Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Diploma Thesis

Predictive modelling in selected database

Ivan Rožman

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Ivan Rožman

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 23. 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 on 30.11.2016 ______________________________

Acknowledgement

I would lid to thank Ing. Tomáš Hlavsa, Ph.D., work colleagues, friends and some of former classmates for their understanding, advices and support during my work on this thesis.

Prediktivního modelování ve zvolené databázi

Souhrn

Diplomová práce se zabývá vyhodnocením chování zákazníka (firmy). Primární cíl práce je identifikovat a analyzovat všechny faktory, ovlivňující chování. Analýza je zpracovaná na základě databáze přímého marketingu banky. S užitím prediktivních analýz a metod dolování dat se sestaví model předpokládaného chování klienta banky, což umožní bance lepší rozhodování. K dosažení cíle práce se použijí statistické metody - explorační analýza dat, vícerozměrné statistické metody a algorytmy strojového učení.

V praktické části bude sestaven prediktivní model, který bude předpovídat, zda by klient měl zájem o založení termínovaného vkladu, na základě získaných dat pomocí přímého marketingu portugalské bankovní instituce. Práce bude vyhotovena na základě doporučené literatury a osobních zkušeností.

Klíčová slova: Prediktivní modelování, dolování dat, zpracování velkých objemů dat, statistická analýza, chování

Predictive modelling in selected database

Summary

This thesis deals with evaluation of customer(company) behaviour. The main goal is to discover and analyse all possible factors influencing and affecting their behaviour. The analysis will be based on bank direct marketing database. There will be used predictive analytics and data mining techniques, which learn from data to predict the future behaviour of individuals in order to make better decisions. To reach the main goal of the thesis there will be employed statistical procedures, such as exploratory data analysis, multivariate statistical methods and machine learning algorithms.

In practical part of this thesis a predictive model for bank database will be made to predict future behaviour of customers, to predict will they subscribe for bank term deposit or not based on data collected through bank direct marketing. The data is related with direct marketing campaigns of a Portuguese banking institution. This thesis will be based on recommended and relevant information sources and authors personal knowledge.

Keywords: Predictive modelling, Data mining, Big data, Statistical analysis, behaviour

Table of Contents

List of Figures

List of Tables

1. Introduction

Growth of information technologies introduced the new era in data analysis because collecting data became much easier and the need for interpretation of large amounts of data became very important for all commercial organizations, for governments, in sports, in medicine and pretty much in all commercial and non-commercial fields. Now, when large amounts of data are so easily accessible to companies they tend to use it to extract information which can be crucial for them and their future business success.

Companies want to predict future events by using historical data, they want to know how much will they sell, which new product can be interesting to old customers, banks want to know does somebody have good or bad credit rating, insurance companies are predicting possible frauds, sport teams are looking for interesting patterns which they cannot see with their own eyes but can help them in tactics and scouting players for example.

Large companies use predictive analytics and data mining to find the way how to target their customers more personally, because nowadays it’s not possible just to push new product or service and expect that everybody will be interested in buying it. So if they want to make it more personal they need to find ways to group certain customers in certain groups based on their preferences, geographical area, gender, age groups for example.

Statistics, machine learning algorithms and now even artificial intelligence are being used to derive information from large amounts of data. Predictive modeling became very important in recent years and the need for it is growing. Data mining helps organizations to find interesting relations and patterns among different factors and different groups which can have influence on customer behavior and preferences. Discovering what is important for customers is the most important for companies and can have great influence on their future business success.

2. Methodology

The diploma thesis deals with discovering possible factors that can influence on customer behavior. The main goal is to predict will customers subscribe a bank term deposit or not and which factors can influence on that. To accomplish the main goal of the thesis its necessary to use data mining techniques, exploratory data analysis and some machine learning algorithms in order to discover interesting patterns and relations in the data and use them to create predictive model.

Exploratory data analysis, multivariate analysis and machine learning algorithms will be used to prepare data for modeling, to find connections between variables as well to discover their influence on the output variable and to develop the predictive model.

To accomplish the main goal of the thesis STATISTICA 13 and Microsoft Excel software will be used.

3. Literature review

3.1. Introduction to predictive analytics and data mining

“Analytics is the process of using computational methods to discover and report influential patterns in data. The goal of analytics is to gain insight and often to affect decisions. Data is necessarily a measure of historic information so, by definition, analytics examines historic data. The term itself rose to prominence in 2005, in large part due to the introduction of Google analytics. Nevertheless, the ideas behind analytics are not new at all but have been represented by different terms throughout the decades, including cybernetics, data analysis, neural networks, pattern recognition, statistics, knowledge discovery, data mining, and now even data science.” (Abbott, 2014)

The popularity of analytics in recent years is understandable because organizations tend to collect a huge amount of data and to summarize it, to help them achieving their goals and their purposes. The amount of the data in today world is exploding and analyzing so called big data became one of the key bases of competition, as groundwork for new waves of productivity growth, innovations and consumer surplus. (McKinsey Global institute, 2011)

Data is nowadays used for forecasting, predicting, helping in making decisions and, of course, to improve efficiency. Because organizations all over the world now can collect large amounts of data they need to find ways to make that data useful, to actually get meaningful and useful information out of it.

“Predictive analytics is the process of discovering interesting and meaningful patterns in data.” (Abbott, 2014) Predictive analytics is related to other disciplines which have been used for discovering data patterns for a long time in history such as pattern recognition, statistics, machine learning, artificial intelligence and, of course, data mining.

Predictive analytics is data driven, which means that algorithms are deriving key characteristics of the models from data itself rather than from assumptions made by data analyst.

"Data-driven algorithms induce models from the data. The induction process can include identification of variables to be included in the model, parameters that define the model, weights or coefficients in the model, or model complexity." (Abbott, 2014)

Predictive analytics as well automate the process of finding the patterns in the data. Powerful induction algorithms not only discover coefficients and weights for the models but also they give us a form of the models. Decision tree algorithms learn which of the candidate inputs best predict a target variable in addition to identifying which values of the variables to use in building predictions. Other algorithms can be modified to perform searches, using exhaustive or greedy searches to find the best set of inputs and model parameters and if variable helps reducing model error, the variable is included in the model. Otherwise, if the variable does not help in reducing model error it will be eliminated. Many software packages and algorithms also contain automation task which automates the process of transforming input variables so that they can be used more efficiently in the predictive models. These software and algorithms don't really do anything which would not be possible to be accomplished without them but if there are hundreds of transformation which should be done it is much easier and faster doing it automatically rather than step by step. (Abbott, 2014)

Also, algorithms can go through all possible combinations of inputs in the data, they can identify which patterns are better so that analyst can focus on them. Going step by step through all possible combinations is possible, of course, but it takes time, sometimes a too much time which can't be afforded. There is also the fact that over the time more and more data is available and without help of powerful software tools and algorithms it would be almost impossible to find which combination of inputs will give the best possible output. Using algorithms in this cases is advantage which just has to be taken so analysts can focus on finding the very best solution. (Abbott, 2014)

But, we should know that automation cannot replace human oversight. Analysts have to be actively involved at every phase of the predictive modelling process. So, instead of asking where humans fit into predictive analytics, we should instead ask how we may design predictive models into the human process of problem solving.

3.1.1. Connection between statistics and predictive analytics

Statistics and predictive analytics are tightly connected, but is predictive modeling only statistics is question for which there is still no agreement between statisticians and predictive modelers. Predictive modeling can be defined as an extension of statistics, which is in the base and core of predictive modeling but predictive modeling is also using techniques and algorithms which are taken from other fields. Predictive modelers have to use statistical tests and algorithms but sometimes without applying all necessary diagnostics which statisticians would apply so they can ensure that models are built properly. One of the biggest differences between statistics and predictive modeling is the fact that statistics is driven by theory while predictive analytics is not, predictive analytics is data driven and everything is about data. Theory is for statistics very important because in statistical analysis at the end everything is based on theory which is very helpful in finding optimum solution which is not always possible in predictive analytics which tend to use many of algorithms based on machine learning and artificial intelligence which don’t have provable optimum solution. Some of key differences between statistics and predictive analytics are given in Table 1. below. (Abbott, 2014)

Statistics	Predictive analytics
Models based on theory: There is an optimum.	Models often based on non-parametric algorithms; no guaranteed optimum
Models typically linear.	Models typically nonlinear
Data typically smaller, algorithms often geared toward accuracy with small data	Scales to big data; algorithms not as efficient or stable for small data
The model is king.	Data is king.

Table 1. Differences between statistics and predictive analytics
Source: Abbott 2014, own processing

3.1.2. Statistical analysis vs. data analysis or analytics

People would often mix two of them as one of the same but even there are very strong similarities between two of them key difference between statistics and analytics is in mindset even though analytics, the same as predictive analytics, uses statistical tests, theory and algorithms.

"Statistics is often used to perform confirmatory analysis where a hypothesis about a relationship between inputs and an output is made, and the purpose of the analysis is to confirm or deny the relationship and quantify the degree of that confirmation or denial." (Abbott, 2014)

Controls are necessary, ensuring us that there is no bias included in the model. Transforming data is essential as well because it's necessary that inputs and outputs are complying with assumptions of the modeling algorithm and coefficients of the model are critically important in understanding what the data is saying. Identifying departure from a Normal distribution is very important as well, so residuals should be examined carefully. If they are not random with constant variance it is necessary correcting this problem by modifying the inputs and outputs until solution is found. (Abbott, 2014)

Predictive modelers usually have to deal with data which was not even collected for that very purpose, they tend to solve less complicated business problems.

3.1.3. Data mining

Data mining is predecessor of predictive analytics. Algorithms used in both, data mining and predictive analytics are basically the same. Approaches as well. So what is the difference between two of them is simply the question you have to ask. Data mining together with business knowledge is what we can call predictive analytics nowadays. A lot of authors tend to say that till recently they were using those two as one of the same, as actual synonyms until predictive analytics didn’t become more popular term. (Abbott, 2014)

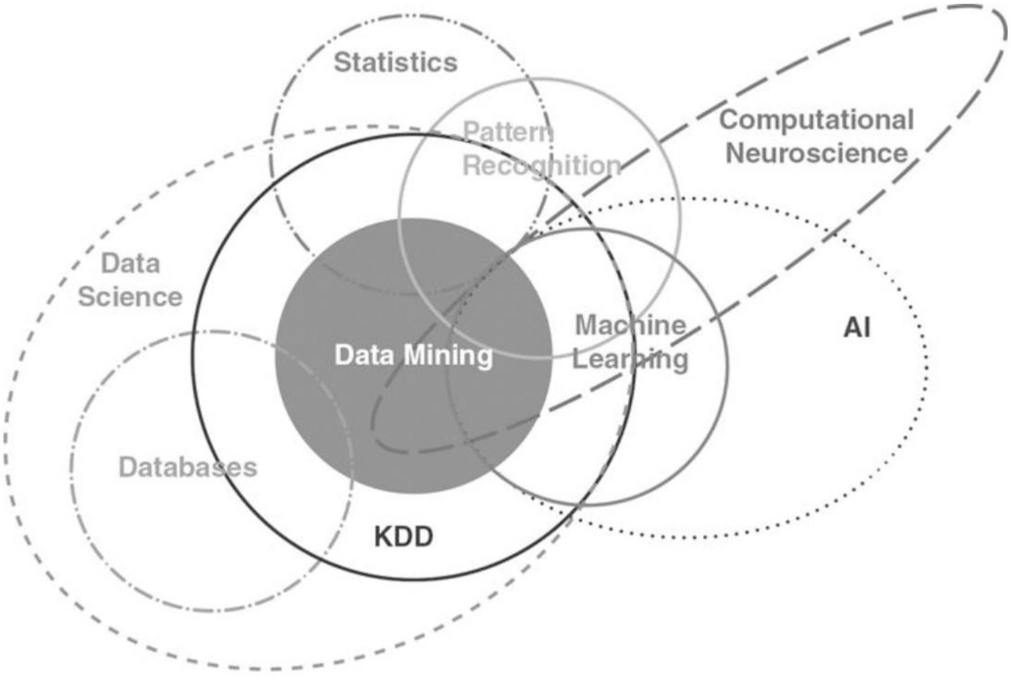
“Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques” (Gartner Group, cited in Larose 2014)

“Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases and visualization to address the issue of information extraction from large data bases” (Evangelos Simoudis in Larose, 2014)

As it can be seen from these definitions of data mining they are almost the same as definitions of predictive analytics and from them we can actually see that techniques, algorithms and approaches for two of them are actually the same.

Two other definitions from Daniel T. Larose (2014) may show us what is the actual difference between two of them. He said that *“Data mining is the process of discovering useful patterns and trends in large data sets.”* and that *“Predictive analytics is the process of extracting information from large data sets in order to make predictions and estimates about future outcomes.”* We can see that he thinks that predictive analytics goes one step further from just discovering interesting and meaningful patterns and trends in data and actually extracts information out of it and provide us estimates about future outcomes and that is why data mining is considered to be predecessor of predictive analytics.

Jared Dean (2014) sees predictive modeling as data mining task and he wrote *“predictive modeling is one of the most common data mining tasks. As the name implies, it is the process of taking historical data (the past), identifying patterns in the data that are seen though some methodology (the model), and then using the model to make predictions about what will happen in the future (scoring new data).”*



Source: SAS Enterprise Miner Training material from 1998 in Dean 2014
Figure 1. Multidisciplinary Nature of Data Mining

Data mining is used in various fields like finance, logistics, engineering, marketing, biotechnology, medicine, production and customer relationship management. It is used by insurance companies, banks, retailers, manufacturing companies and many other. It became popular in sports as well, used for example by multiple NBA teams, and not just them, for various reasons like scouting players and discovering some tactical advantages and disadvantages which are not obvious or possible to see that easy. It is widely spread, and widely used in all commercial fields but as well in scientific researches, medicine, governments and police in crime detection and prevention and it is still growing because, mentioned before, more and more data has been collected and more and more data is now available and data mining and predictive analytics are the best way making actual information and knowledge out of the data.

That growth in the field of data mining and knowledge discovery has been influenced by combination of variety of factors:

- The explosive growth in data collection
- The storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database
- The availability of increased access to data from Web navigation and intranets
- The competitive pressure to increase market share in a globalized economy
- The development of off-the-shelf commercial data mining software suites
- The tremendous growth in computing power and storage capacity (Larose, 2014)

Leading companies nowadays are closely looking to their data, they examine and process it because they understand real power of data and that value they can get from it can help them in outperforming their competition which will lead to better market place. They can find out who to target with advertises, who will most likely be keen in buying their new products, would new product or technology be something that is actually needed and wanted at moment.

3.1.4. Data mining in CRM and marketing

Like it was already mentioned, data mining is widely spread and widely used but focus of this thesis will be on data mining in customer relationship management and marketing.

Small service oriented businesses developed one to one relationships with their customers and that is something what bigger companies are trying to do, instead of trying to access all customers in one, same way they tend to recognize which of them will be more likely interesting in their new product or service. This very process is not easy, focusing on the small group of customers for small businesses is not as challenge as it is for huge corporations which will have a lot of different groups of customers. In order to discover very fact how to make more personal relations with different groups of customers, and to discover how to target people based on their preferences instead of targeting all of them in the same way, big corporations need to approach their customers more personally and data mining can help them in achieving that. Data mining can have important role in improving customer relationship management by improving their ability to learn from relationships with their customers. (Linoff & Berry, 2011)

The goal is simple, you want to understand your customers more individually so you can offer them something which competitors cannot, to make for them conducting business with you easier and more profitable than with competitors. This type of conducting business, customer based approach or customer relationship is revolutionary change for most of companies which tended to be product focused organizations. Data mining will not solve all their problems in transiting from old way of doing business to new one, it takes much more than just data mining, but it will definitely help them.

"In narrow sense, data mining is a collection of tools and techniques. It is one of several technologies required to support a customer-centric enterprise. In broader sense, data mining is an attitude that business actions should be based on learning, that informed decisions are better than uninformed decisions, and that measuring results is beneficial to the business." (Linoff & Berry, 2011)

To make data mining techniques and tools effective, there are other requirements for analytic CRM which have to be done.

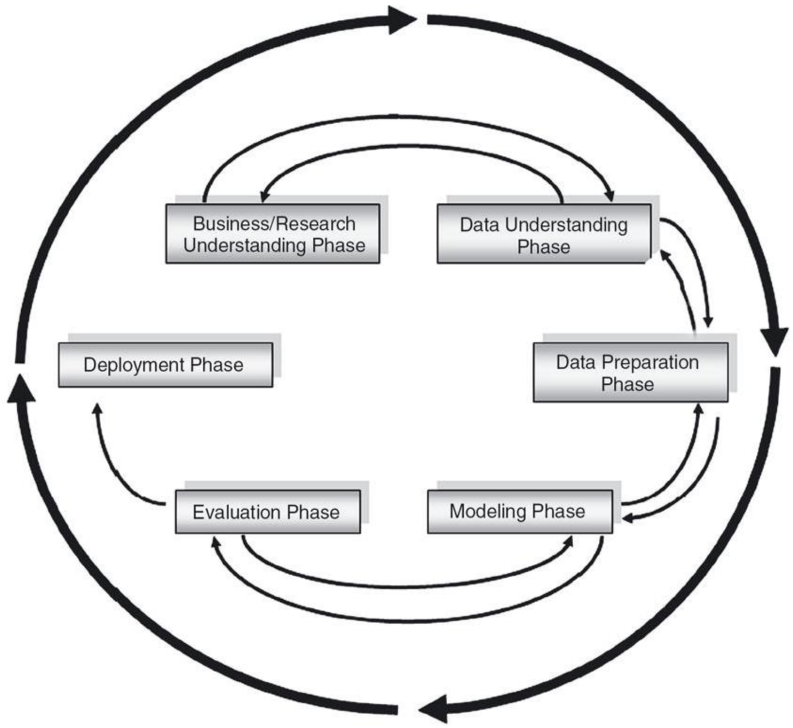
A company must be able to:

- Notice what its customers are doing
- Remember what it and its customers have done over time
- Learn from what it has remembered
- Act on what it has learned to make customers more profitable (Linoff & Berry, 2011)

Transaction processing system has to capture interactions, data warehouses will be used to store historical customer behavior data, data mining will be used to translate all that historical data to plans and customer relationship management to take advantages from those plans by putting them to practice. (Linoff & Berry, 2011)

3.1.5. CRISP-DM

CRISP-DM is cross industry standard process for data mining and it was developed by analysts from Daimler-Chrysler, SPSS and NCR. Idea was to create standard process which will make data mining a part of general problem solving strategy of a business. They agreed that data mining project life cycle should consist from six phases as shown on Figure 2. below. Important thing is that these phases are dependent on outcomes of the previous phase and on figure below it is possible to see those phases and their connection with each other. Significant dependencies between them are indicated by arrows. Sometimes is necessary to return to previous phase before continuing to the next one, for example to prepare data in different way before moving on to evaluation phase. Outer circle arrows are showing iterative nature of CRISP-DM. (Larose D & C, 2014)



Source: Larose D and C, 2014
Figure 2. CRISP-DM is an iterative and adaptive process

Outline and description of six phases of CRISP-DM are:

1. Business/research understanding phase
    a. First project objectives and requirements should be defined in terms of business
    b. Second step is to translate this objectives and requirements into a form of a data mining problem
    c. Preparing a preliminary strategy for accomplishing these objective is a final phase of this phase
2. Data understanding phase
    a. First step is data collection
    b. Second step is getting familiar with the data set by using exploratory data analysis

   c. Evaluation of data quality
   d. If desired select subsets which may contain actionable patterns
3. Data preparation phase
   a. This phase is phase of cleaning and adjusting raw data, this is labor intensive and the most time consuming phase
   b. Selection of variables which will be used for analysis
   c. If necessary, transform some variables
   d. Complete cleaning the raw data so it becomes suitable for the modeling phase
4. Modeling phase
   a. Selection and application of appropriate modeling techniques
   b. Calibrate model settings to optimize results
   c. Apply several different techniques to the same problem
   d. If necessary, loop back to data preparation phase to change the data so it can suit better to specific requirements of a specific data mining technique
5. Evaluation phase
   a. One or more models are generated and now they should be evaluated for effectiveness and quality before deployment
   b. Determine if the model meets goals and objectives set in phase 1
   c. Establish whether some important fact of the business or research problem has not been sufficiently accounted for.
   d. Make decision whether to use data mining results
6. Deployment phase
   a. Model(s) is/are created but now it/they should be used and applied
   b. Make example of simple deployment and generate report
   c. Complex deployment, implement data mining process in another department
   d. For businesses, the customer often carries out the deployment based on your model. (Larose D & C, 2014)

3.2. Description and preparation of data

Data preprocessing or data preparation and data understanding is first and probably the most important step. Raw data cannot be used for data mining so data need to be prepared and analyzed before we get to further steps, choosing algorithms and making actual predictive model. Raw data usually contains so called noisy data, data which is not preprocessed and should not be used. Raw data will most likely have outliers, missing values, data which should be transformed because in the current form is not suitable for data mining. Those are the things which have to be dealt before we start making model because they can cause problems which will make model not good enough or actually not usable.

Most of predictive modelers and data miners agree that this part is the most important part and because of that fact this is also the most time consuming part and they agree that data preprocessing will take around 60% of time in making predictive model.

Much of raw data contained in databases is not preprocessed, usually incomplete, and tends to be noisy. For example, the databases may contain:

- Fields that are absolute or redundant
- Missing values
- Outliers
- Data in form not suitable for data mining models
- Values non consistent with policy or common sense (Larose 2005, 27)

So if want to get the best from our data, data needs to be cleaned and transformed. The goal is to minimize “bad” data coming into the model so we can reduce “bad” information coming out of the model.

3.2.1. Data description - the first step to understand the data

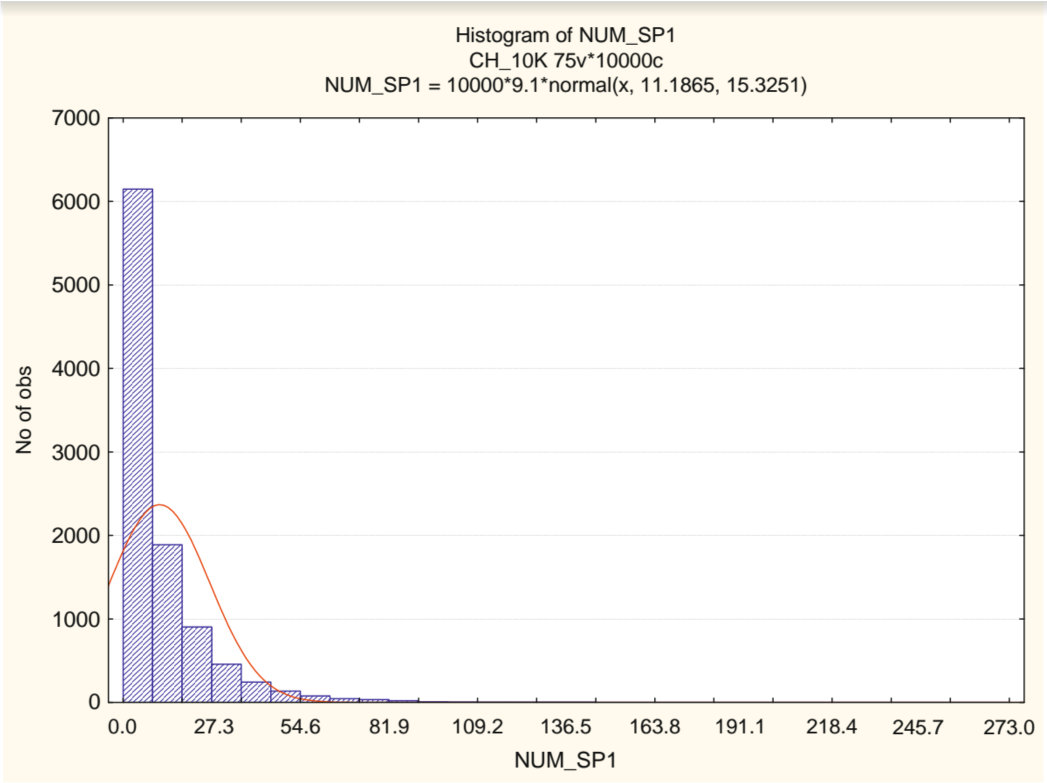
Descriptive statistics is very important because it will give us first insight in our data. Data description can be done by univariate analysis for individual variables, bivariate analysis for pair of variables and graphical and visual techniques for viewing some more complex relationships between variables. Basic descriptive statistics measures and techniques which will give us first insight in the data are: mean, standard deviation, minimum, maximum, frequency tables and histograms. (Nisbet & Elder & Miner, 2009)

Mean value which tells us average value and shows us central tendency of data set can be very useful used together with standard deviation, minimum and maximum. It can help us discovering outliers if maximum or minimum are too far from mean value or if variable has relatively low mean and high standard deviation it tells us that variable is most likely not suitable to be used because of low potential for predicting the target. Frequency tables and histograms can give us more details about distribution. (Nisbet & Elder & Miner, 2009)

N	Mean	Min	Max	StDev
10000	9,769700	0.00	454.00	15.10153

Table 2: Descriptive statistics
Source: (Nisbet & Elder & Miner, 2009), own processing

In this Table 2. we can see that maximum value and mean value are too far from each other also standard deviation is around 15 and this creates a very suspicious situation. But to determine is this value outlier or not we should first look at histogram.



Source: Nisbet & Elder & Miner, 2009
Figure 3: Histogram of the distribution of NUM_SP1 variable, compared to a normal distribution.

Looking at histogram can confirm us that this maximum value of 454 is certainly outlier. Except that this histogram is showing us that this variable cannot be used for regression analyses, for example, because the distribution is skewed to the right and it's not normal distribution which is required for linear regression. (Nisbet & Elder & Miner, 2009)

Descriptive statistics can show us a lot about our data set and the model we will use. It's very important to do this step properly because it will tell us what do our data look like. For basic descriptive statistics MS Excel Analysis Tool Pack add-on will do the job, but for deeper and more robust descriptive analyses software packages like SAS, SPSS or STATISTICA should be used. More skilled predictive modelers and data scientist will most probably use one of programming languages used for statistics and predictive modeling like R or Python.

