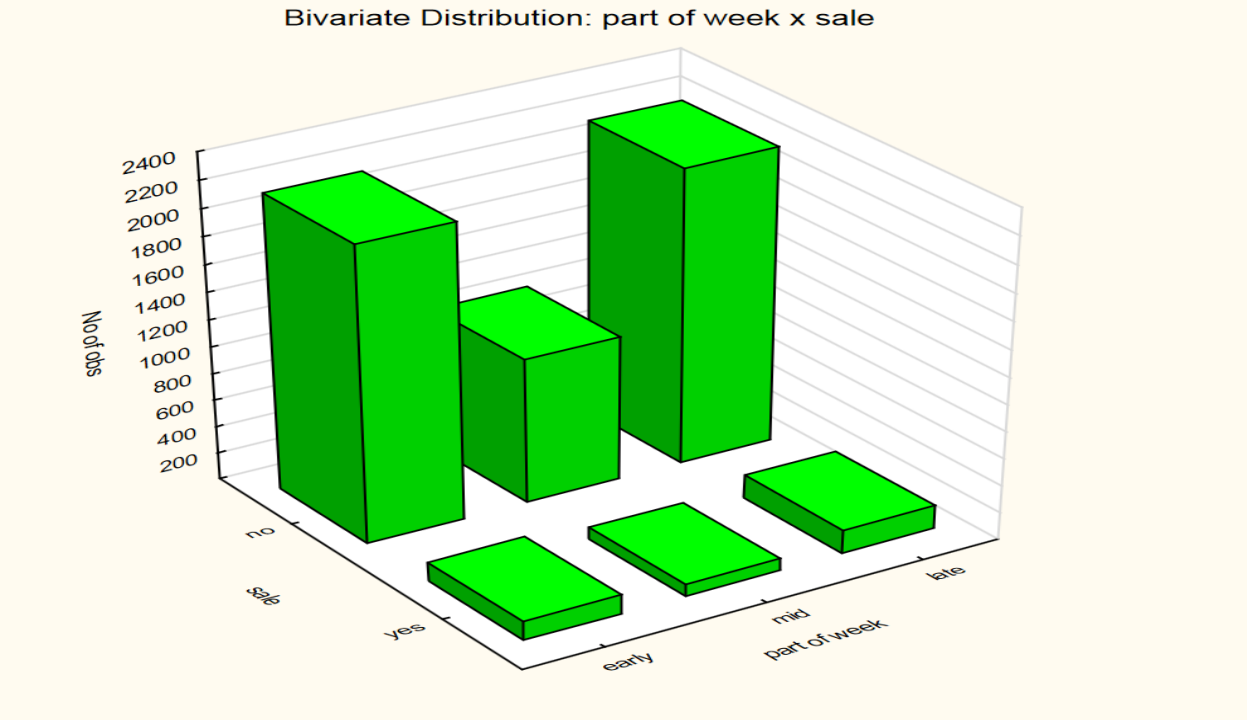
Table 17.2-Way Summary Table Part of the week x Sales [source: own]

Statistic	Statistics: part of week(3) x sale(2) (Spreadsheet246.sta) Chi-square	df	p
Pearson Chi-square	4.051541	df=2	p=.13189
M-L Chi-square	4.091777	df=2	p=.12927
Uncertainty coefficient	X=.0003322	Y=.0013782	X|Y=.00054

Table 18.Statistics Pearson Chi –square Part of the week x Sales [source: own]

Table 18 shows that Chi-squared test statistic is 4.051 with an associated p =0.49947> 0.05 so we can conclude that the Part of the week and Sales are not associated. Part of the week will be removed from the model as sales are not affected by the part of week that customers are contacted whether it is early, in the middle of the week or the late week. Bar chart 42 shows that the sales and non-sales are almost the same for the early part of the week and for the late part of

the week. And we can clearly notice that on middle of the week there is a lower level of non-sales and sales.



Bar Chart 42.Bivariate Distribution: part of week x sale [source: own]

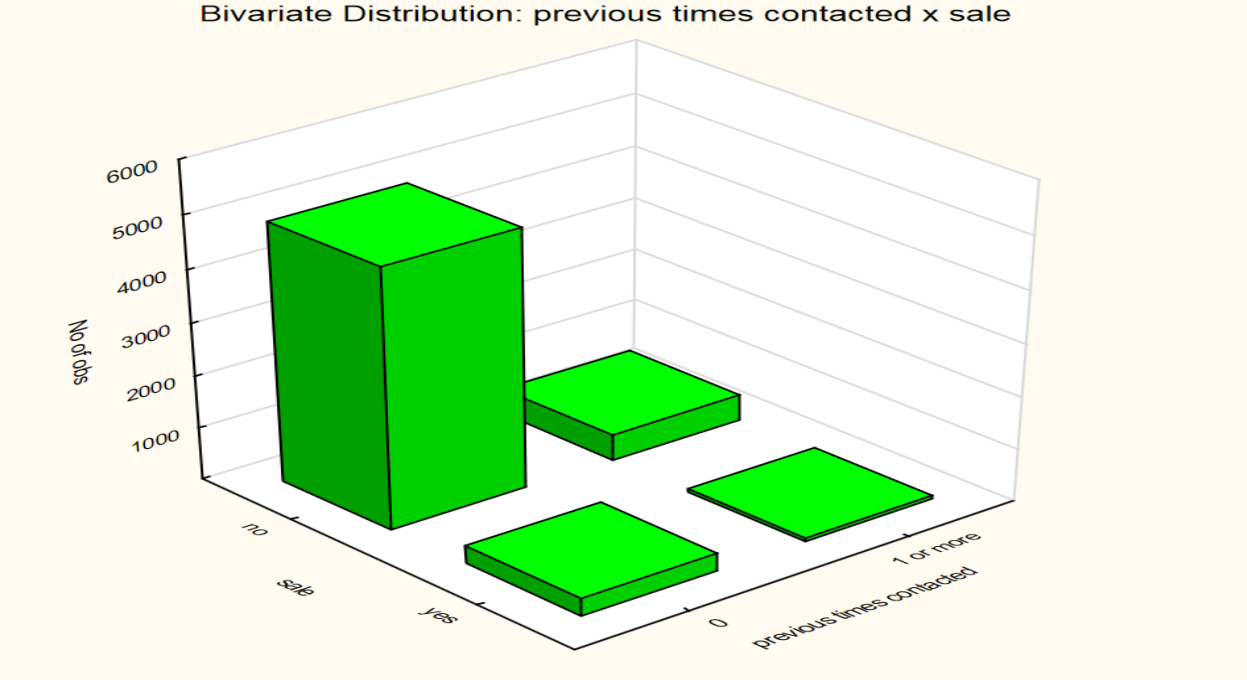
sale	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta) previous times contacted 0	previous times contacted 1 or more	Row Totals
no	4903.800	527.2003	5431.000
yes	370.200	39.7997	410.000
Totals	5274.000	567.0000	5841.000

Table 19.2-Way Summary Table Previous time contacted x Sales [source: own]

Statistic	Statistics: previous times contacted(2) x sale(2) (Spreadsheet246.sta) Chi-square	df	p
Pearson Chi-square	19.00581	df=1	p=.00001
M-L Chi-square	16.47731	df=1	p=.00005
Uncertainty coefficient	X=.0044271	Y=.0055501	X|Y=.00493

Table 20.Statistics Pearson Chi –square previous time contacted x Sales [source: own]

Table 20 shows that Chi-squared test statistic is 19.0058 with an associated p=0.00001< 0.05 so we can conclude that the “previous times contacted” and Sales are associated. The level of sales will be affected by whether the customers were previously contacted or not. And from Bart chart 43 we can notice that when people where not contacted previously they had considerably higher level of sales and non-sales compared to when they were contacted at least one time before.



Bar Chart 43.Bivariate Distribution: previous times contacted x sale [source: own]

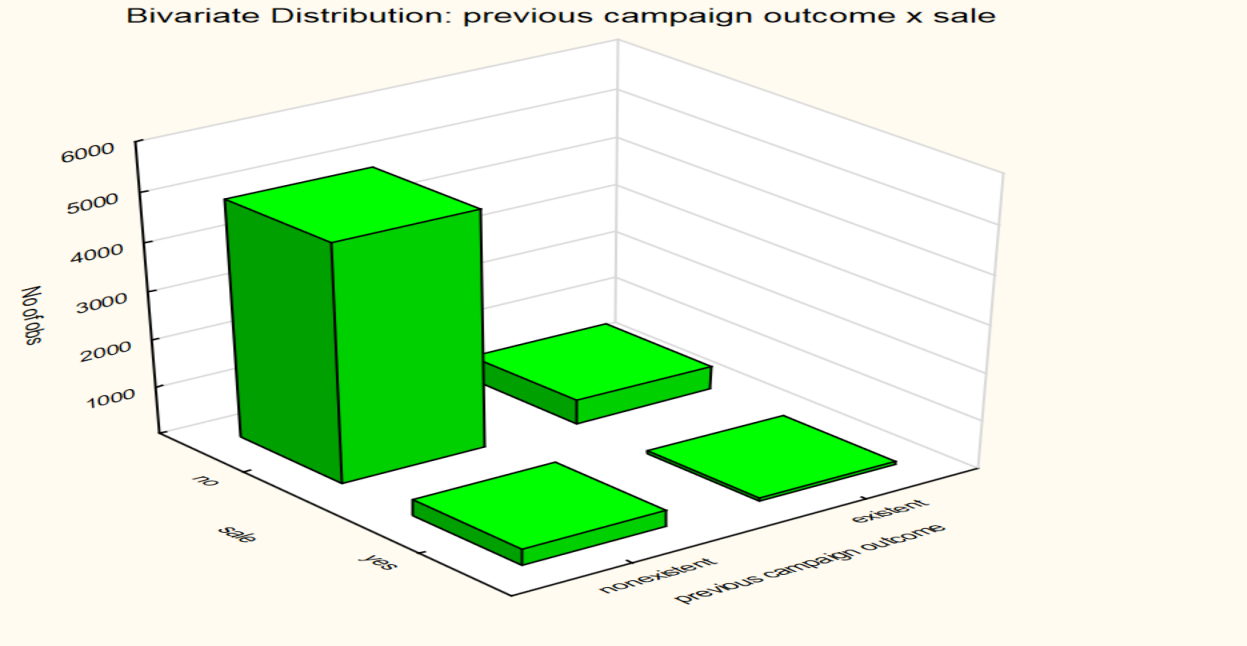
	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
sale	previous campaign outcome nonexistent	previous campaign outcome existent	Row Totals
no	4903.800	527.2003	5431.000
yes	370.200	39.7997	410.000
Totals	5274.000	567.0000	5841.000

Table 21.2-Way Summary Table Previous campaign outcome x Sales [source: own]

	Statistics: sale(2) x previous campaign outcome(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	19.00581	df=1	p=.00001
M-L Chi-square	16.47731	df=1	p=.00005
Uncertainty coefficient	X=.0055501	Y=.0044271	X|Y=.00493

Table 22.Statistics Pearson Chi –square previous campaign outcome x Sales [source: own]

Table 22 shows that Chi-squared test statistic is 19.0058 with an associated p=0.00001< 0.05 so we can conclude that the "previous campaign outcome" and Sales are associated. This means that whether there was a previous campaign existent company or not, will affect the level of sales in the company.



Bar Chart 44.Bivariate Distribution: previous campaign outcome x sale [source: own]

Bar Chart 44 shows that if there was not a previous campaign than the sales were higher comparing to one there was and existing campaign but also the non-sales were higher comparing to one there was an existing campaign.

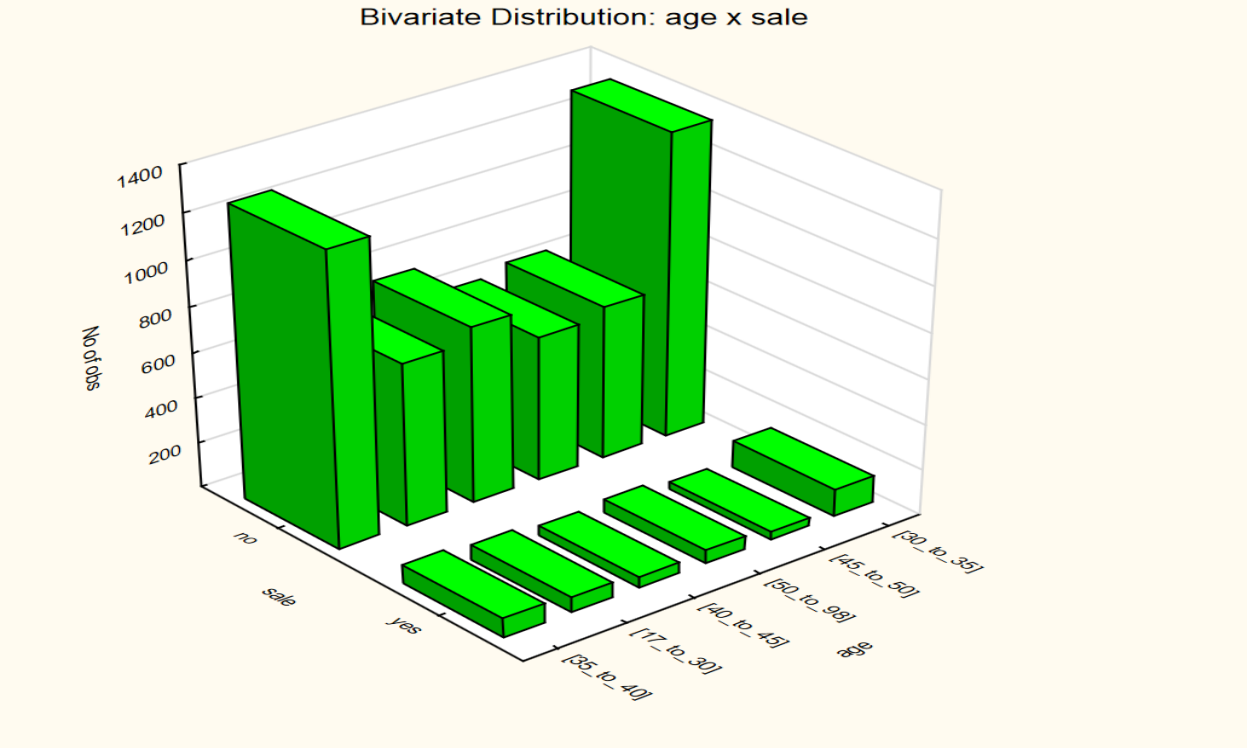
	2-Way Summary Table: Expected Frequencies (Spreadsheet430)		
age	sale no	sale yes	Row Totals
[35_to_40]	1276.624	96.3756	1373.000
[17_to_30]	727.109	54.8913	782.000
[40_to_45]	764.301	57.6990	822.000
[50_to_98]	647.145	48.8546	696.000
[45_to_50]	659.233	49.7672	709.000
[30_to_35]	1356.588	102.4123	1459.000
Totals	5431.000	410.0000	5841.000

Table 23.2-Way Summary Table Age x Sales [source: own]

	Statistics: age(6) x sale(2) (Spreadsheet430)		
Statistic	Chi-square	df	p
Pearson Chi-square	18.41308	df=5	p=.00247
M-L Chi-square	18.95312	df=5	p=.00196
Uncertainty coefficient	X=.0009316	Y=.0063840	X|Y=.00163

Table 24.Statistics Pearson Chi –square Age x Sales[source: own]

Table 24 shows that Chi-squared test statistic for the test of relationship among age and sale is 18.413 and an associated p =0.00247 < 0.05 so we can conclude that Sale and Age are associated and therefore the age of the customer affects the level of sales in the company.



Bar Chart 45.Bivariate Distribution: age x sale [source: own]

Bar Chart 45 shows that people belonging to category [30_to_35] have made the highest purchase, followed by people belonging to category [35_to_40]. This makes us think that people will already a certain economic stability are more willing to make purchases comparing to a younger and older category of customers.

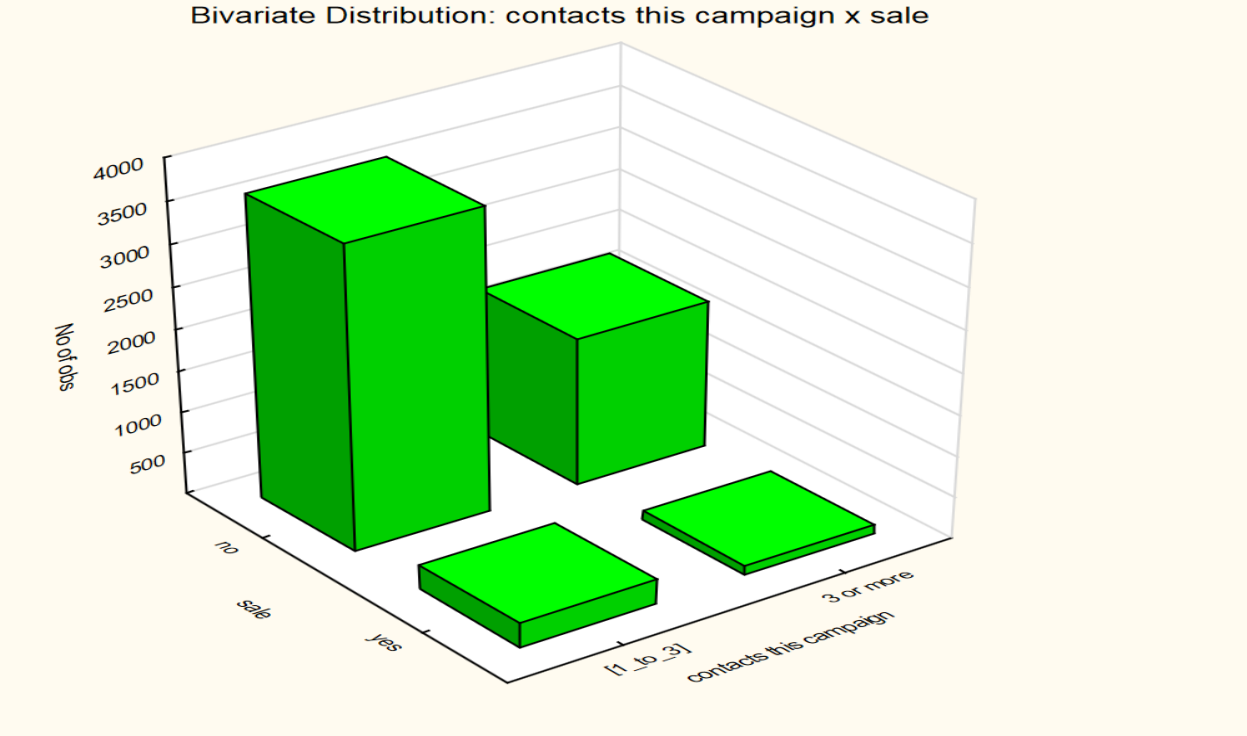
	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
contacts this campaign	sale no	sale yes	Row Totals
[1_to_3]	3659.719	276.2815	3936.00
3 or more	1771.281	133.7185	1905.00
Totals	5431.000	410.0000	5841.00

Table 25.2-Way Summary Table Contacts this campaign x Sales [source: own]

	Statistics: contacts this campaign(2) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	4.640849	df=1	p=.03122
M-L Chi-square	4.762397	df=1	p=.02909
Uncertainty coefficient	X=.0006456	Y=.0016041	X|Y=.00092

Table 26.Statistics Pearson Chi –square Contacts this campaign x Sales [source: own]

As seen from Table 26 Chi-squared test statistic for "contact campaign x sales" is 4.64 with an associated p=0.0312< 0.05 so we can conclude that the Contacts this campaign and Sales are associated and therefore we can say that the level of sales is affected by the times the customer contacts the existing campaign.



Bar Chart 46.Bivariate Distribution: contacts this campaign x sale [source: own]

Bar Chart 46 shows that when this campaign was contacted less than three times both sales and non-sales where higher compare to when the campaign was contacted 3 times or more.

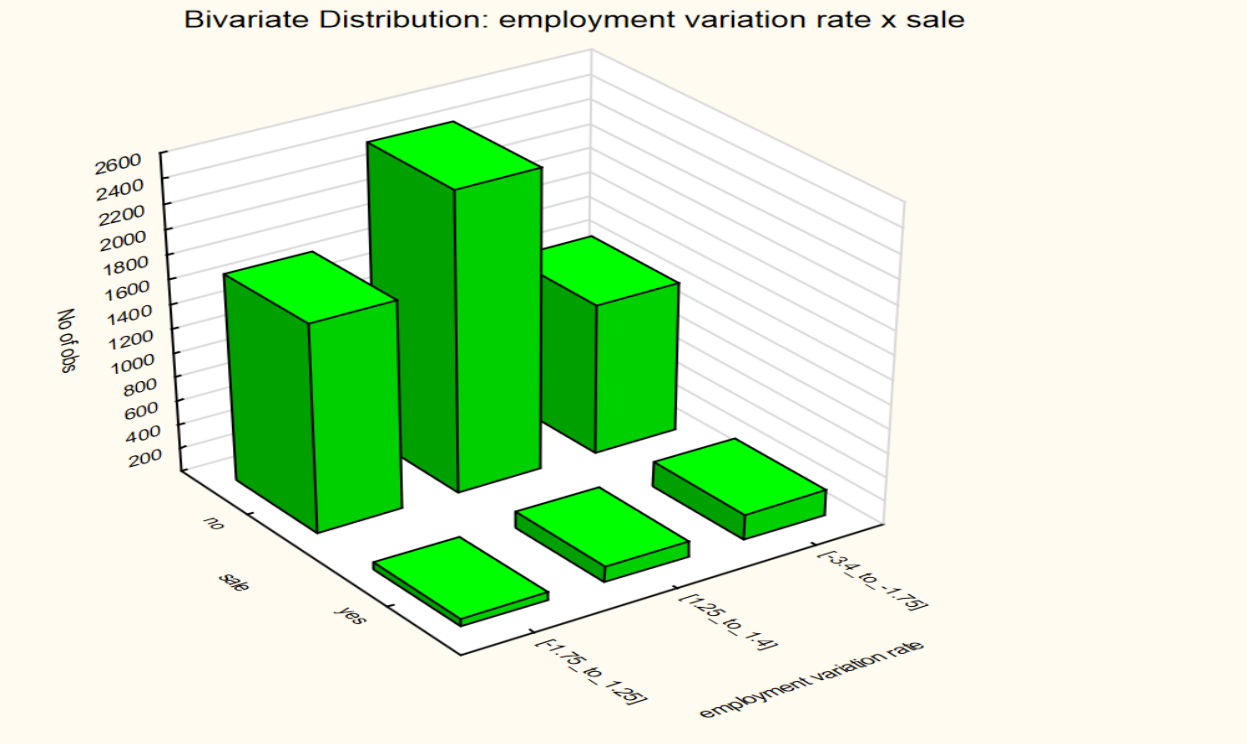
	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
employment variation rate	sale no	sale yes	Row Totals
[-1.75_to_1.25]	1659.705	125.2953	1785.000
[1.25_to_1.4]	2416.567	182.4328	2599.000
[-3.4_to_-1.75]	1354.728	102.2719	1457.000
Totals	5431.000	410.0000	5841.000

Table 27.2-Way Summary Table Employment Variation rate x Sales [source: own]

	Statistics: employment variation rate(3) x sale(2) (Spreadsheet246.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	171.7896	df=2	p=0.0000
M-L Chi-square	151.3125	df=2	p=0.0000
Uncertainty coefficient	X=.0121171	Y=.0509667	X|Y=.01958

Table 28.Statistics Pearson Chi –square Employment Variation rate x Sales [source: own]

Table 28 shows that Chi-squared test statistic is 171.79 with an associated $p < 0.05$ so we can conclude that the Employment Variation rate and Sales are associated. Therefore, the extent to which the labor resources are used will affect the level of sales in the company. From Bart Chart 47 show that the higher the employment rate, thus the use of the labor resources" the higher were the sales of the company but at well the non-sales.