3.2.2. Data cleaning

First step of preprocessing data is data cleaning. In this step, which is also called data scrubbing, goal is to identify incorrect, incomplete, inaccurate or irrelevant data and modify, replace or delete this so called bad or dirty data. Typical errors which can be found in raw data are shown in table below.

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction amount
1001	10048	M	75000	C	M	5000
1002	J2S7K7	F	-40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	F	999999	30	D	3000

Source: Larose 2005
Table 3. Example of data set with problematic data

In this table as we can see there are some inconsistencies, as well missing values also some possible not correct values. For example, if we take a look at column representing Zip code data we can notice that one of them is completely different than rest of values in the same column. Usually zip codes contain five numeral value and this zip code of J2S7K7 use both letters and numbers. On the first look this may seem as error and the first thought is to remove this customer. But this zip code is actual zip code from Canada so it is valid value and that has to be checked. The other problem is that customer 1004 has four-digit zip code so that can be an error as well. It can be error but also we should be aware that in some cases there are zip codes starting with zero and if data was not formatted as text or non-numerical value most of the software's will just remove that zero from first position so that as well has to be checked and confirmed.

Next thing which is easily noticeable and we can say strange is in age variable we have two problems. First one is most likely categorical value C in numerical variable, so we need to decide how to deal with this problem. It can be an error but it can be as already mentioned categorical value. Data mining software will not like this value in this numerical field. Options are contacting the person who has more experience with actual database which maybe can help with this issue. If not, it has to be removed. (Larose 2005, 29)

The other thing is value zero for customer 1004. This is most likely missing value replaced with zero because, even though it is possible, it highly unlikely that newborn person made transaction of 1000 dollars.

"*Of course, keeping an age field in a database is a minefield in itself, since the passage of time will quickly make the field values obsolete and misleading. It is better to keep date-types fields (such as birthdate in a database, since these are constant and may be transformed into ages when needed.*" (Larose 2005, 29)

In field income we can see clear example of outlier. Customer 1003 with income of 10 000 000 dollars per year can be also wrong value but it is possible for somebody to have that income and if value is true it is outlier and most of data mining modeling techniques are not functioning properly in presence of outliers or extreme values. Handling outliers and missing data will be explained in next part of this chapter.

3.2.3. Data imputation - handling missing data problem

Missing data is not that rare especially in huge databases. Handling missing data is important because algorithms will manage to give better results and because of fact there is no software which will appreciate blank cell more than having some value instead. So, the main goal is replacing missing data with intuitive data if possible. Adding a reasonable estimate of missing value instead of leaving it blank is always a good choice because the absence of any kind of information is rarely good or beneficial.

Two types of missing data are:

- MCAR - *"Values in a data set are missing completely at random (MCAR) if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random."* (Polit 2012 cited on wikipedia.org 2016) So if probability of missing value in one variable is not related to the value of that variable or to values of other variables then that missing value is missing completely at random. Missing values are rarely MCAR.

- MAR - *"The assumption of Missing at Random (MAR) is satisfied when the probability of a value's being missing in one variable is unrelated to the probability of missing data in another variable, but may be related to the value of the variable itself."* (Nisbet & Elder & Miner, 2009, 60)

A very common approach in handling missing value is just deleting or omitting records which contain blank fields. This can be dangerous because we are not just removing that missing field but also all other fields for that record which can be very useful. Therefore, data analysts tend to replace missing values instead of just omitting the whole row of data.

Deleting the whole record of data which contains a missing value is called list-wise deletion.

Pair-wise deletion means that all cases that contain value for that variable will be used to calculate the covariance of that variable. Using pairwise deletion is especially useful in some statistical packages regression algorithms (PROC CORR in SAS and napredict in R), because linear regression can be estimated from sample means and covariance matrix and this will preserve inclusion of all cases. This can be used only if missing data is MCAR. (Nisbet & Elder & Miner, 2009)

Reasonable value imputation is mostly imputation of missing values with the mean of the non-missing cases and it is also known as mean substitution. But if there is a possibility to safely apply some decision rule to supply a specific value to the missing value, that value can be closer to the true value than mean of other values. Good example for this would be that replacing missing value for number of children with zero is more reasonable than replacing it with mean or median of other records.

“The technique of maximum likelihood imputation assumes that the predictor variables are independent. It uses a function that describes the probability density map (analogous to a topographic map) to calculate the likelihood of a given missing value, using cases where the value is not missing. A second routine maximizes this likelihood, analogous to finding the highest point on the topographic map. “(Nisbet & Elder & Miner, 2009, 61)

Case-wise deletion	**Pair-wise deletion**	**Substitution**	**ML imputation**	**Expectation Maximization**	**Simple random imputation**	**Multiple random imputation**
Simplest and easiest	Preserves cases	Good when a decision rule is known	Relatively unbiased with large samples	An iterative process	Tends to overestimate correlations	Best for nonlinear algorithms
Sacrifices cases			Consistently estimates under a wide range of conditions	Uses all other variables to predict missing values		Not good for determining interaction effects
Acceptable if the number of cases is large and the event to be modeled is not rare			Data should be MAR	Data should be MAR		Must be matched to the analysis model
Most valid statistically			Best when data is monotonic	Assumption of a normal distribution		Appropriate if data deleted by case-wise deletion is intolerable
Safe for any kind of data mining analysis			Appropriate if number of cases deleted by case-wise deletion is intolerable			
Good for data sets where the variables are completely independent						

Source: Nisbet & Elder & Miner, 2009, own processing
Table 4. Guidelines for Choosing the Right Data Imputation Technique

Multiple imputation can be simple random imputation or multiple random imputation and it uses other variables to predict which values for missing data are most likely or probable.

"Simple random imputation is technique that calculates a regression on all the non-missing values in all of the variables to estimate the value that is missing. This approach tends to underestimate standard error estimates. A better approach is to do this multiple times." (Nisbet & Elder & Miner, 2009, 61)

Multiple random imputation is technique where a simple random imputation is repeated multiple or n-times. *"This method is more realistic because it treats regression parameters (e.g., means) as sample values within a data distribution. An elaboration of this approach is to perform multiple random imputation m-times with different data samples. The global mean imputed value is calculated across multiple samples and multiple imputations."* (Nisbet & Elder & Miner, 2009, 62)

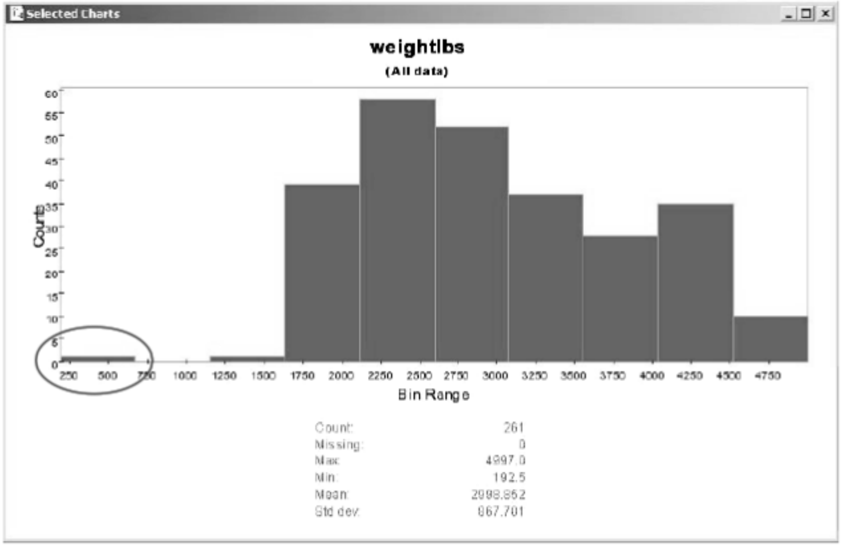
In Table 4. are shown some of guidelines for choosing right data imputation technique with their pros and cons.

3.2.4. Outliers - identifying and dealing with them

Filtering data is also much needed step, and one of the most common kind of filtering is removal of outliers. Depends on what kind of model you want to make, sometimes it's necessary to remove outliers or extreme values. Outliers can be useful in fraud detection or credit risk prediction and in those situations removing them is not a good idea. But, for example in models for normal response removing outliers can be good idea and smart decision. In that case outliers might interfere with model algorithms and create noise which can eventually lead to reduced predictability of the model. (Nisbet & Elder & Miner, 2009, 62)

“Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data.” (Larose 2005, 34)

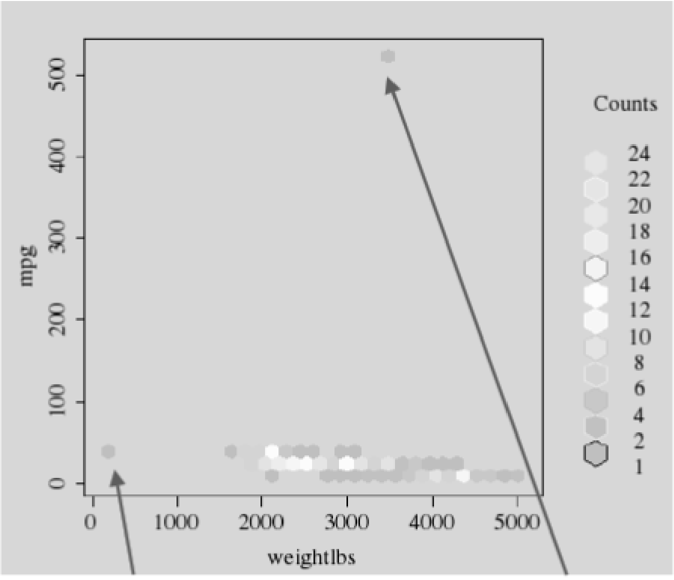
Identifying them can be very important because they can represent an error in data as well. In case that outlier value is valid and not an error still it can be very useful removing them because some statistical methods and algorithms can be sensitive to their presence in data set and may deliver unstable results. One of the ways for identifying outliers graphically for numeric variables is checking histogram of variable. For cases with more than one variable detecting outliers can be done with two dimensional scatter plots. Beside using graphical methods there are some numeric methods for identifying and dealing with outliers. For numerical methods data should be transformed first.



```
Count:        261
Missing:        0
Max:       4997.0
Min:        192.5
Mean:    2998.862
Std dev:  867.701
```

Source: Larose 2005

Figure 4. Histogram of vehicle weights shows outliers



Source: Larose 2005
Figure 5. Scatter plot of mpg against weightlbs shows two outliers

If outliers prove to be correct values and not just an error, there are five typical approaches to deal with them.

1. Removing outliers from data set - This approach is commonly used when there are assumptions that outliers can make problems and harm model. It is usually case with numeric algorithms such as linear regression, Principal Component Analysis, k-nearest neighbor and K-Means clustering. Outliers might even create such a bias that will lead to algorithms ignore most of the data. Problems which can occur after removing outliers from model is that model will be deployed without outliers which can lead to compromised results when outliers appear in data set. Records with outliers will have to be removed prior scoring or they will be scored based on data algorithms are not used to which can lead to unreliable results. As already mentioned before in some cases, like fraud detection, outliers or extreme values should not be removed because they represent the value which represents unusual behavior we are looking for.
2. Separate them and create model just from outliers - With this approach, which is an extension of the first one, outliers will be moved from data and new data set will be

created just from them. New models will be made just for them as well. It is used so analysts can overcome problem from the first approach where models will be deployed without outliers. "*Some predictive modelers who build separate models for the outliers will relax the definition of outliers to increase the number of records available for building these models. One way to achieve this is by relaxing the definition of outlier from three standard deviations from the mean down to two standard deviations from the mean*" (Abbott, D. 2014, 87)

3. Third way is transforming outliers so that they stop being outliers. Transforming data which contains outliers can be helpful in avoiding the bias which is caused by them. Idea is to use techniques which will reduce the distance between the outliers and the main body of the distribution.
4. Bin the data. If outliers stay extreme even after transformation alternative way is to convert numeric variable to categorical one through binning. Later these variables will have to be converted to dummy variables so they can be used in numeric algorithms.
5. Last approach is leaving the outliers without any modifications. This approach limits modeler to use only algorithms which are not affected by outliers such as decision trees. (Abbott, D. 2014, 87-88)

"The business objective will often dictate which of the five approaches in the preceding bullets should be used to treat the outliers" (Abbott, D. 2014, 88)

3.2.5. Data transformation

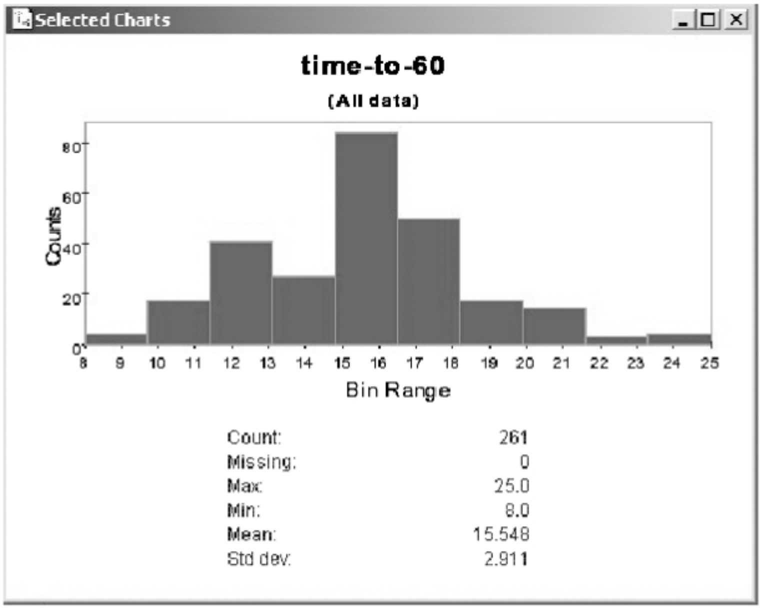
Because of nature of numerical data which can have variety of ranges, some variables will have range between zero and one while others between zero and one hundred, sometimes numerical data has to be transformed because some data mining algorithms tend to be influenced more by variables with greater range. So, data miners should normalize numerical variables to standardize the scale of effect which every numerical variable will have on final results. Even though, there are several techniques to normalize data two are the most common.

3.2.5.1. Min-Max normalization

This technique works by seeing how much bigger value of the field is than the minimum value of that variable and scaling the difference by the range.

Formula is:

$$X^* = \frac{X - min(X)}{range(X)} = \frac{X - min(X)}{max(X) - min(X)} \tag{3.1}$$



```
Count:        261
Missing:        0
Max:         25.0
Min:          8.0
Mean:      15.548
Std dev:    2.911
```

Source: Larose 2005

Figure 6. Histogram of time to 60 mph with summary statistics

For example, if we take cars and there speeding from 0-60 mph some of them the range can go from 8 to 25 seconds. For three cars with time to 60 mph with values of 8, 15.548 and 25 seconds if we do min-max normalization results will be:

- For the fastest car which can speed from 0-60 mph for 8 seconds when we normalize this we will get

$$X^* = \frac{X - min(X)}{range(X)} = \frac{8 - 8}{25 - 8} = 0 \tag{3.2}$$

So normalized value for the fastest car will be zero.

- For car with time value in between minimum and maximum, which needs 15.548 to reach 60 mph

$$X^* = \frac{X - min(X)}{range(X)} = \frac{15.548 - 8}{25 - 8} = 0.44 \tag{3.3}$$

This value is telling us that we can expect that values near the center of distribution to have min max normalization value around 0.5.

- And, of course, for the slowest car which needs 25 seconds to reach 60 mph result will be 1. So from this we can see that all values from min-max normalization will be between 0 and 1.

$$X^* = \frac{X - min(X)}{range(X)} = \frac{25 - 8}{25 - 8} = 1 \tag{3.4}$$

(Larose 2005, 35-36)

3.2.5.2. Z-Score Standardization

“Z-Score standardization, which is very widespread in the world of statistical analysis, works by taking the difference between the field value and the field mean value and scaling this difference by the standard deviation of the field values.” (Larose 2005, 37)

Equation is:

$$X^* = \frac{X - mean(X)}{SD(X)} \tag{3.5}$$

- Now, for the fastest car which needs only 8 seconds to reach 60 mph, Z-score standardization will be

$$X^* = \frac{X - mean(X)}{SD(X)} = \frac{8 - 15.548}{2.911} = -2.593 \tag{3.6}$$

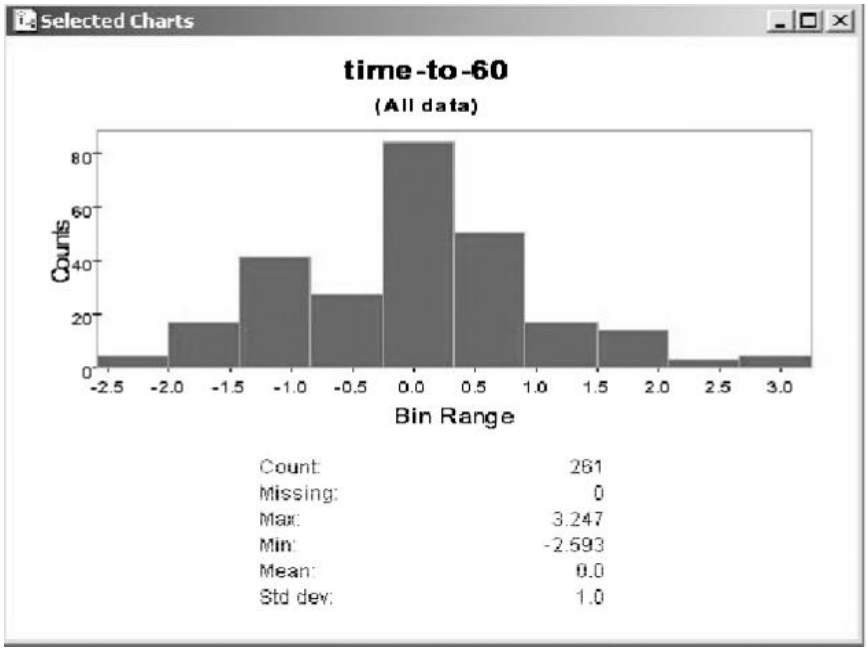
- For vehicle in between min and max value will be zero, of course.

$$X^* = \frac{X - mean(X)}{SD(X)} = \frac{15.548 - 15.548}{2.911} = 0 \tag{3.7}$$

- And for the car that needs 25 seconds to reach 60 mph, result will be

$$X^* = \frac{X - mean(X)}{SD(X)} = \frac{25 - 15.548}{2.911} = 3.247 \tag{3.8}$$

(Larose 2005, 35-36)



Count:	261
Missing:	0
Max:	3.247
Min:	-2.593
Mean:	0.0
Std dev:	1.0

Source: Larose 2005

Figure 7. Histogram of time to 60 mph after Z score standardization

3.2.5.3. Categorical data transformation

Categorical variables sometimes as well have to be transformed. Some of them contain numbers from 1-5 for example and they will be assumed to be continues numbers by the parametric algorithm and such variables can be used safely. The problem can be with categorical variables containing text values. In this kind of situation, it might be required to transform them into dummy variables. Dummy variables are binary variables, containing values 0 or 1, and those values are showing presence or absence of particular value of categorical variable. So, for example if we have colors and we want to transform them to numbers we will need to use dummy variables. Instead of having one variable with more color categories dummy variables will be added for every possible color we have. Now in case there are five colors possible five dummy variables will be added and all the cases of the data set will be coded with 0 or 1 depending on presence or absence of particular color. This has to be done when algorithms which depend on calculations of covariance, like regression, or require other numerical operations, like neural nets, are planned to be used because they can only operate with numbers.

Adding dummy variables can be very useful and helpful, especially by creating a better fit of the model. But, adding them comes with some issues and problems. *"Each raw variable that you represent by using a group of dummies causes you to lose 1 degree of freedom in the analysis. The number of degrees of freedom represents the number of independent items of information available to estimate another item of information (the target variable). Therefore, the more tightly you fit your model (the more precise your model is), the more degrees of freedom you lose. Consequently, you have less information to work with, and you are left with less ability to apply the model successfully on other data sets, which may have a slightly different target pattern than the one you fit tightly with the model. This situation is called reducing the generality of the model. Generality is just as important as (maybe even more so than) the accuracy of the model."* (Nisbet & Elder & Miner, 2009, page 58)

3.2.6. Data reduction

Data reduction can be reduction of records also known as data sampling, reduction of dimensionality or reducing number of variables and discretization of values. Data sampling is used for four main purposes in data mining and predictive modelling:

1. Simple random sampling. Because of the fact that in most of the cases it's possible to make predictive model on 10-20% of the data, predictive modelers tend to make samples. Random sampling means that each sample record selected from dataset has equal chance to be selected as any other record.
2. Stratified random sampling or selecting just those records in which response patterns are homogenous. First partitioning has to be done and after it's finished and we have clear partitions of data set then its needed to make random samples for each partition.
3. Sampling for balancing the data set is used for analysis by machine learning tools.
4. Sampling data set in 3 types of random sets.

    - Training set used for creating training model.
    - Testing set which is used to assess the predictability of the model before refining it.
    - Validation set which is used to test the final performance after all modeling is done. (Nisbet & Elder & Miner, 2009, page 69)

3.2.6.1. Reduction of dimensionality

Reduction of dimensionality is based on removing unnecessary variables. There are more ways of determining which variables are unnecessary for the model. Some of them are:

1. Correlation coefficients: Using correlation coefficients is one of the simplest ways to examine the relationships among variables. Correlation matrix can show us which of the variables have high and significant correlation coefficients and which have low and insignificant. Also if correlation coefficient between two variables is higher than 0.9 then they can cause too collinear effect so one of them should be excluded.
2. CHAID (Chi-Square Automatic Interaction Detection) algorithms is most commonly used for variable screening for the purpose of dimensionality reduction. Problems with this method is it can cause bias toward variables with more levels for splits.
3. PCA or Principal Component Analysis *"is used to identify some of the strong predictor variables in a data set. PCA is a technique for revealing the relationships between variables in a data set by identifying and quantifying a group of principal components."* (Nisbet & Elder & Miner, 2009, 71)
4. Gini index, developed by Corrado Gini in 1912, is used to examine evenness of values distributed. Gini index has values between 0 and 1 where zero represents complete evenness in distribution and maximum value 1 complete unevenness. *"This index measures the degree of unevenness in the spread of values in the range of a variable. The theory is that variables with a relatively large amount of unevenness in the frequency distribution of values in its range (a high Gini Index value) have a higher probability to serve as a predictor variable for another related variable."* (Nisbet & Elder & Miner, 2009, 71)
5. Graphical methods are jumping to help us when we have categorical variables instead of numerical ones and it's not possible to use correlation coefficients because of that very fact. They can give us insight which of categorical variables have strong relationship between each other. One of the graphs used for this is Web diagram in SPSS Clementine.
6. Other techniques used for data dimensionality reduction are: factor analysis, multidimensional scaling, singular value decomposition and others. (Nisbet & Elder & Miner, 2009, page 70-72)

3.3. Predictive models

"Predictive model—A mechanism that predicts a behavior of an individual, such as click, buy, lie, or die. It takes characteristics of the individual as input, and provides a predictive score as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior." (Siegel, E. 2013)

There are variety of ways how to create predictive models, they come in all shapes and sizes but when we talk about types of predictive models there are only few of them. Types of predictive models, the most used ones, are:

1. Linear models
2. Decision trees (Classification and Regression Trees, Boosted trees, Random Forest)
3. Neural networks
4. Support vector machines
5. Cluster models
6. Expert systems (Finlay, S. 2014)

These are the most common types of models but as mentioned before there is more than one way in creating them, there is more algorithms which can be used in creating decision trees, more than one way generating linear model and so on.

All of these types of models share one thing in common, they all need data and there are two types of data necessary to derive predictive model from it:

1. Predictor data (predictor variables) - The type of data which will be used in making prediction, data which can explain outcome data better.
2. Behavioral (outcome) data or behavior we actually want to predict (Finlay, 2014)

Both types of data are necessary for creating predictive model. Statistical and mathematical techniques are used to help in determining what is relationship between predictor variables and outcome variable, how strong that relationship is and which predictor variables will have bigger influence or impact on outcome variable. Depends on technique which is used for generating predictive model there are different methods and different logic in determining which variable will feature in the final model and how much it will contribute to the final score. (Finlay, S. 2014)

3.3.1. Predictive models: Parametric and non-parametric

Algorithms for predictive modelling can be parametric or non-parametric algorithms. When we talk about parametric algorithms they assume that distribution in data is known. Many of statistical tests, not all of them, assume normal distributions and find linear relationships in the data. On the other hand, machine learning algorithms do not make assumptions about distribution and they are considered to be non-parametric or distribution free models. (Abbott, 2014)

"The advantage of parametric models is that if the distributions are known, extensive properties of the data are also known and therefore algorithms can be proven to have very specific properties related to errors, convergence and certainty of learned coefficients." (Abbott, 2014) When those assumptions are clear and known analysts often spend considerable time transforming the data so they can take advantage of them.

Machine learning algorithms are more flexible because there are no any assumptions about distribution of data and therefore they can save considerable amount of time for analysts in preparing the data sets since they are not sensitive and dependent on distribution of data. Of course, there are also disadvantages of this types of algorithms because far less is known about the data and these algorithms which are typically iterative, cannot guarantee that optimal solution has been found.

3.3.1.1. Supervised and Unsupervised Learning

There are two main groups of algorithms: supervised learning methods and unsupervised learning methods. In first group, the supervisor is our target variable, in other words that is the column in the data which value we want to predict using other columns in the data. That target variable is supposed to be answer on the question company or organization wants to know. To main supervised learning algorithms are regression and classification, first one used for continuous target variables and second one for categorical target variables. Target variables could be will customer buy some product, the amount of a purchase, will transaction be fraudulent, will customer take a loan from bank and so on. Any missing value records have to be removed because they can't be used in the model. Supervised learning is sometimes also called predictive modeling. (Abbott, 2014)

"Unsupervised learning, sometimes called descriptive modeling, has no target variable. The inputs are analyzed and grouped and clustered based on the proximity of input values to one another. Each group or cluster is given a label to indicate to indicate which group a record belongs to." (Abbott, 2014) In customer analytics it's just called segmentation because it is used for segmenting customers in the groups.

Further in this chapter types of predictive models and data mining techniques and algorithms will be more thoroughly explained.

3.3.2. Linear models

Linear models are most used models in predictive analytics. There are some key characteristics related to linear models:

- The relationship between each predictor variable and behavior is represented by a weight.

- The contribution that each predictor variable makes to the model score is calculated by multiplying the predictor variable by the weight.
- The final model score is calculated by adding up the contribution made by each predictor variable (the sum of the predictor variables multiplied by the weights). (Finlay, 2014)

Linear models are really useful when relationship in data is linear but with non-linear data they will not create accurate prediction. There are a lot of techniques which can help in overcoming this problem, if there is no linear connection between raw variables maybe there is between square root or square of one of variables. Discretization or binning of the data is also used in this purpose.

Most commonly used linear models are linear and logistic regression. Linear regression works with numerical data so both target and input variables have to be numeric because linear regression equation is describing an arithmetic relationship between them. Linear regression looks for the best relationship by minimizing the sum of squares of the vertical distances from the data points to the line. This type of linear regression is also called ordinary least square regression. (Linoff & Berry, 2011)

When we talk about classification models, where the goal is to predict will someone do something or not, most used method when we talk about linear models is Stepwise logistic regression. Logistic regression represents a linear classification technique used for binary classification problems because linear regression cannot be used for modeling binary outcomes. Stepwise regression is a semi-automated process of building a model by sequent adding or removing variables based solely on the *t*-statistics of their estimated coefficients. With this method creation of model is done in stages. At initial stage model is created just with one variable which is determined by algorithm as the most significant predictor. At each next stage new predictor variables are being added or removed depends on how significant influence they have on final outcome. Process is done when there are no more predictor variables which can improve model. (Finlay, 2014) (Abbott, 2014)

For both logistic and linear regression issue about selection process is that it tends to be harder when number of predictor variables increases, there is always a possibility of the increase of effects and interactions between variables. So, two things should be considered when we develop linear models, models should be complex enough to fit the data and that they should be relatively simple to interpret tending to smooth to data rather than over fit it. (Agresti, 2013)

Beside logistic and linear regression other popular methods for creating linear (classification) models include Discriminant Analysis (the first popular method for developing Predictive models), Integer Programming (a method for building credit scoring models), Genetic Algorithms, probit and Tobit analysis (Both similar to logistic regression) and Bayesian methods. (Finlay, 2014)

3.3.3. Decision trees

Decision trees are probably the second most popular type of predictive models nowadays. Three key characteristics of decision trees are:

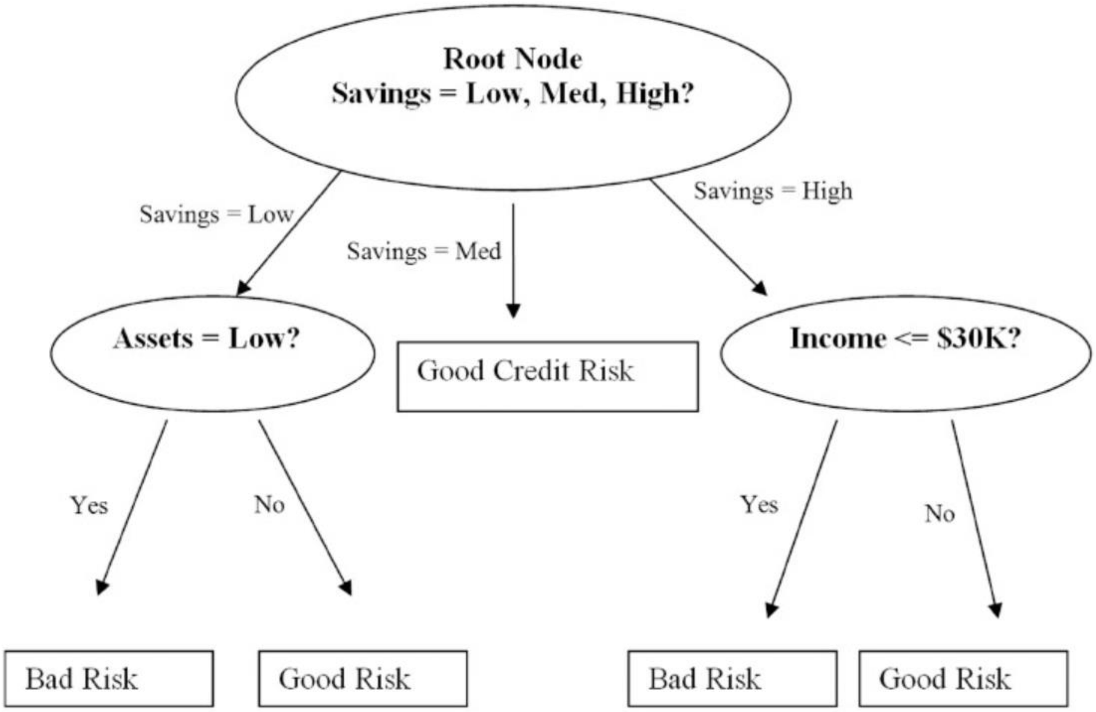
- The model is created by segmenting a population into smaller and smaller segments.
- The model can be represented as a "Tree diagram."
- The model score is determined by the end node into which an observation falls after passing through the tree. (Finlay, 2014)

Eric Siegel (2013) said that decision trees are simple, elegant, precise and practically math less. In order to use decision trees we have to start from the top, the root and answer yes or no questions until we reach a leaf. The leaf indicates prediction for that individual.

Decision tree is a collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes. Tree begins at the root node, which by convention is placed at the top of the decision tree diagram, attributes are tested at the decision nodes and possible outcomes are resulting in a branch. Branches then can lead to another decision nodes or to an end leaf node. (Larose, D & C. 2014)

As Siegel (2013) said decision trees are extending far beyond just the business world, they are used to make decisions almost everywhere, in medical, legal, governmental, astronomical, industrial and all other spheres. Learning process is very adaptable because decision tree is making decisions solely by the data upon which it grows. "*A decision tree grows upon the rich soil that is data, repeatedly dividing groups of individuals into subgroups. Data is a recording of prior events, so this procedure is learning from the way things turned out in the past.*" (Siegel, 2013)

When building of a tree is finally done and decision tree algorithm has finally built it, leaf or terminal nodes are the ones to collect records which are matching the rule of the tree above and every record can end up in one and strictly one of the terminal nodes. (Abbott, 2014)



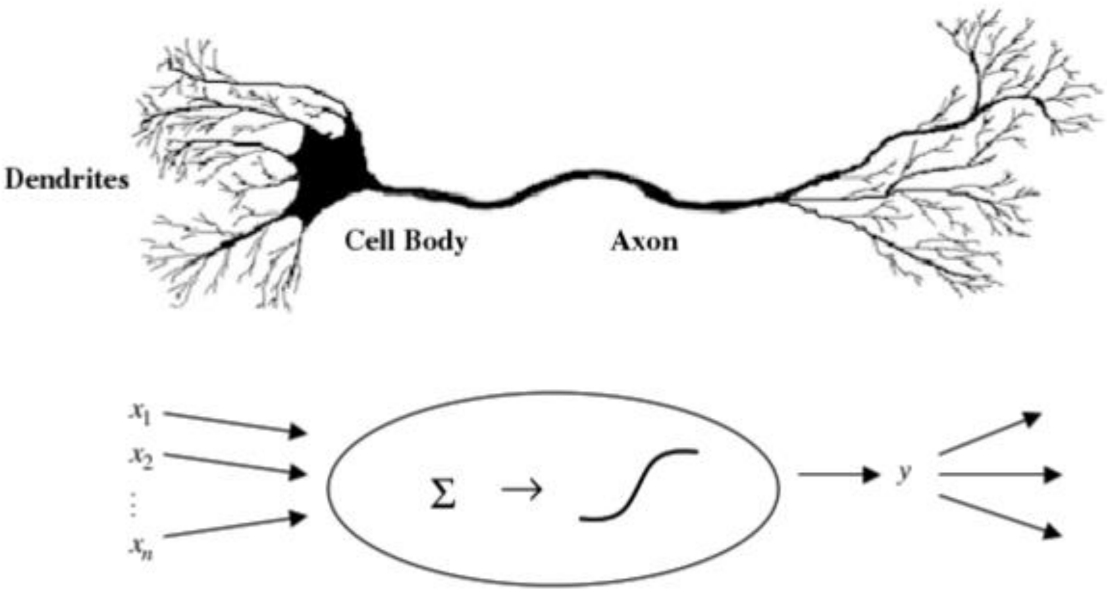
Source: Larose 2005
Figure 8. Example of simple decision tree

Decision trees unlike linear models are not using weights and everything is based on the logic for segmenting population and knowing how far segmentation process should go. Decision trees are non-linear form of model and they are able to represent non-linear relationships between variables and there is no the same need in transforming data like in linear models. One of the best things about decision trees is that they are relatively easily understandable and easy to explain to non-technical audience. There is a lot algorithms used for generating decision trees but the most used algorithms for making decision trees are C4.5/C5.0, CART and CHAID and even though that all of algorithms have their pros and cons these four are most used in standard statistical software packages. (Finlay, 2014)

Building the decision tree which belongs to a class of recursive partitioning algorithms there are few steps which should be done. First, for each candidate input variable we should find the best way to split the data in subgroups. Data will be divided to subgroups which are defined by the best split. Next step is to select one of the subgroups and repeat the first step and then repeat it to all subgroups previously created. Third step is to continue splitting until all records belong to the same target variable value or until one of the stop condition is applied, which can be a statistical significance test or a minimum record count. (Abbott, 2014)

3.3.4. Neural networks

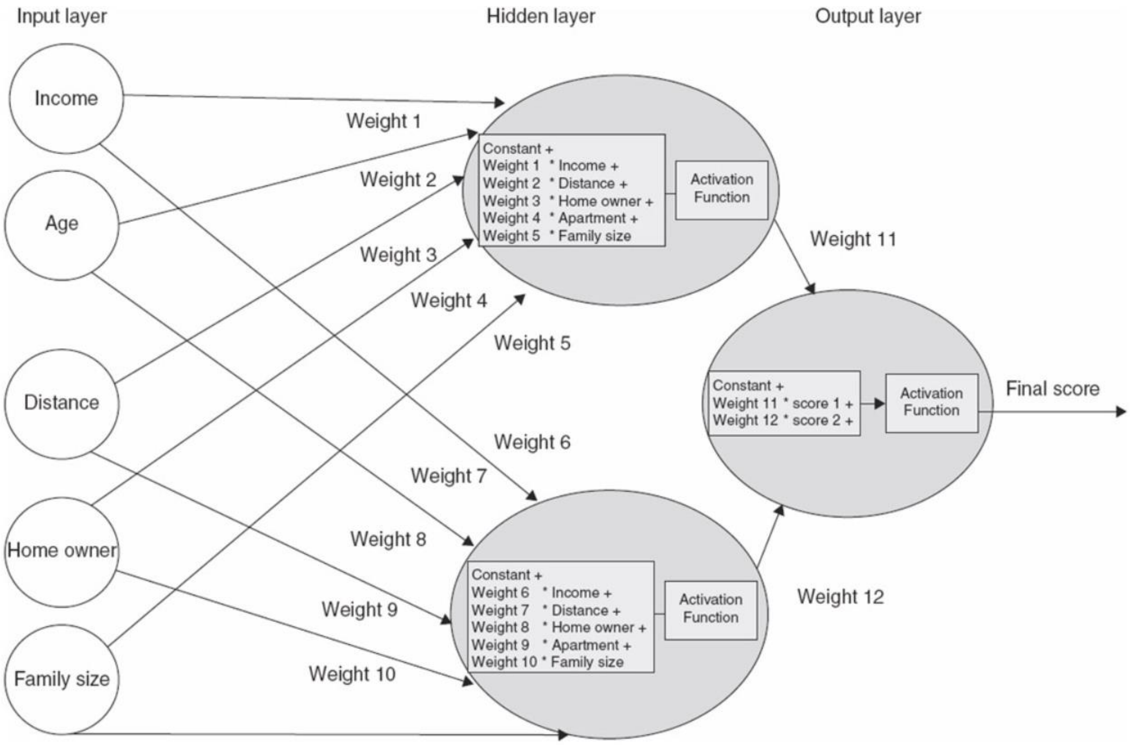
"Artificial neural networks represent an attempt at a very basic level to imitate the type of nonlinear learning that occurs in the networks of neurons found in nature." (Larose, 2005,128)



Source: Larose 2005 page 128
Figure 9. Similarities between neuron in neural networks and nature

It would be wrong to look at neural networks like on some kind of artificial human brain, because they don't have any artificial intelligence and they are simply not artificial brains. Similarity is purely and only in idea.

Neuron is the key part of a neural network and neuron is operating in two stage process. First every predictor variable is multiplied by the weight and then summed to generate a score. This part of neural network is linear model, like linear models described earlier. After linear stage it comes nonlinear stage known as activation function which then generates the output. Commercially most used neural networks consist from input layer, hidden layer and an output layer and it's shown on figure below. Hidden layer contains linear model with weights and, usually nonlinear, activation function. (Finlay, 2014)



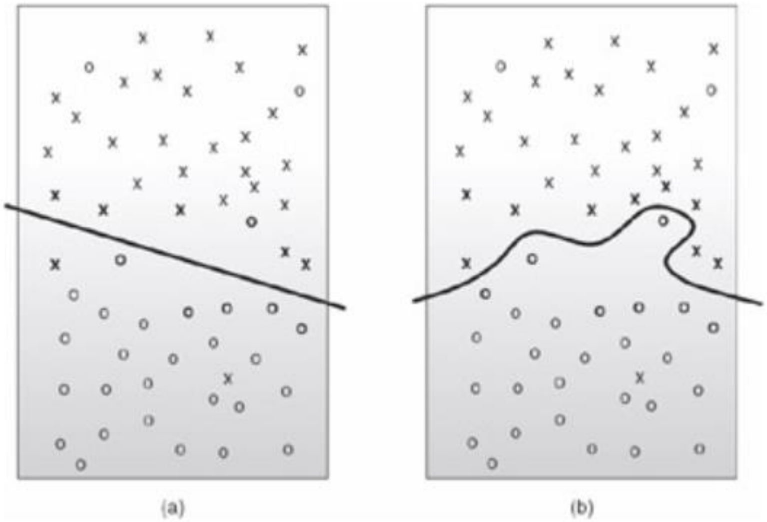
Source: Finlay 2014
Figure 10: Example of neural network

Number of neurons in hidden layer depends on problem complexity and varies from two to up to twice the number of inputs while output layer contains a single neuron which has same features as neuron in hidden layer. (Finlay 2014)

One of the biggest problem of neural networks is that they tend to be very complex and because of that complexity it can be very hard to explain them as well as to choose which input will go where, how to determine the weights and how many neurons we will have in hidden layer. Even though they tend to have better predictive accuracy than other more commonly used models because of that very complexity sometimes is just easier and better to use some other easy explicable predictive models.

3.3.5. Support vector machines

Support vector machines are the most difficult to explain among all types of models used for predictive modeling. Even though they share similarities with neural networks understanding and explaining them without proper mathematical knowledge is almost impossible. Support vector machines are nonlinear models, using machine learning algorithms. Beside similarities with neural networks they still operate in different way. They operate in the way of finding equation of the line which will maximize the margin and that equation will be the model. The score will be a measure of probability of the case lying on one or the other side of that line. High score will represent probably event while low score will represent probably not an event. (Finlay, 2014)



Source: Finlay 2014
Figure 11: Maximizing the margin

The Figure 11. above shows us two examples, where the first one is representing perfect straight line dividing two groups clearly, and this type is more similar to linear models. On the (b) part of the wavy line is not restricted to being straight and because of that fact it does better job in dividing two groups. This is more similar to nonlinear models like neural networks. (Finlay, 2014)

With first part of the figure we can tell that it looks simple enough using support vector machines but the (b) part of figure shows us how exactly complicated this can be. In this situation we need to perform transformations of data before we even think about finding equation for maximizing the margin. These transformations tend to be very complex and interpreting them in any understandable way is hardly possible, it's very hard to describe results and describe what is actual contribution of each variable. Big difference between SVMs and all other types of models (linear, decision trees or neural networks) is lying in that all other types of models are using all the data which is used in creating sample to model development as well. SVMs will use only those cases which are close to the line because cases which are far from the margin are highly unlikely to tell us where the line should be drawn. So the most important part of the SVM algorithms is deciding which of cases will be support vector and which not, after that we can do needed transformations to support vector cases and then create the model. (Finlay, 2014)

3.3.6. Cluster models

When we have to deal with the data which have similar attributes like same income, family size or education level grouped together, clustering is the data mining technique which should be used. Clustering is divided in two groups, first which is not used to predict the behavior but only to group together cases which are similar and the second predictive analytics technique called K-nearest neighbor, widely used technique when we want to know something about behavior of the cases. For the first technique there is an assumption that cases, people, in the same cluster or group will tend to behave in certain similar way like buying the same products, have the same health problems or have the same mortality rate for example. This type of clustering is widely used in marketing but of course usable in other areas as well. In marketing this type of clustering is used to help companies to decide how will they target certain groups of people. People grouped by income helps companies to decide whether to offer people expensive or cheaper products. Two algorithms which are most often used in this type of clustering are K-means clustering and Hierarchical clustering. Advantage of these technique of clustering is we don't need to know anything about behavior so we can use it, the only need is sample of people with some data describing them to do this type of clustering. (Finlay, 2014)

K-nearest algorithm is most often used classification technique and it's also used for making predictions and estimations. This technique is so called instance-based learning, where samples stored in training data set is used to classify new unclassified records by comparing them with already classified records from training data set by comparing them and finding which are the most similar records with ones already clustered. (Larose, 2005)

This type of clustering algorithm is a non-parametric algorithm, the training data is actual model and its used as a look up table where new records should be grouped based on similarities or how close they are to neighbors. The "k" refers to number of neighbors, which are near the data point, that are needed to make prediction. The number of neighbors you should choose is not anyhow related to theory, more neighbors you include will result in smoother prediction. For binary classification suggestion is to use odd number of neighbors to stay out of the possible ties in voting. Instead of choosing by theory most likely the number of nearest neighbors used will be discovered and determined through an iterative process where you start with small number of nearest neighbors and then increase them step by step until the moment testing data error starts to increase. (Abbott, 2014)

3.3.7. Choosing the best model

There are more criteria which have to be considered when we talk about choosing the right model for predicting. The most common criteria which should be considered are:

1. Predictive accuracy of the model
2. Explicability of the model
3. Simplicity
4. Stability
5. Business acceptability (Finlay, 2014)

Predictive accuracy is criteria on which predictive modelers tend to focus the most. So when we talk about predictive accuracy the most studies shown that important thing to know is that most of them will achieve the similar predictive accuracy with slightly better accuracy gotten from support vector machine and neural network models. But these better results are usually not that better and different from using less complicated algorithms and models, the difference in their accuracy is most of the time really small. Also we should know that there is no assurance that SVM or neural networks will outperform way simpler linear models or decision trees which can derive us better prediction accuracy depending on situation and data used. (Finlay, 2014)

Neural networks and decision trees are usually prone to overfitting to the data, in other words they will give better results on the data they have been developed than in real world usage. And that can be a problem. Also, models should be simple and explicable but neural networks, support vector machines and clustering have the problem to be highly complex and not easy to understand. Their problem is also that you cannot know which variable contributed more and there is always possibility that some of them will not contribute to final outcome at all. One more problem related to clustering is that clustering model is not model by its nature, yes data has been clustered and yes you can add new data but when you want to add new data you will have to run algorithm again so the new data will be just placed to one of the clusters. Second problem of clustering models is that they tend to have problems with large quantities of data, regardless we talk about number of records or variables. (Finlay, 2014)

Stability of the model is the problem by itself, every model has its own lifespan and some can be useful for years while others have predictive ability measured in days and even hours. So rebuilding models which will be used in longer period of time is much needed action. Sometimes the most important thing is will it be possible to apply that model in business and two things should be considered when we talk about this criterion. First thing addresses a physical infrastructure required and second addresses the people and culture of the business or will the business agree with the model developed by data scientist. (Finlay, 2014)

Steven Finlay (2014) suggests that when it comes to choosing the right model these things should be considered:

- We should always use linear models as a benchmark. Stepwise linear and logistic regression should be used in this case to develop benchmark models. Developing them with nowadays software tools should be easy and quick and it will give us a baseline to which we will compare the other types of models.
- If there is space and time we should develop different models and use them on independent sample which was not used in generating models and then compare the results.
- And third thing is that even though we don’t want to use neural networks and support vector machines due to their complexity we should develop them just to see what is the benefit we can gain by using them.

So, we can say that there is no perfect model and all of them have pros and cons. It should be carefully decided which to use based on situation, data and when time allows we should develop more models and compare them to choose the right one, or the one we consider as such.

4. Practical part

4.1. Dataset and variables description

The main goal of this part of the thesis will be to make predictive model to predict will the client subscribe a bank term deposit or not, so there is a binary classification goal. In order to predict will clients subscribe or not there are 20 possible predictors, input variables which will be described in Table 5. below.

Nr.	Name	Description
1.	Age	Numeric
2.	Job	Categorical: type of job (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)
3.	Marital	Categorical: marital status (married, single, divorced, unknown)
4.	Education	Categorical: type of education (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)
5.	Default	Categorical: has credit in default or not (yes, no, unknown)
6.	Housing	Categorical: has housing loan (yes, no, unknown)
7.	Loan	Categorical: has personal loan (yes, no, unknown)
8.	Contact	Categorical: contact communication type (cellular, telephone)
9.	Month	Categorical: which month was last contact performed (jan, feb, … , nov, dec)
10.	Day_of_week	Categorical: which day was last contact performed (mon, tue, … , sun)
11.	Duration	Numeric: last contact duration in seconds
12.	Campaign	Numeric: number of contacts performed during this campaign and for this client, last contact included
13.	Pdays	Numeric: number of days that passed by after the client was last contacted from a previous campaign (999 means the client was not previously contacted)
14.	Previous	Numeric: number of contacts performed before this campaign and for this contact
15.	Poutcome	Categorical: outcome of previous campaign (nonexistent, failure, success)
16.	Emp.var.rate	Numeric: employment variation rate – quarterly indicator
17.	Cons.price.idx	Numeric: consumer price index – monthly indicator
18.	Cons.conf.idx	Numeric: consumer confidence index – monthly indicator
19.	Euribor3m	Numeric: euribor 3 month rate – daily indicator
20.	Nr.employed	Numeric: number of employees – quarterly indicator
21.	Y	Binary: has the client subscribed a term deposit (yes, no)

Table 5. Description of variables in data set
Source: Moro et al., 2014, own processing

Dataset contains 41188 cases (or instances) and 20 input and one output variable. The data is related with direct marketing campaigns of banking institution from Portugal. The data is collected through marketing campaign which was based on phone calls and usually more than one contact with the same client was required.

There are several missing values in several categorical variables, all missing values are labeled as “unknown” and they can be treated as new category or using deletion or imputation techniques. After initial exploring of the data decision about dealing with missing data will be made.

4.2. Initial graphical review and descriptive statistics

In order to explore the data STATISTICA 13 and Microsoft Office Excel software packages will be used. For initial graphical exploration which should show how data looks like Interactive drill down tool in STATISTICA 13 will be used, this tool allow user to automatically create bar charts or any other type of charts or summary statistics for all variables just with one click. After initial graphical review of variables by bar charts descriptive statistics will be used for numeric type of data.

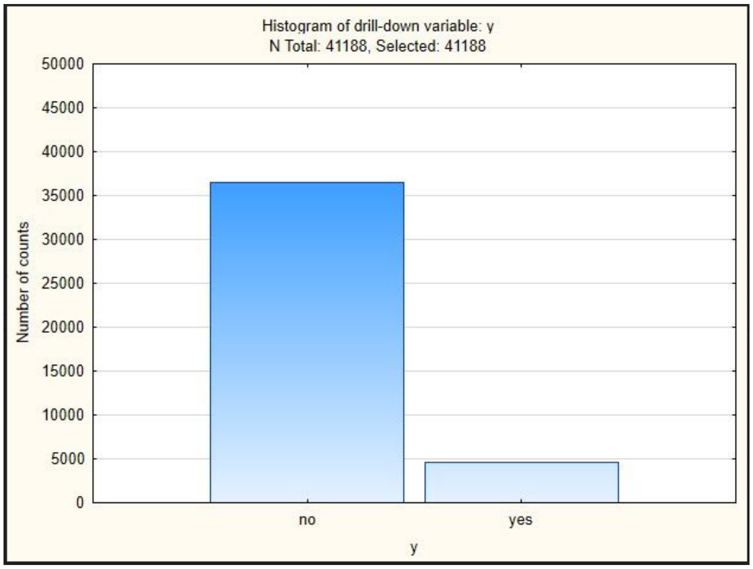


Figure 12. Bar chart of outcome variable Y
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Bar chart of our output variable shows that there are slightly more cases than 10% of whole data set which will subscribe a bank term deposit and that vast majority will not. This means that for validation dataset most likely stratified random sample will be used.