Bar Chart 47.Bivariate Distribution: employment variation rate x sale [source: own]

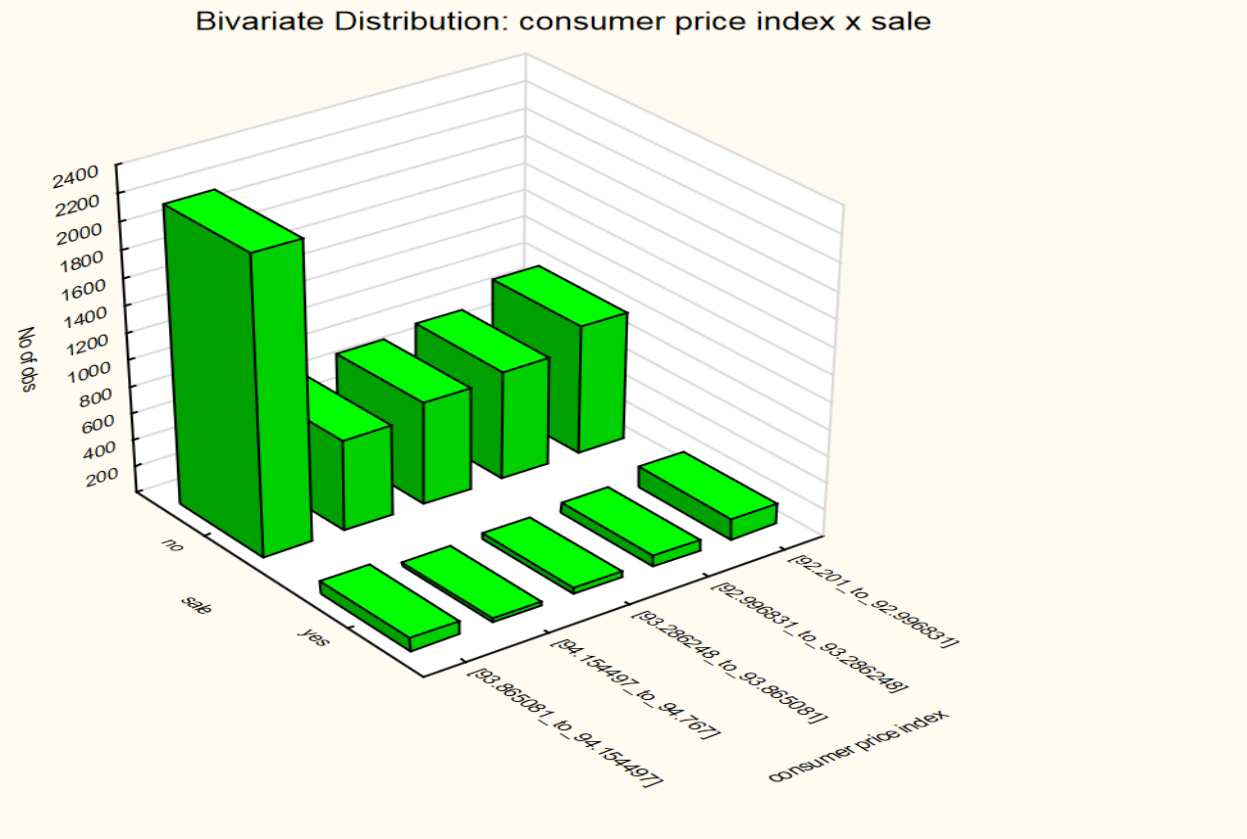
consumer price index	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta) sale no	sale yes	Row Totals
[93.865081_to_94.154497]	2142.274	161.7257	2304.000
[94.154497_to_94.767]	656.443	49.5566	706.000
[93.286248_to_93.865081]	755.933	57.0673	813.000
[92.996831_to_93.286248]	830.317	62.6828	893.000
[92.201_to_92.996831]	1046.032	78.9676	1125.000
Totals	5431.000	410.0000	5841.000

Table 29.2-Way Summary Table Consumer Price Index x Sales [source: own]

Statistic	Statistics: consumer price index(5) x sale(2) (Spreadsheet246.sta) Chi-square	df	p
Pearson Chi-square	123.2127	df=4	p=0.0000
M-L Chi-square	112.2436	df=4	p=0.0000
Uncertainty coefficient	X=.0064004	Y=.0378070	X|Y=.01095

Table 30.Statistics Pearson Chi –square Consumer Price Index x Sales [source: own]

The Chi-squared test statistic is 123.2127 as shown from Table 30 with an associated $p < 0.05$ so we can conclude that the Consumer Price Index and Sales are associated. So the cost of living as is sometimes called Consumer Price Index will affect the result of sales. From Bar Chart 48 we can see that the second highest Consumer Price Index (cost of living) has the highest level of sales and non-sales followed by the lowest Consumer Price Index. The highest Consumer Price Index has on the other hand the lowest level of sales and non-sales.



Bar Chart 48.Bivariate distribution: consumer price index x sale [source: own]

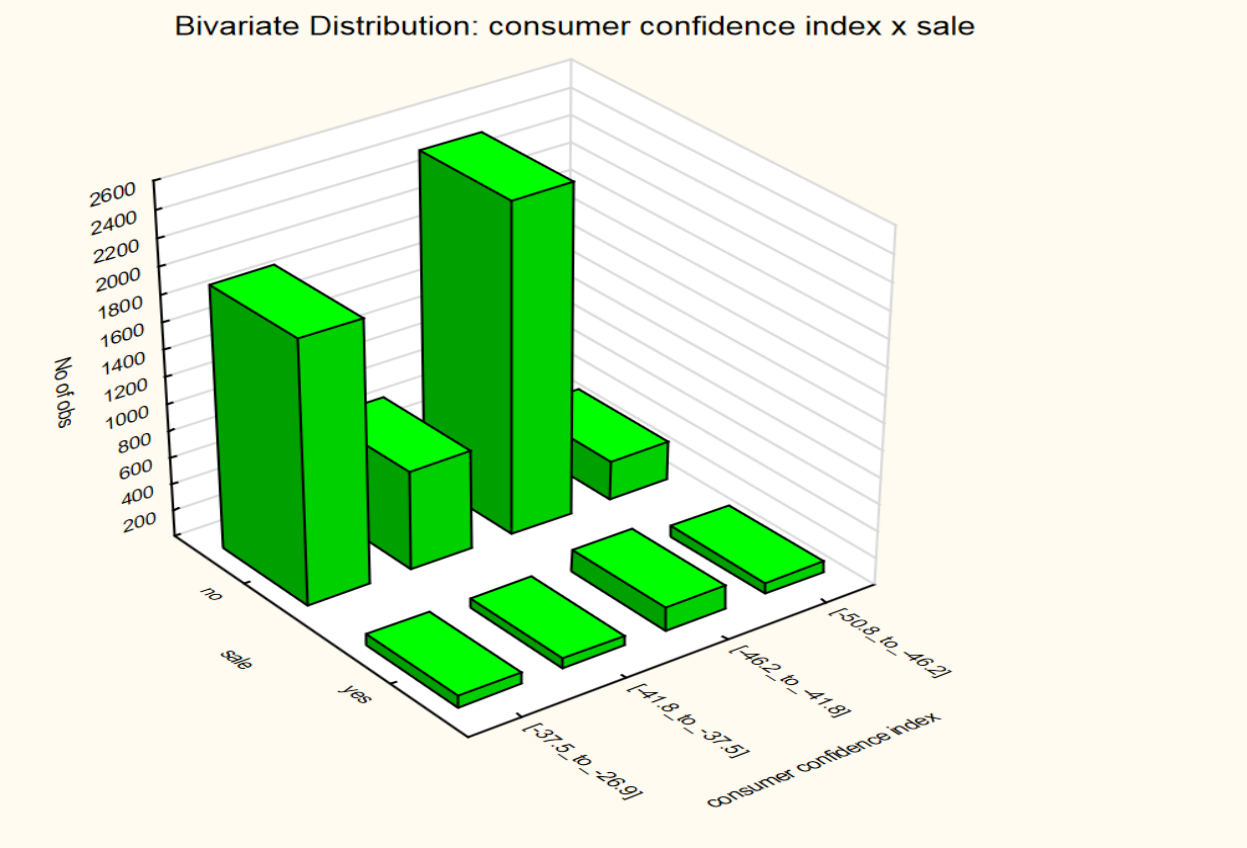
consumer confidence index	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta) sale no	sale yes	Row Totals
[-37.5_to_-26.9]	1894.016	142.9841	2037.000
[-41.8_to_-37.5]	755.003	56.9971	812.000
[-46.2_to_-41.8]	2434.234	183.7665	2618.000
[-50.8_to_-46.2]	347.748	26.2524	374.000
Totals	5431.000	410.0000	5841.000

Table 31.2-Way Summary Table Consumer Confidence Index x Sales [source: own]

Statistic	Statistics: consumer confidence index(4) x sale(2) (Spreadsheet246.sta)		
	Chi-square	df	p
Pearson Chi-square	145.6337	df=3	p=0.0000
M-L Chi-square	111.9705	df=3	p=0.0000
Uncertainty coefficient	X=.0081411	Y=.0377151	X|Y=.01339

Table 32.Statistics Pearson Chi –square Consumer Confidence Index x Sales [source: own]

The Chi-squared test statistic is 145.6337 as seen in Table 32 with an associated $p < 0.05$ so we can conclude that the Consumer Confidence index and Sales are strongly associated. So the willingness of consumer to spend, borrow and save will have an effect on the level of sales. Bar Chart 49 shows that the second lowest consumer confidence index interval has the highest level of sales and non-sales for the company, followed by the highest consumer confidence index and the lowest consumer confidence index has the lowest level of sales and non-sales. Therefore, we can say that when people have very high willingness to spend, borrow and save and not so low willingness to spend, borrow and save they do the highest purchase.



Bar Chart 49.Bivariate distribution: consumer confidence index x sale [source: own]

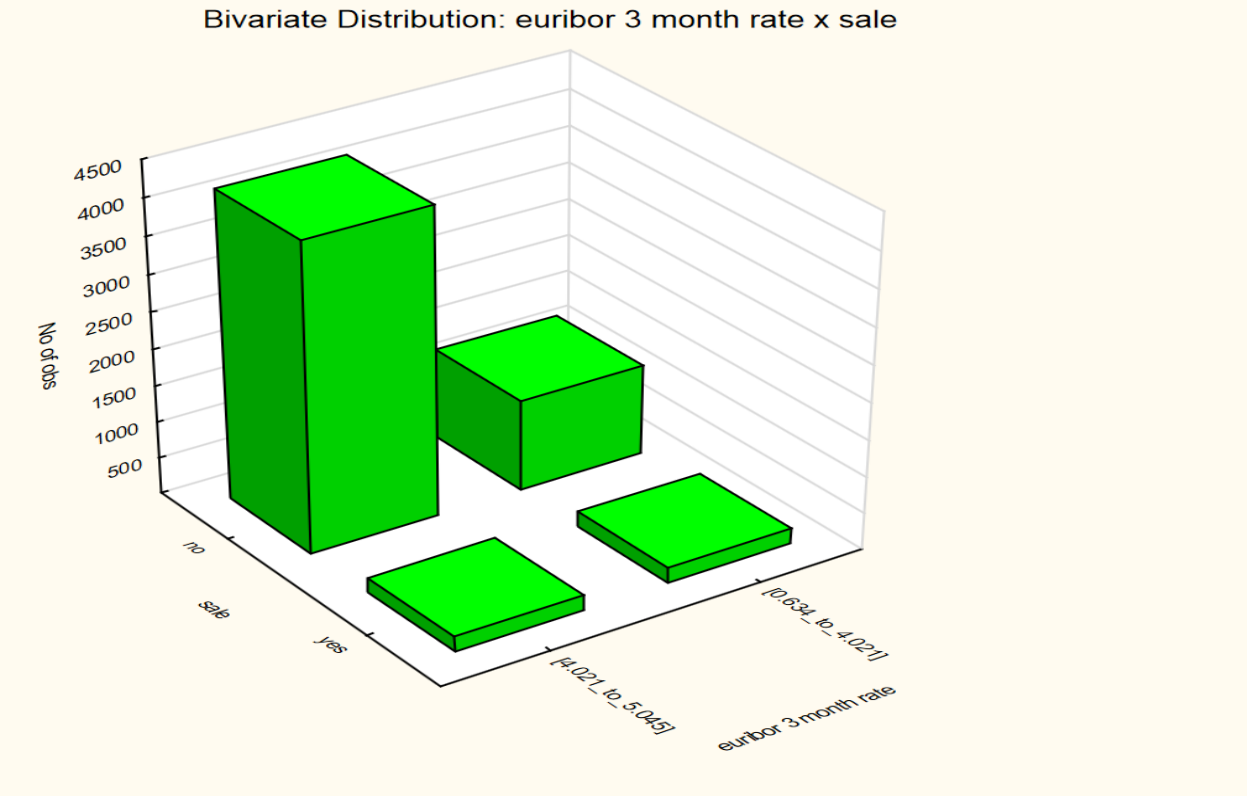
euribor 3 month rate	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
	sale no	sale yes	Row Totals
[4.021_to_5.045]	4074.412	307.5877	4382.000
[0.634_to_4.021]	1356.588	102.4123	1459.000
Totals	5431.000	410.0000	5841.000

Table 33.2-Way Summary Table Euribor 3 month rate x Sales [source: own]

Statistic	Statistics: euribor 3 month rate(2) x sale(2) (Spreadsheet246.sta)		
	Chi-square	df	p
Pearson Chi-square	168.1103	df=1	p=0.0000
M-L Chi-square	146.0812	df=1	p=0.0000
Uncertainty coefficient	X=.0222466	Y=.0492046	X|Y=.03064

Table 34.Statistics Pearson Chi –square Euribor 3 month rate x Sales [source: own]

The Chi-squared test statistic is 168.1103 as shown in Table 34 with an associated $p < 0.05$ so we can conclude that the Euribor 3 month rate and Sales are associated. So the rate at which euro interbank term deposits are offered by one prime bank to another within the EMU (The economic and monetary union) affect sales of this company. And from Table 33 and Bar chart 50 we can see that the higher eurobor 3 month rate so the higher the rate at which bank offer deposit, the higher both sales and non-sales.



Bar Chart 50.Bivariate distribution: euribor 3 month rate x sale [source: own]

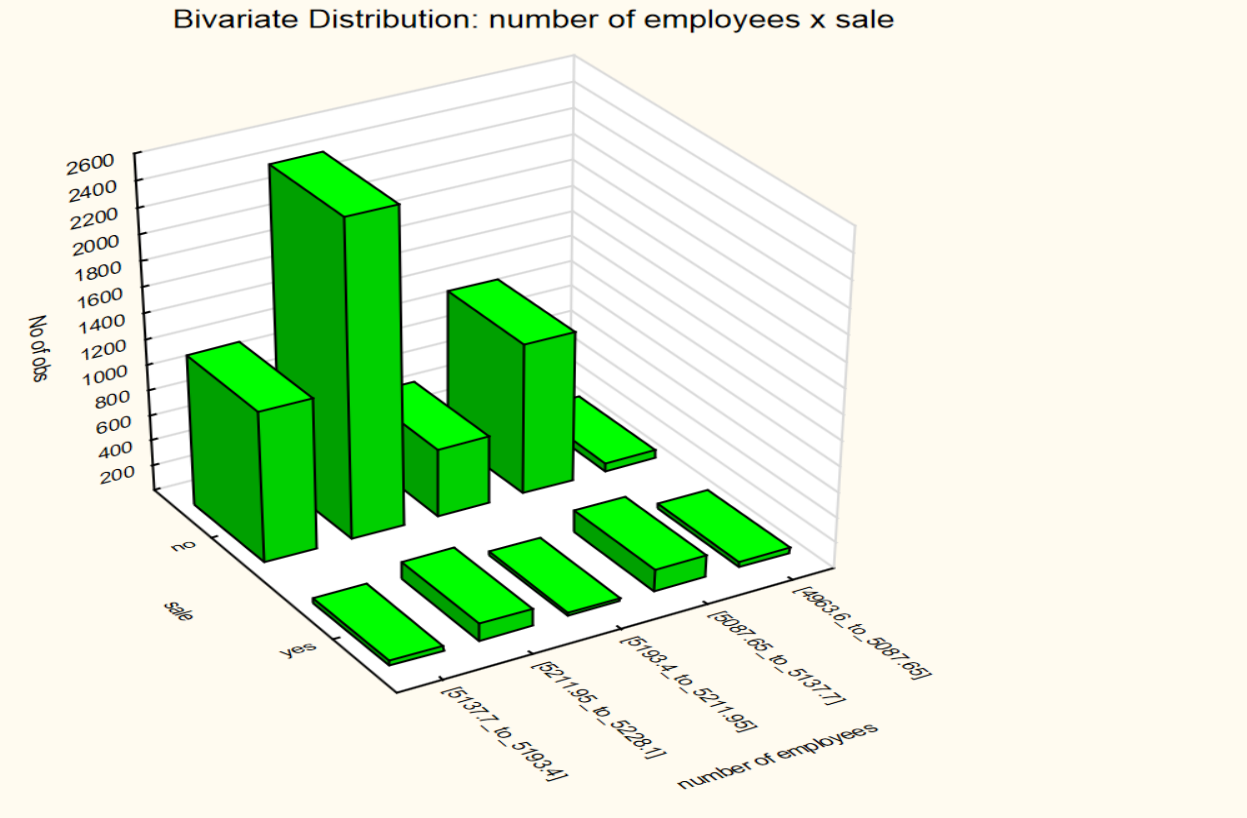
Statistic	Statistics: number of employees(5) x sale(2) (Spreadsheet246.sta)		
	Chi-square	df	p
Pearson Chi-square	299.8381	df=4	p=0.0000
M-L Chi-square	202.9610	df=4	p=0.0000
Uncertainty coefficient	X=.0131227	Y=.0683635	X|Y=.02202

Table 35.Statistics Pearson Chi –square Number of employees x Sales [source: own]

From Table 36 we can see that Chi-squared test statistic is 299.8381 with an associated $p < 0.05$ so we can conclude that the Number of employees and Sales are associated and therefore the number of employees in the company affects the levels of sales of the company.

number of employees	2-Way Summary Table: Expected Frequencies (Spreadsheet246.sta)		
	sale no	sale yes	Row Totals
[5137.7_to_5193.4]	1139.943	86.0572	1226.000
[5211.95_to_5228.1]	2416.567	182.4328	2599.000
[5193.4_to_5211.95]	519.762	39.2381	559.000
[5087.65_to_5137.7]	1256.169	94.8314	1351.000
[4963.6_to_5087.65]	98.559	7.4405	106.000
Totals	5431.000	410.0000	5841.000

Table 36.2-Way Summary Table Number of employees x Sales [source: own]



Bar Chart 51.Bivariate distribution: number of employees x sale [source: own]

Table 35 which is visually represented in Bar Chart 51 shows that the category with the range" [5211.95_to_5228.1] which represent also the range with highest number of employees "has the highest level of sales and non-sales. And the companies with the lowest range of employees have very considerable low level of sales comparing to the ones that have higher range of number of employees.

Based on the data exploration we will remove all the variables that are not associated with sales as they don't affect the depend variable (sales) and may mislead our final analysis and then on the next section we will start modeling the model.

5.3 Modeling

Predictive modeling is the process in which we create a model in order to predict an outcome. If the outcome is categorical, like in the case of our dataset than it is called *classification* and if the outcome was numerical it is called *regression*. Classification is that task of data mining that predicts the value of a categorical variable (called target or class) by building a model based on one or more numerical and /or categorical variables (called predictors or attributes). In our case we will be building a model based on categorical variable to predict the value of a categorical variable which in our case is SALES. For the purpose of modeling the data we are going to divide data into train, test and validation samples. The train sample is used to create models and to find patterns. Test sample prevents the models from learning only the train data and helps the model to generalize to new cases. And the validation sample estimates and compares the performance of the models. The sampling command in Statistica will split the data into two samples data for model building with train and test data and data for validation. After the cleaning of the dataset, dataset was reduced to 5841 cases which based on the models that are going to be used to model the categorical data set will be separated into training and sample data. Three different approaches will be used to build the model *Random Forest, Boosted Trees* and *C&RT*. At the end of every model the *Rapid Deployment* tool will be used to check the accuracy of the model and to help which the best model.

5.3.1 Random Forest Tree

Random forest is a tool that builds a series of classification trees, and then uses the prediction from each tree in the series. Each tree on its own makes a prediction good or bad. These predictions vote to make the random forest prediction. The purpose of this analyzes is to correctly identify the class labels of each data by using Random Forest model that will be built in

this analysis. So for the given set of predictors which are all the variables that were associated with the sales on the data exploration, we want to correctly categorize the sales as either No or Yes. As our data are categorical on the Random Forest menu we will choose the classification analysis. By default, Random Forest partitions, the data into training and testing samples by using random selection of cases from the data set. The training sample is used to build the model (add the simple trees) and the testing set is used to validate the performance of the model. This performance of the model is used as goodness of the model which in our case that is a classification task is simply defined as misclassification rate. And by default 30% of the data set is selected as test cases in the case of Random Forest. The frequency table before starting the Random forest model shows that most of the responses based on the observation are negative sales.