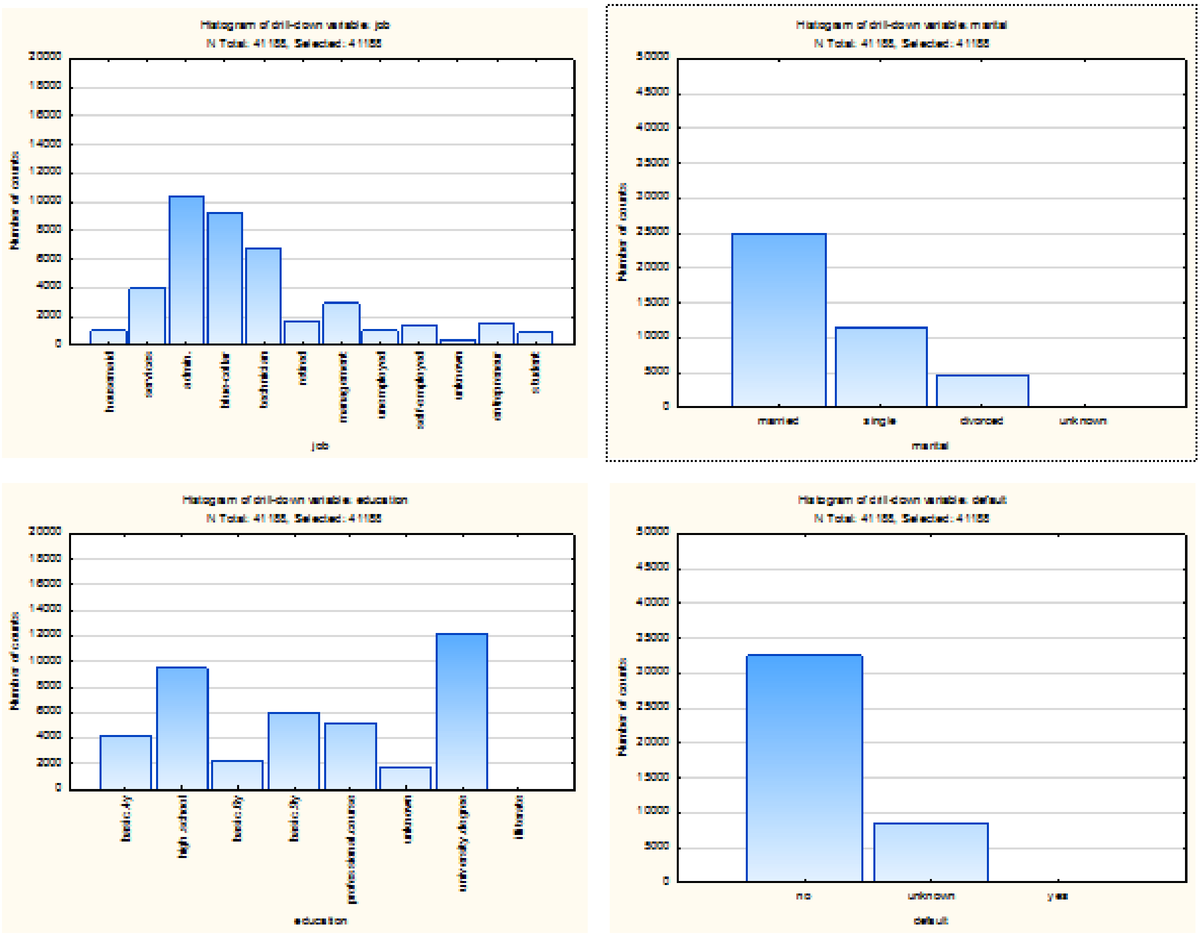


Figure 13. Bar charts of variables job, marital, education and default
Source: Data from (Moro et al., 2014) STATISTICA output own processing

On bar chart for variable “education” it’s possible to see that basic education with 4, 6 and 9 years are significantly less than high school and university education so possibility is to group 3 groups of basic education to one but more about that in data transformation chapter. Bar chart also shows us that frequency of class “illiterate” is very low so there is possibility that this class is outlier so that has to be checked as well. The bar chart also shows that number of missing values is much higher than in variables “job” and “marital” and this will have to be inspected in later chapter focused on dealing with missing values.

Variable “default” which should show does client have credit in the bank or not contains more than 20 % of missing data also category “yes” is significantly low with only 3 cases out of 40+ thousands of cases, it barely exists so possibilities to deal with this problem will be omitting this variable from model because of high number of missing data and low number of category “yes” or treating missing data as new category. There is twice as much married than single people so there is possibility in combining “single” and “divorced” in one class “single”.

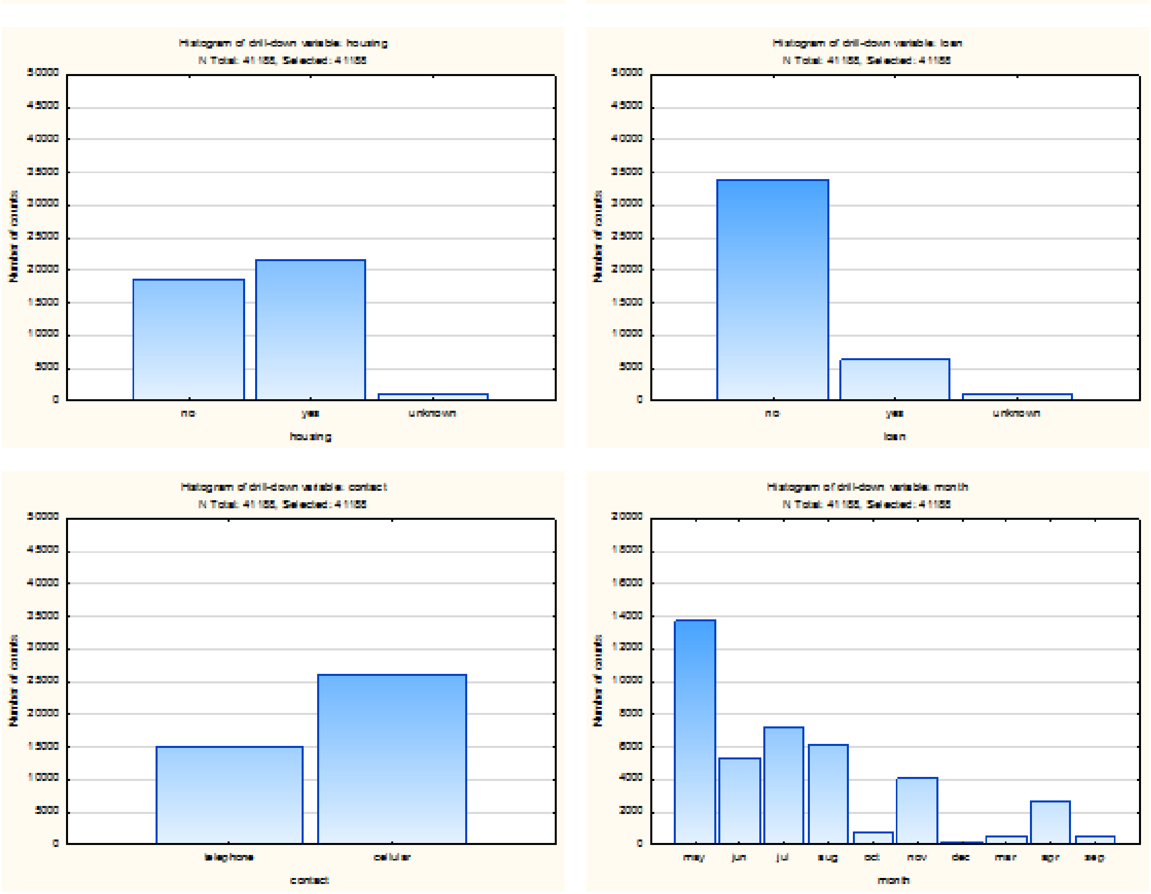


Figure 14. Bar charts of variables housing, loan, contact and month
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Variables housing loan and personal loan have similar amount of missing data but very different frequencies of classes “yes” and “no”. Bar chart of months when clients were contacted shows that clients were more frequently contacted from May to August than in the other months of the year. One of possibilities is to group months to quartiles because some months have very low frequencies. Also in months January and February there were not contacts at all.

4.3. Cleaning data - variable transformation and recoding, outliers and missing data

The very first step will be recoding and transforming variables which have classes with very low frequencies in order to try to resolve that problem before searching for outliers. Bar charts in initial graphical exploration showed that some frequencies are very low so the decision is to recode some of classes of categorical variables.

First one will be variable “job” and the decision is to put category “students” into the category “unemployed” since students are not employed. As well combining categories “self-employed” and “entrepreneur” into same category because even though they are different they share the same characteristic of not working for somebody else. Also category “housemaid” will be added to category “blue-collar” because they share same characteristic of low paid job.

After transformations bar chart is showing us that there is no more classes with very low frequencies in this variable.

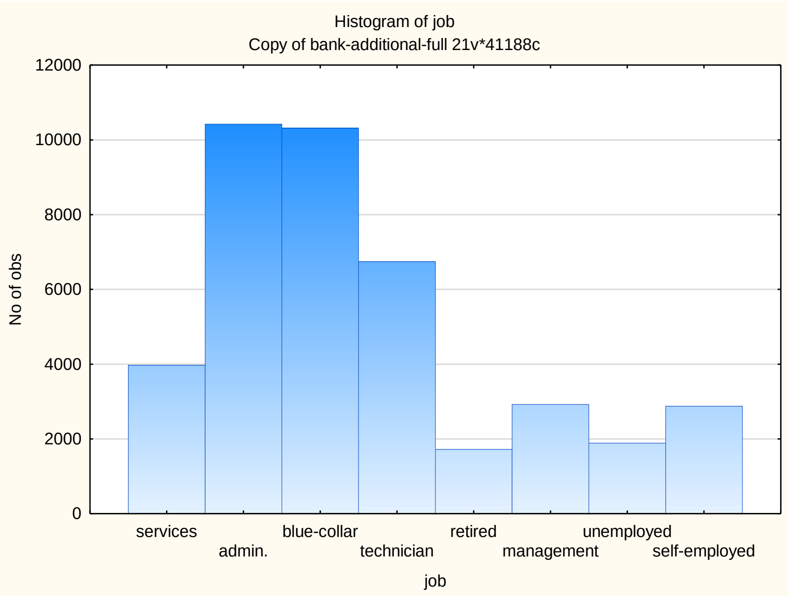


Figure 15. Bar chart of transformed variable “job”
Source: Data from (Moro et al., 2014) STATISTICA output own processing

In variable “education” all three levels of basic education were combined into one named “basic” and frequencies of education are more even now.

Variable months was transformed as well, first idea of making quarters instead of months was not good because frequency for first quarter of the year was extremely low. So, instead of making quartiles months April and March were combined in one group as well as months from September till December because otherwise they would be considered as outliers.

Bar charts of variables “euribor.3m” and “nr.employed” showed that these variables have really low frequencies for some values and since none of numerical transformations didn’t help decision was to transform them to categorical variables. From same reasons variable “emp.var.rate” was transformed as well.

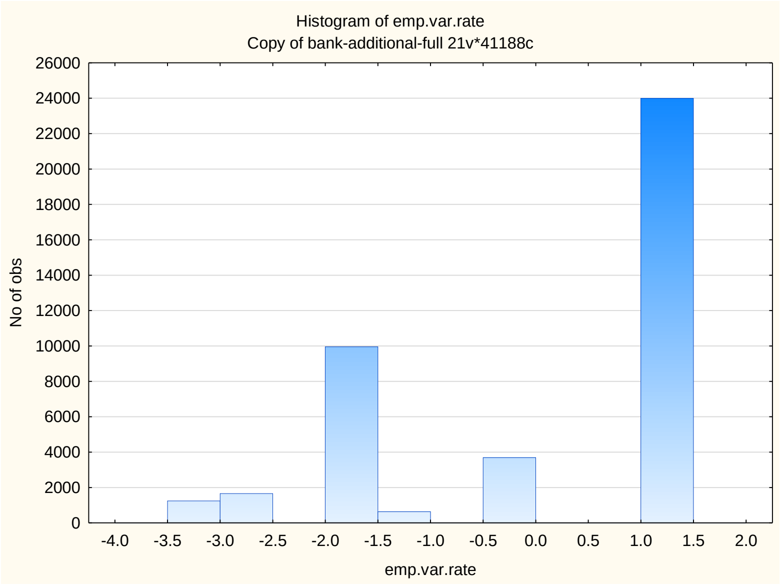


Figure 16. Bar chart of variable "emp.var.rate" before transformation
Source: Data from (Moro et al., 2014) STATISTICA output own processing

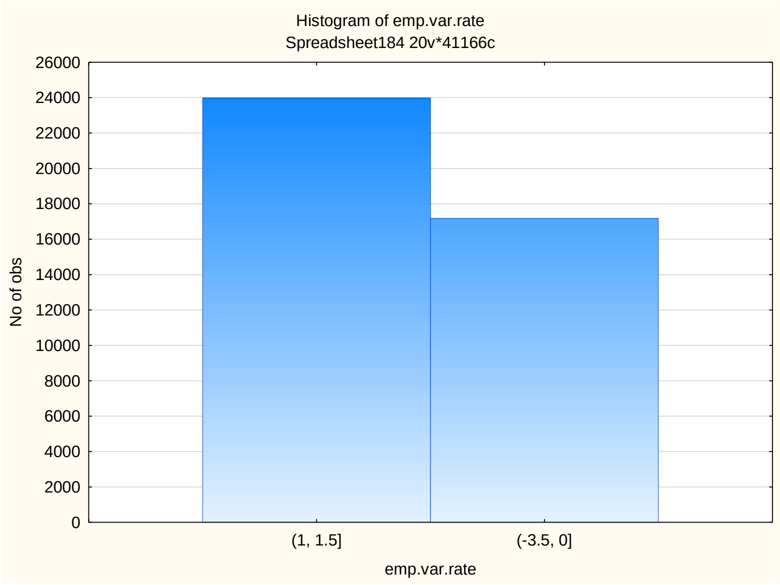


Figure 17. Bar chart of variable "nr.employed" after transformation
Source: Data from (Moro et al., 2011) STATISTICA output own processing

Variable "campaign" was transformed to categorical as well because of extreme values showed on bar chart. Classes of variable are now "one" for one contact and "two_or_more" for more than two contacts. Bar chart below shows how variable frequencies look like after transformation.

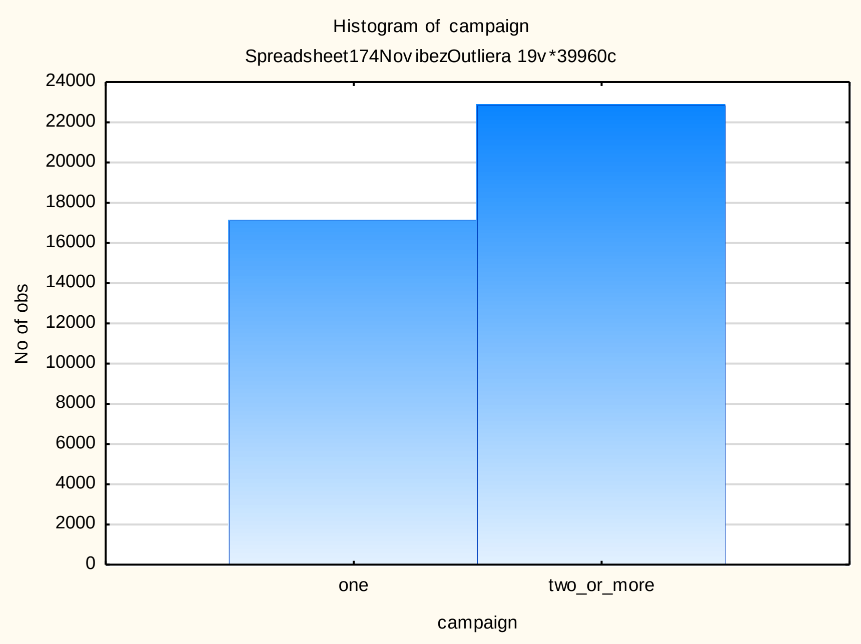


Figure 18. Bar chart of variable “campaign” after transformation
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Variable “pdays” is removed from data set because vast majority of cases showed that there was no previous contact and none of numerical transformations were helpful as well as transforming it to categorical, and every other value than 999 which means no contact at all would be considered as outlier.

4.3.1. Dealing with outliers

After transformations, second step is to detect and recode outliers. For detecting outliers in numerical (continuous) variables box plot charts were used.

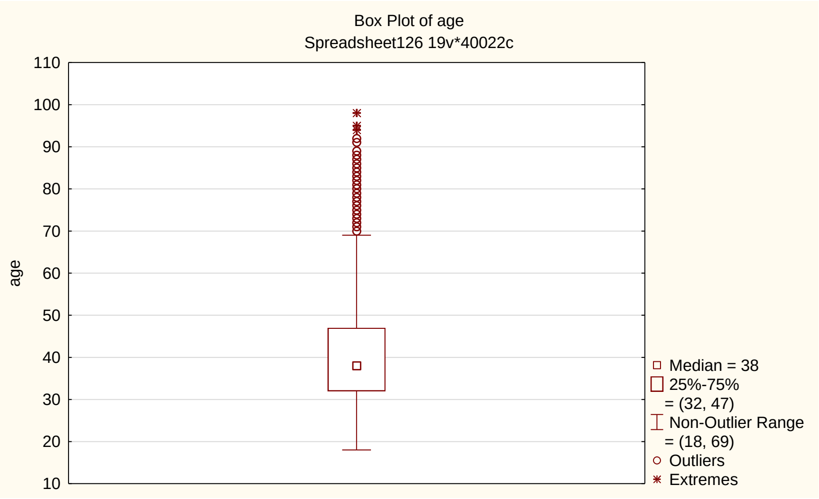


Figure 19. Box plot of variable “age” showing outliers and extreme values
Source: Data from (Moro et al., 2014) STATISTICA output own processing

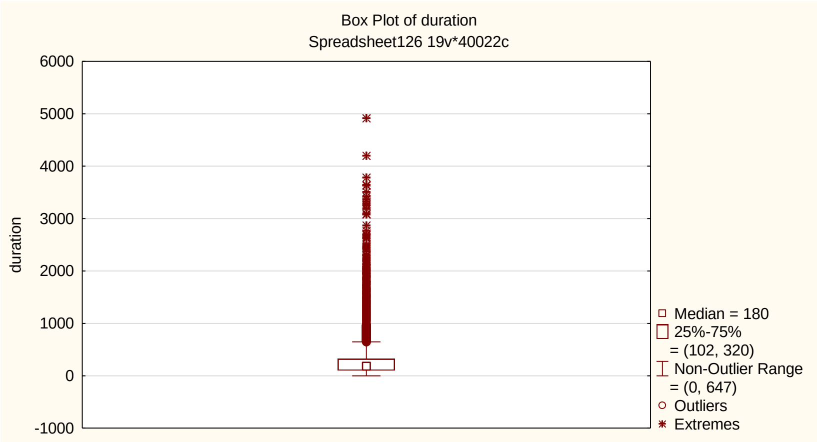
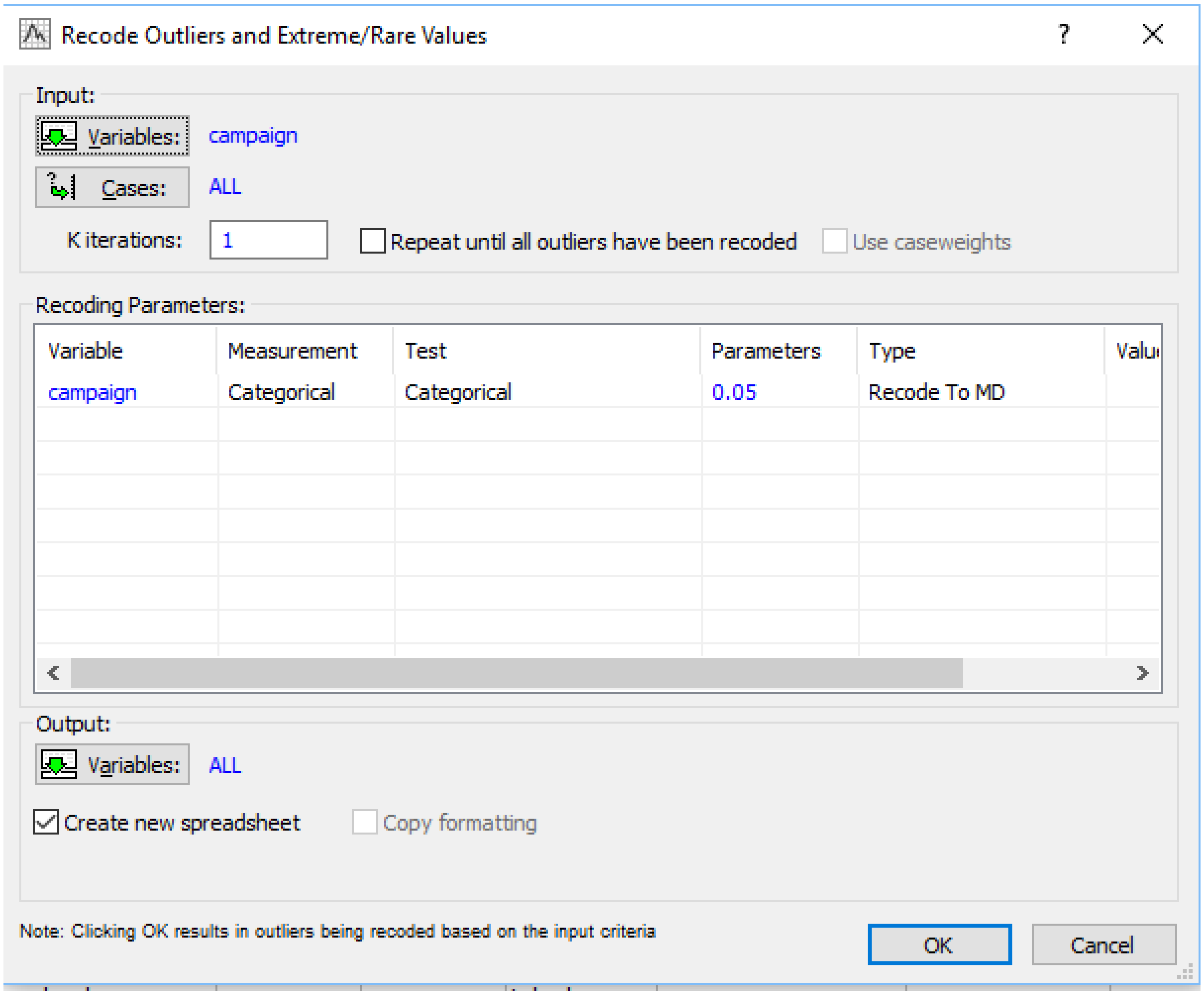


Figure 20. Box plot of variable “duration” showing outliers and extreme values
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Box plots of variables “age” and “duration” showed outliers and using feature for recoding outliers they were recoded to missing values and later in next chapter decision about treating missing data in those two variables will be made. Variable consumer price index didn’t have outliers.

For detecting outliers in categorical data first idea is to look at bar charts of those variables (shown in previous chapter) and after that tool which STATTISTICA 13 software offers for finding and treating outliers in categorical data found outliers and recoded them to missing data.

Outliers were class "illiterate" in variable "education" as well as class "yes" in variable "default". Other variables, because of previous transformations, didn't have outliers. All outliers were recoded to missing data.



Recode Outliers and Extreme/Rare Values

Input:

Variables: campaign

Cases: ALL

K iterations: 1

□ Repeat until all outliers have been recoded □ Use caseweights

Recoding Parameters:

Variable	Measurement	Test	Parameters	Type	Valu
campaign	Categorical	Categorical	0.05	Recode To MD	

Output:

Variables: ALL

☑ Create new spreadsheet □ Copy formatting

Note: Clicking OK results in outliers being recoded based on the input criteria

Figure 21. Recode outliers tool in STATISTICA
Source: Data from (Moro et al., 2014) STATISTICA output own processing

4.3.2. Dealing with missing data

First step in dealing with missing data is filtering sparse data and STATISTICA offers the tool for that. Sparse data which represents cases and variables which contain some amount of missing data (usually more than 10% in cases or variables) have to be filtered out from data set since they can harm model.

After using tool for filtering sparse data and removing duplicate cases, number of cases is reduced to 39947 cases. Also variable “default” was removed from data set since it had more than 10% of missing values.

Since there is still some missing data left in couple of categorical and numerical variables some of techniques for data imputation have to be used because most of algorithms do not do well when there is missing data in data set. Missing data can be recoded with mean, median, mode, using more advanced k-nearest neighbor algorithms or with optional value decided by analyst.

First variable “age” had outliers which were recoded to missing values and all of those values considered as extreme or outlier values were for people who were already retired. Since the number of missing values was not high and significant and since none of techniques of imputation didn’t help because imputing mean value (value of mean is 40,024) does not really represent real age for most retired people and k-nearest neighbor also didn’t give results good enough (imputing values around 50) decision is to impute optional value of 62 which is the mean value for retired people age in this data set.

For missing data in variables “job” k-nearest neighbor algorithm was used. For predicting each of their missing values variable “age” was also used in combination with variable “education“ and combination of all three variables can give us the best possible imputation of missing data. Age is highly related to class “retired” in variable “job” and education level is also related to other classes of variable job. By default settings number of k-nearest neighbors was 3. Results didn’t make significant change of frequencies of classes in variables which can be seen in bar charts for that variable and gave much better results than mode imputation for example.

For variable “marital” mode imputation was used since there was only few cases with missing data.

Models which will be explained in later chapters gave the best results on data set cleaned in this way in all other versions of cleaned data models had lower accuracy and less success.

4.5. Finding relations between variables

To find relations between input variables and output variable its necessary to use some of graphs and cross tabulation statistics, because graphical exploration and statistical analysis can show us which of variables are possibly good predictors of output variable. Depending on type of variable, numerical or categorical, mean plot graphs and 3D bivariate bar charts were used.

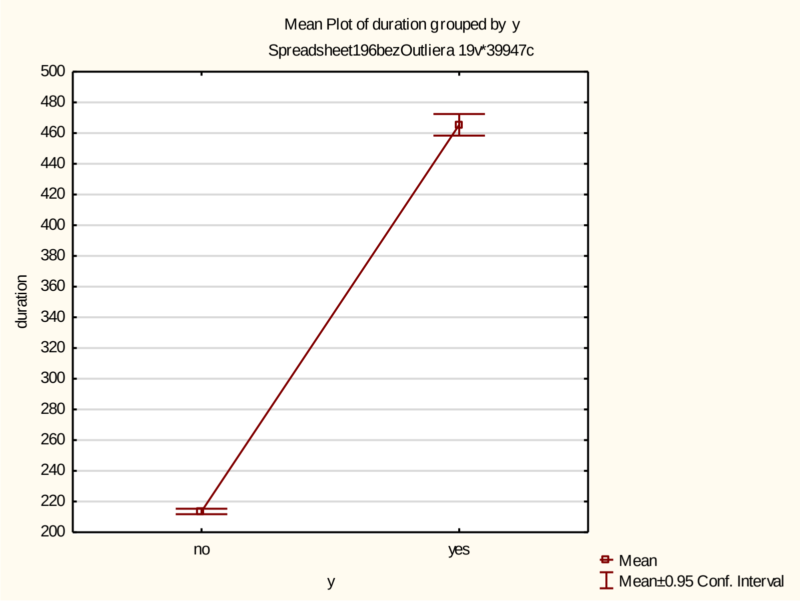


Figure 22. Mean plot of variable “duration”
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Mean plot of variable “duration” grouped by output variable “y” shows that there is relation between mean value of duration of call and output variable “y”. Shorter duration of call will most likely lead to not subscribing a bank term deposit while with longer duration of call there are bigger chances that person will subscribe. This is because people who are not interested in subscribing a bank term deposit will not talk for a long time with bank personal.

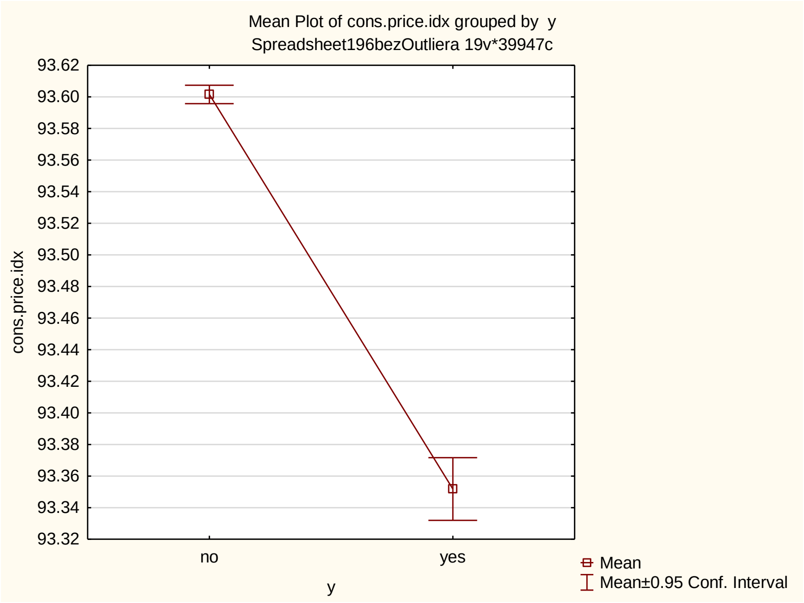
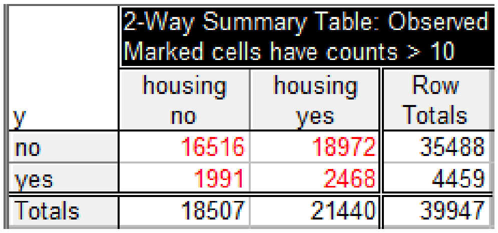


Figure 23. Mean plot of variable “cons.price.idx”
Source: Data from (Moro et al., 2014) STATISTICA output own processing

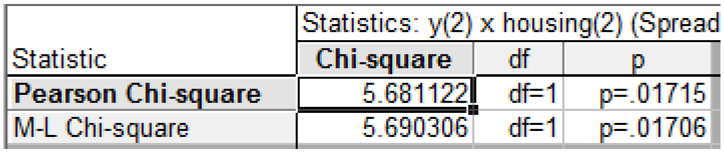
Next mean plot shows us relations between variables consumer price index and outcome variable and with lower values for consumer price index there are bigger chances for subscribing bank term deposit than with higher values of consumer price index. Lower values of consumer price index indicate better purchasing power for customer so there is bigger chance for them to subscribe a bank term deposit as well.



	2-Way Summary Table: Observed Marked cells have counts > 10		
y	housing no	housing yes	Row Totals
no	16516	18972	35488
yes	1991	2468	4459
Totals	18507	21440	39947

Table 6. 2-Way summary table for variables “y’ and “housing”
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Cross tabulation statistics and bivariate 3D bar chart show that there are similar chances of subscribing or not subscribing deposit with both having or not having housing loan. P-value which can be seen in table 7. below is less than 0.05 so this variable can be used as predictor. In next chapter using Feature selection tool in STATISTICA decision will be made is this variable is going to be part of model or not.



	Statistics: y(2) x housing(2) (Spread		
Statistic	Chi-square	df	p
Pearson Chi-square	5.681122	df=1	p=.01715
M-L Chi-square	5.690306	df=1	p=.01706

Table 7. Chi-square test for variables "y' and "housing"
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

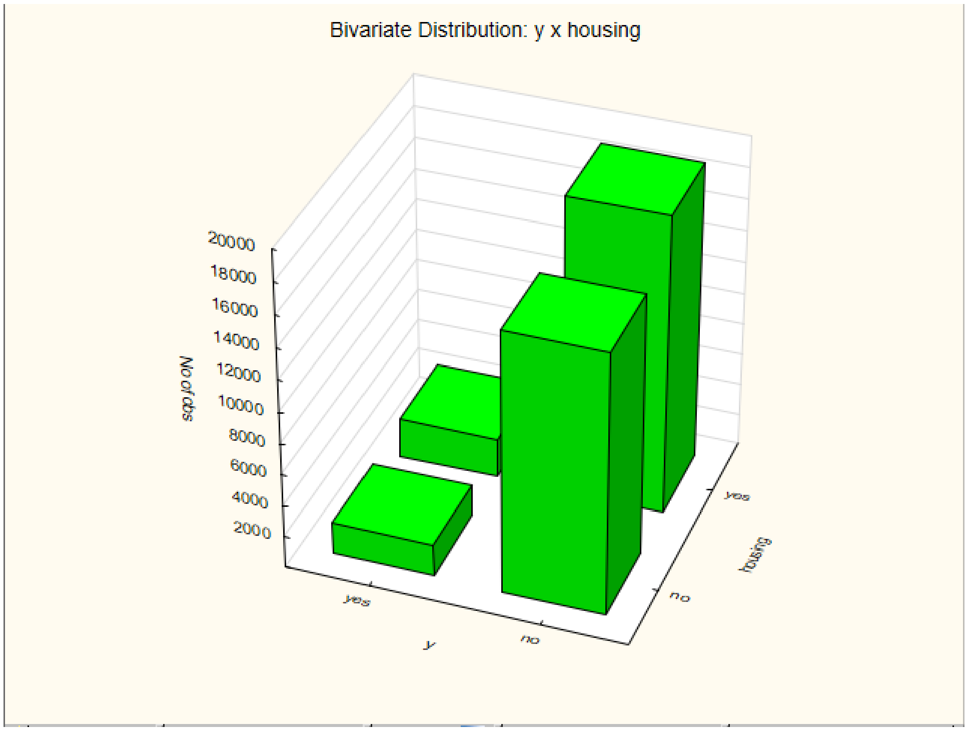
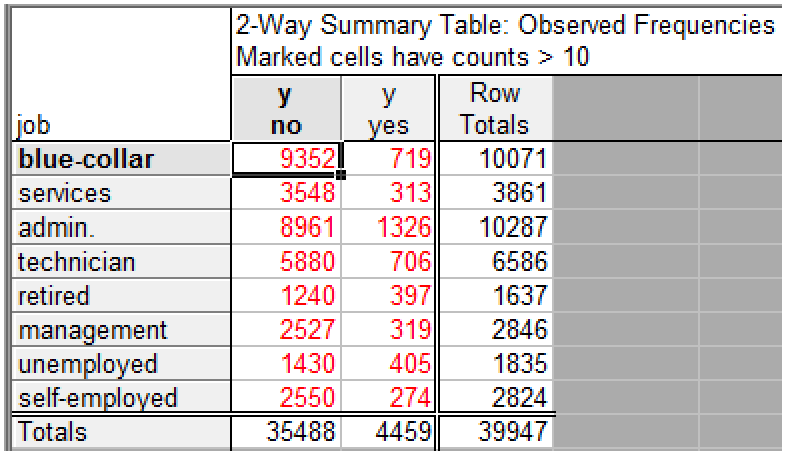


Figure 24. 3D bivariate bar chart for variables "housing" and "y"
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Next variable is occupation or type of job and results are shown in tables and graph below.

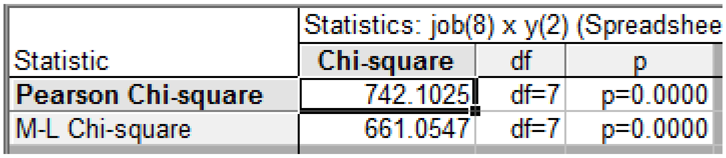


2-Way Summary Table: Observed Frequencies
Marked cells have counts > 10

job	y no	y yes	Row Totals
blue-collar	9352	719	10071
services	3548	313	3861
admin.	8961	1326	10287
technician	5880	706	6586
retired	1240	397	1637
management	2527	319	2846
unemployed	1430	405	1835
self-employed	2550	274	2824
Totals	35488	4459	39947

Table 8. 2-Way summary table for variables "y' and "job"
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Lower paid jobs have less chances of subscribing a bank term deposit than better paid jobs. The highest percentage of subscribing a bank term deposit have retired people. And that can be explained by the fact that retired people generally tend to save money.



Statistics: job(8) x y(2) (Spreadshee

Statistic	**Chi-square**	df	p
Pearson Chi-square	742.1025	df=7	p=0.0000
M-L Chi-square	661.0547	df=7	p=0.0000

Table 9. Chi-square test for variables "y' and "job"
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

P-value also indicates that this variable is good for predicting dependent variable.

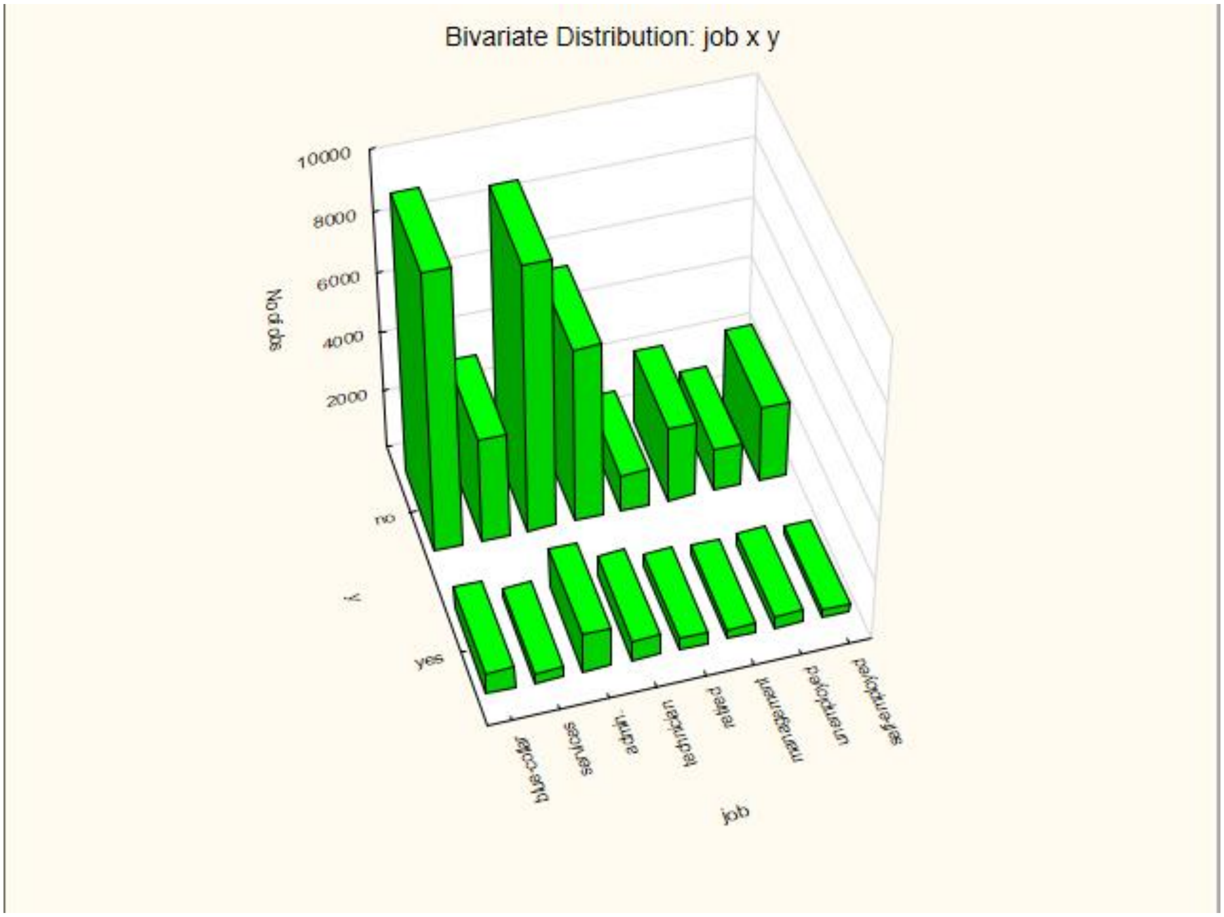
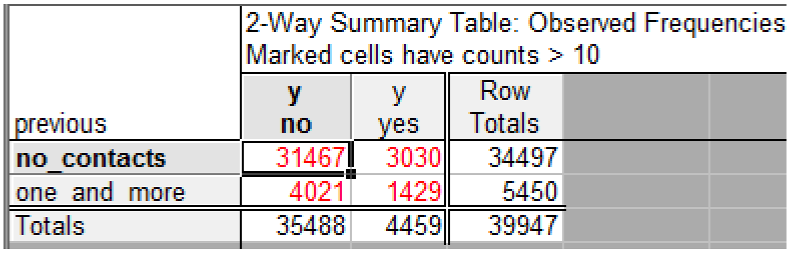


Figure 25. 3D bivariate bar chart of variables "job" and "y"
Source: Data from (Moro et al., 2014) STATISTICA output own processing

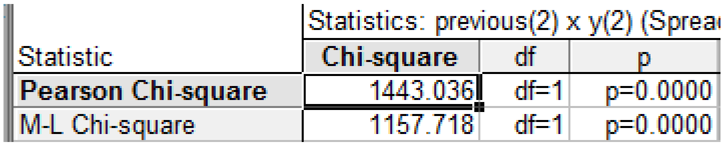
For variable "previous" which represents number of previous contacts in previous campaigns results show that if there was contact in previous campaigns there are bigger chances of success in this campaign as well.



	2-Way Summary Table: Observed Frequencies Marked cells have counts > 10		
previous	y no	y yes	Row Totals
no_contacts	31467	3030	34497
one and more	4021	1429	5450
Totals	35488	4459	39947

Table 10. 2-Way summary table for variables "y' and "previous"
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

And p-value is also suggesting that this variable should be very good predictor variable for predictive model.



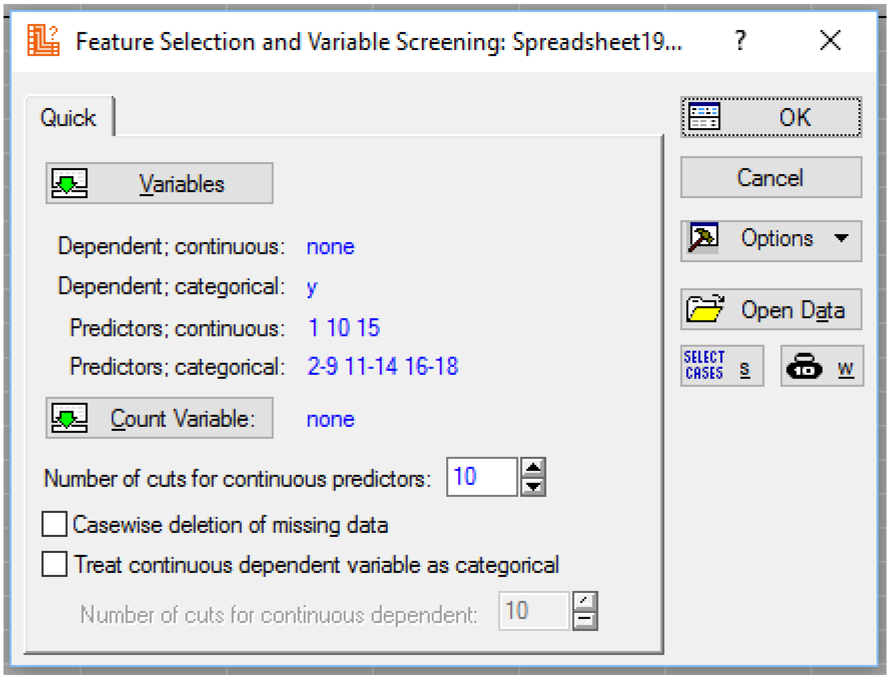
	Statistics: previous(2) x y(2) (Sprea		
Statistic	Chi-square	df	p
Pearson Chi-square	1443.036	df=1	p=0.0000
M-L Chi-square	1157.718	df=1	p=0.0000

Table 11. Chi-square test for variables "y' and "previous"
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

For all other variables outputs of analysis can be seen in appendices. In next chapter of thesis feature selection tool will be used to select best predictors for making predictive model.

4.6. Choosing best predictors – Feature selection tool

Feature selection and Variable screening tool is data mining tool in STATISTICA software which is used for finding best predictors for dependent variable.

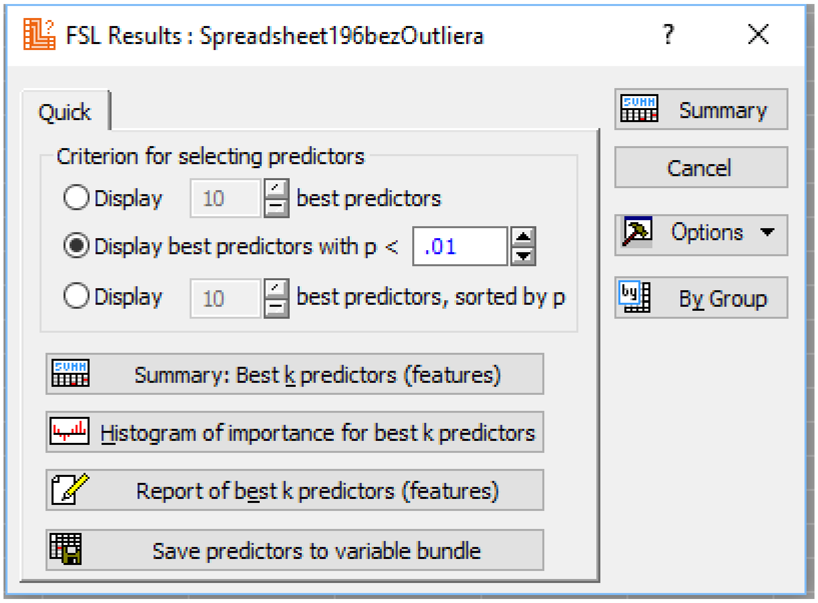


Feature Selection and Variable Screening: Spreadsheet19...

Dependent; continuous: none
Dependent; categorical: y
Predictors; continuous: 1 10 15
Predictors; categorical: 2-9 11-14 16-18
Count Variable: none
Number of cuts for continuous predictors: 10
☐ Casewise deletion of missing data
☐ Treat continuous dependent variable as categorical
Number of cuts for continuous dependent: 10

Figure 26. Feature selection and variable screening tool in STATISTICA
Source: Data from (Moro et al., 2014) STATISTICA output own processing

First step is to choose dependent categorical (or continuous) variable and then to select adequate continuous and categorical predictors. After that, feature selection tool allows analyst to choose criterion of selecting predictors. As it can be seen on picture below criterions are: best 10 (optionally more or less predictors), best predictors with p-value less than 0.01 (also optional and it can be changed) and third option is displaying certain number of best predictors sorted by p-value. There is option for making bar charts of best predictors and also saving bundles of best predictors both continuous and categorical which can help later for quick selection of these predictors when time for creating models come.



FSL Results : Spreadsheet196bezOutliera

Quick

Criterion for selecting predictors

○ Display 10 best predictors

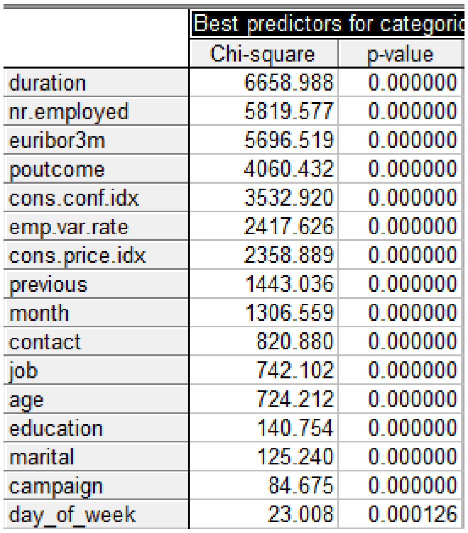
● Display best predictors with p < .01

○ Display 10 best predictors, sorted by p

?
Summary
Cancel
Options
By Group
Summary: Best k predictors (features)
Histogram of importance for best k predictors
Report of best k predictors (features)
Save predictors to variable bundle

Figure 27. Feature selection tool options
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Results from choosing best predictors with p-value less than 0.01 can be seen in table below.



	Best predictors for categoric	
	Chi-square	p-value
duration	6658.988	0.000000
nr.employed	5819.577	0.000000
euribor3m	5696.519	0.000000
poutcome	4060.432	0.000000
cons.conf.idx	3532.920	0.000000
emp.var.rate	2417.626	0.000000
cons.price.idx	2358.889	0.000000
previous	1443.036	0.000000
month	1306.559	0.000000
contact	820.880	0.000000
job	742.102	0.000000
age	724.212	0.000000
education	140.754	0.000000
marital	125.240	0.000000
campaign	84.675	0.000000
day_of_week	23.008	0.000126

Table 12. Table of best predictors with p-value less than 0.01
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Out of 18 possible predictors 16 had p-value less than 0.01 and the best predictor is variable “duration”. Variables “loan” and “housing” are not chosen to be predictors for this model. These group of predictors will be tested for creating model.

Also, group of best 10 predictors will be tested to compare results with models made on 16 predictors. Best 10 predictors sorted by Chi-square are shown in bar chart below.

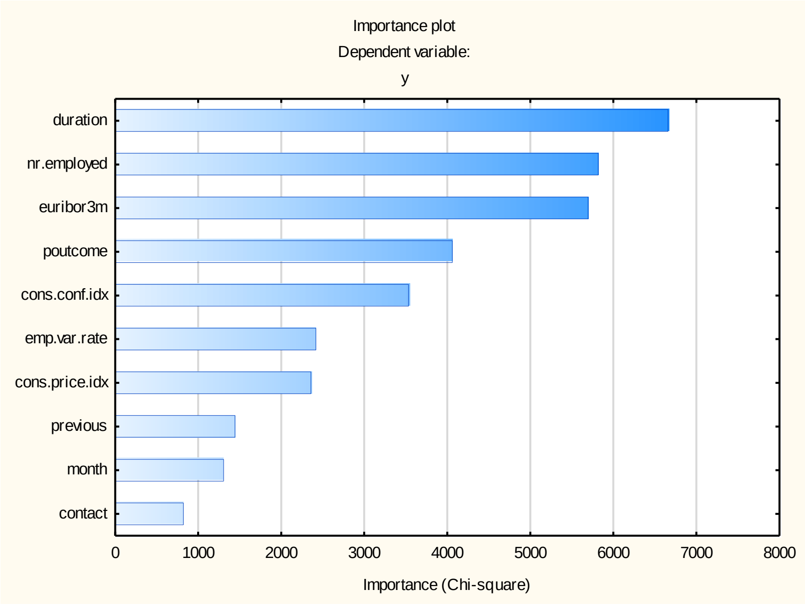


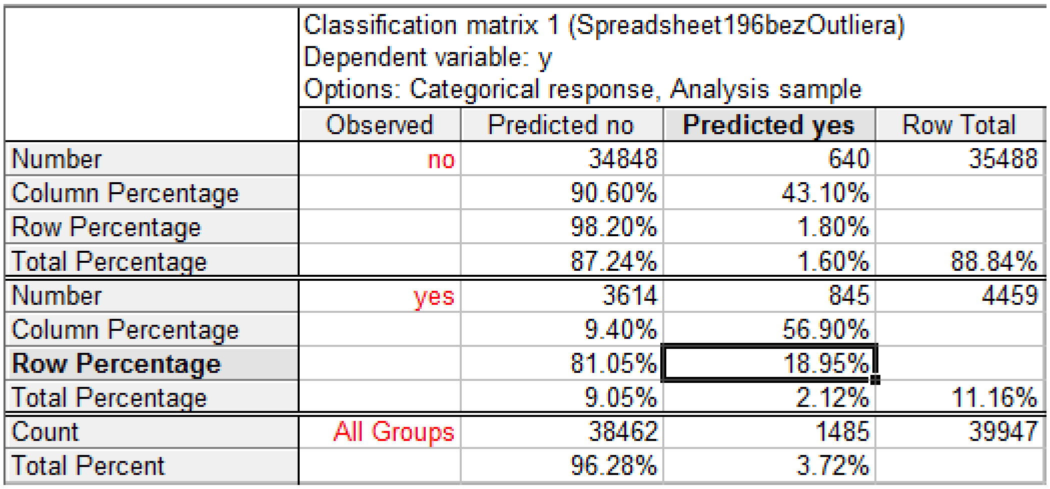
Figure 28. Best ten predictors, importance by Chi-square
Source: Data from (Moro et al., 2014) STATISTICA output own processing

4.7. Building predictive models

4.7.1. CART (Classification and Regression Trees)

In this chapter couple of predicted models will be made so it can be decided which one suits best for this data set. Since data set contains most of categorical variables with only three continuous variables algorithms for classification will be used. First one to be tested is CART (Classification and Regression Tree).

Results can be seen in table below. And from results it can be seen that model showed really good accuracy in predicting "no" but very low percentage in predicting "yes". This can be influenced by the fact that more than 80% of cases had observed "no".