	Frequency table: sale (Spreadsheet246.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
no	5431	5431	92.98065	92.9807
yes	410	5841	7.01935	100.0000
Missing	0	5841	0.00000	100.0000

Table 37.Frequency table for Sale [source: own]

Table 37 shows the observed result for each category of variable sales. These results we will be compared with the predicted values from the models which will be built in this practical section of the thesis.

The summary of Random Forest for a number of trees 100 shows that train data model and test data model have similar trend. As the train data are used to build the model and the test data to validate the performance we can see from Figure 13 than the performance of the test data model is very satisfactory towards the train model and Table 38 shows the standard error of 0.004513 and 0.005218 for respectively train and test model.

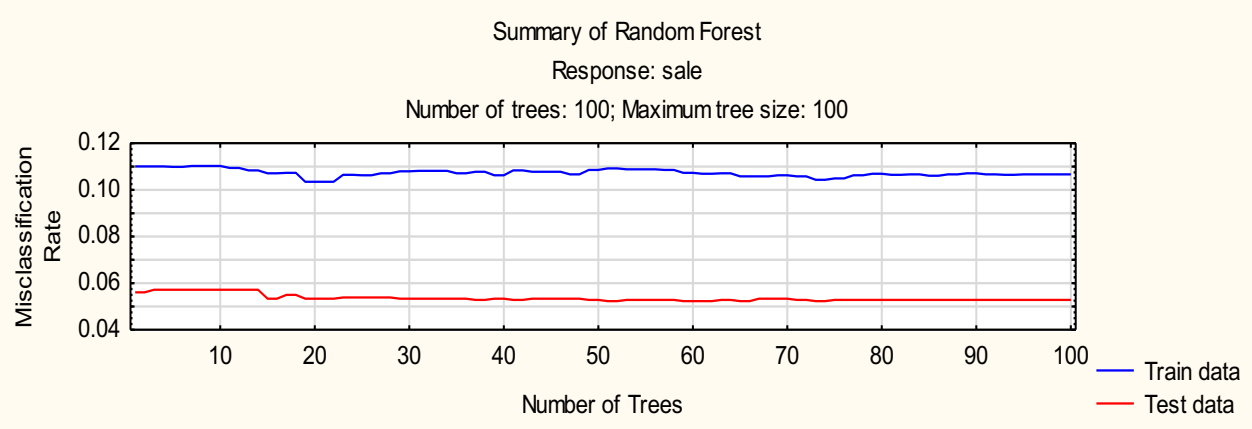
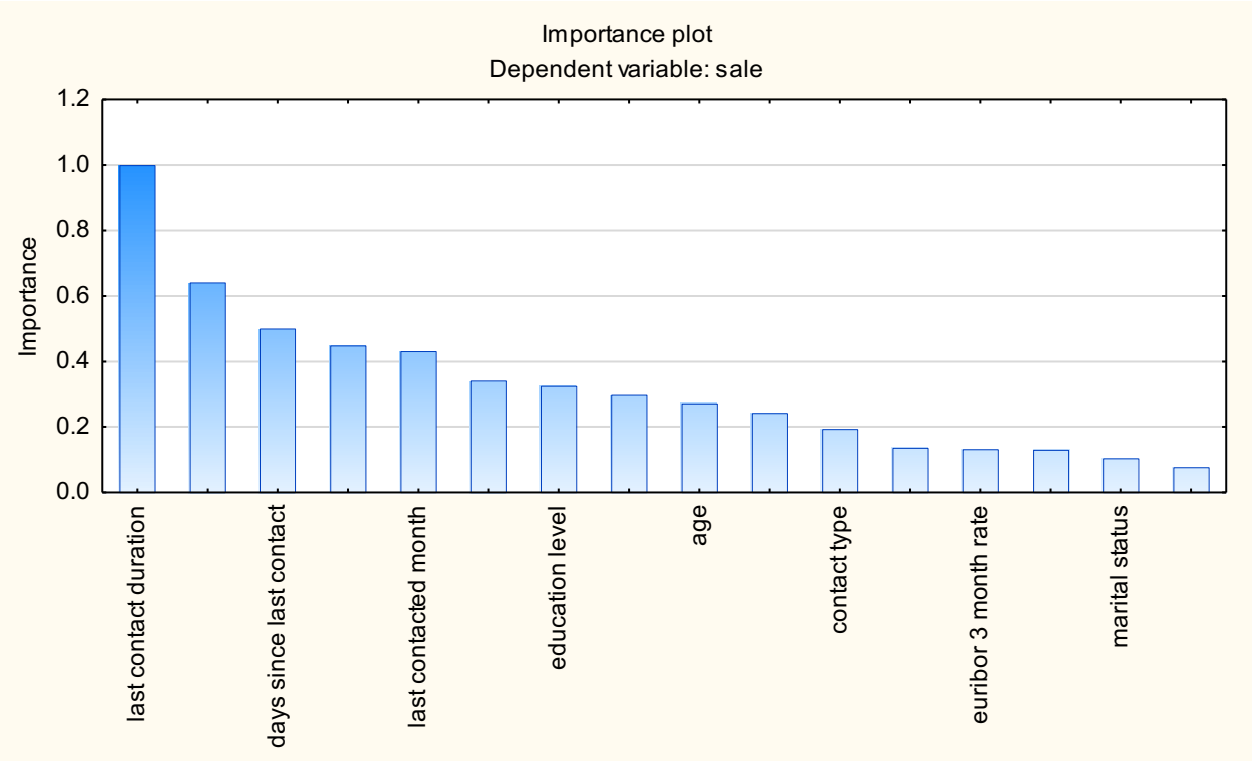


Figure 12.Summary of Random Forest [source: own]

	Risk estimates (Spreadsheet233.sta) Response: sale	
	Risk Estimate	Standard error
Train	0.106891	0.004513
Test	0.052803	0.005218

Table 38. Random forest Risk Estimates Random Forest [source: own]



Bar Chart 52.Random Forest Importance plot [source: own]

	Predictor importance (Spreadsheet233.sta) Response: sale	
	Variable Rank	Importance
last contact duration	100	1.000000
number of employees	64	0.641095
days since last contact	50	0.500308
consumer confidence index	45	0.448758
last contacted month	43	0.431595
consumer price index	34	0.342102
education level	33	0.325749
employment variation rate	30	0.298557
age	27	0.270524
job type	24	0.241598
contact type	19	0.192898
previous campaign outcome	14	0.135571
euribor 3 month rate	13	0.131481
previous times contacted	13	0.129370
marital status	10	0.103279
contacts this campaign	8	0.076632

Table 39. Random Forest Predictor Importance [source: own]

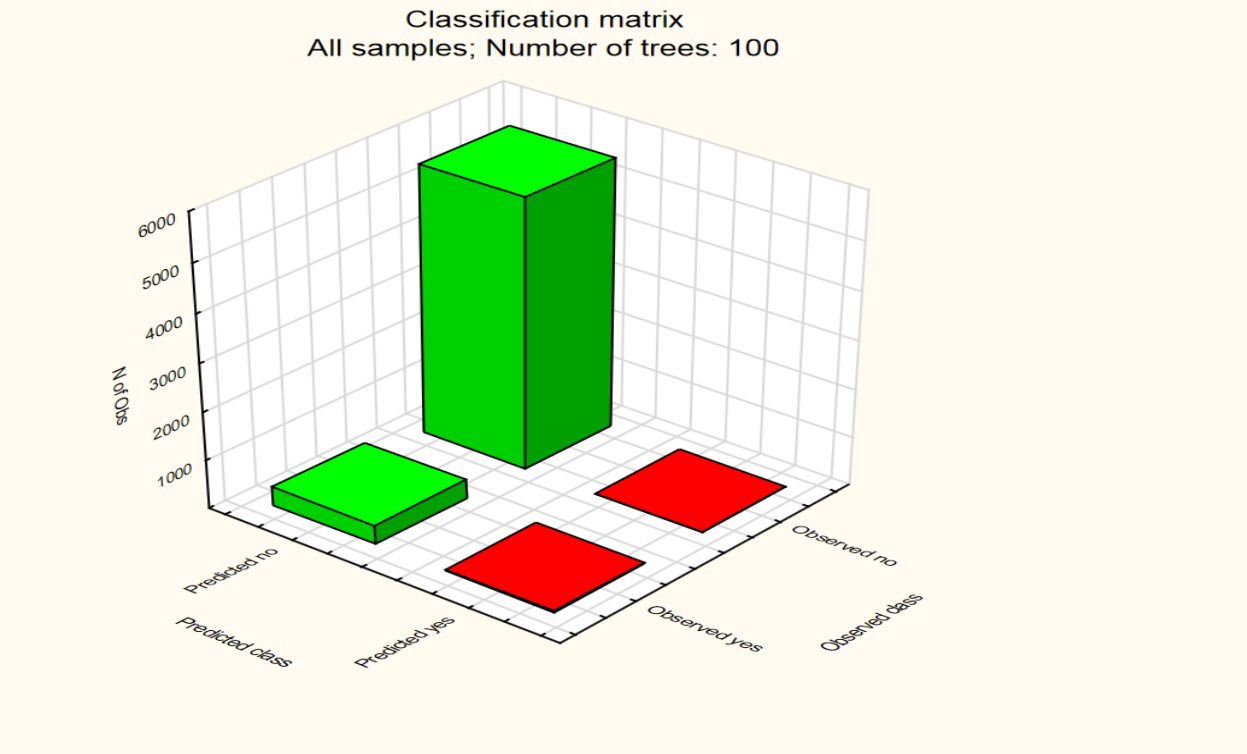
From the Bar Chart 52 and Table 39 we can see that the predictor that has the highest effect in the sales is 'last contact duration' followed by 'number of employees' and the 'days since last contact' which has half the rank of last contact duration and the predictor that has less effect in sales is the 'contacts this campaign'. Last contact duration based on the *Random Forest Model* has the highest and the most considerable importance in the model which means in the prediction of the final result of sales whether the prediction will be negative or positive. Bar Chart 52 is a graphical display of the Table 39 where there is the ranking of the importance of the predictors based on *Random Forest.* And from first predictor "last contact duration" to the last predictor "marital status", the variable rank changes considerably.

	Classification matrix (Spreadsheet233.sta) Response: sale All samples; Number of trees: 100			
	Observed	Class Predicted no	Class Predicted yes	Row Total
Number	no	5572	3	5575
Column Percentage		93.40%	10.00%	
Row Percentage		99.95%	0.05%	
Total Percentage		92.93%	0.05%	92.98%
Number	yes	394	27	421
Column Percentage		6.60%	90.00%	
Row Percentage		93.59%	6.41%	
Total Percentage		6.57%	0.45%	7.02%
Count	All Groups	5966	30	5996
Total Percent		99.50%	0.50%	

Table 40. Random forest Classification matrix percentage [source: own]

	Classification matrix (Spreadsheet233.sta) Response: sale All samples; Number of trees: 100	
	Class Predicted no	Class Predicted yes
Observed no	5572.000	3.00000
Observed yes	394.000	27.00000

Table 41. Random Forest Classification Matrix [source: own]



Bar Chart 53.Random Forest Classification Matrix [source: own]

The classification matrix both as histogram in Bar Chart 53 and as numbers shown in Table 40 and Table 41 shows us the accuracy of prediction. 5572 cases against 3 are predicted correctly for negative sales category and 394 cases against 27 are predicted correctly for the positive category of sales. The "yes" category does not have a very satisfactory result and therefore we will check other decision trees model to make a comparison of the result in order to find the most satisfactory model.

The lift chart provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting a binomial (categorical) outcome variable (dependent variable); for multinomial (multiple-category) outcome variables, lift charts can be computed for each category. (StatSoft, 2015)

In this way, in the chart of "no" category in Figure 13 we can see that by taking the top 10 percent (shown on the x axis) of cases classified into the respective category with the greatest certainty (classification probability), you would end up with a sample that had almost 1.1 times as many cases belong to the respective category when compared to the baseline random selection (classification) model.

Figure 14 shows the chart of "yes" category and we can see that by taking the top 10 percent (shown on the x axis) of cases classified into the respective category with the greatest certainty (classification probability), you would end up with a sample that had almost 5 times as many cases belong to the respective category when compared to the baseline random selection (classification) model.

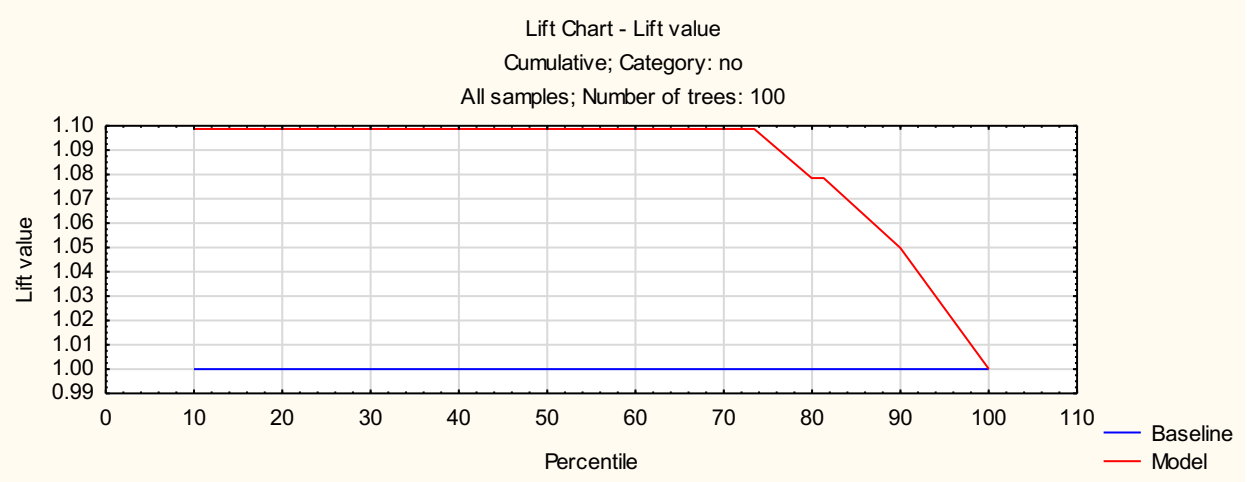


Figure 13.Random Forest Lift Chart for Category No [source: own]

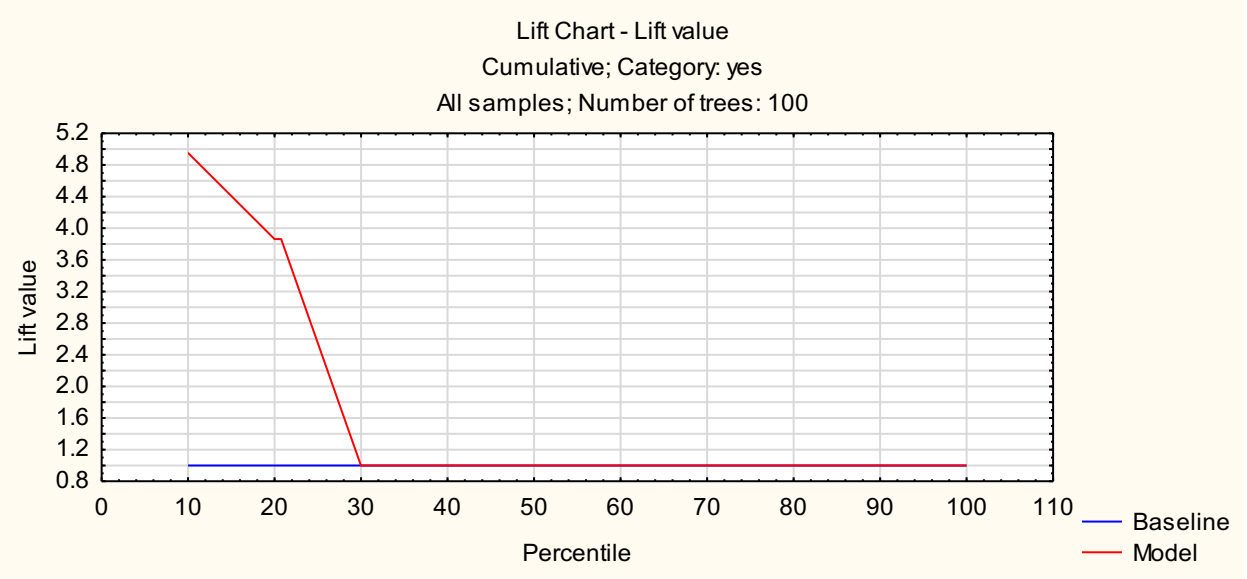


Figure 14.Random Forest Lift Chart for Category Yes [source: own]

	Misclassification cost (Spreadsheet233.sta) Response: sale	
	Class no	Class yes
no		1.000000
yes	1.000000	

Table 42. Random Forest misclassification cost [source: own]

The misclassification cost by default in all decision trees is set as 1, as we will see in other decision trees example included in this practical part (*Boosted Trees and C&RT*).

Color maps of predicted category frequencies relative to the total observed class frequency for sale
File Name: 1randomforest.xml
Color band of percentages = 0% >0 and <10 % 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

(Actual)	1 no (Predicted) RandomForestModel	2 yes (Predicted) RandomForestModel
no	95.07%	0.05%
yes	59.43%	4.07%
	Overall Accuracy = 85.82%	

Table 43. Color maps of predicted category frequencies for Random Forest model [source: own]

Rapid deployment tool was used after the model was created based on the code generated by the model. Table 43 shows the color maps of predicted category frequencies we can see the that the overall accuracy is 85.82%.

5.3.2 Boosted Trees

As well as Random Forest, boosted trees are used to build a prediction model based on several categorical predictor variables. The graph in Figure 19 indicates the particular number of trees (boosting steps) that resulted in the lowest average squared error. That solution of optimal number of trees 171 is likely near the prediction model with the best predictive validity.

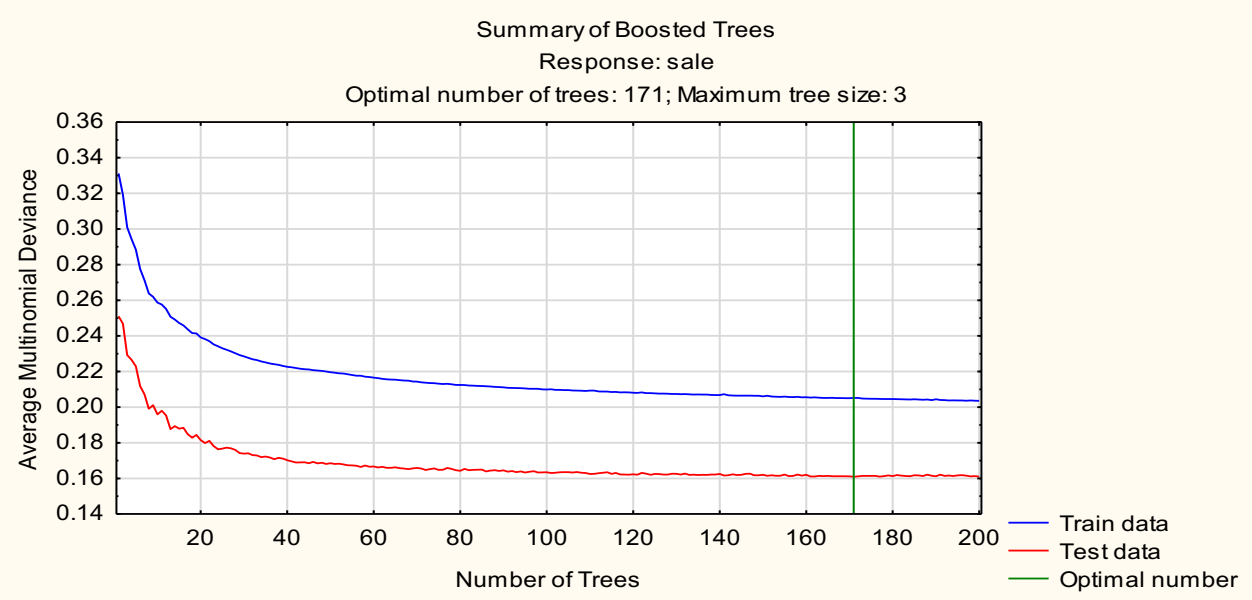


Figure 15.Summary of Boosted Trees [source: own]

	Risk estimates (Spreadsheet233.sta) Response: sale	
	Risk Estimate	Standard error
Train	0.185044	0.005644
Test	0.151397	0.008472

Table 44.Boosted Trees Risk Estimates [source: own]

Risk estimates shown in Table 44 are calculated as the proportion of cases incorrectly classified by the trees (in the respective type of sample) and in the case of the model build on the train sample is 0.185 and the standard error is equal to 0.0056. In the case of test sample, the standard error is shown to be higher with a value of 0.0085.Figure 17 shows the chart of "no" category for the boosted tree model and we can see that by taking the top 10 percent (shown on the x axis) of

cases classified into the respective category with the greatest certainty (classification probability), you would end up with a sample that had almost 1.12 times as many cases belong to the respective category when compared to the baseline random selection (classification) model. Instead on Figure 18 we have the chart of "yes" category and we can see that by taking the top 10 percent (shown on the x axis) of cases classified into the respective category with the greatest certainty (classification probability), you would end up with a sample that had almost 5 times as many cases belong to the respective category when compared to the baseline random selection (classification) model.