Classification matrix 1 (Spreadsheet196bezOutliera)
Dependent variable: y
Options: Categorical response, Analysis sample

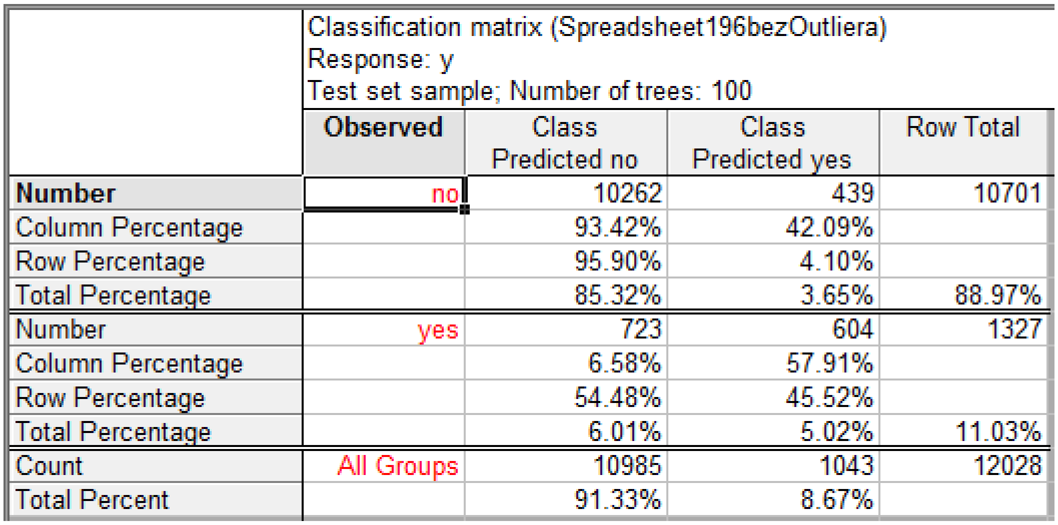
	Observed	Predicted no	**Predicted yes**	Row Total
Number	no	34848	640	35488
Column Percentage		90.60%	43.10%	
Row Percentage		98.20%	1.80%	
Total Percentage		87.24%	1.60%	88.84%
Number	yes	3614	845	4459
Column Percentage		9.40%	56.90%	
Row Percentage		81.05%	18.95%	
Total Percentage		9.05%	2.12%	11.16%
Count	All Groups	38462	1485	39947
Total Percent		96.28%	3.72%	

Table 13. Classification matrix for CART model prediction
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Very similar results were gotten by using only best 10 predictors so this model is maybe not the best choice for this kind of problem. Next step will be to test random forest.

4.7.2. Random Forest

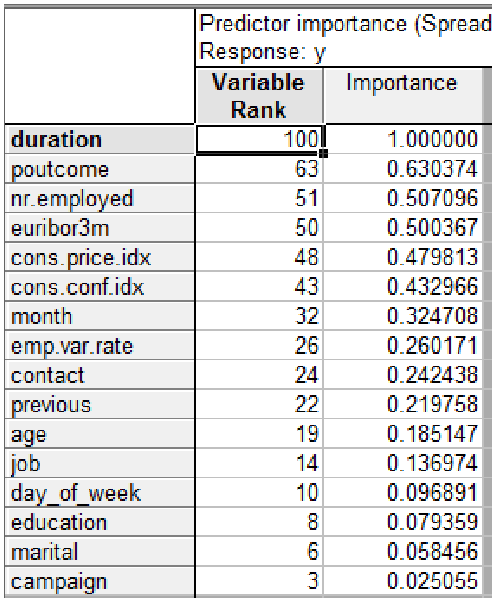
Random forest with 16 predictors and 100 trees achieved significantly better results in predicting "yes" with accuracy of 45.52% and in predicting "no" accuracy was 95.9% which is really good result. Results can be seen in table below. Adding number of trees didn't give any better results.



Classification matrix (Spreadsheet196bezOutliera)
Response: y
Test set sample; Number of trees: 100

	Observed	Class Predicted no	Class Predicted yes	Row Total
Number	no	10262	439	10701
Column Percentage		93.42%	42.09%	
Row Percentage		95.90%	4.10%	
Total Percentage		85.32%	3.65%	88.97%
Number	yes	723	604	1327
Column Percentage		6.58%	57.91%	
Row Percentage		54.48%	45.52%	
Total Percentage		6.01%	5.02%	11.03%
Count	All Groups	10985	1043	12028
Total Percent		91.33%	8.67%	

Table 13. Classification matrix for Random Forest model prediction
Source: Data from (Moro et al., 2014) STATISTICA output, own processing



Predictor importance (Spread
Response: y

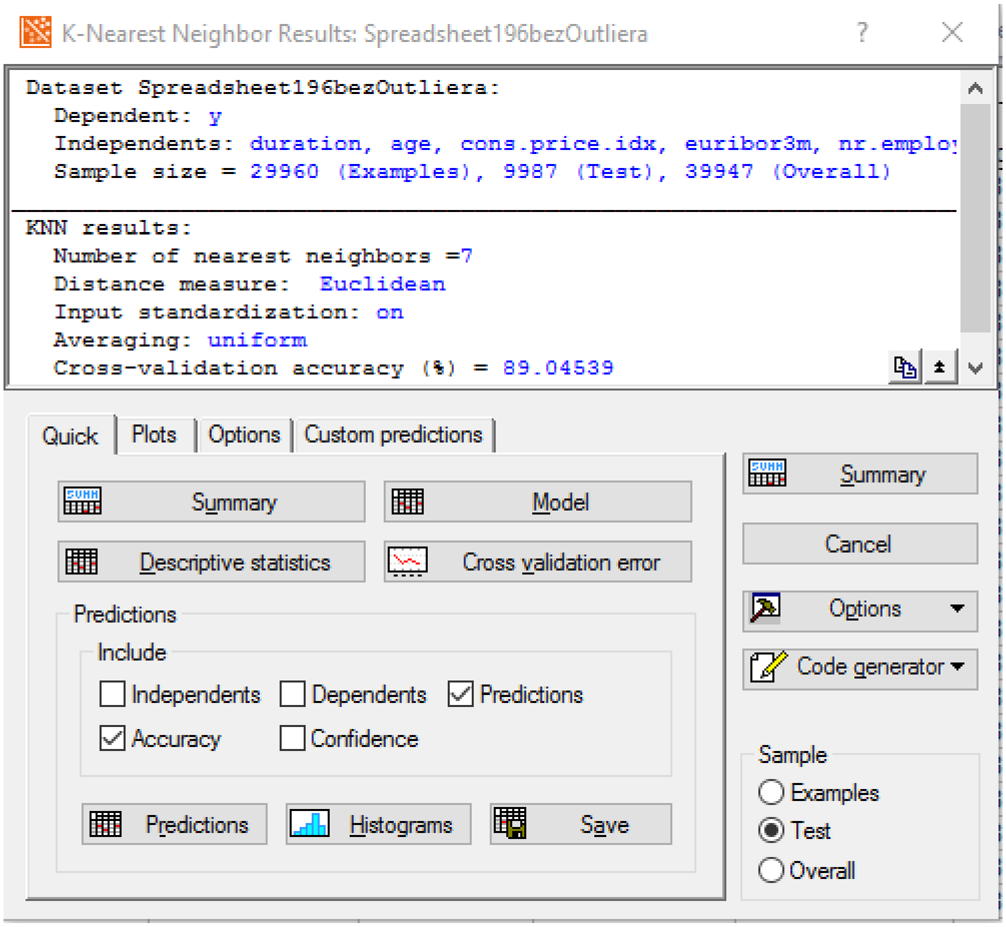
	Variable Rank	Importance
duration	100	1.000000
poutcome	63	0.630374
nr.employed	51	0.507096
euribor3m	50	0.500367
cons.price.idx	48	0.479813
cons.conf.idx	43	0.432966
month	32	0.324708
emp.var.rate	26	0.260171
contact	24	0.242438
previous	22	0.219758
age	19	0.185147
job	14	0.136974
day_of_week	10	0.096891
education	8	0.079359
marital	6	0.058456
campaign	3	0.025055

Table 14. Predictor importance for Random Forest model
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

In Table 14. are given the best predictors for Random forest model and the best predictor is variable “duration” while the second best is variable is “poutcome” (Outcome of previous campaigns) which is different from the list of best predictors used for all models.

4.7.3. K-Nearest Neighbor

Next model which will be tested is K-Nearest neighbor model. Results are shown in graphs and tables below.



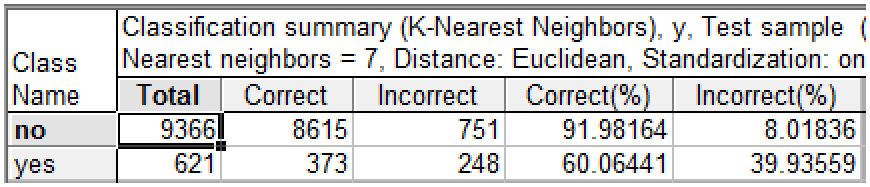
```
Dataset Spreadsheet196bezOutliera:
  Dependent: y
  Independents: duration, age, cons.price.idx, euribor3m, nr.emplo
  Sample size = 29960 (Examples), 9987 (Test), 39947 (Overall)

KNN results:
  Number of nearest neighbors =7
  Distance measure:  Euclidean
  Input standardization: on
  Averaging: uniform
  Cross-validation accuracy (%) = 89.04539
```

☐ Independents ☐ Dependents ☑ Predictions
☑ Accuracy ☐ Confidence

○ Examples
● Test
○ Overall

Figure 29. K-Nearest neighbor results
Source: Data from (Moro et al., 2014) STATISTICA output own processing

K-Nearest Neighbor with 16 predictors and number of k=7 achieved the best results from all tested models so far for predicting “no”. Though misclassification matrix shows that predicting "yes" when it was observed was in around 40% of cases which makes it similar to Random Forest results while predicting "no" gave slightly better results than Random Forest. This model as well as Random Forest outperformed CART model. Classification summary in table below shows correct classification for "no" in 91.98% cases and 60.06% for "yes".



Class Name	Classification summary (K-Nearest Neighbors), y, Test sample (Nearest neighbors = 7, Distance: Euclidean, Standardization: on				
	Total	Correct	Incorrect	Correct(%)	Incorrect(%)
no	9366	8615	751	91.98164	8.01836
yes	621	373	248	60.06441	39.93559

Table 15. Classification summary for k-nearest neighbor model
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

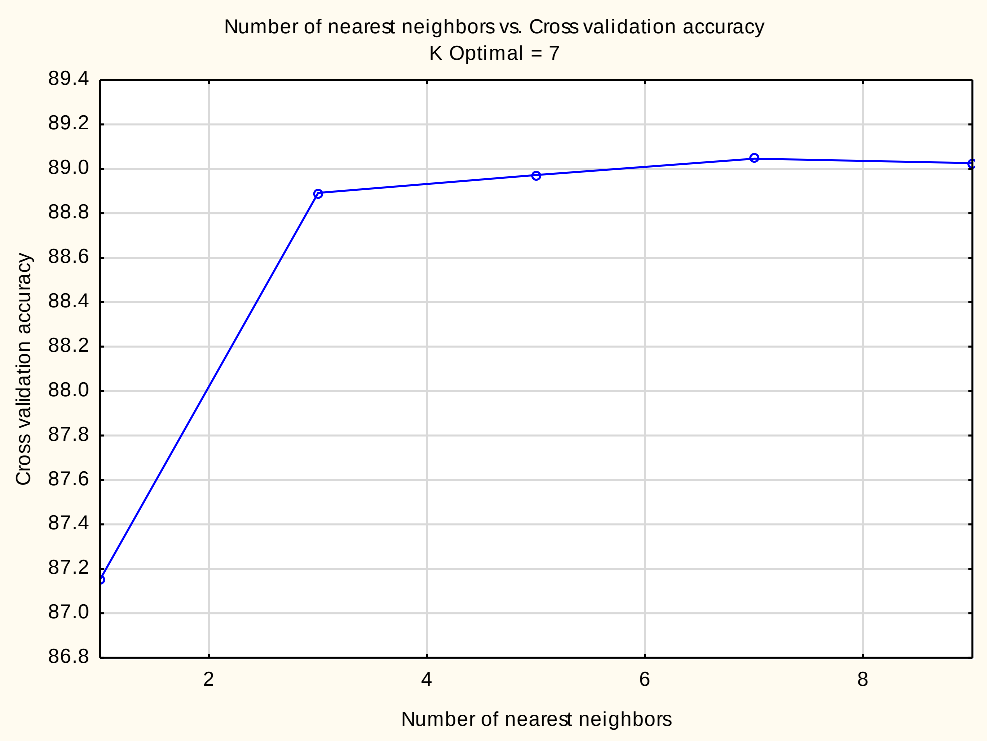
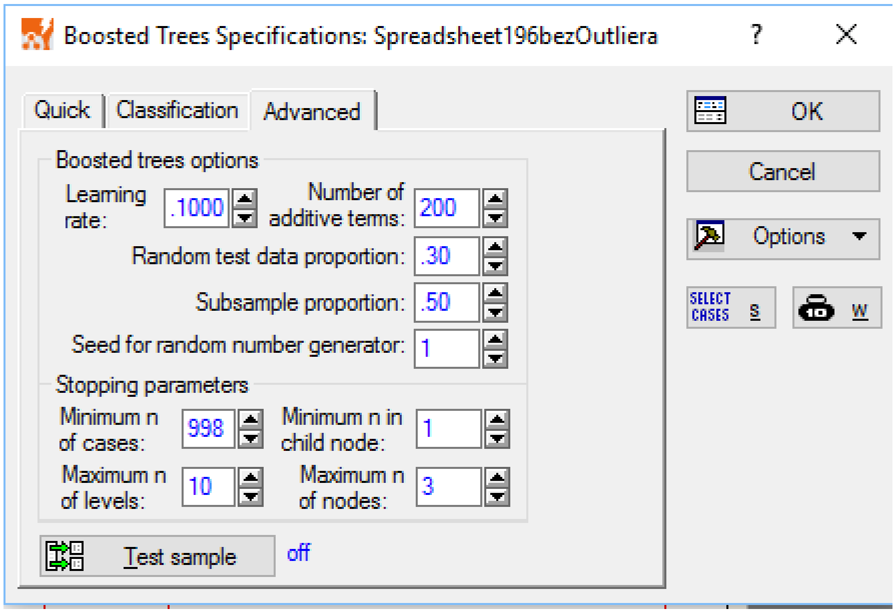


Figure 30. Number of nearest neighbors vs. cross validation accuracy
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Optimal number of k is 7, and with this number of k result was the best. Since it is recommended for binary outcome variables to use odd number of values for "k", which is already explained in theoretical part of thesis, possible values for k were 1,3,5,7 or more and the best result is achieved with k=7. This can be seen in figure 30. which is showing how cross validation accuracy is changing with growing number of neighbors.

4.7.4. Boosted trees

Last one to be tested will be Boosted trees. Boosted trees use weak predictors together in order to make strong predictors and prediction comes as classification of simple trees as whole. The whole idea of boosting is to use weak and simple trees together to make good predictions.



Boosted Trees Specifications: Spreadsheet196bezOutliera

Boosted trees options

Learning rate: .1000

Number of additive terms: 200

Random test data proportion: .30

Subsample proportion: .50

Seed for random number generator: 1

Stopping parameters

Minimum n of cases: 998

Minimum n in child node: 1

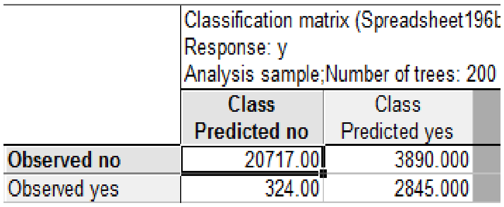
Maximum n of levels: 10

Maximum n of nodes: 3

Test sample off

Figure 31. Boosted trees specifications
Source: Data from (Moro et al., 2014) STATISTICA output own processing

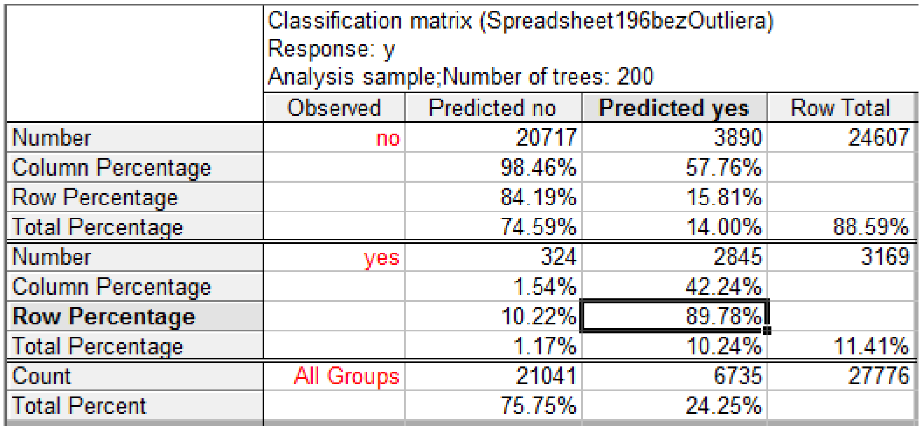
Learning rate is set to 0.1 by default because it's proven that with learning rate of 0.1 or less Boosted trees achieve the best results. Maximum number of nodes is 3 because of binary outcome.



Classification matrix (Spreadsheet196b
Response: y
Analysis sample;Number of trees: 200

	Class Predicted no	Class Predicted yes
Observed no	20717.00	3890.000
Observed yes	324.00	2845.000

Table 16. Classification matrix for Boosted Trees model
Source: Data from (Moro et al., 2014) STATISTICA output, own processing



Classification matrix (Spreadsheet196bezOutliera)
Response: y
Analysis sample;Number of trees: 200

	Observed	Predicted no	Predicted yes	Row Total
Number	no	20717	3890	24607
Column Percentage		98.46%	57.76%	
Row Percentage		84.19%	15.81%	
Total Percentage		74.59%	14.00%	88.59%
Number	yes	324	2845	3169
Column Percentage		1.54%	42.24%	
Row Percentage		10.22%	89.78%	
Total Percentage		1.17%	10.24%	11.41%
Count	All Groups	21041	6735	27776
Total Percent		75.75%	24.25%	

Table 16. Classification matrix for Boosted Trees model with percentages
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

It the table it can be seen that Boosted trees achieved the best results so far with correct prediction for “yes” of 89.78%. but had the worst results for predicting “no” with 84.19% accuracy. In graphs below simple tree structure for “no” and “yes” is shown.

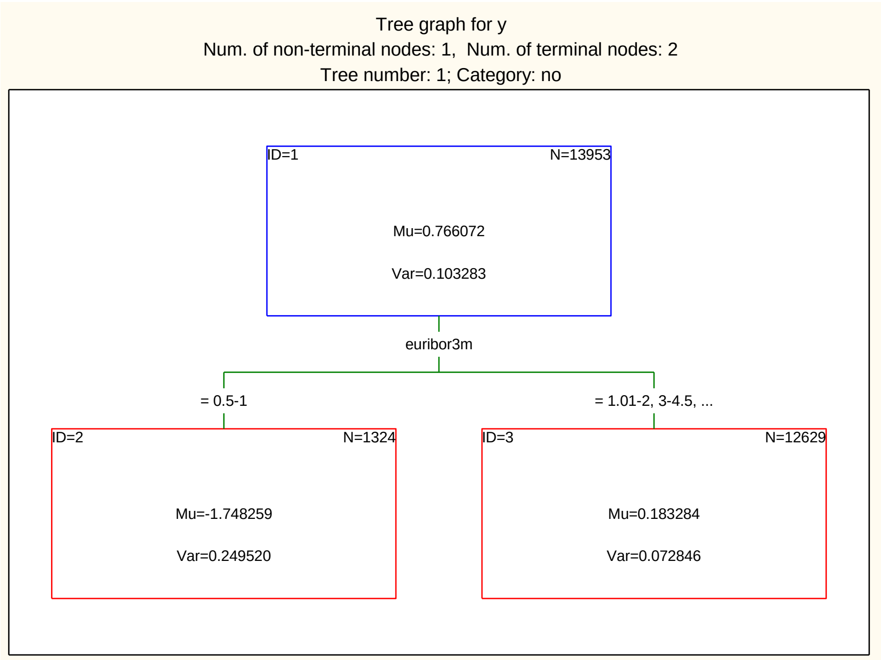


Figure 32. Boosted trees tree nr. 1 graph for category “no”
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Tree number 1 for category “no” uses “euribor.3m” variable for the split while tree number 1 for “yes” category uses “duration” for the first split.

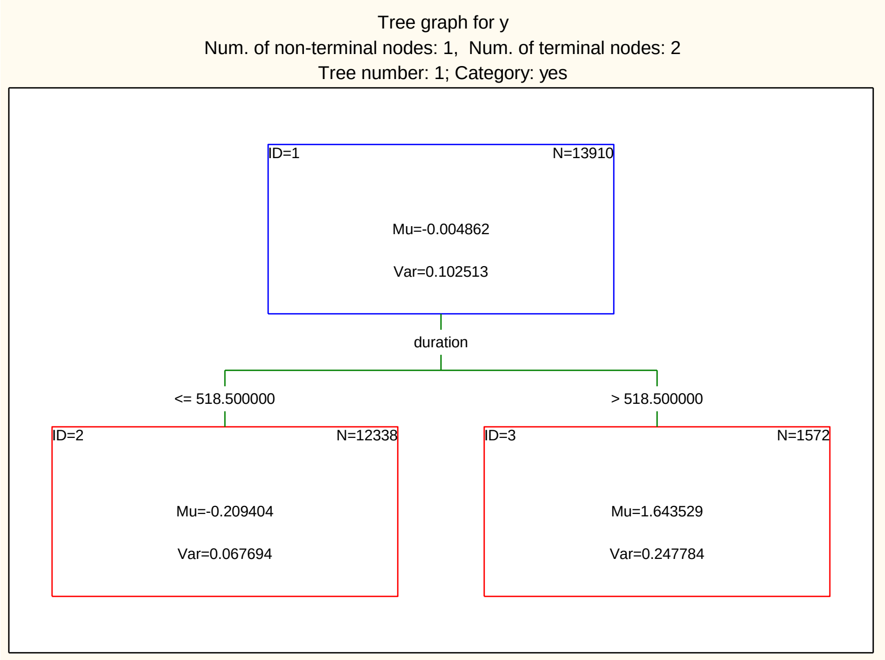
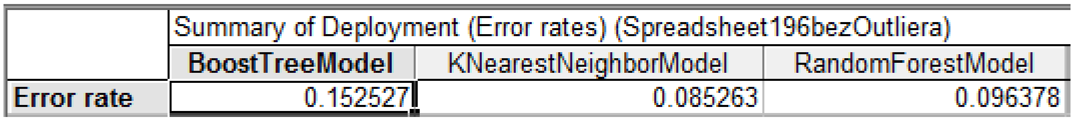


Figure 33. Boosted trees tree nr. 1 graph for category “yes”
Source: Data from (Moro et al., 2014) STATISTICA output own processing

5. Results and discussion

5.1. Comparison of models

Rapid deployment tool is used to compare results obtained from models tested so far. To use rapid deployment tool, it is necessary to create PMML (The Predictive Model Markup Language) codes from all models used. After saving PMML codes, those codes were loaded into rapid deployment tool. Rapid deployment tool shows overall prediction accuracy for all models, error rate for all models and accuracy for predicting “yes” and “no”. It also calculates voted predictions by comparing results obtained by loaded models. Because of CART’s very low accuracy in predicting “yes” and insufficient processing power of computer only Random Forest, K-Nearest Neighbor and Boosted Trees models will be compared since they achieved significantly better results.

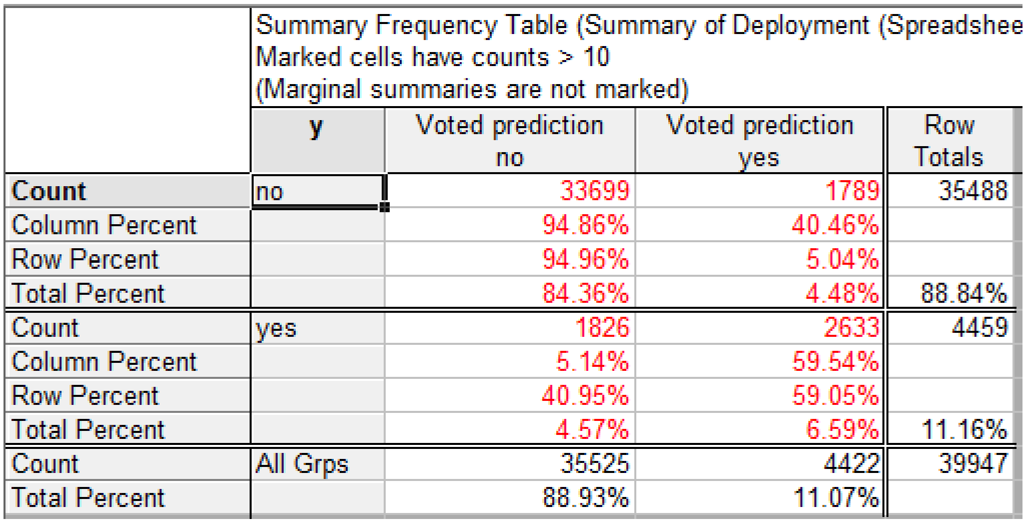


	Summary of Deployment (Error rates) (Spreadsheet196bezOutliera)		
	BoostTreeModel	KNearestNeighborModel	RandomForestModel
Error rate	0.152527	0.085263	0.096378

Table 17. Error rates for three best models
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

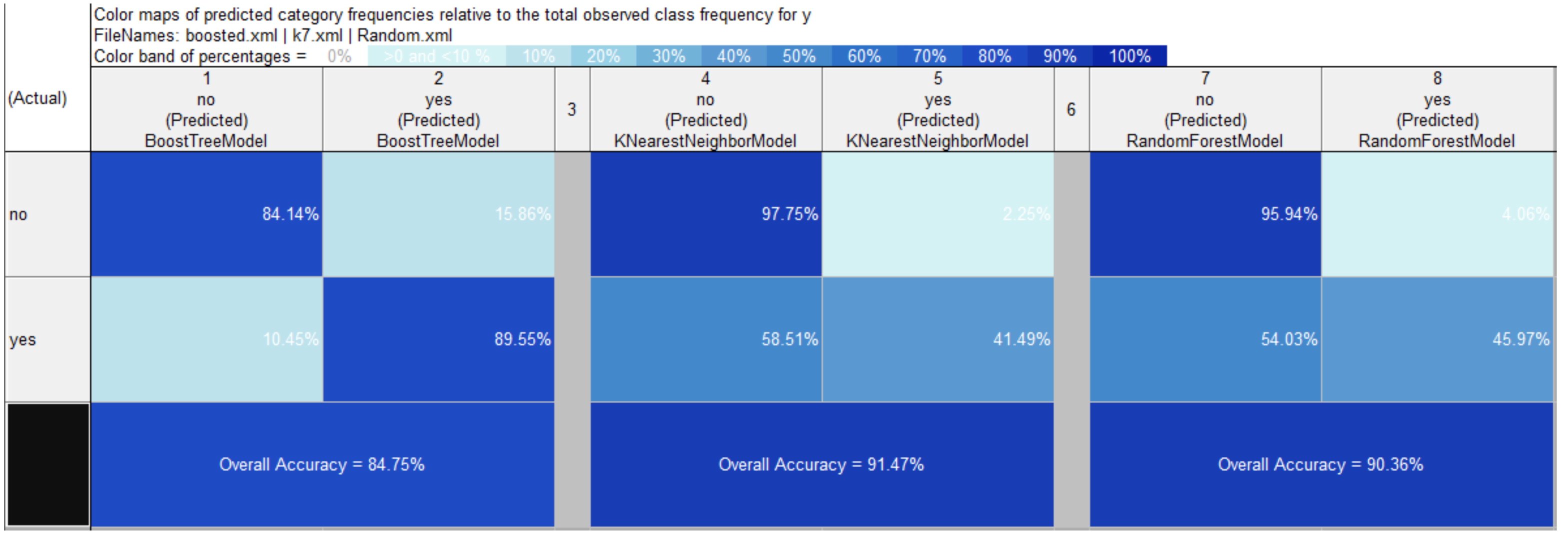
According to results from table the best error rate has K-Nearest Neighbor model and the worst, biggest error rate is for Boosted Trees model. On next page in Figure 33. its shown overall accuracy of all three models and K-Nearest Neighbor model had the best overall accuracy of 91.47%, second best was Random Forest model with overall accuracy of 90.36% while Boosted Trees model had the worst overall accuracy even though it achieved by far the best accuracy with predicting “yes”.