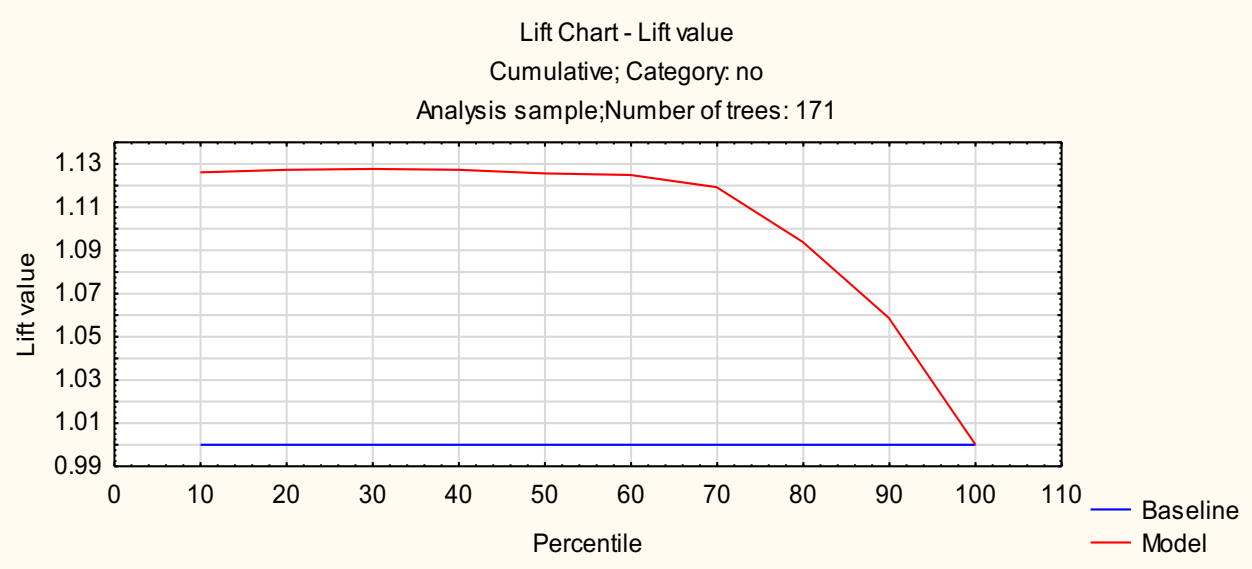


Figure 16. Boosted Trees Lift Chart Category No [source: own]

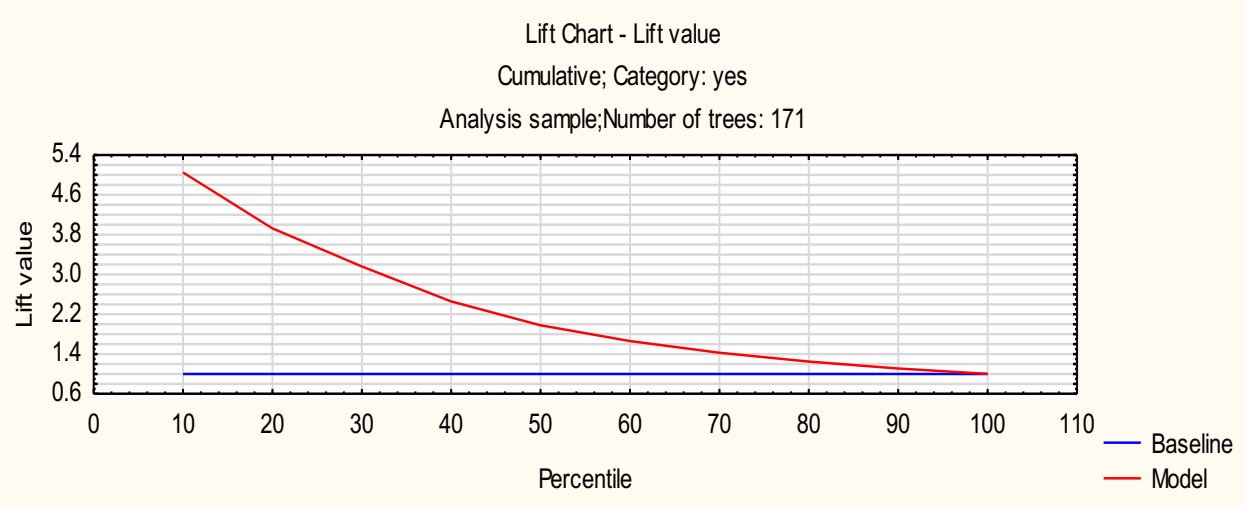
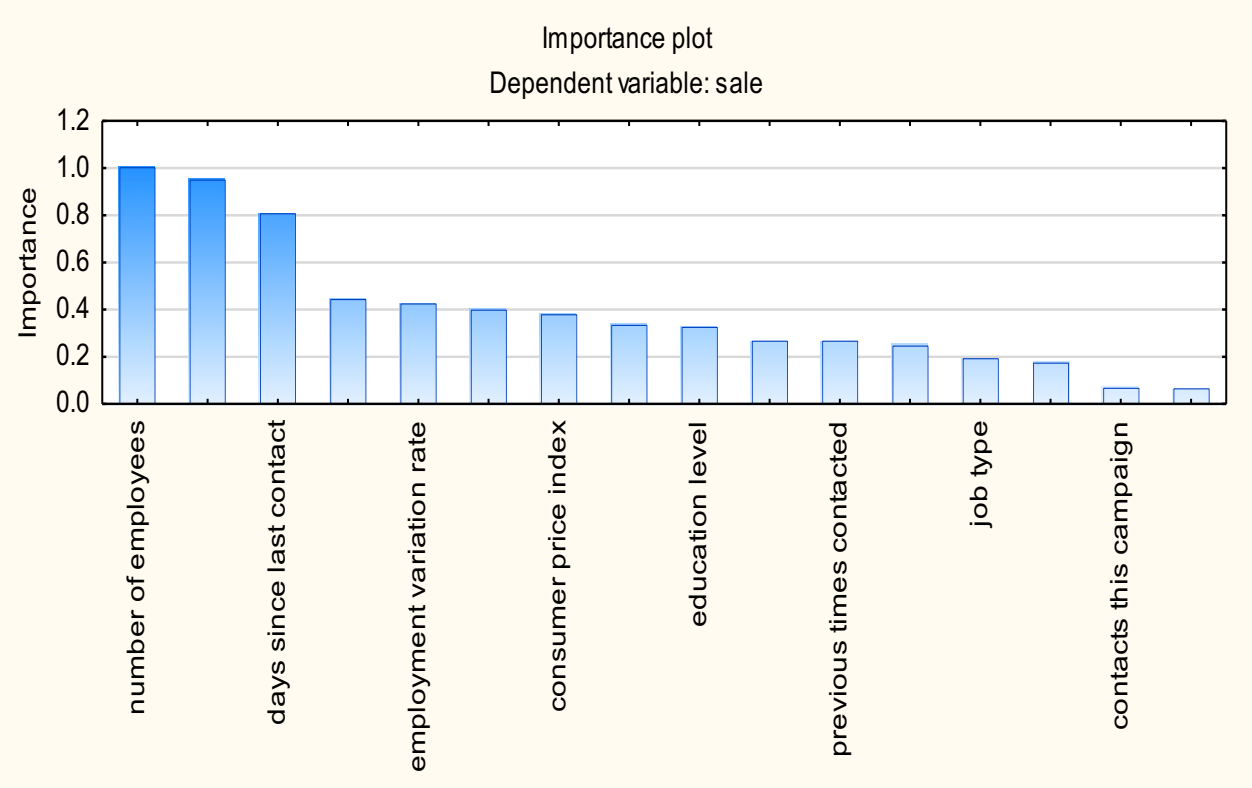


Figure 17.Boosted Trees Lift Chart Category Yes [source: own]

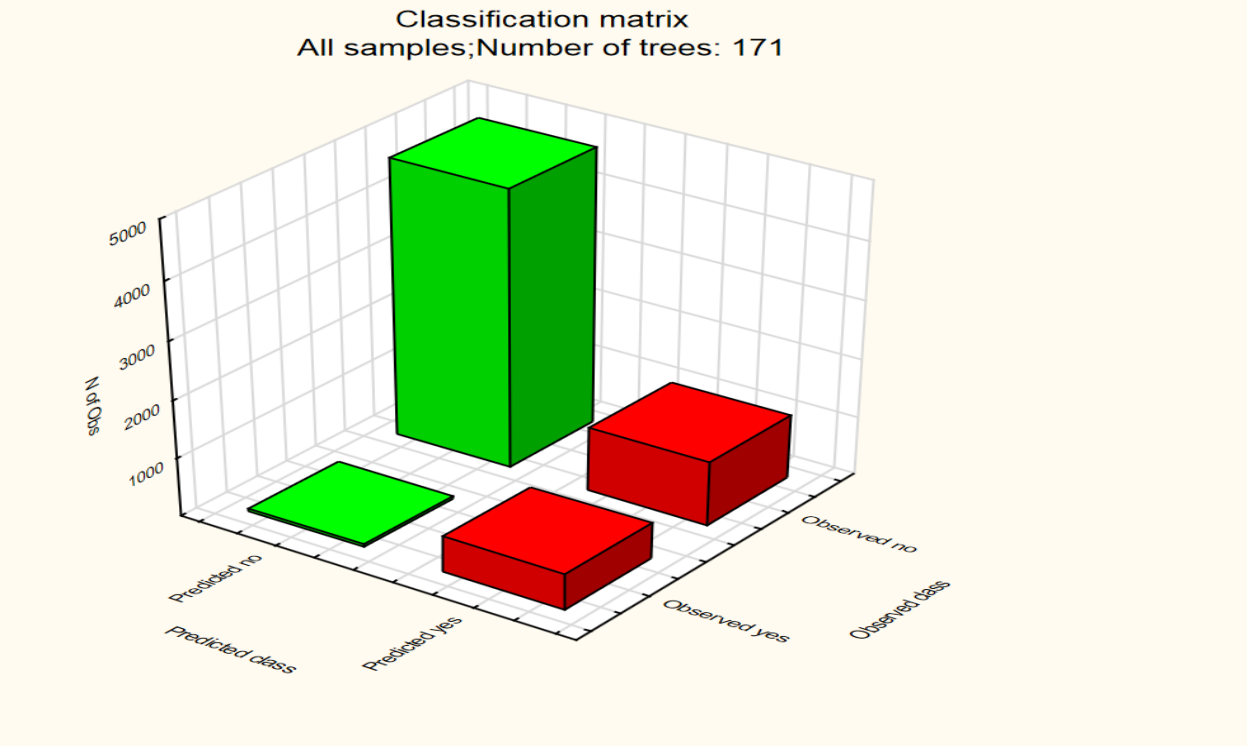


Bar Chart 54.Boosted Trees Importance Plot [source: own]

	Predictor importance (Spreadsheet233.sta) Response: sale	
	Variable Rank	Importance
number of employees	100	1.000000
last contact duration	95	0.947550
days since last contact	80	0.804498
last contacted month	44	0.442094
employment variation rate	42	0.422463
consumer confidence index	40	0.396615
consumer price index	38	0.377141
contact type	33	0.332440
education level	32	0.323623
previous campaign outcome	26	0.264816
previous times contacted	26	0.264816
euribor 3 month rate	24	0.244519
job type	19	0.190635
age	17	0.172308
contacts this campaign	7	0.065300
marital status	6	0.063628

Table 45. Boosted Trees Predictor Importance [source: own]

Bar Chart 54 and Table 45 gives as an overlook of the predictor importance. As well as in the Random Forest the higher contribution is given by predictor "last contact duration" and the least by "marital status" with the same variable ranks. But the variable rank changes considerably for the other predictors between *Boosted Trees* and *Random Forest*. We can notice that in the *Boosted Trees* there is a small difference between the first and the second and third rank comparing to *Random Forest*.



Bar Chart 55.Boosted Trees Classification Matrix [source: own]

	Classification matrix (Spreadsheet233.sta) Response: sale All samples;Number of trees: 171	
	Class Predicted no	Class Predicted yes
Observed no	4761.000	1100.000
Observed yes	47.000	616.000

Table 46. Boosted Trees Classification Matrix [source: own]

Classification matrix either in form of tables in Table 46 and Table 47 or in the form of histogram in Bar Chart 55 shows another form of accuracy of prediction. We can see that comparing to *Random Forest*, *Boosted Trees* shows a higher level of accuracy for the "YES"

category where 616 case against 47 are predicted correctly and 4761 case against 1100 are predicted correctly for the "No" category. Figure 18 and Figure 19 show the general look of the structure of the tree graph build by Boosted Trees Model where the tree as we can see its build based on the variable that has the second highest importance in the rank of variable importance.

	Classification matrix (Spreadsheet233.sta) Response: sale All samples;Number of trees: 171			
	Observed	Predicted no	Predicted yes	Row Total
Number	no	4761	1100	5861
Column Percentage		99.02%	64.10%	
Row Percentage		81.23%	18.77%	
Total Percentage		72.98%	16.86%	89.84%
Number	yes	47	616	663
Column Percentage		0.98%	35.90%	
Row Percentage		7.09%	92.91%	
Total Percentage		0.72%	9.44%	10.16%
Count	All Groups	4808	1716	6524
Total Percent		73.70%	26.30%	

Table 47. Boosted Trees Classification Matrix Percentage [source: own]

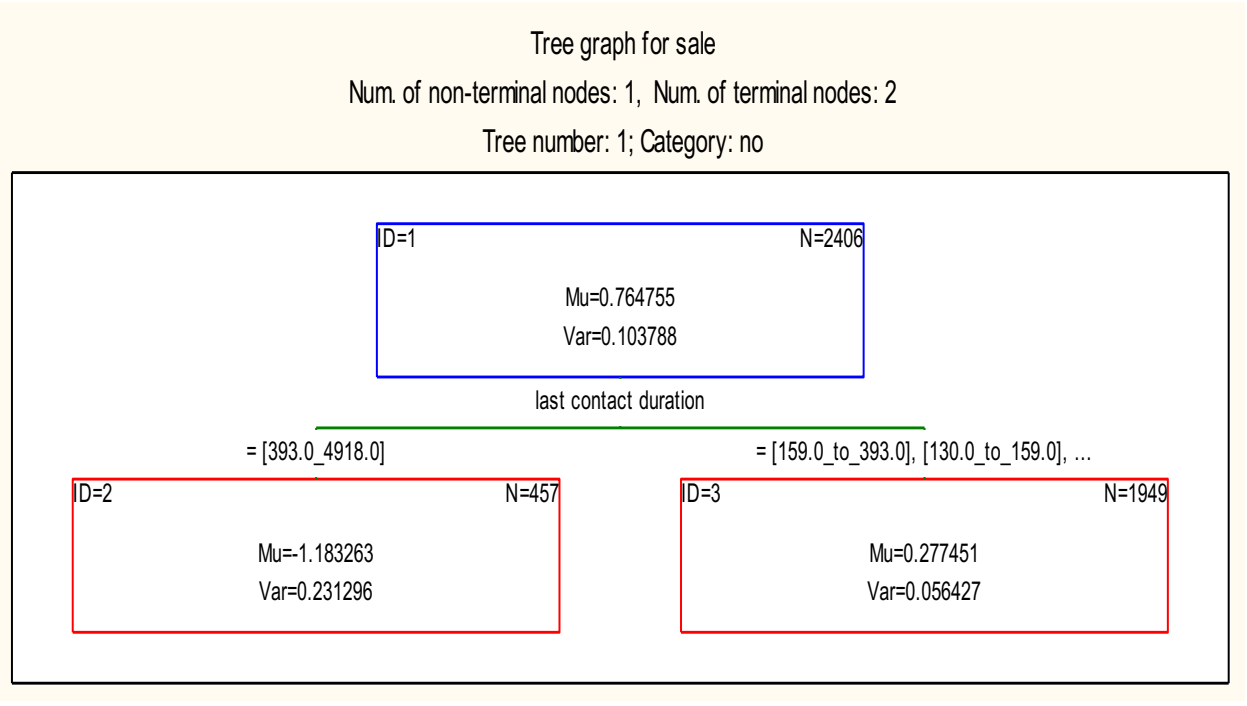


Figure 18. Boosted Trees,Tree graph for sales for Category No [source: own]

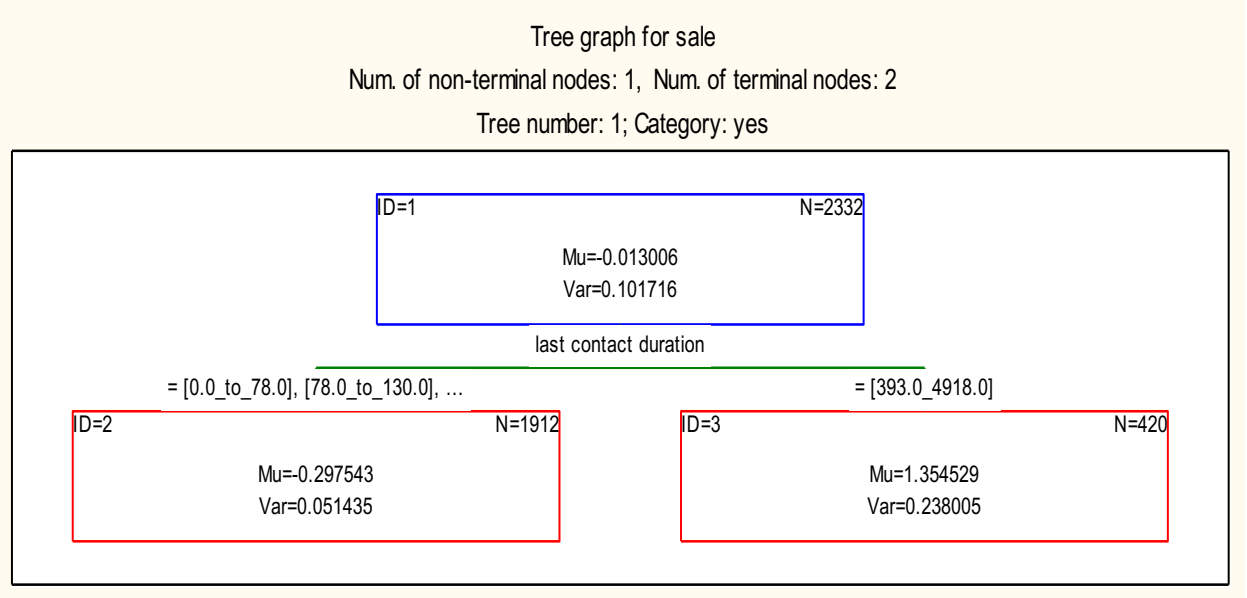


Figure 19. Boosted Trees, Tree graph for sales for Category No [source: own]

Color maps of predicted category frequencies relative to the total observed class frequency for sale
File Name: 1boostedtree.xml
Color band of percentages = 0% >0 and <10 % 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

(Actual)	1 no (Predicted) BoostTreeModel	2 yes (Predicted) BoostTreeModel
no	80.16%	14.96%
yes	5.58%	57.92%
	Overall Accuracy = 77.90%	

Table 48.Color maps of predicted category frequencies for Boosted Trees model [source: own]

As well as for *Random Forest* even for *Boosted Trees* after the model was built and the code generated from the report we used *Rapid deployment* tool. And based on Table 48 that shows the color maps of predicted category frequencies we can see that the overall accuracy is 77.9% which is less than the *Random Forest* model accuracy.

5.3.3 C&RT decision trees

C&RT stands for classification and regression trees but as our dataset is categorical we will focus in the classification problem. It is a non-parametric approach which means that no distribution assumptions are made about the data compared to generalized linear model which assumes that the depend variable follows a specific distribution such as binomial. Splits for the decision trees are made with the variable that best differentiate the target variable.

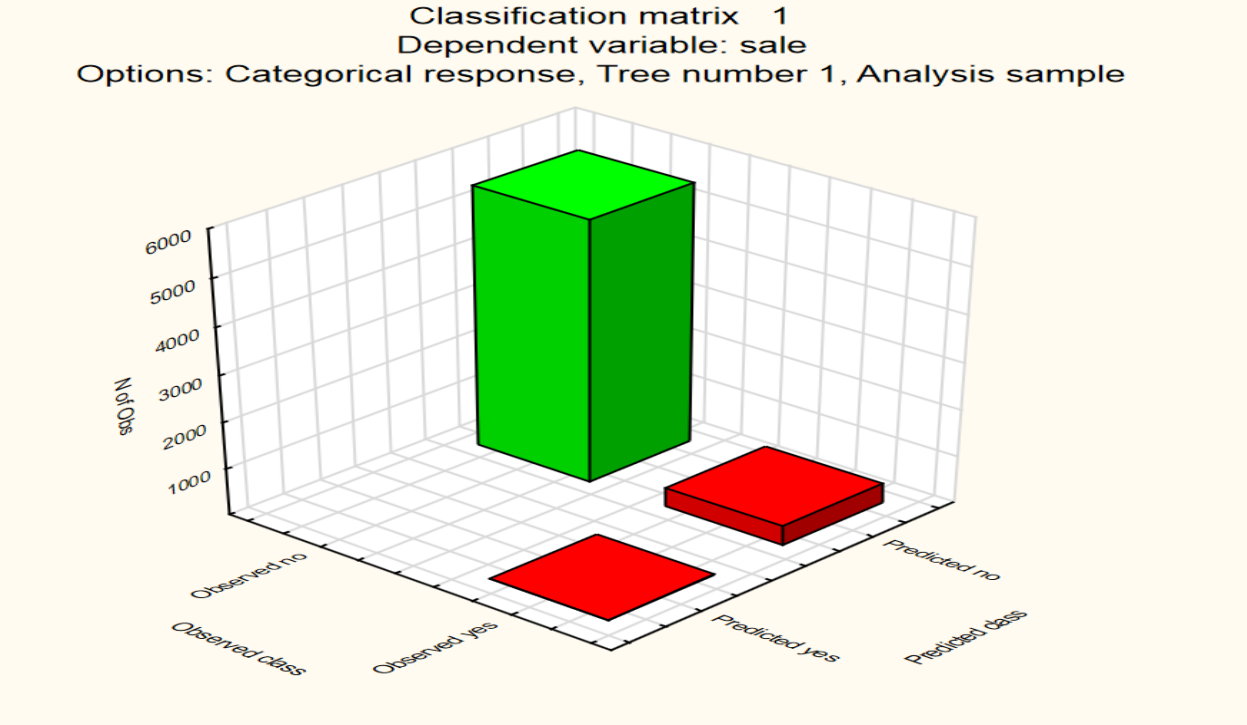
Table 49 and 50 represents the classification matrix which show the number of cases that are correctly classified and we can notice that 144 cases against 88 are predicted correctly for the category "yes" and 5575 cases are predicted correctly for the category "no" and there is no prediction for the bad prediction of the category "no". The same results we can see even from Bar Chart 56.

	Classification matrix 1 (Spreadsheet233.sta) Dependent variable: sale Options: Categorical response, Tree number 1, Analysis sample	
	Observed no	Observed yes
Predicted no	5575.000	417.0000
Predicted yes		4.0000

Table 49.Classification matrix [source: own]

	Classification matrix 1 (Spreadsheet233.sta) Dependent variable: sale Options: Categorical response, Analysis sample			
	Observed	Predicted no	Predicted yes	Row Total
Number	no	5575		5575
Column Percentage		93.04%	0.00%	
Row Percentage		100.00%	0.00%	
Total Percentage		92.98%	0.00%	92.98%
Number	yes	417	4	421
Column Percentage		6.96%	100.00%	
Row Percentage		99.05%	0.95%	
Total Percentage		6.95%	.7%	7.02%
Count	All Groups	5992	4	5996
Total Percent		99.93%	.7%	

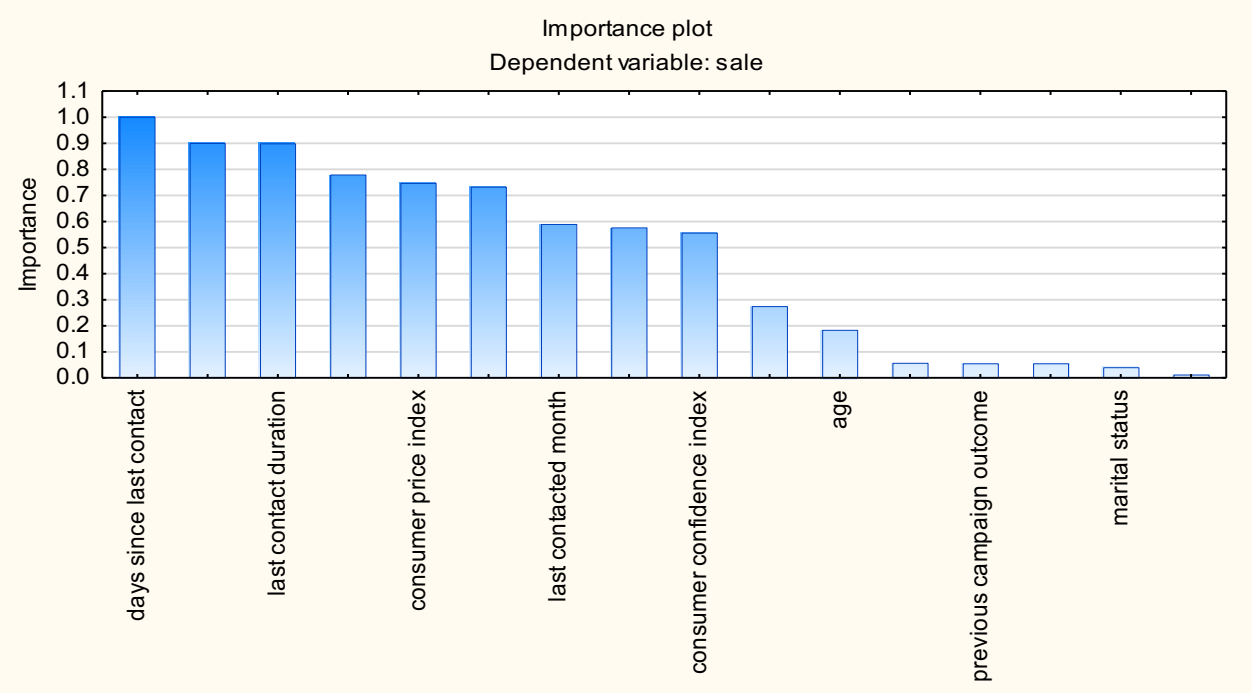
Table 50.Classification matrix percentage [source: own]



Bar Chart 56.Classification matrix [source: own]

	Predictor importance 1 (Spreadsheet233.sta) Dependent variable: sale Options: Categorical response, Tree number 1	
	Variable rank	Importance
days since last contact	100	1.000000
number of employees	90	0.899262
last contact duration	90	0.897902
employment variation rate	78	0.779114
consumer price index	75	0.746196
euribor 3 month rate	73	0.731756
last contacted month	59	0.589129
contact type	58	0.575644
consumer confidence index	56	0.555431
education level	27	0.274133
age	18	0.182911
job type	6	0.056130
previous campaign outcome	5	0.054516
previous times contacted	5	0.054516
marital status	4	0.039595
contacts this campaign	1	0.011307

Table 51.Predictor Importance [source: own]



Bar Chart 57.Importance plot [source: own]

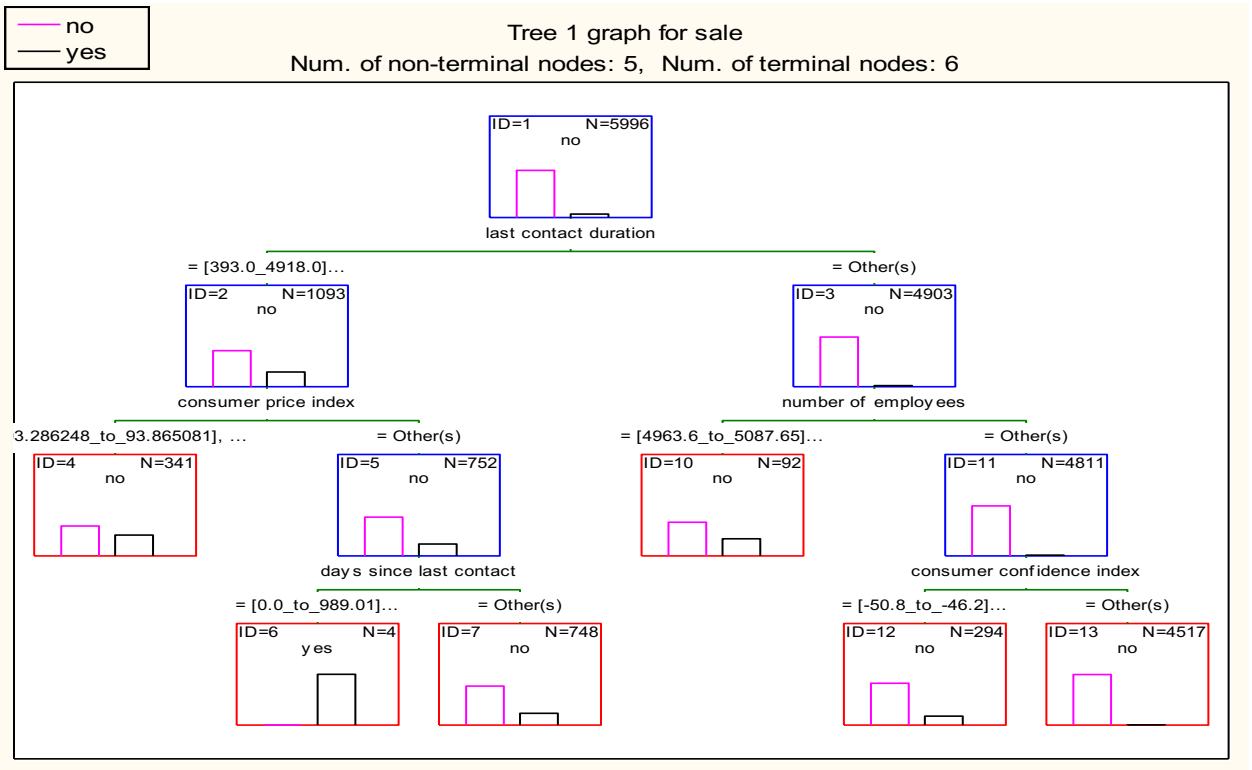


Figure 20. Tree 1 graph for sale [source: own]

Table 51 and Bar Chart 57 like in the previous models show the importance of predictors in case of the model built by *C&RT*. As in the previous two models the highest effect comes from the "last contacted duration" and different from the previous model the least important variable is contacts in campaign and marital status is the second last. As in the case of Boosted Trees there is not too much difference between the first variable and the second one on contrary to Random Forest model.

Figure 20 shows the Tree graph for sale which is build based on the predictor that has the highest influence in the model "last contact duration". Considering the importance analysis above and the difference in the variable rank among "last contact duration" and other variables, the tree as well shown figure is split in nodes. Each node then will be analyzed below. Figure 20 shows the weight of each predictor in variable sale for node 1 and the density, the higher N value the denser is the graph. We can see the density difference from node 1 Figure 20 which has the higher density with a N=5996 and node 6 which has the lowest density with a N=4. And the detailed graphs and bar chart for each node in the tree can be found in the Appendix section.

Table 52 shows the nodes included in the tree.

	Tree sequence (Spreadsheet233.sta) Dependent variable: sale		
	Terminal node	Resubstitution cost	Node complexity
Tree 1	6	0.069546	0.000000
Tree 2	5	0.069546	0.000000
Tree 3	4	0.069546	0.000000
Tree 4	1	0.070213	0.000222

Table 52.Tree Sequence [source: own]

Table 52 shows the number of terminal nodes for each tree and resubstitution cost for each tree. We can notice as well that the Resubstitution cost for the sample from which the splits were determined increases as the pruning proceeds (and as the tree number increases from 1 to 4, the number of terminal nodes decreases, i.e., consecutive tree numbers are increasingly "pruned-back"); this is to be expected since the fit for the data from which the tree was computed will become worse the fewer terminal nodes are included.

As in the case of two previous models we need to use the *Rapid Deployment* to check the accuracy of the model built by C&RT. Table 53 will the color maps of predicted category

frequencies shows us an accuracy of 85.52% close to the *Random Forest* model which shows the higher accuracy.

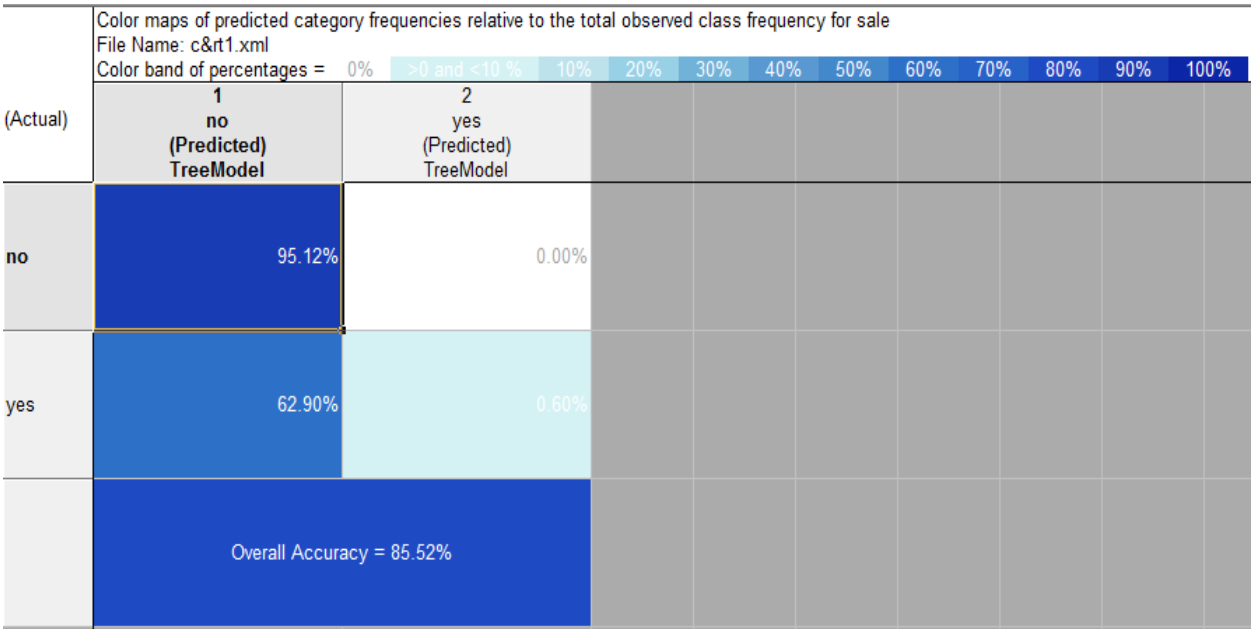


Table 53.Color maps of predicted category frequencies for C&RT model [source: own]

5.3.4 Comparative evaluation of the three models

From the tables of Color maps of all the models (Table 43, Table 48 and Table 53) we noticed that Random Forest model has higher overall accuracy of 85.82%, followed by C&RT model with an overall accuracy of 85.52% and Boosted Tree with 77.90% but from the same table we can notice that Boosted Tree has the higher accuracy for the correct predictions (no-no and yes-yes)

	Summary of Deployment (Error rates) (Spreadsheet233.sta)		
	BoostTreeModel	RandomForestModel	TreeModel
Error rate	0.152435	0.066211	0.069546

Table 54.Summary of Deployment [source: own]

The rapid deployment tool allows you to load multiple data mining at once. This allows us to compare the Random Forest model, Boosted Trees and C&RT model that we built before, simultaneously. And Table 54 shows the standard error for all three model and we can notice that Random Forest has the lowest value of standard error followed by C&RT model and Boosted Tree model which has the highest value of standard error.

The gains chart provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting categorical dependent variable. This charts show the percentage of correctly classified observations for a given category. Specifically, the chart summarizes the utility that one can expect by using the respective predictive models, as compared to using baseline information only. In Figure 21 for the case of "no" category we can see that all three models have a very high accuracy but there is a considerable change between them in prediction of the "yes" category shown in Figure 22. Figure 22 shows that Boosted Tree model is the best for the prediction of the "yes" category. For Boosted Tree model, if we consider the top two deciles (after sorting based on the confidence of prediction), we will correctly classify approximately 90 percent of the cases in the population belonging to category "yes." The baseline model serves as a comparison to measure the utility of the respective models for classification.

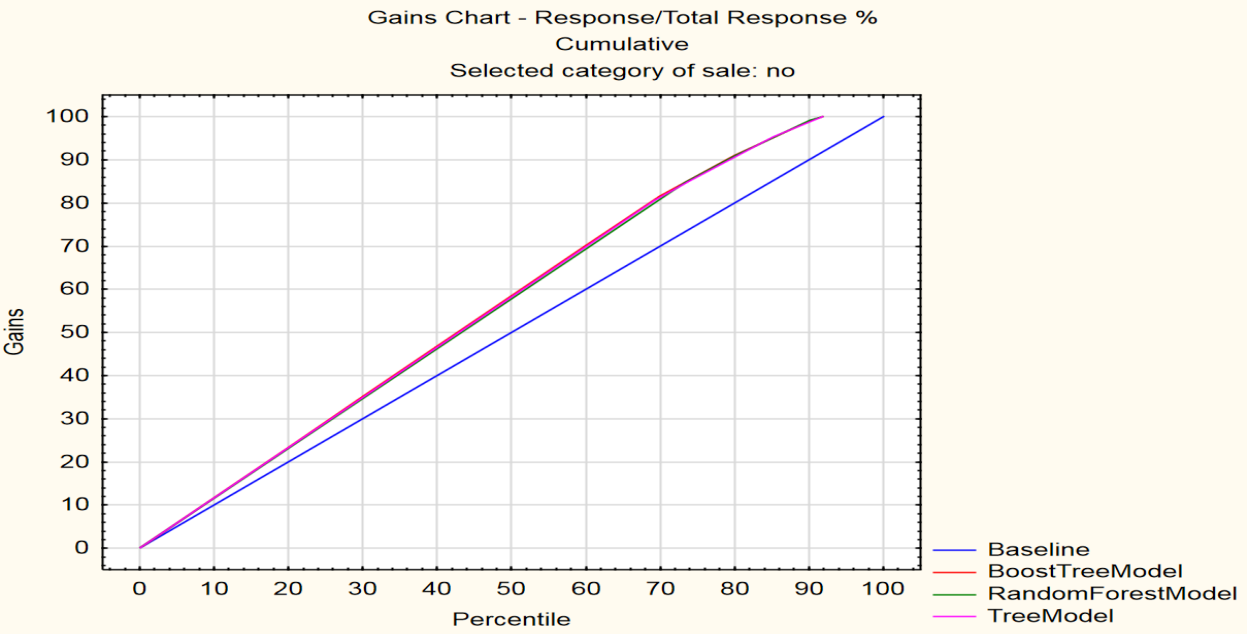


Figure 21.Gains Chart 'no' category [source: own]

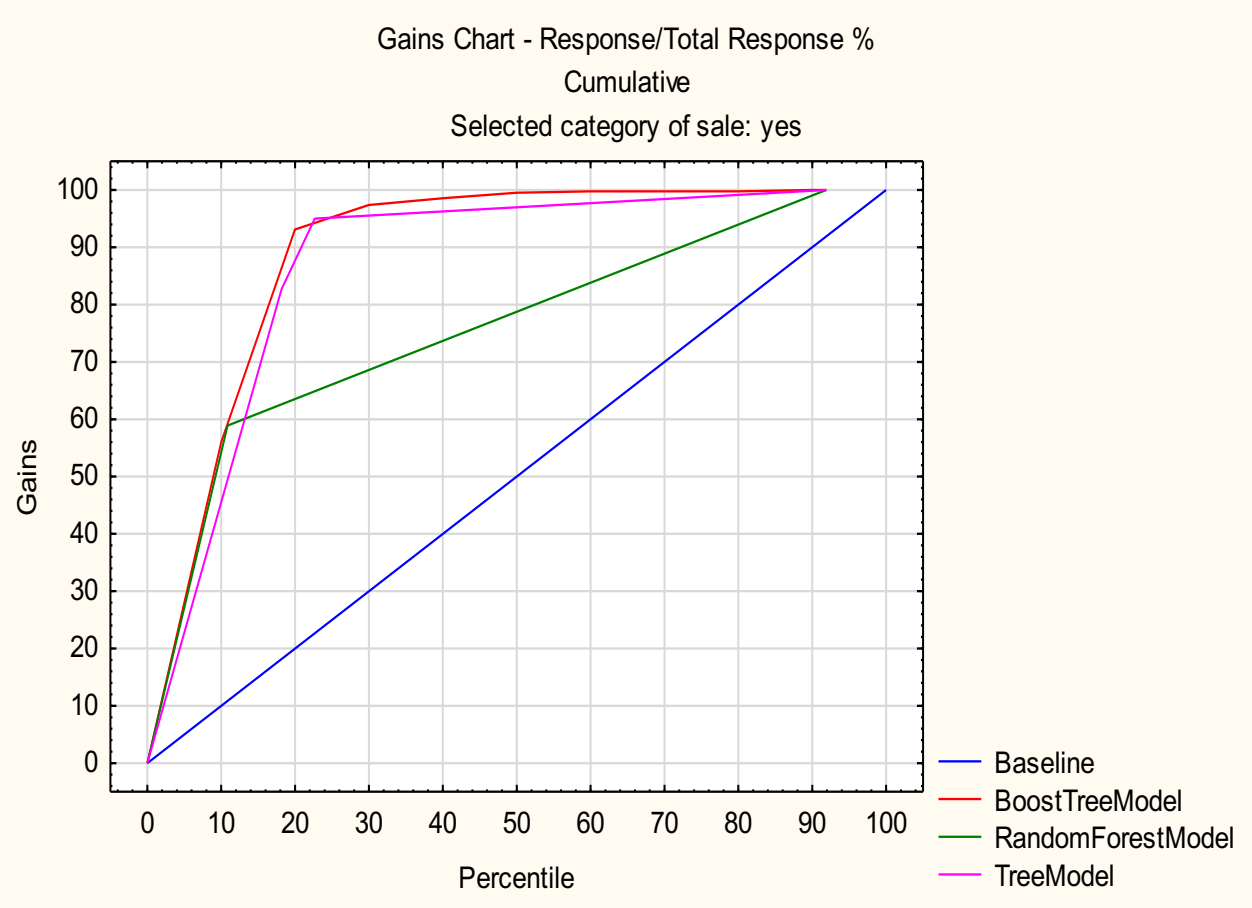


Figure 22. Gains Chart 'yes' category [source: own]

Figure 23 and 24 show the Lift chart for both categories and comparison of the accuracy of the prediction among the three models. We can notice than in both figures Boosted Trees model is the best among Random Forest and C&RT model for prediction purposes. If you consider the top two deciles, you would end up with a sample that has almost 99.9% the number of 'no' customers and 30 % the number of "yes customers when compared to the baseline model. Figure 23 shows us how well we did for predictive accuracy when we were raking them in terms of how strongly the model feels that that customer is a non-sale customer.