Voted predictions from all three models compared to observed values gave result of 94.96 % for predicting “no” and 59.05% of predicting “yes”. Those results can be seen in the table 18. below.



	Summary Frequency Table (Summary of Deployment (Spreadshee Marked cells have counts > 10 (Marginal summaries are not marked)			
	y	Voted prediction no	Voted prediction yes	Row Totals
Count	no	33699	1789	35488
Column Percent		94.86%	40.46%	
Row Percent		94.96%	5.04%	
Total Percent		84.36%	4.48%	88.84%
Count	yes	1826	2633	4459
Column Percent		5.14%	59.54%	
Row Percent		40.95%	59.05%	
Total Percent		4.57%	6.59%	11.16%
Count	All Grps	35525	4422	39947
Total Percent		88.93%	11.07%	

Table 18. Summary frequency table for observed values and voted prediction
Source: Data from (Moro et al., 2014) STATISTICA output, own processing



Color maps of predicted category frequencies relative to the total observed class frequency for y
FileNames: boosted.xml | k7.xml | Random.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	3	4 no (Predicted) KNearestNeighborModel	5 yes (Predicted) KNearestNeighborModel	6	7 no (Predicted) RandomForestModel	8 yes (Predicted) RandomForestModel
no	84.14%	15.86%		97.75%	2.25%		95.94%	4.06%
yes	10.45%	89.55%		58.51%	41.49%		54.03%	45.97%
	Overall Accuracy = 84.75%			Overall Accuracy = 91.47%			Overall Accuracy = 90.36%	

Figure 33. Overall accuracy for all three models
Source: Data from (Moro et al., 2014) STATISTICA output own processing

In addition to this results, all three models were tested on stratified random sample, strata variable was outcome variable “y”. In stratified random sample were similar amounts of both possible outcomes “yes” and “no” (Figure 34) and the same three models were tested so we can see how much results will change and differ from results obtained from testing models on original dataset.

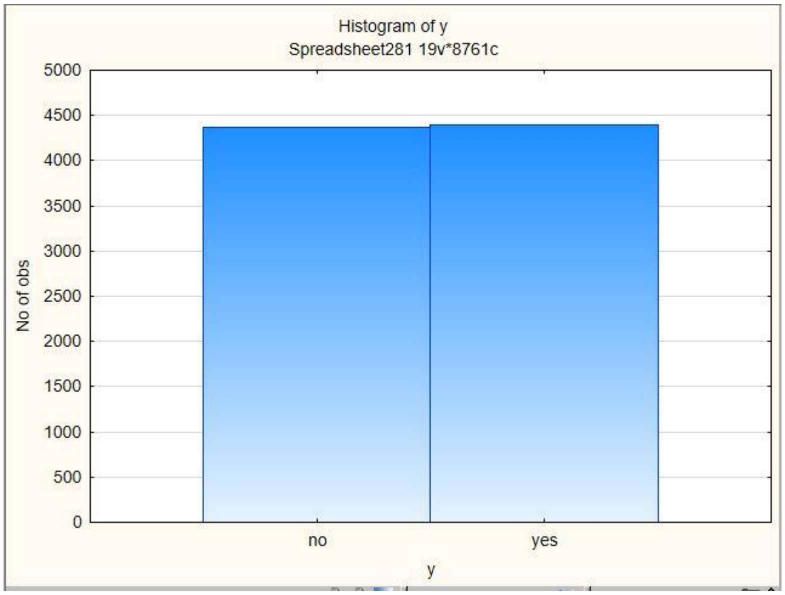
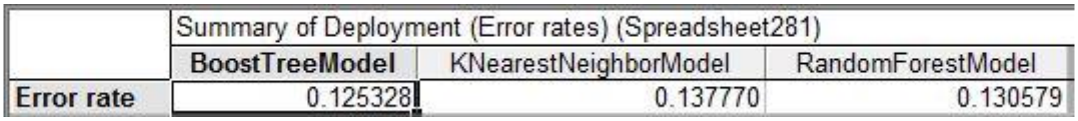


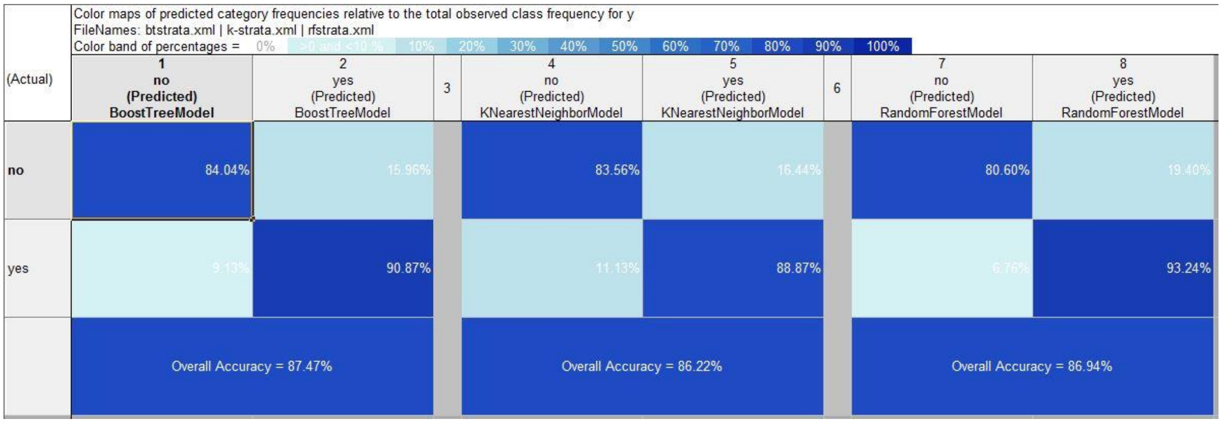
Figure 34. Bar chart of outcome variable Y (stratified random sample)
Source: Data from (Moro et al., 2014) STATISTICA output own processing

Overall accuracy and error rates changed and the best result in both having best overall accuracy and the lowest error rate was achieved by the Boosted trees model, second best result was achieved by Random Forest model and the worst result was obtained by K-Nearest Neighbor model. Results can be seen in Figure 35. and Table 19. below.



	Summary of Deployment (Error rates) (Spreadsheet281)		
	BoostTreeModel	KNearestNeighborModel	RandomForestModel
Error rate	0.125328	0.137770	0.130579

Table 19. Error rates for three best models (stratified random sample)
Source: Data from (Moro et al., 2014) STATISTICA output, own processing



Color maps of predicted category frequencies relative to the total observed class frequency for y
FileNames: btstrata.xml | k-strata.xml | rfstrata.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	3	4 no (Predicted) KNearestNeighborModel	5 yes (Predicted) KNearestNeighborModel	6	7 no (Predicted) RandomForestModel	8 yes (Predicted) RandomForestModel
no	84.04%	15.96%		83.56%	16.44%		80.60%	19.40%
yes	9.13%	90.87%		11.13%	88.87%		6.76%	93.24%
	Overall Accuracy = 87.47%			Overall Accuracy = 86.22%			Overall Accuracy = 86.94%	

Figure 35. Overall accuracy for all three models (stratified random sample)
Source: Data from (Moro et al., 2014) STATISTICA output own processing

As it can be seen, overall accuracy of Boosted Trees model was 87.47% with accuracy of predicting “no” of 84.04% and predicting “yes” accurately in 90.87% percent of cases. Error rate was 0.125328. Second best overall result was achieved by Random Forest model with overall accuracy of 86.94 % with accurately predicting “no” in 80.60% percent of cases where “no” was observed and “yes” in 93.24% of cases where it was observed with error rate of 0.130579. The worst results were achieved by K-Nearest Neighbor model with overall accuracy of 86.22%, accurately predicting “no” in 83.56% of cases and “yes” in 88.87% of cases with error rate slightly worse than Random Forest model 0.137770. Predicting “yes” when it was rare occasion was a problem for Random Forest model and K-Nearest model and in stratified random sample they achieved much better results, Random Forest model even outperformed Boosted Trees model in accurately predicting “yes”. But, Boosted Trees model achieved the best overall accuracy (and lowest error rate) in stratified random sample and also gave the closest results to ones achieved on original dataset which makes it the best model for this kind of problem because two other models showed to be highly affected by number of observed values.

6. Conclusion

In this thesis all phases of CRISP-DM process were thoroughly theoretically explained and later on practically applied. Data preparation\cleaning phase took most of the time in process of creating predictive models. A lot of different methods for recoding outliers, dealing with missing data and especially transforming and recoding variables were applied and each of them gave different results, some were similar to final results some very different. As, mentioned, CRISP-DM process has six phases but those phases are highly dependent on outcome of previous phases and returning to data preparation phase and after that to data exploration phase was essential until acceptable results were achieved.

Throughout reading theory and learning about predictive modeling it became more and more clear why most of authors agree that data mining process is mixture of statistical rules and analysis, which have to be applied, and creativity and personal experience from analyst which make the whole process of creating appropriate model way easier. Making predictive model, as we could see, mostly depends on quality of data and how well data cleaning process was done. Data cleaning process and generally whole process of creating predictive model depends in high percentage of predictive modelers ability to apply adequate combination of data preparation and transformation techniques, which mostly comes with experience. Yes, theoretical knowledge is essential and necessary but without proper experience and creativity to follow the theory creating predictive model can be very, very hard.

In addition to that, also choosing appropriate model depends on a lot of things not only on statistics. Choosing the right model is highly related to business objectives and what is the actual purpose of data mining project. The good example of that are results obtained from this thesis. Three different types of models gave different results with different prediction accuracy and the decision which one will be used is not that simple. Even though best overall performance was achieved by K-Nearest Neighbor model and it also had lowest error rate it didn’t give the best result for predicting “yes” which can be crucial for bank management. The second overall best model, Random Forest, had the second lowest error but again like K-Nearest Neighbor model failed in predicting “yes”. Bank loses potential money in predicting “no” for those who would subscribe a bank term deposit. Predicting “yes” where it was observed “no” is also not good but in terms of profit it’s much more dangerous predicting “no” where it should be “yes”. And the best job in predicting “yes” was done by Boosted Trees model which had third best overall accuracy and the biggest error rate and also the worst accuracy when it comes to predicting “no”. So, in this specific situation where predicting “yes” is really important, because losing potential money is not good for any commercial organization but especially for banks, suggestion would be to use Boosted Trees model. In some other situations it would be probably smarter to use models with better overall accuracy. Also Boosted trees model showed the best performance in predicting rare occasion, because data set contained more than 85% of cases with observed value of “no”. Maybe data can be transformed differently and some models can be tested again but in this situation where all three models had a fairly good overall accuracy suggestion would be to use Boosted Trees model because the other two failed in predicting category “yes” accurately. Statistics wise it would be best to use either K-Nearest Neighbor model or Random Forest model but business wise decision falls on Boosted Trees.

Information sources:

1. Abbott, D. *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst.* USA, NJ, Somerset: Wiley, 2014. ISBN 978-1-118-72793-5.
2. Agresti, A. *Categorical data analysis.* Hoboken: John Wiley & Sons, 2013. ISBN 978-0-470-46363-5.
3. Atzmueller, M. & Oussena, S. & Roth-Berghofer, T. *Enterprise Big Data Engineering, Analytics, and Management.* United States: IGI Global, 2016. ISBN:9781522502937
4. Bari, A. & Chaouchi, M. & Jung, T. *Predictive analytics for dummies.* Hoboken: John Wiley & Sons, 2014. ISBN:9781118728963
5. Batista, G. & Monard, M C. *A Study of K-Nearest Neighbor as an Imputation Method*. University of Sao Paulo - USP
6. Brown, S M. *Data mining for dummies.* Hoboken: John Wiley & Sons, 2014. ISBN:9781118893173
7. Cichosz, P. *Data Mining Algorithms: Explained Using R.* John Wiley & Sons, Ltd, 2015. ISBN:9781118332580.
8. Dean, J 2014, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley, 2014. ISBN: 978-1-118-61804-2
9. Finlay, S. *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods.* 1st pub. New York; Basingstoke; Palgrave Macmillan, 2014. ISBN 9781137379276
10. Larose, D T. *Discovering knowledge in data: An introduction to data mining.* Hoboken, N.J.: Wiley-Interscience, 2005. ISBN 0471666572.
11. King, R. S. *Cluster Analysis and Data Mining—An Introduction.* Dulles, Virginia; Mercury learning and information, 2015. ISBN: 978-1-938549-38-0
12. Larose, D & Larose C, Discovering Knowledge in Data: An Introduction to Data Mining, 2nd Edition. Hoboken, N.J.: Wiley, 2014. ISBN: 978-0-470-90874-7
13. Linoff, G S. & Berry, M J A. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (3)*. Wiley, Hoboken, US, 2011.

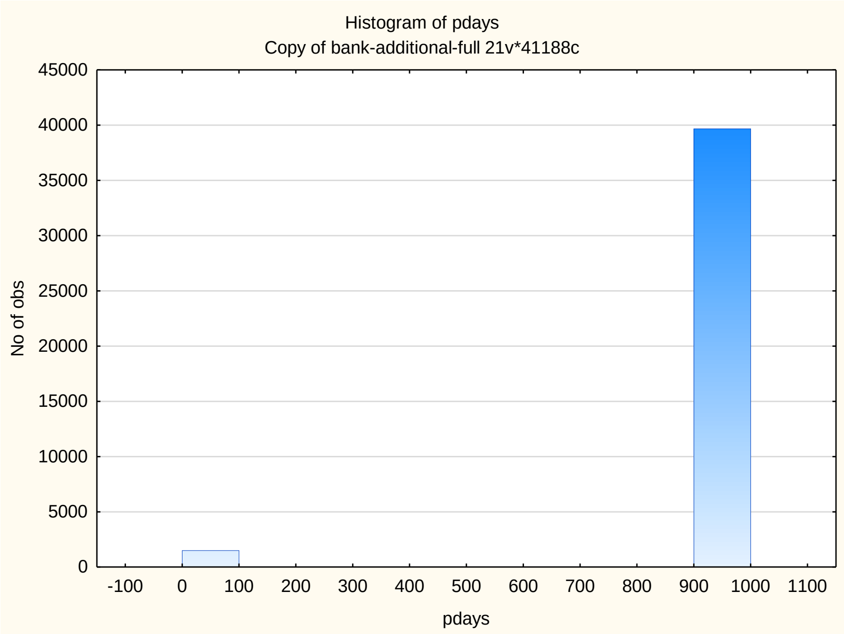
14. [Moro et al., 2011] Moro, S. & Laureano, R. & Cortez, P. *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.* In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.
15. [Moro et al., 2014] Moro, S. & Cortez, P. & Rita, P. *A Data-Driven Approach to Predict the Success of Bank Telemarketing.* Decision Support Systems, Elsevier, 2014. Available from: http://www.sciencedirect.com/science/article/pii/S016792361400061X
16. Myatt, J G. & Johnson, P W., *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Second Edition, John Wiley & Sons, 2014. ISBN:9781118407417
17. Nisbet, R. & Elder, J. & Miner, G. *Handbook of statistical analysis and data mining applications [online]*. Burlington: Academic Press, 2009. ISBN 978-0-12-374765-5.
18. Olson, D. L. *Data mining models*. Business Expert Press, 2016. ISBN:9781631575488
19. Piegorsch, W. *Statistical Data Analytics—Foundations for Data Mining, Informatics, and Knowledge Discovery*. John Wiley & Sons, Ltd, 2015. ISBN: 9781118619650
20. Ratner, B. *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, Second Edition.* Taylor and Francis Group, LLC, 2012. ISBN:9781439860915
21. Robson, L. F. C & Christos F. & Caetano T. Jr. *Data Mining in Large Sets of Complex Data [e-book]*, Springer, 2013, ISBN 978-1-4471-4890-6
22. Siegel, E. *Predictive Analytics.* Hoboken: John Wiley & Sons, 2013. ISBN 978-1-118-35685-2.

Internet sources:

23. McKinsey Global institute, Big data: *The next frontier for innovation, competition, and productivity* Available from: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation, May 2011
24. STATISTICA Data Mining tutorial available from: https://www.youtube.com/playlist?list=PLB804A810436AFB03

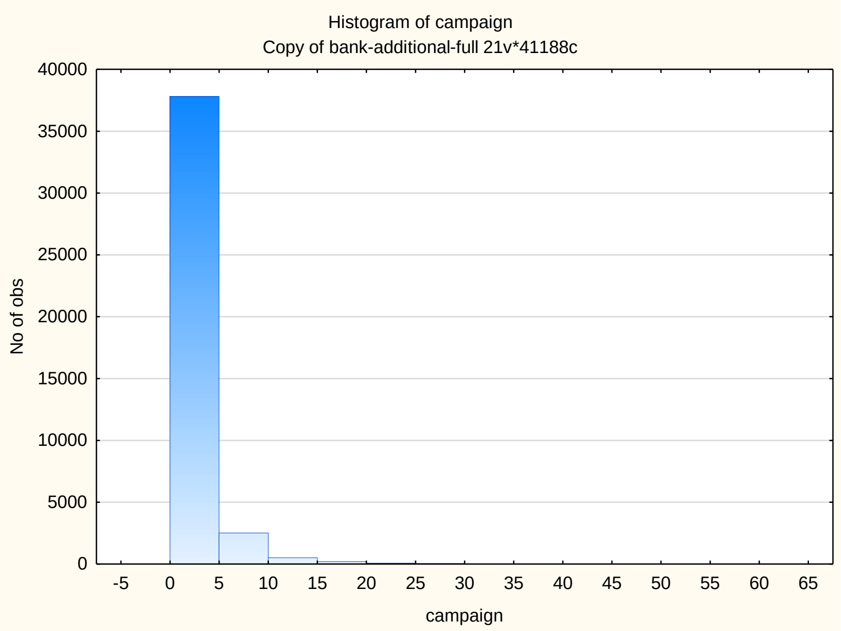
Appendices

Appendix 1. Bar chart of removed variable "pdays"



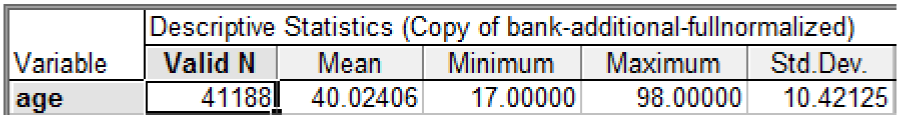
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 2. Bar chart of variable "campaign" before transformation



Source: Data from (Moro et al., 2014) STATISTICA output, own processing

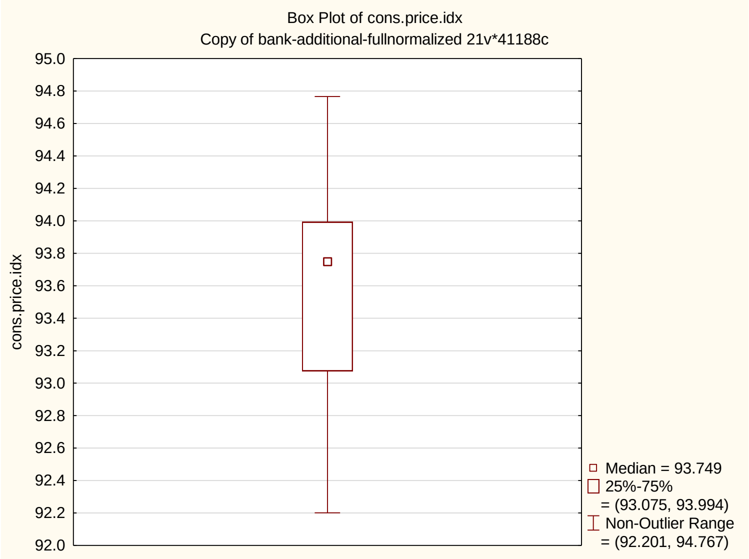
Appendix 3. Summary statistics for variable “age”



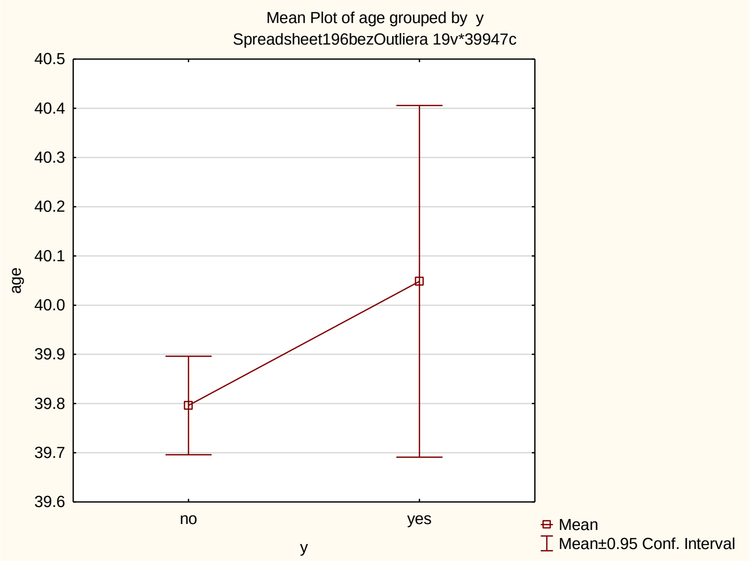
	Descriptive Statistics (Copy of bank-additional-fullnormalized)				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
age	41188	40.02406	17.00000	98.00000	10.42125

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 4. Box plot of variable “cons.price.idx”
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

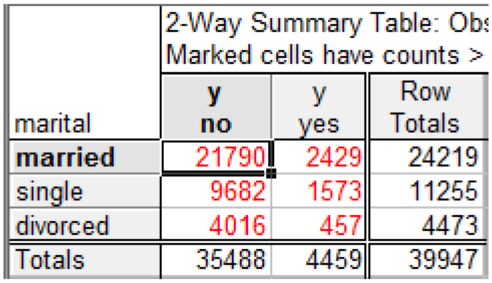


Appendix 5. Mean plot of variables “age” and “y”



Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 6. 2-way summary tables for variables marital and y

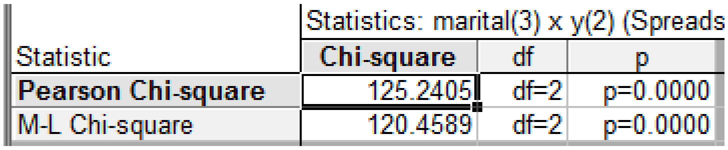


2-Way Summary Table: Obs
Marked cells have counts >

marital	y no	y yes	Row Totals
married	21790	2429	24219
single	9682	1573	11255
divorced	4016	457	4473
Totals	35488	4459	39947

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 7. Chi-square test results for marital and y

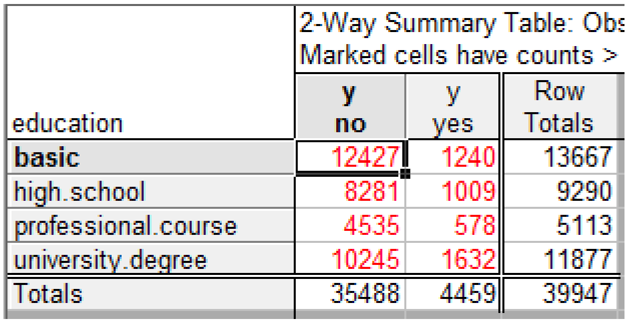


Statistics: marital(3) x y(2) (Spreads

Statistic	Chi-square	df	p
Pearson Chi-square	125.2405	df=2	p=0.0000
M-L Chi-square	120.4589	df=2	p=0.0000

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 8. 2-way summary tables for variables education and y

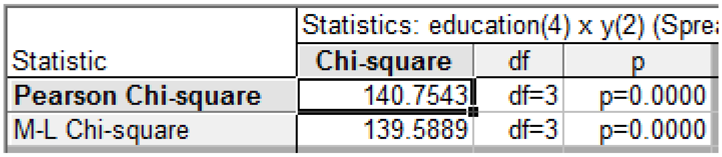


2-Way Summary Table: Obs
Marked cells have counts >

education	y no	y yes	Row Totals
basic	12427	1240	13667
high.school	8281	1009	9290
professional.course	4535	578	5113
university.degree	10245	1632	11877
Totals	35488	4459	39947

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 9. Chi-square test results for education and y

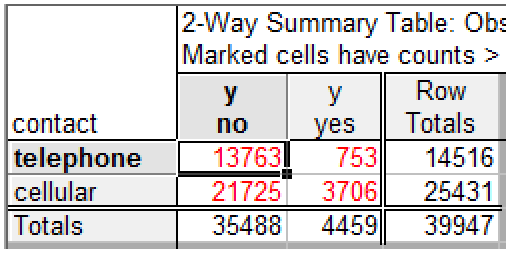


Statistics: education(4) x y(2) (Spre

Statistic	Chi-square	df	p
Pearson Chi-square	140.7543	df=3	p=0.0000
M-L Chi-square	139.5889	df=3	p=0.0000