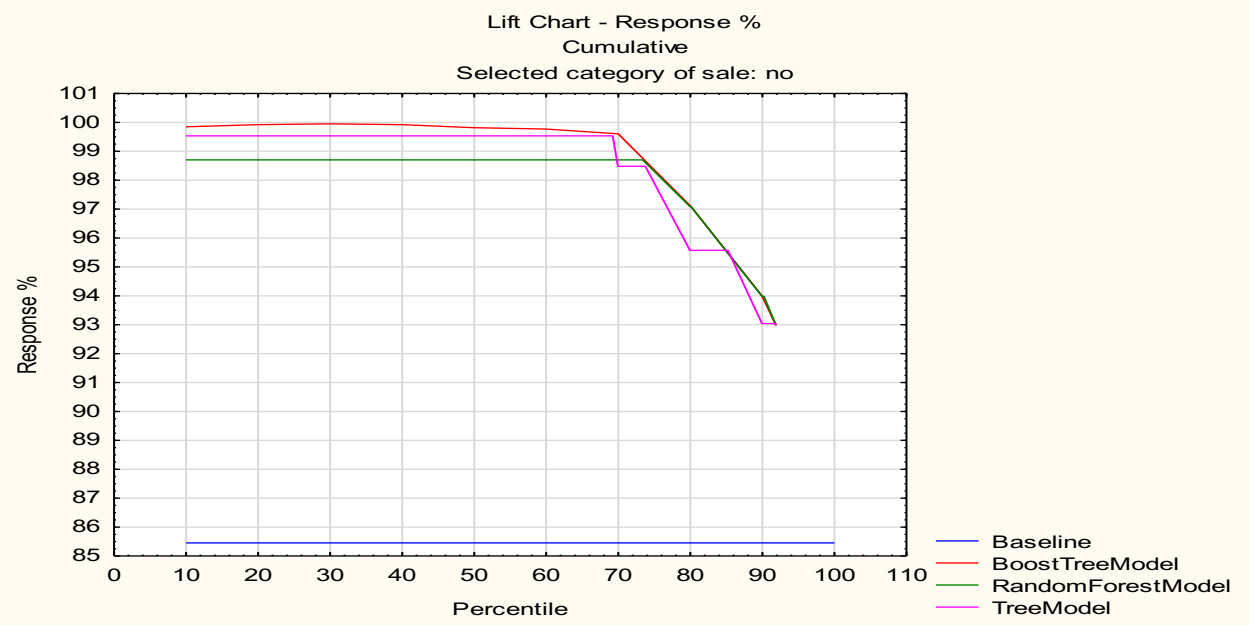


Figure 23. Lift Chart 'no' category [source: own]

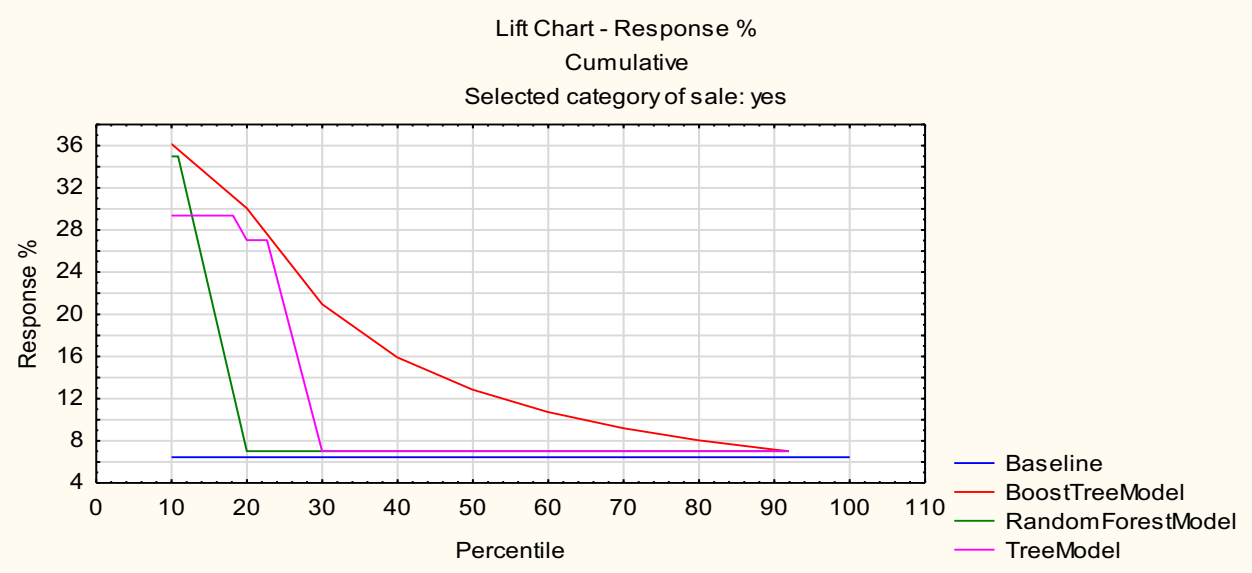


Figure 24. Lift Chart 'yes' category [source: own]

6. Conclusion

Big Data is the topic of nowadays. Advanced technologies and tools are used to analyze the huge amount of data produced in day to day basis. And on this thesis after a theoretical research conducted on the first part of it then a concrete dataset was analyzed based on some techniques

pointed out from the research in order to reach the objectives pointed out in the Objectives chapter and the tools used were three data mining algorithms for predictive modeling. The objectives were to take a concrete dataset and apply all steps of a data mining analysis so to predict which of the variables or factors effected most 'sales' . We started with data preparation which was the first section of the practical part that dealt with cleaning the data from extreme values in order to construct a good dataset for the next phases following: data exploration and data modeling. As the data set is made of categorical variables it was checked out for extreme values and extreme values are considered all categories that have less than 5% of the data. In case of extreme they were either recoded into new variables or dismissed from the data set completely. After the dataset was clean of the extreme values in the next section we started exploring the relationship among all the variables and the dependent variable 'sales' in order to be able to find and remove from the dataset the variables that where not associated to variable 'sales' and which would not contribute to the analysis and might as well lead to wrong analysis. From the second section of Data Exploration we realized that several variables including 'housing loan', 'personal loan' and 'Part of the week' are not associated so we removed them from the dataset to reach a better structure of the dataset ready to apply in it the models for prediction. So the last section of the practical part applied predictive models in cleaned dataset from extreme values and unimportant variables. The three models applied were "Random Forest", "Boosted Trees" and "C&RT" and three different model were built that were predicting the importance of the variables in "sales" and predicting some customer behavior as well, meaning if the customer were going to have more sales or more non-sales in the future. As our original dataset had very higher level of non0sales comparing to sales all three models did well in predicting the behavior of customers. The Gain Chart and Lift Chart showed as that Boosted Tree had better accuracy in the predicting the customer. Even though the overall accuracy for the Boosted Tree is smaller (77.90%) compare to other models we clearly noticed that the accuracy for the correct prediction (no-no and yes-yes) was better for the Boosted Tree model which will be our chosen model. Based on this model we can now choose the rank of the variable based on the level of effectiveness they had on 'sales'. And based on Boosted Trees the variable that mostly effected 'sales' is the 'number of employees' of the company followed by 'last contact duration' and 'days since last contact' and the variable that least effected 'sales' is the marital status of the customers.

Bibliography

Adrian, Merv. 2011. Teradata Magazine. *Teradata Magazine.* [Online] 2011. [Cited: 07 17, 2016.] http://www.teradatamagazine.com/v11n01/Features/Big-Data/.

Andritsos, Periklis. 2002. *Data Clustering Techniques Qualifying Oral Examination Paper.* March 2002.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning.* s.l. : Springer, 2006. 0-387-31073-8.

Brown, Martin. 2012. Data mining techniques. *IBM.* [Online] December 11, 2012. [Cited: July 20, 2016.] https://www.ibm.com/developerworks/library/ba-data-mining-techniques/.

Chapman, Pete, et al. 2000. CRISP-DM 1.0. *The Modeling Agency .* [Online] 2000. [Cited: July 21, 2016.] https://the-modeling-agency.com/crisp-dm.pdf.

Data Cleaning: Problems and Current Approaches. **Rahm, Erhard and Do, Hong-Hai. 2000.** 4, December 2000, IEEE Bulletin of the Technical Committee on Data Engineering, Vol. 23, p. 11.

Data mining: A conceptual overview. **Jackson, Joyce. 2002.** Claremont : Communications of the Association for Information Systems, 2002, Vol. 8, pp. 267-296. 1529-3181 .

European Money Market Institute. 2014. Frequently Asked Questions about Euribor. *European Money Market Institure.* [Online] 2014. [Cited: October 24, 2016.] http://www.emmi-benchmarks.eu.

Franks, Bill. 2012. *Wiley and SAS Business Series : Taming The Big Data Tidal Wave : Finding Opportunities in Huge Data Streams with Advanced Analytics (1).* Hoboken : Wiley, 2012. p. 334. 9781118208786.

Gendron, Michael S. 2014. *Wiley and SAS Business Series : Business Intelligence and the Cloud : Strategic Implementation Guide (1).* Hoboken : Wiley, 2014. p. 242. 9781118631720.

—. 2014. *Wiley and SAS Business Series: Business Intelligence and the Cloud: Strategic Implementation Guide(1).* Sormset : Wiley, 2014. p. 242. 9781118631720.

IBM. IBM analytics. *IBM analytics.* [Online] [Cited: July 27, 2016.] http://www.ibm.com/analytics/us/en/technology/data-integration/.

Institute for Research on Poverty. 2014. What is the consumer price index and how is it used? *Institute for Research on Poverty.* [Online] University of Wisconsin Madison, 2014. [Cited: October 24, 2016.] http://www.irp.wisc.edu/faqs/faq5.htm.

Jain, Rajni. 2012. Introduction to data mining techniques. [Online] July 26, 2012. [Cited: July 28, 2016.] http://www.iasri.res.in/ebook/expertsystem/datamining.pdf.

Kabacoff, Robert I. 2014. Statmethos. *Quick-R accessing the power of R.* [Online] 2014. [Cited: July 22, 2016.] http://www.statmethods.net/advstats/cluster.html.

Linoff, Gordon S. and Berry, Michael J.A. 2011. *Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management.* Hoboken : Wiley, 2011. p. 885. Vol. 3. 9781118087503.

Manyika, James, et al. 2011. *Big data: The next frontier for innovation, competition and productivity.* s.l. : McKinsey&company, 2011.

Minelli , Michael, Chambers, Michele and Dhiraj, Ambiga. 2012. *Wiley CIO : Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends for Today's Businesses (1).* Somerset : Wiley, 2012. p. 215. 9781118147603.

NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH. *Engineering Statistics Handbook.* [Online] NIST. [Cited: July 29, 2016.] http://www.itl.nist.gov/div898/handbook/.

Organization for Economic Co-operation and developmnet. 2016. OECD Data. *Organization for Economic Co-operation and Development.* [Online] 2016. [Cited: October 24, 2016.]

Pang-Ning , Tan, Michael , Steinbach and Vipin , Kumar. Introduction to data mining . *Introduction to data mining.* [Online] [Cited: July 19, 2016.] http://www-users.cs.umn.edu/~kumar/dmbook/index.php.

SAS Institute Inc. . 2012. SAS. *SAS.* [Online] 2012. [Cited: 07 18, 2016.] http://www.sas.com/en_th/insights/big-data/what-is-big-data.html.

Sayad, Saed. An introduction to data mining. *An introduction to data mining.* [Online] [Cited: July 25, 2016.] http://www.saedsayad.com/clustering_kmeans.htm.

The Clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses. **Schonlau , Matthias. 2002.** 3, 2002, The Stata Journal, pp. 316-327.

Tufféry, Stéphane. 2011. *Wiley Series in Computational Statistics : Data Mining and Statistics for Decision Making .* [trans.] Rod Riesco. West Sussex : Wiley, 2011. p. 717. Vol. 1. 9780470688298.

Appendix

Below are listed has all the figures and bar charts related to each node of the tree build by C&RT model.

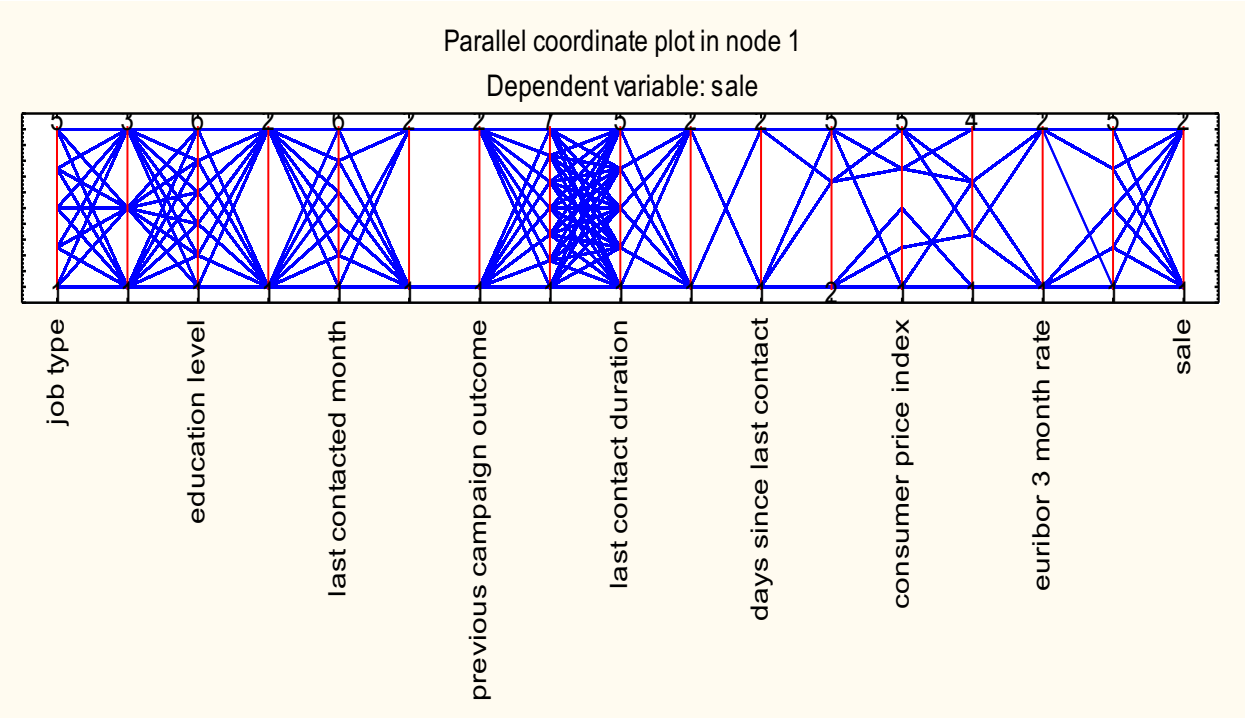
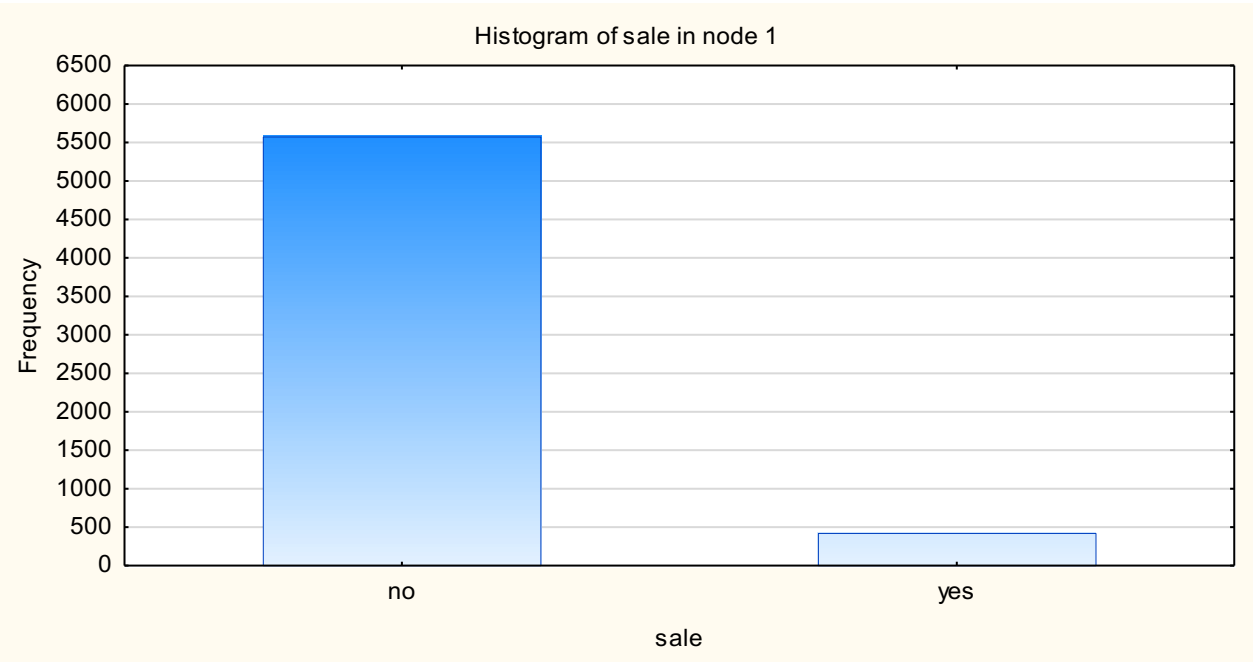


Figure 1. Parallel coordinate plot in node 1 [source: own]



Bar Chart 1. Sale in node 1[source: own]

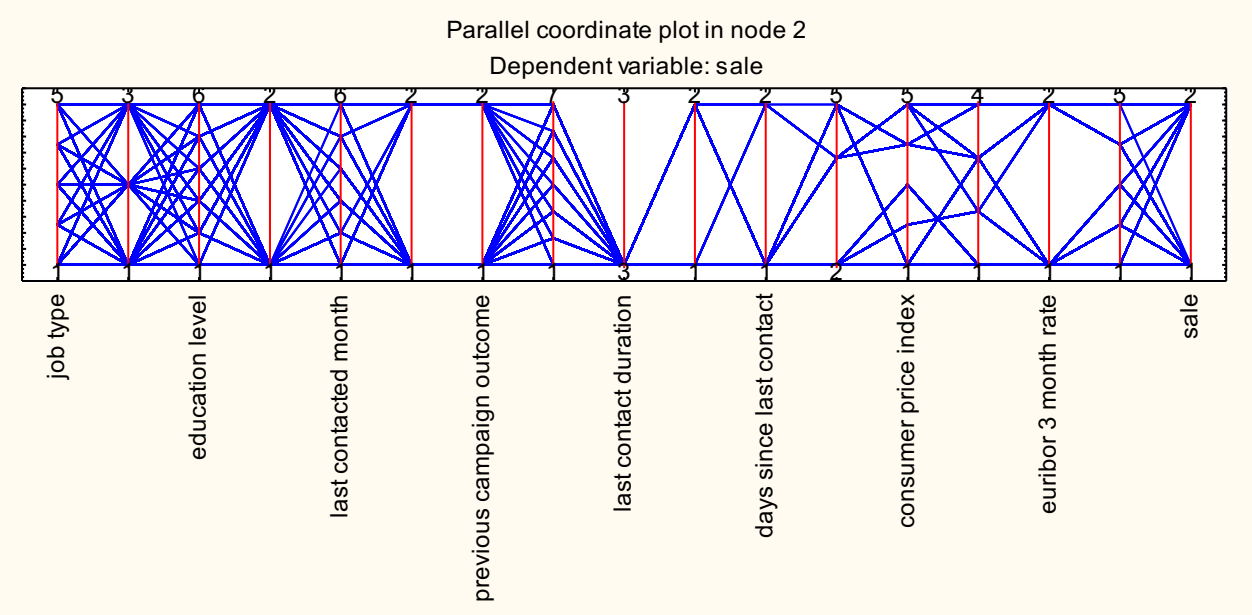
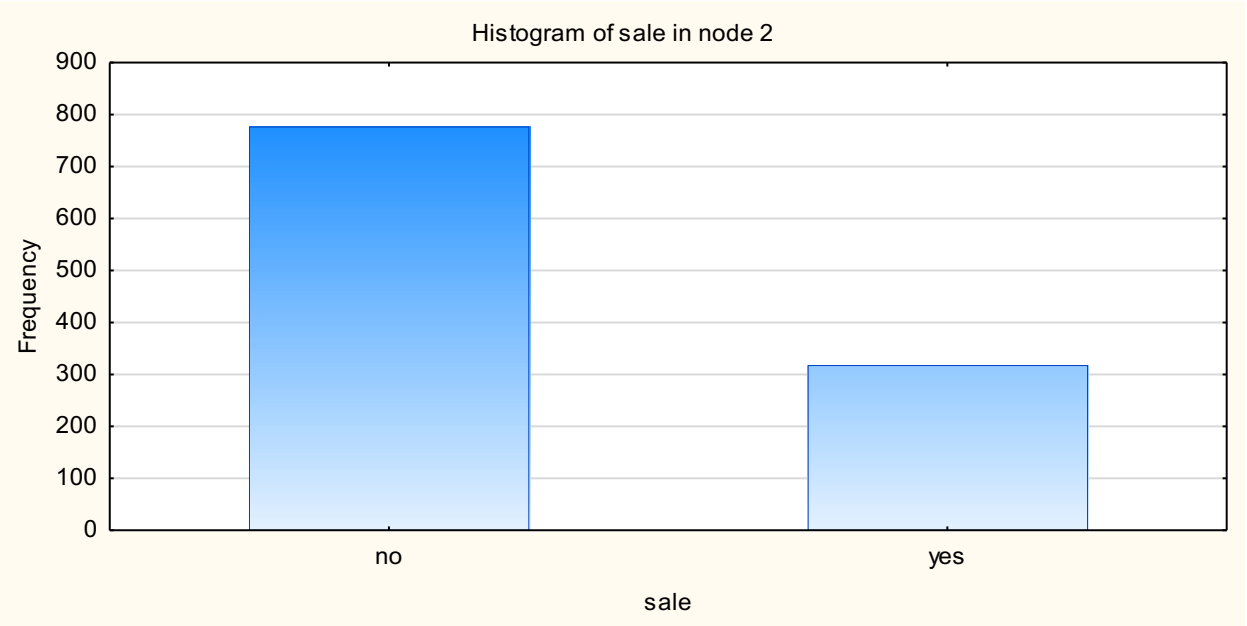


Figure 2. Parallel coordinate plot in node 2 [source: own]



Bar Chart 2. Sale in node 2[source: own]

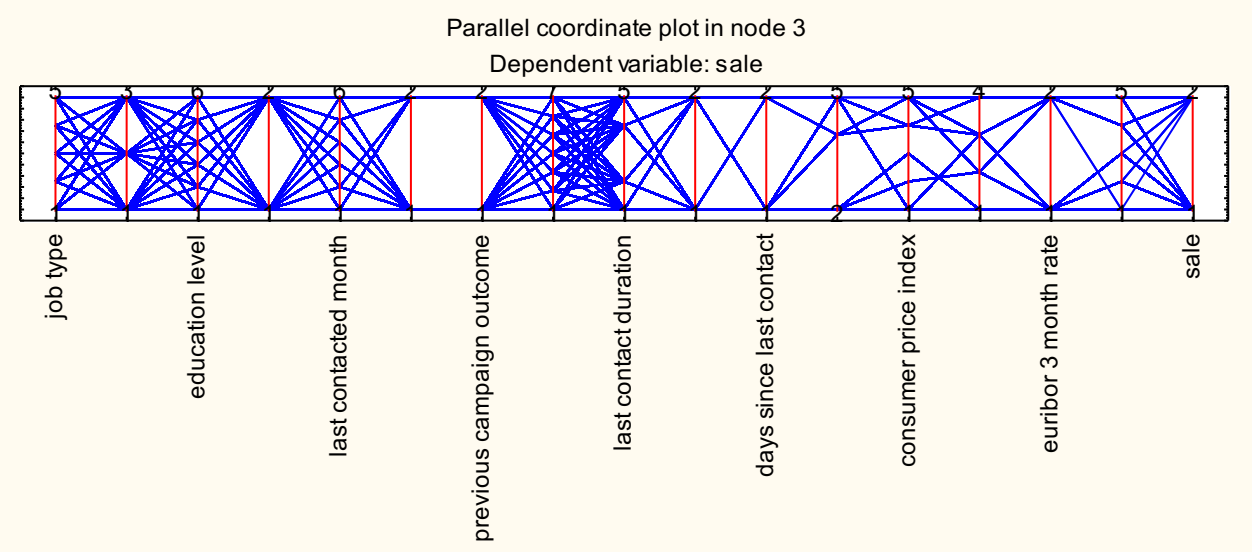
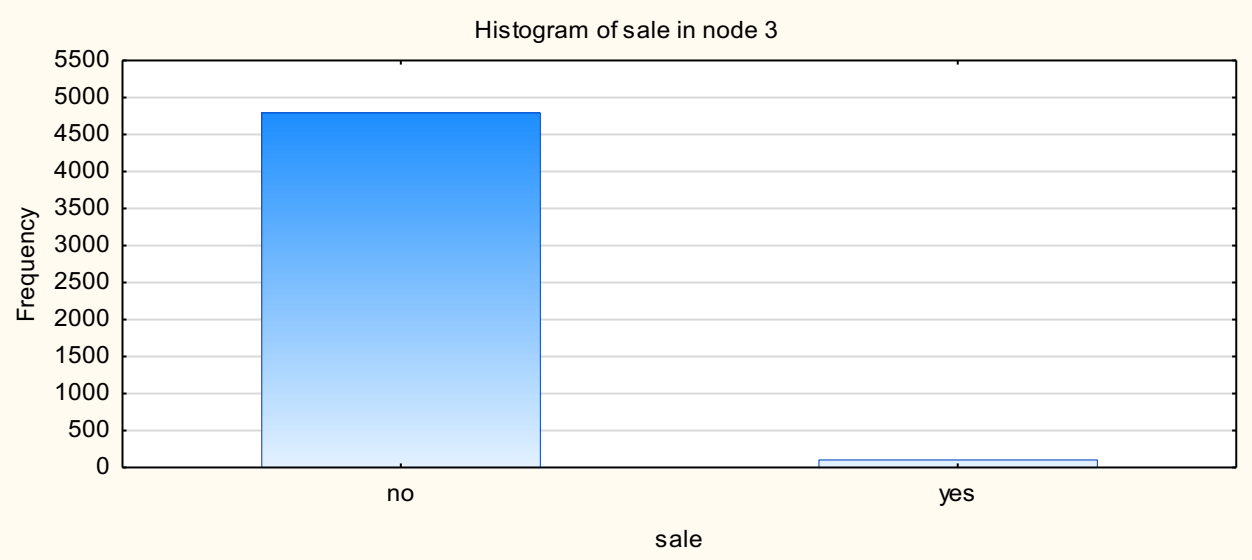


Figure 3. Parallel coordinate plot in node 3 [source: own]



Bar Chart 3. Sale in node 3 [source: own]

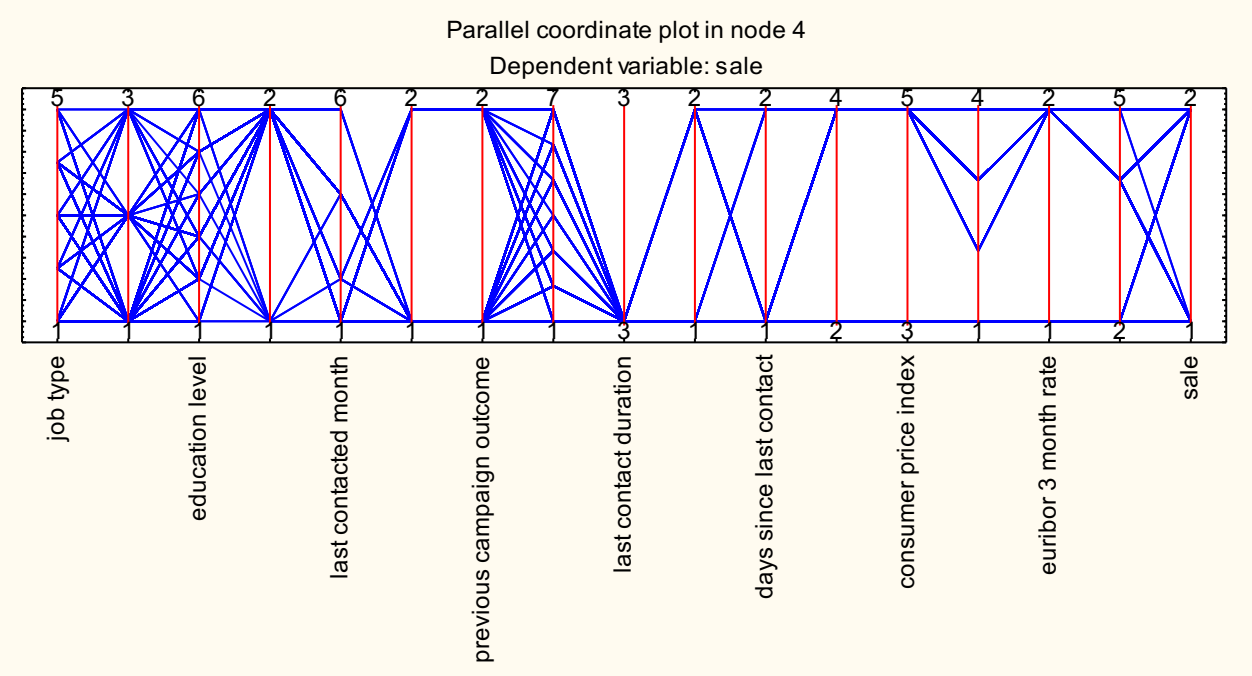
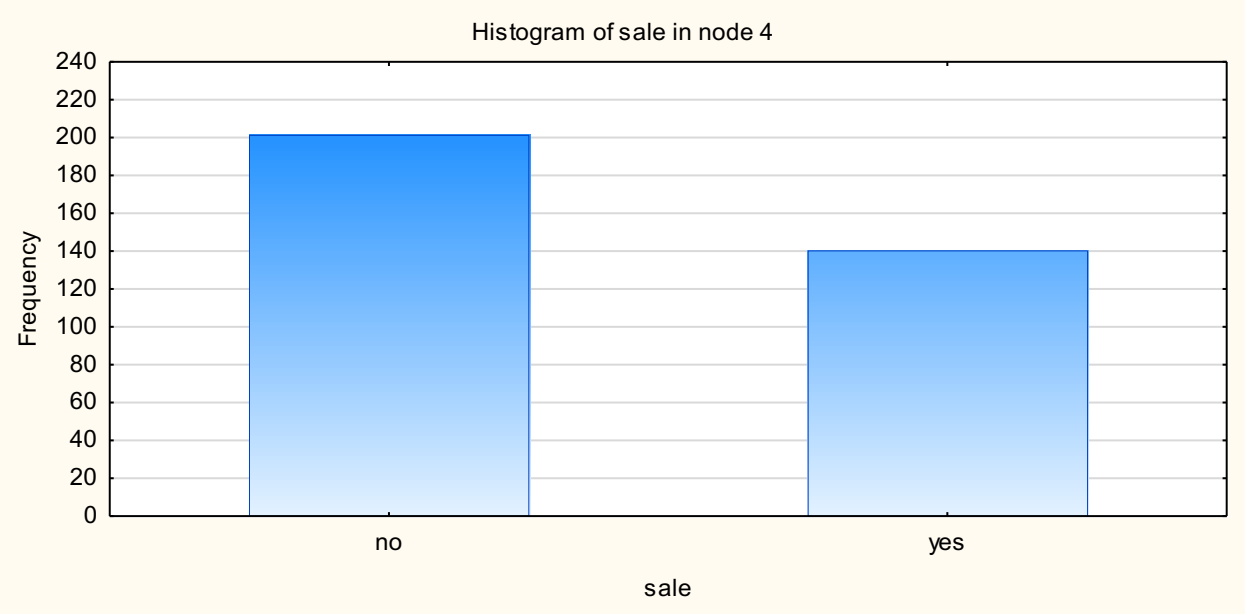


Figure 4. Parallel coordinate plot in node 4[source: own]



Bar Chart 4. Sale in node 4[source: own]

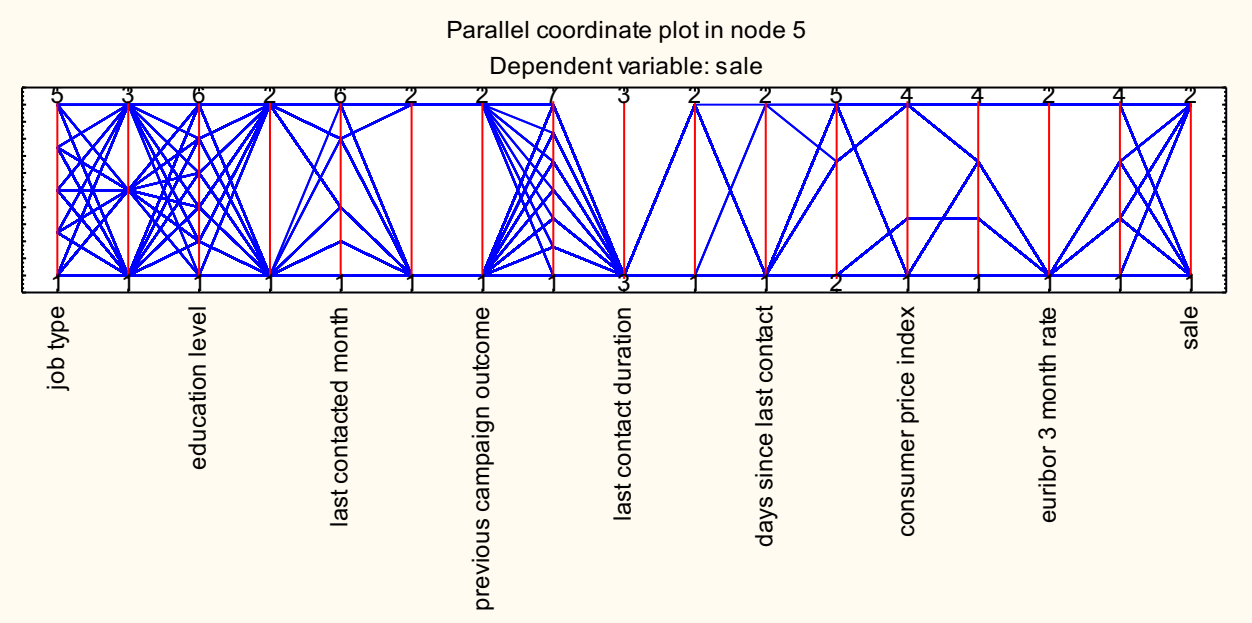
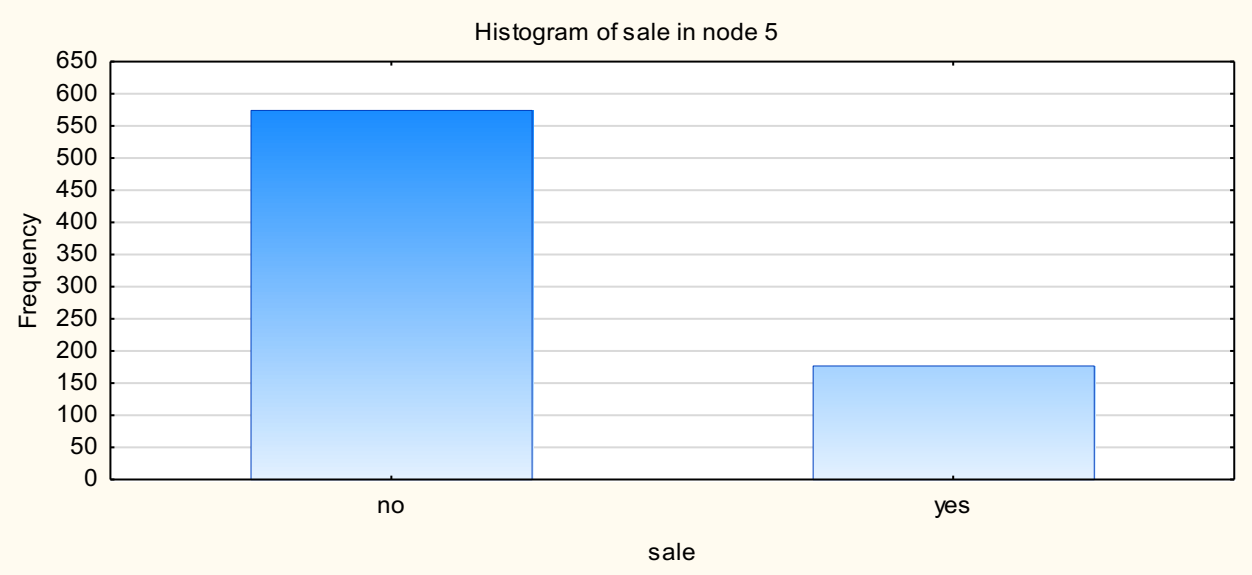


Figure 5. Parallel coordinate plot in node 5[source: own]



Bar Chart 5. Sale in node 5[source: own]

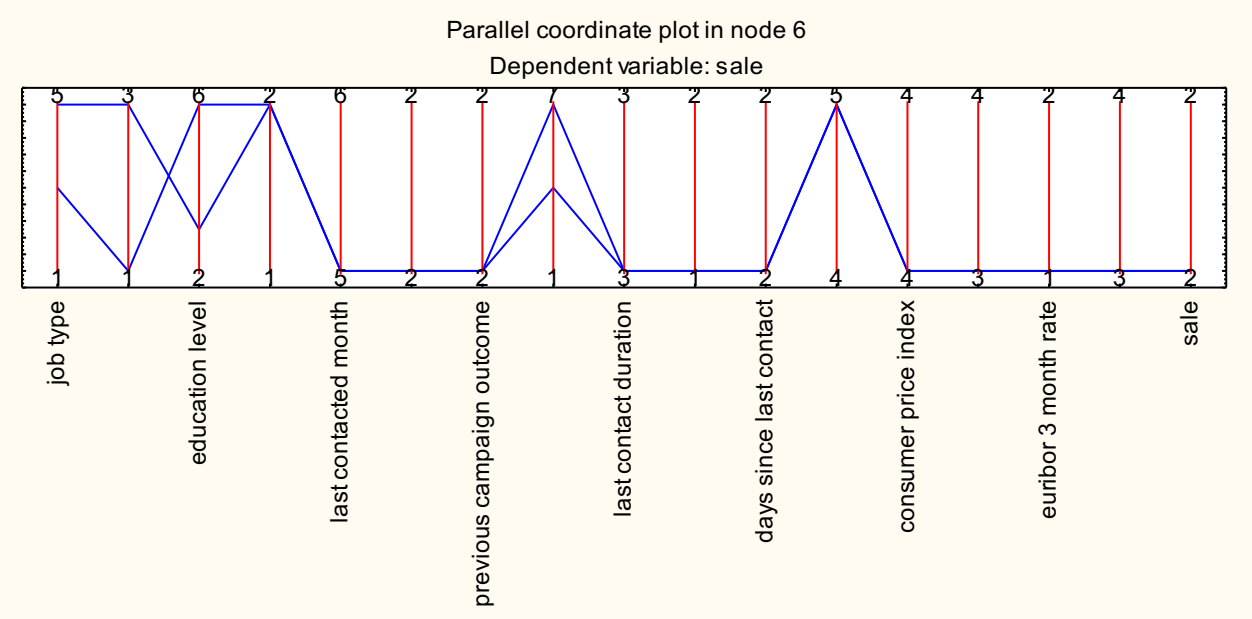
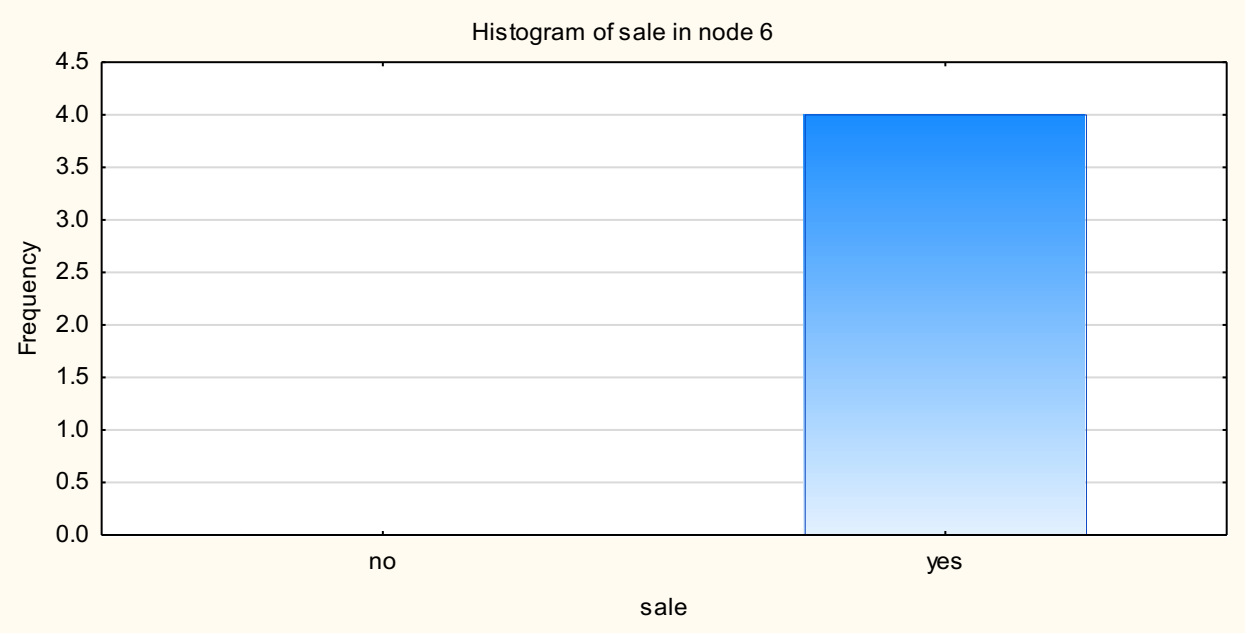


Figure 6. Parallel coordinate plot in node 6[source: own]



Bar Chart 6. Sale in node 6[source: own]

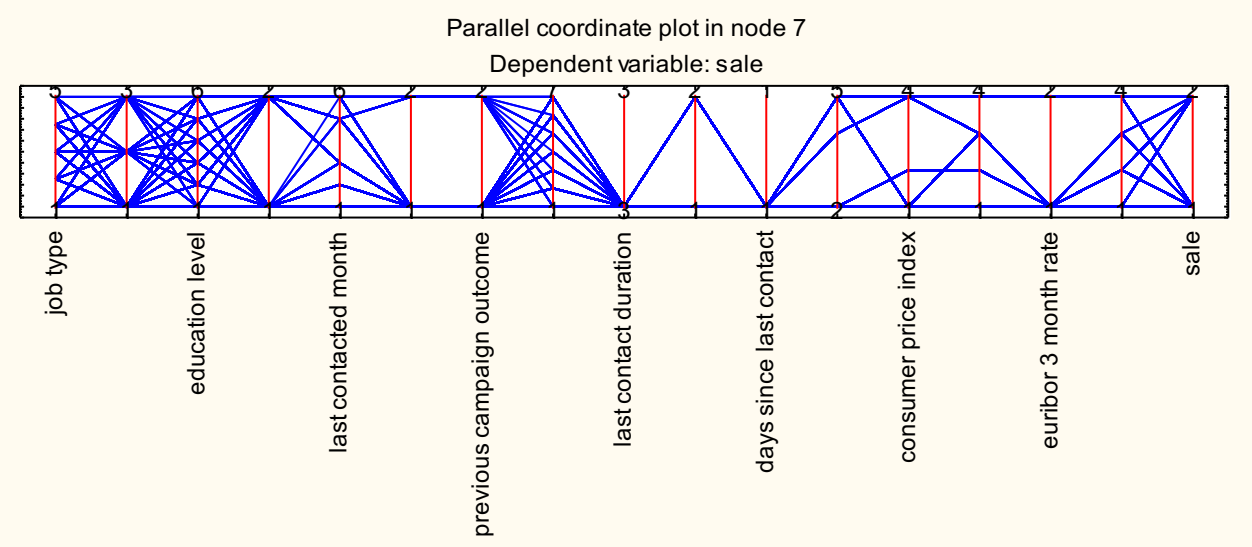
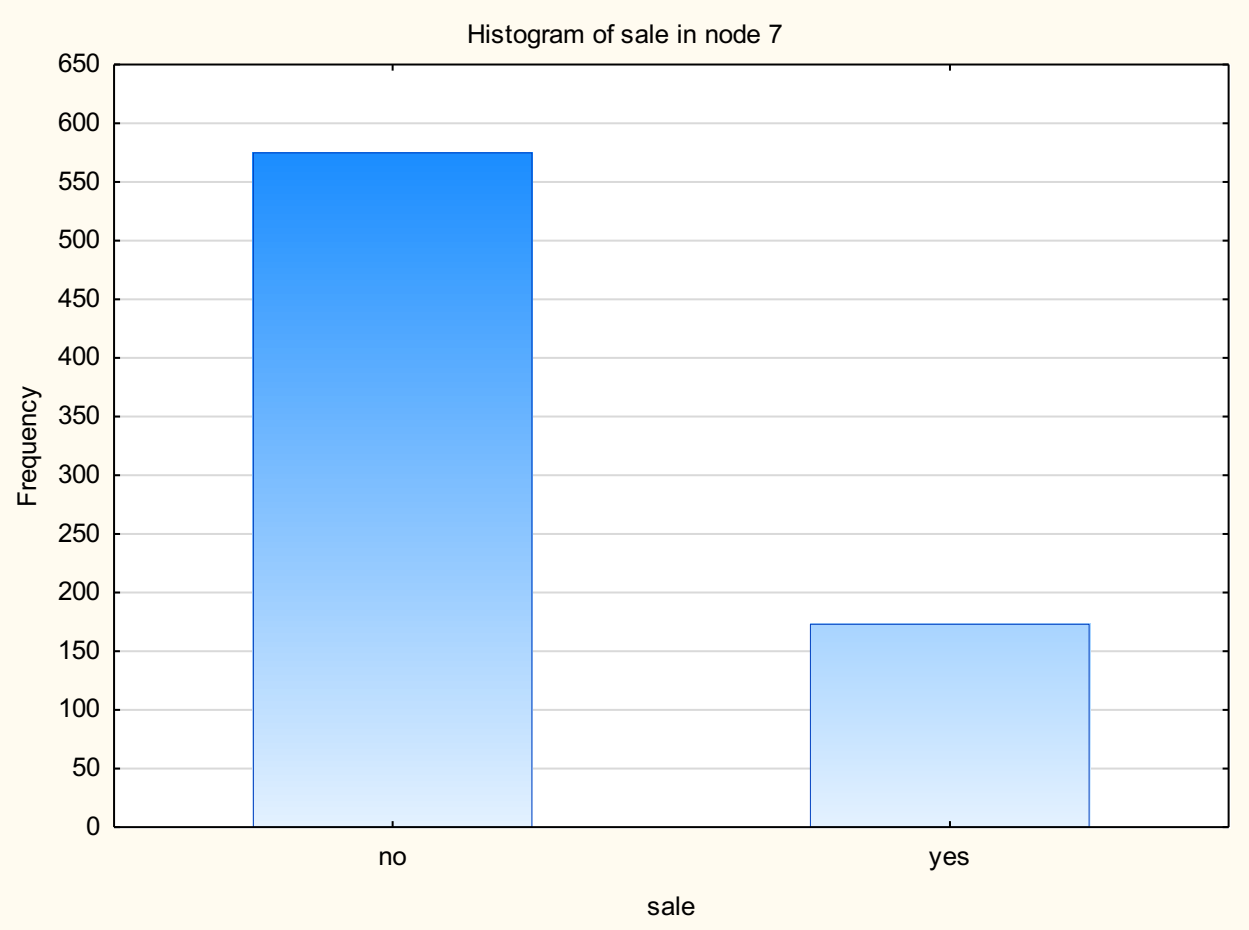


Figure 7. Parallel coordinate plot in node 7[source: own]



Bar Chart 7. Sale in node 7[source: own]

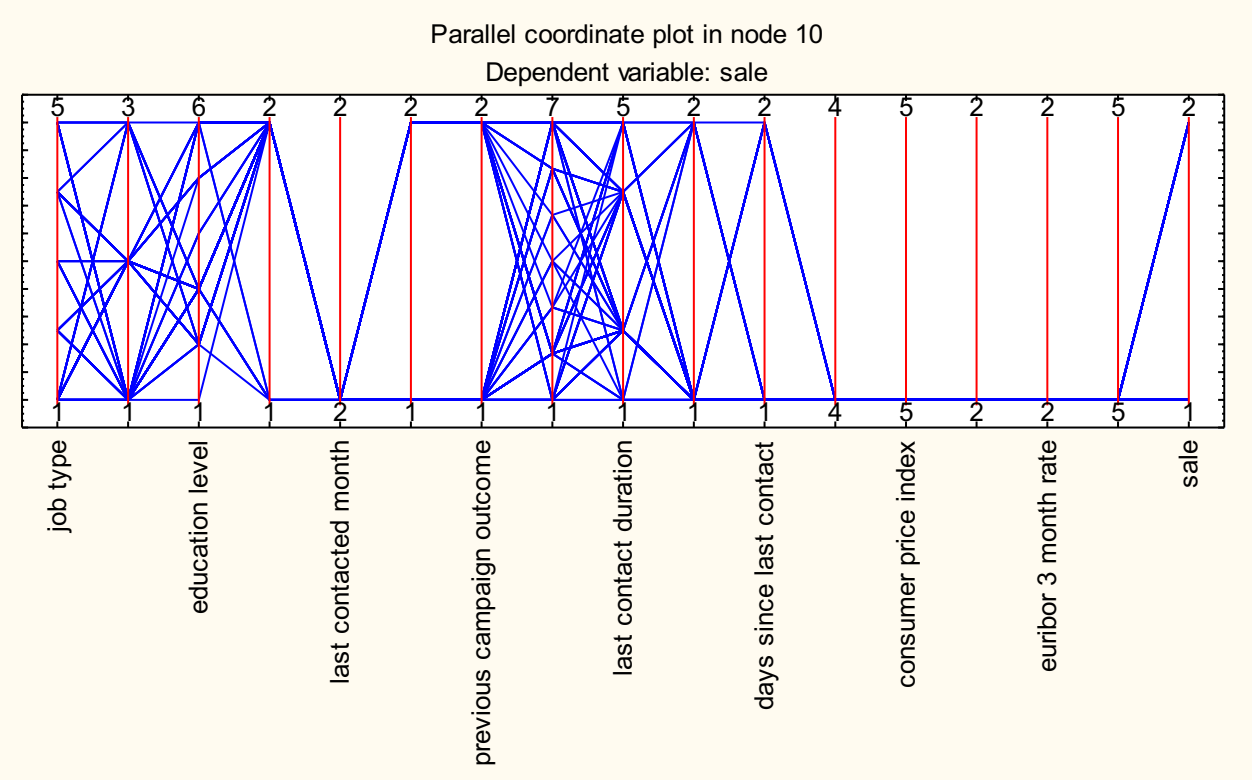
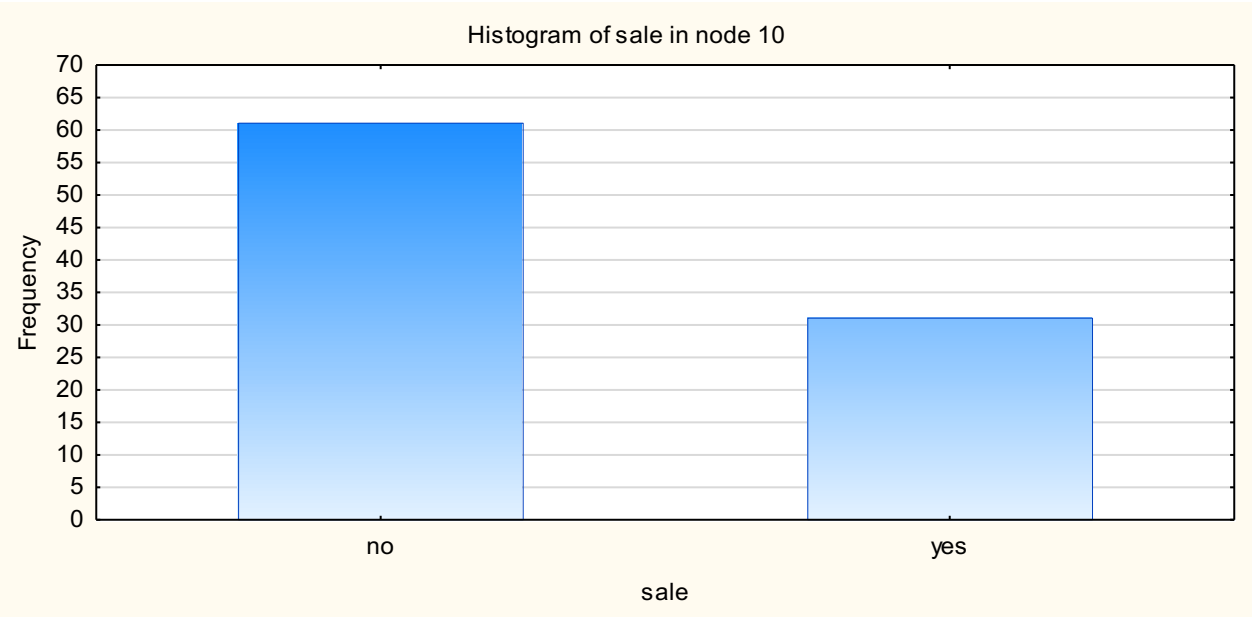


Figure 8. Parallel coordinate plot in node 10[source: own]



Bar Chart 8. Sale in node 10[source: own]

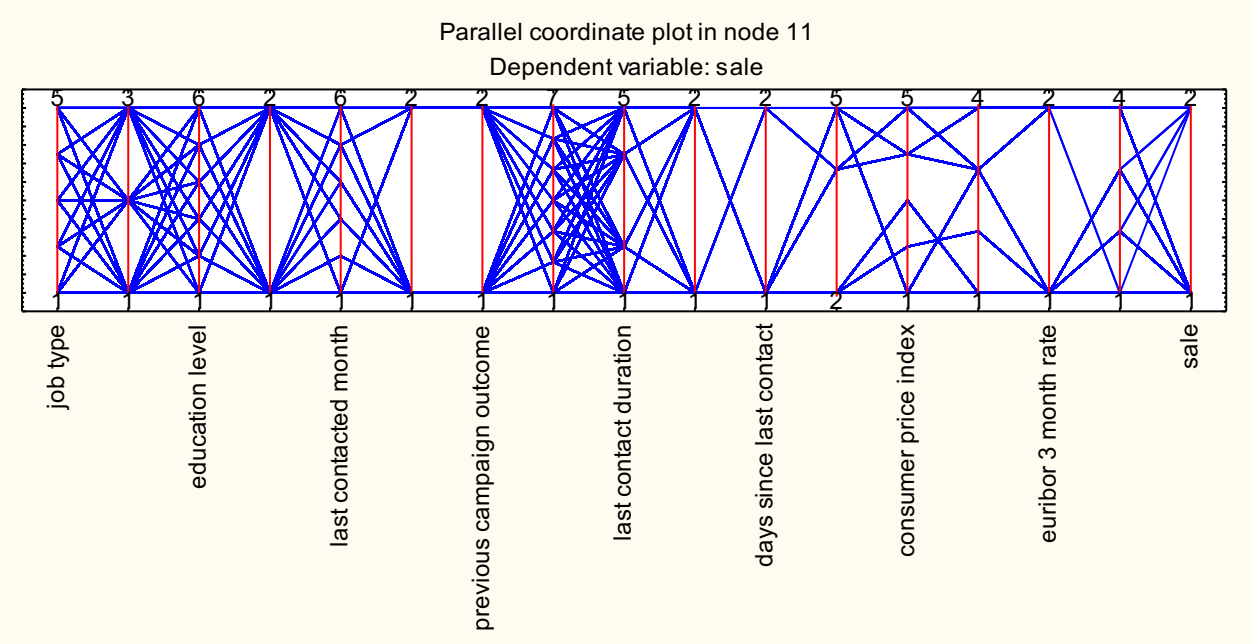
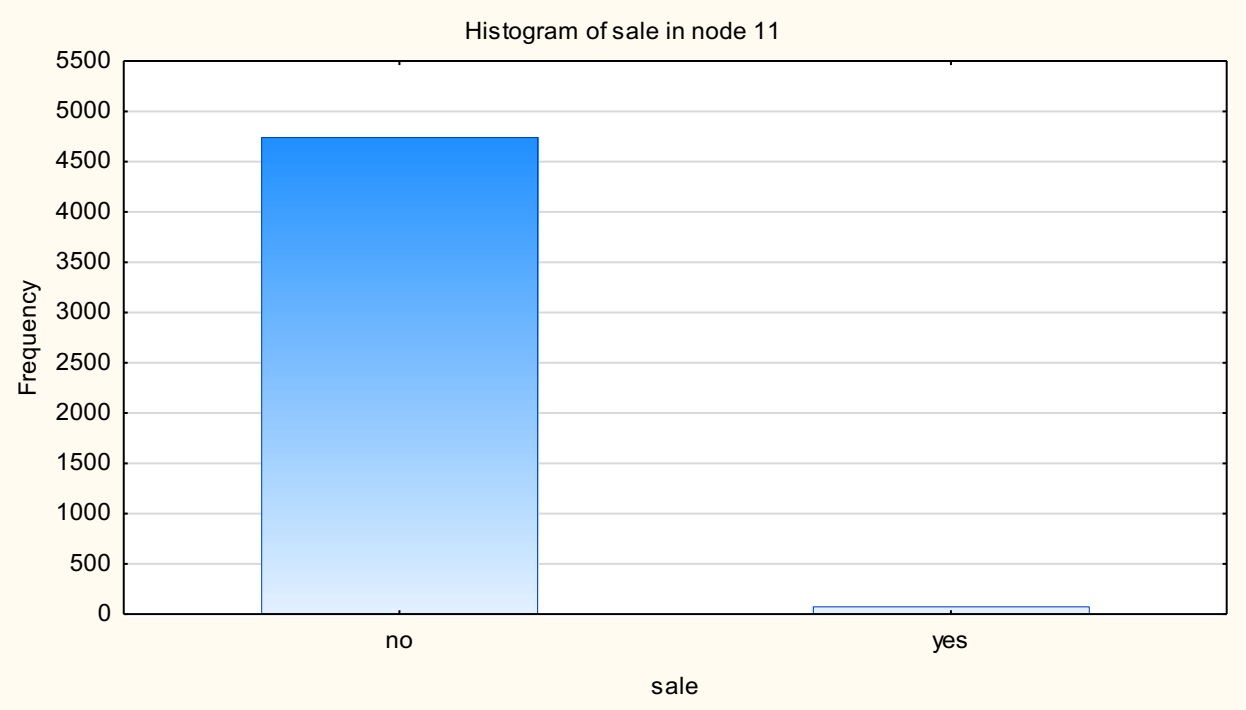


Figure 9. Parallel coordinate plot in node 11[source: own]



Bar Chart 9. Sale in node 11[source: own]

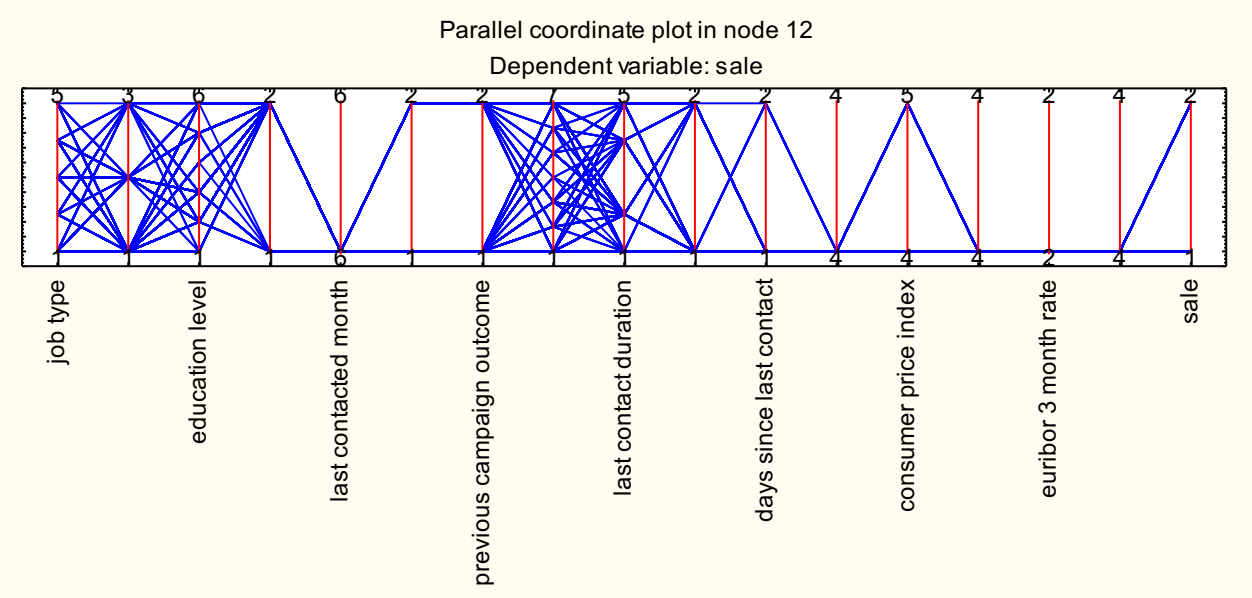
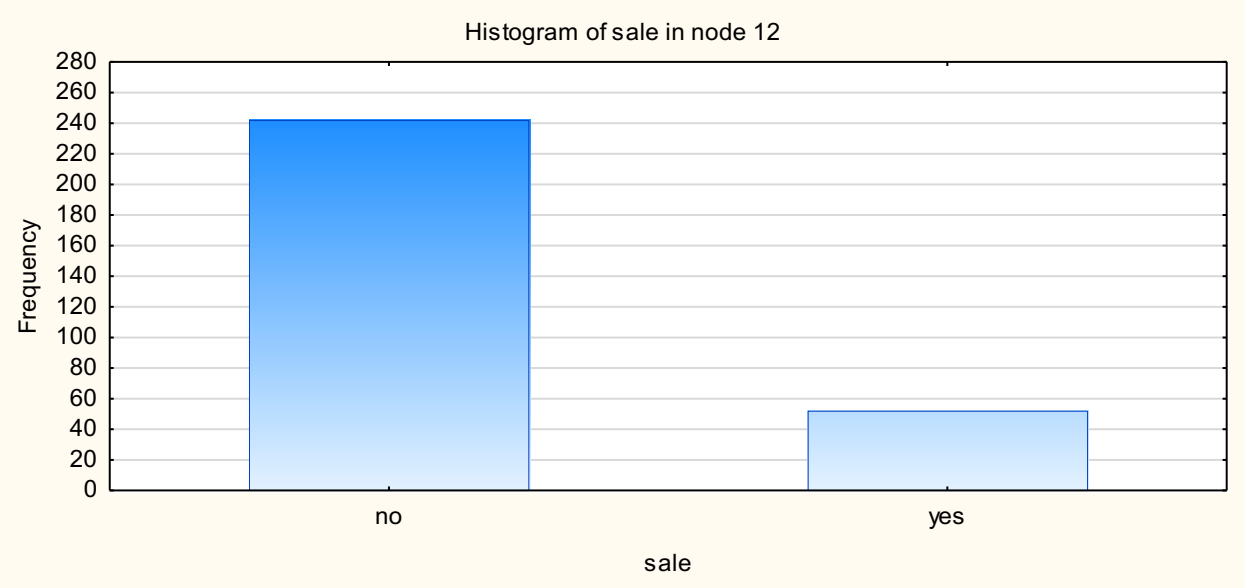


Figure 10. Parallel coordinate plot in node 12[source: own]



Bar Chart 10. Sale in node 12[source: own]

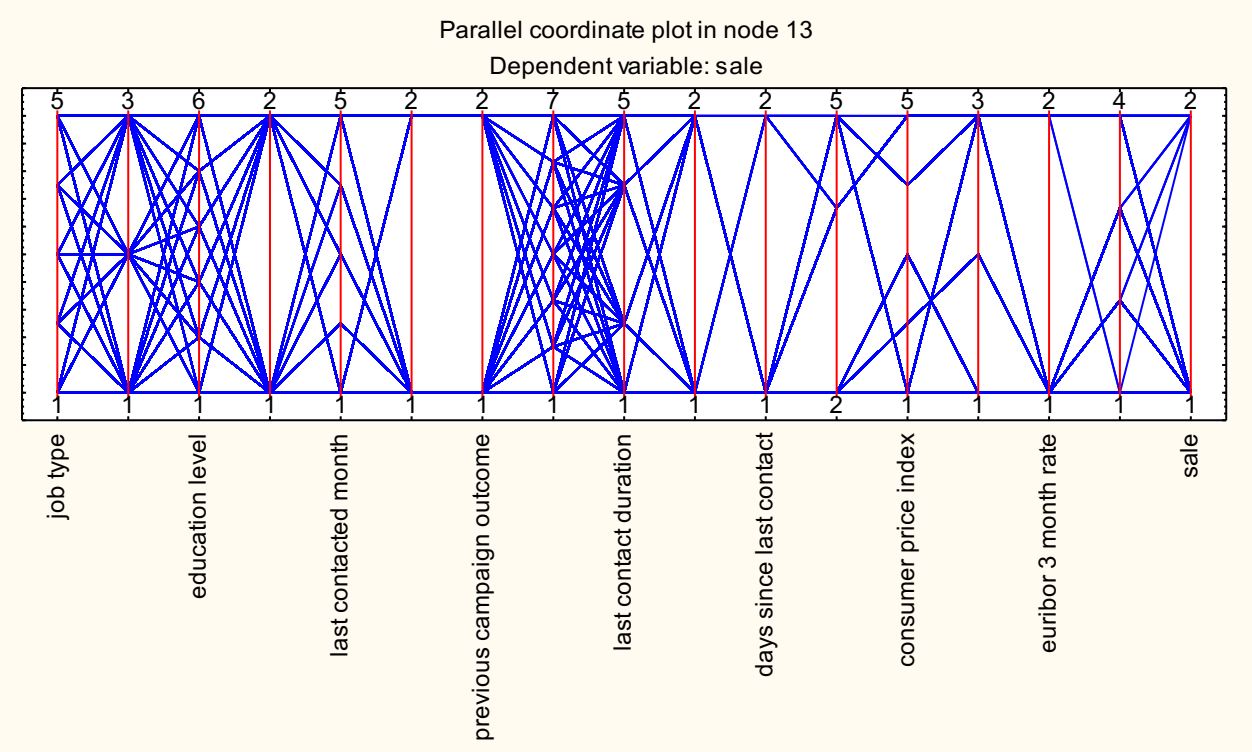
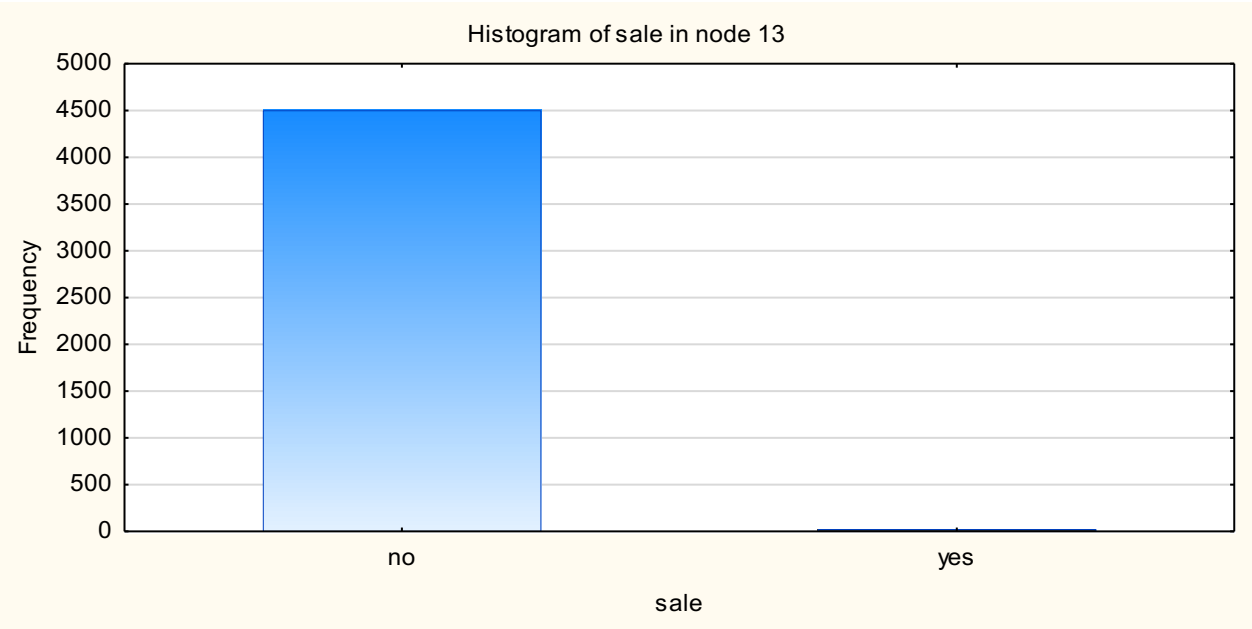


Figure 11. Parallel coordinate plot in node 13[source: own]



Bart Chart 11. Sale in node 13[source: own]