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

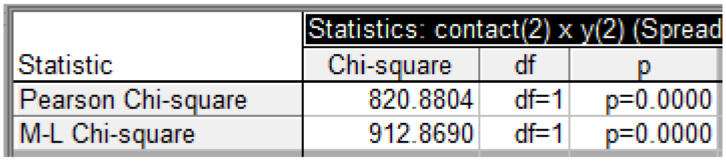
Appendix 10. 2-way summary tables for variables contact and y



	2-Way Summary Table: Obs Marked cells have counts >		
contact	y no	y yes	Row Totals
telephone	13763	753	14516
cellular	21725	3706	25431
Totals	35488	4459	39947

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

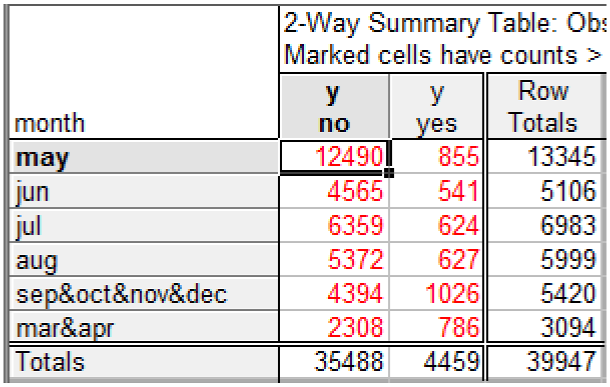
Appendix 11. Chi-square test results for contact and y



	Statistics: contact(2) x y(2) (Spread		
Statistic	Chi-square	df	p
Pearson Chi-square	820.8804	df=1	p=0.0000
M-L Chi-square	912.8690	df=1	p=0.0000

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

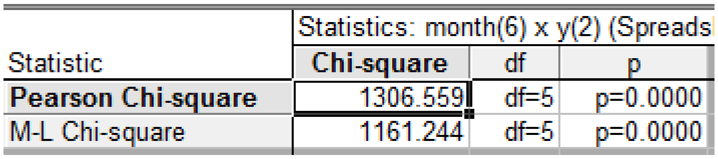
Appendix 12. 2-way summary tables for variables month and y



	2-Way Summary Table: Obs Marked cells have counts >		
month	y no	y yes	Row Totals
may	12490	855	13345
jun	4565	541	5106
jul	6359	624	6983
aug	5372	627	5999
sep&oct&nov&dec	4394	1026	5420
mar&apr	2308	786	3094
Totals	35488	4459	39947

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

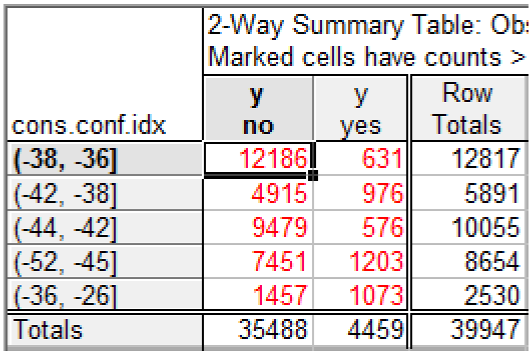
Appendix 13. Chi-square test for variables month and y



	Statistics: month(6) x y(2) (Spreads		
Statistic	Chi-square	df	p
Pearson Chi-square	1306.559	df=5	p=0.0000
M-L Chi-square	1161.244	df=5	p=0.0000

Source: STATISTICA output, own processing

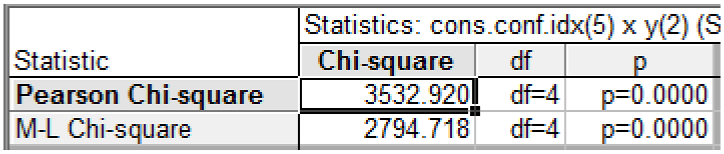
Appendix 14. 2-way summary tables for variables CCI and y



	2-Way Summary Table: Ob Marked cells have counts >		
cons.conf.idx	y no	y yes	Row Totals
(-38, -36]	12186	631	12817
(-42, -38]	4915	976	5891
(-44, -42]	9479	576	10055
(-52, -45]	7451	1203	8654
(-36, -26]	1457	1073	2530
Totals	35488	4459	39947

Source: Data from (Moro et al., 2014) STATISTICA output, own processing

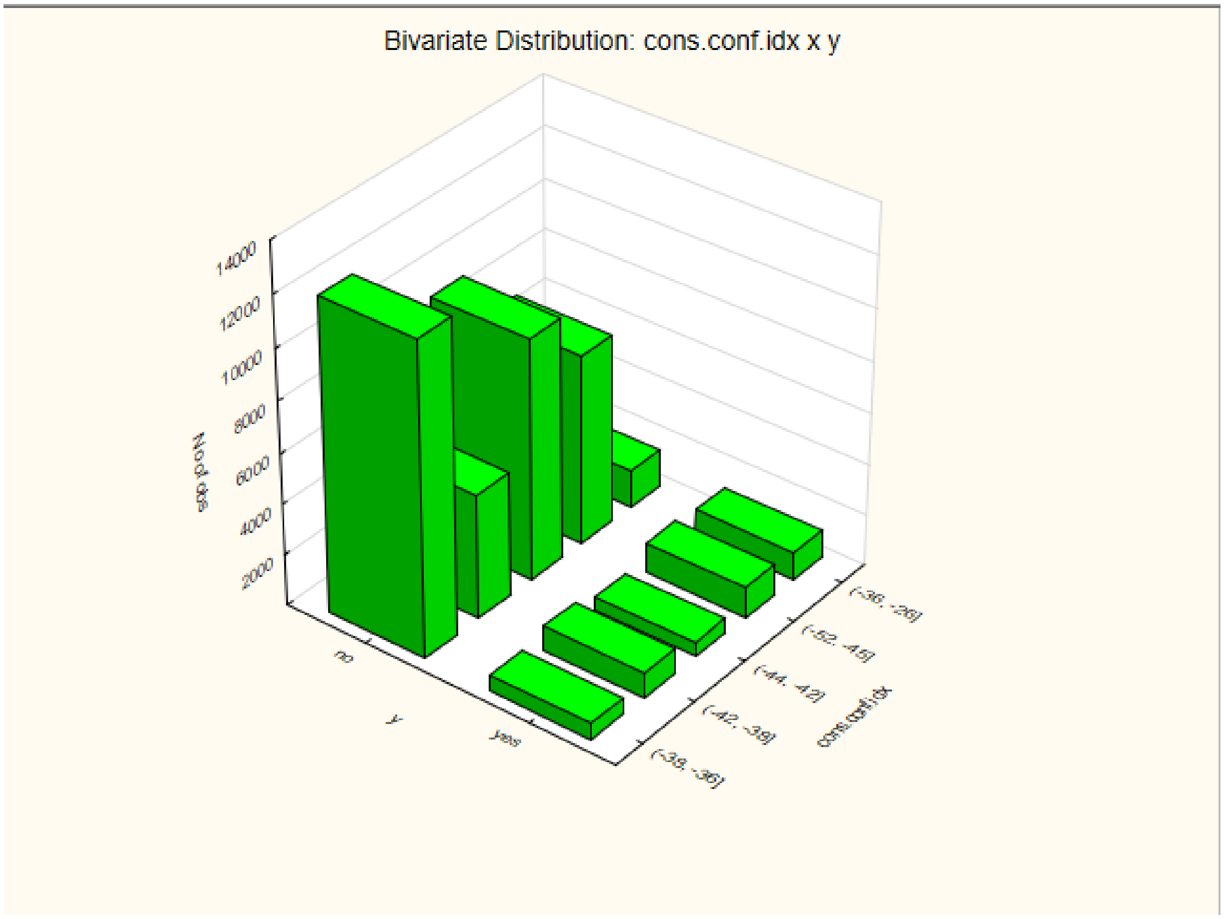
Appendix 15. Chi-square for CCI and y



	Statistics: cons.conf.idx(5) x y(2) (S		
Statistic	**Chi-square**	df	p
Pearson Chi-square	3532.920	df=4	p=0.0000
M-L Chi-square	2794.718	df=4	p=0.0000

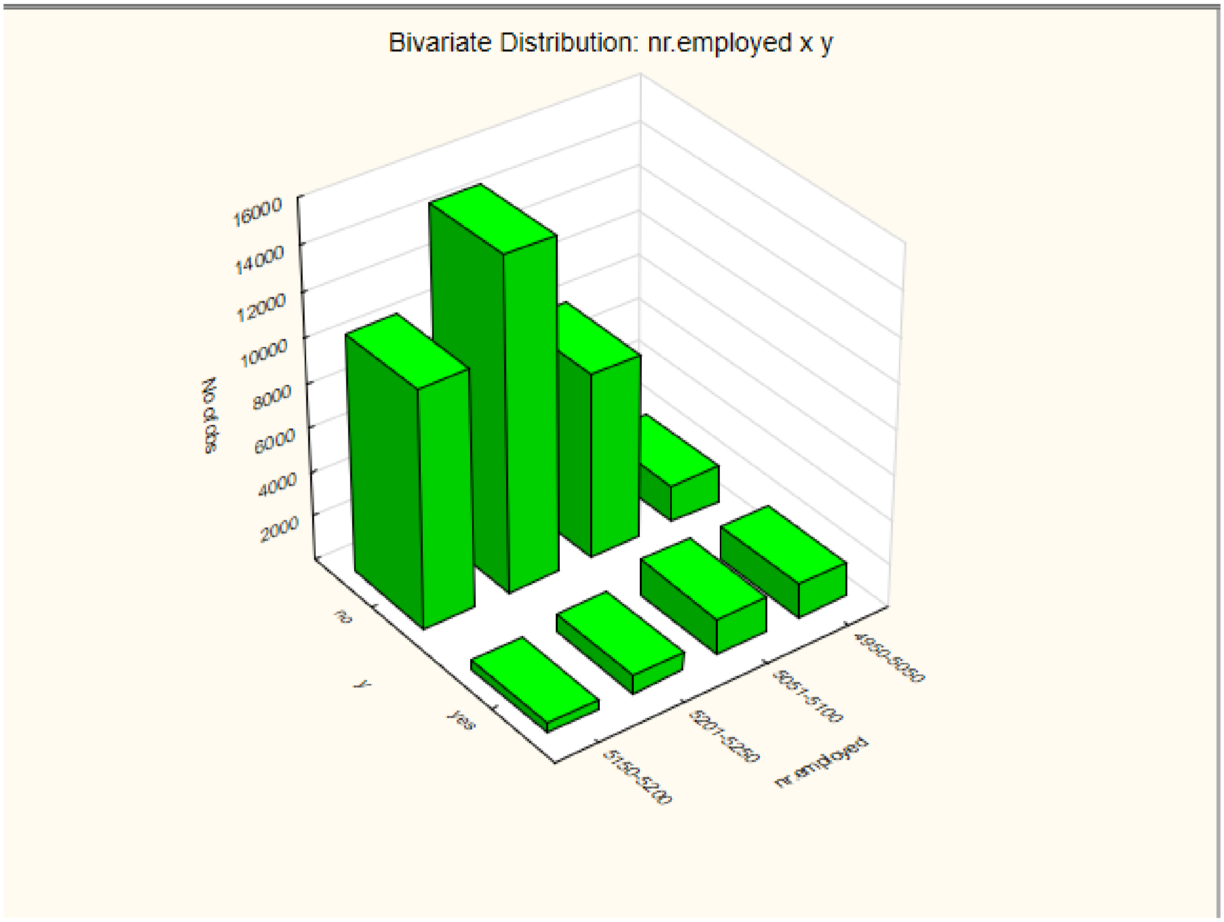
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 16. Bivariate distribution of variables consumer confidence index and y



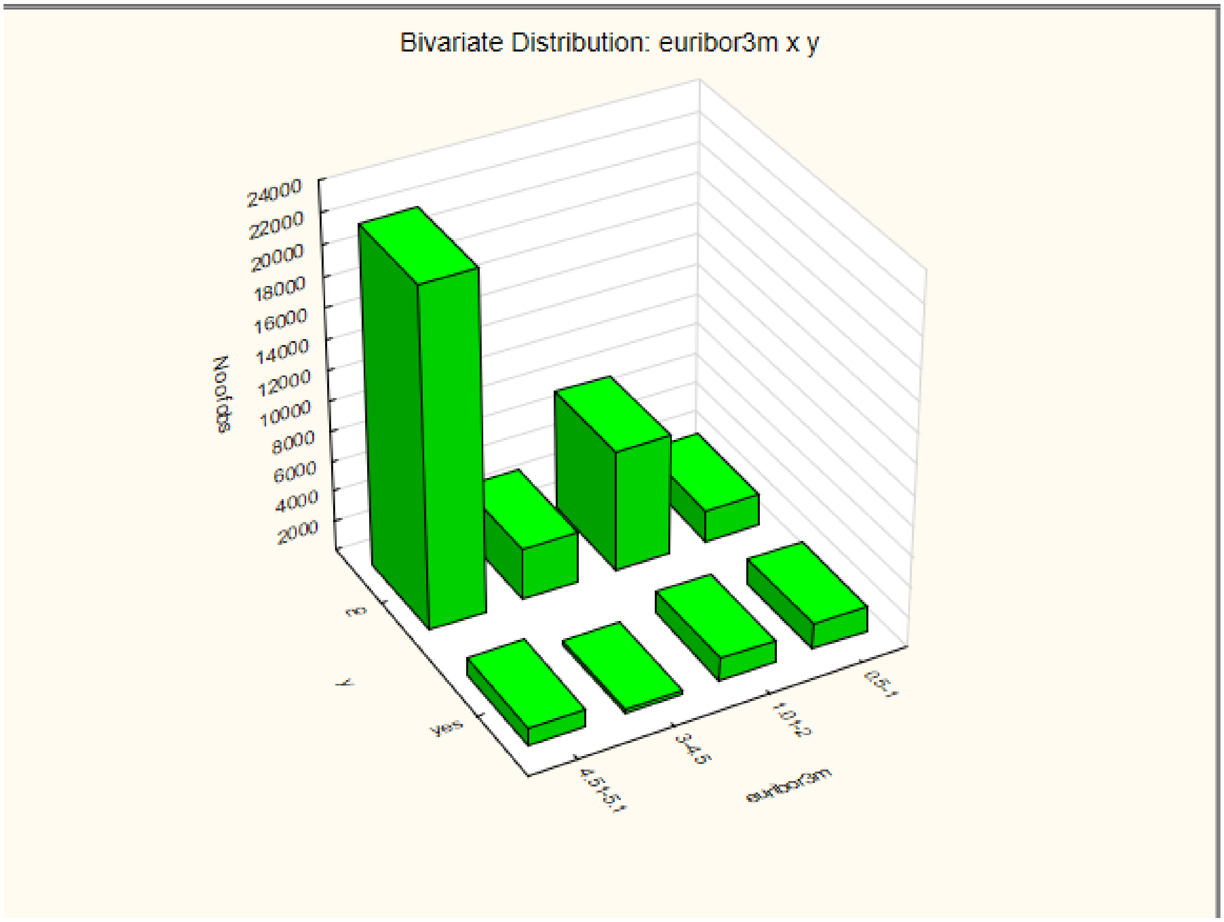
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 17. Bivariate distribution number employed and y



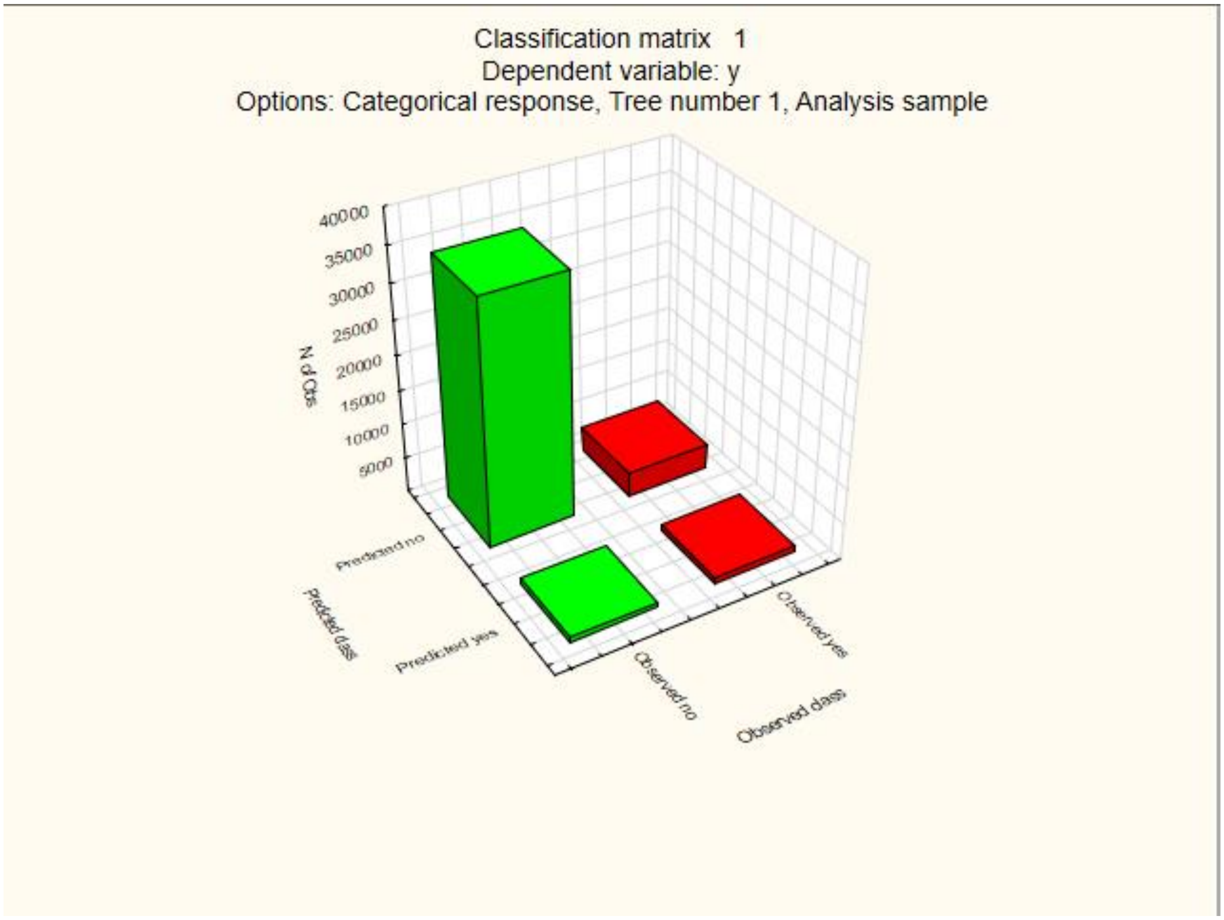
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 18. Bivariate distribution of variables “euribor.3m” and “y”



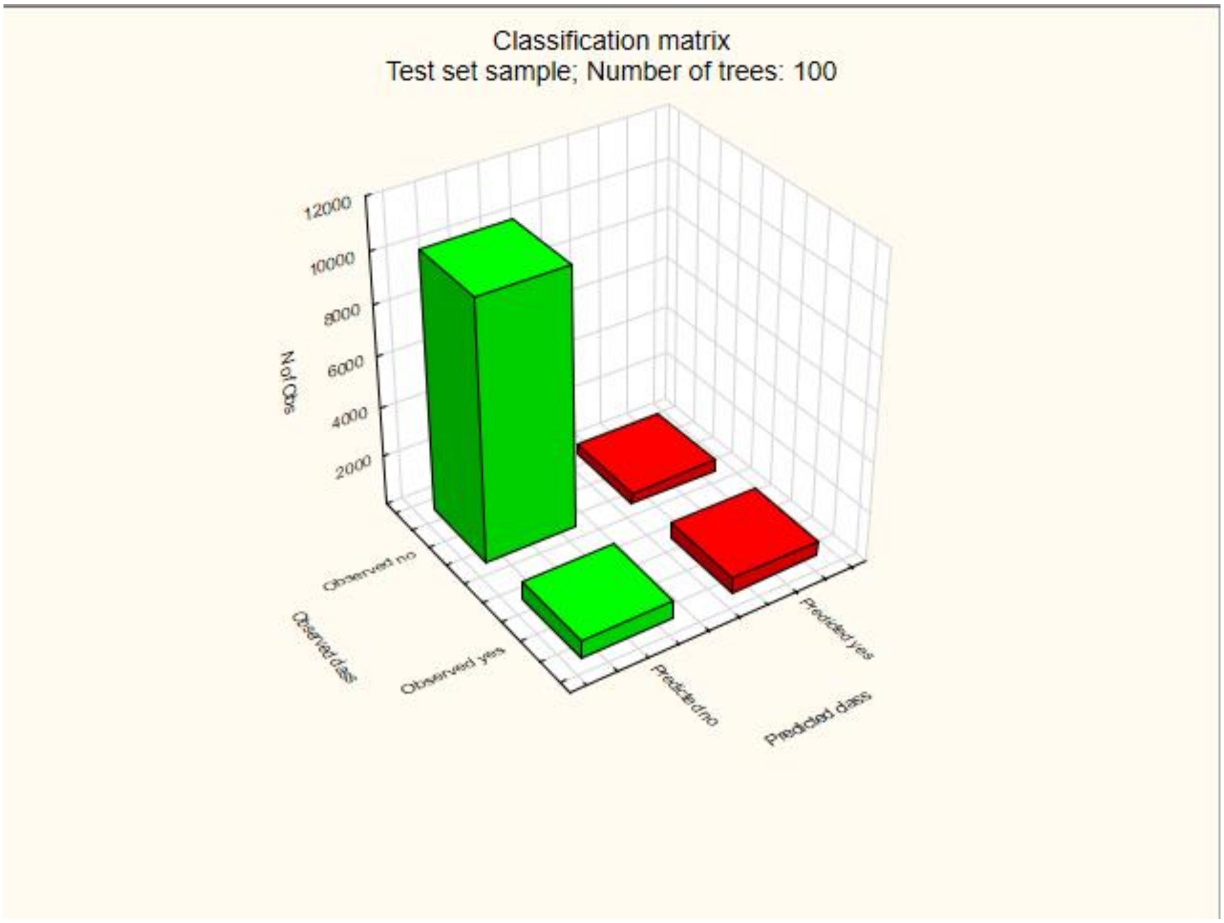
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 19. Classification matrix chart for CART model



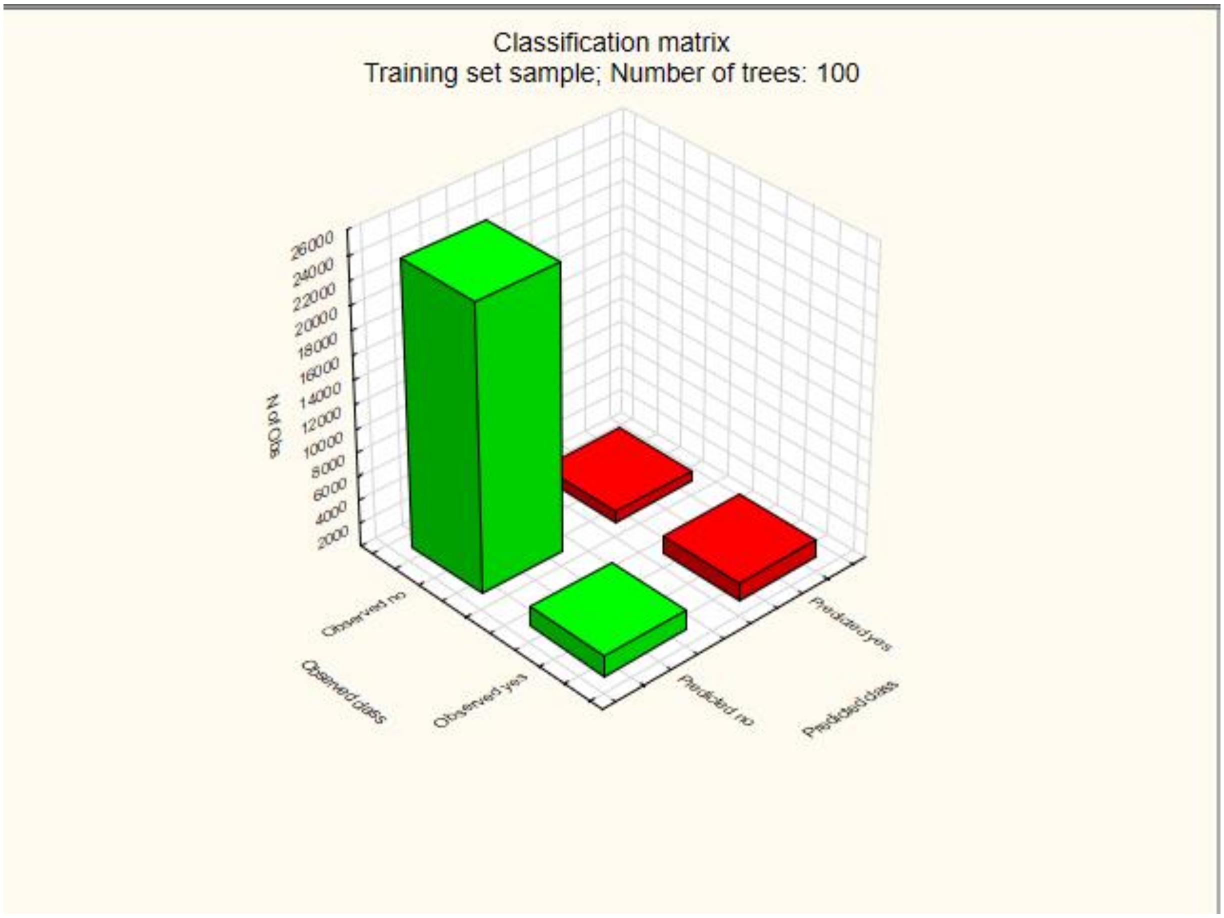
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 20. Classification matrix chart for Random Forest model test set sample



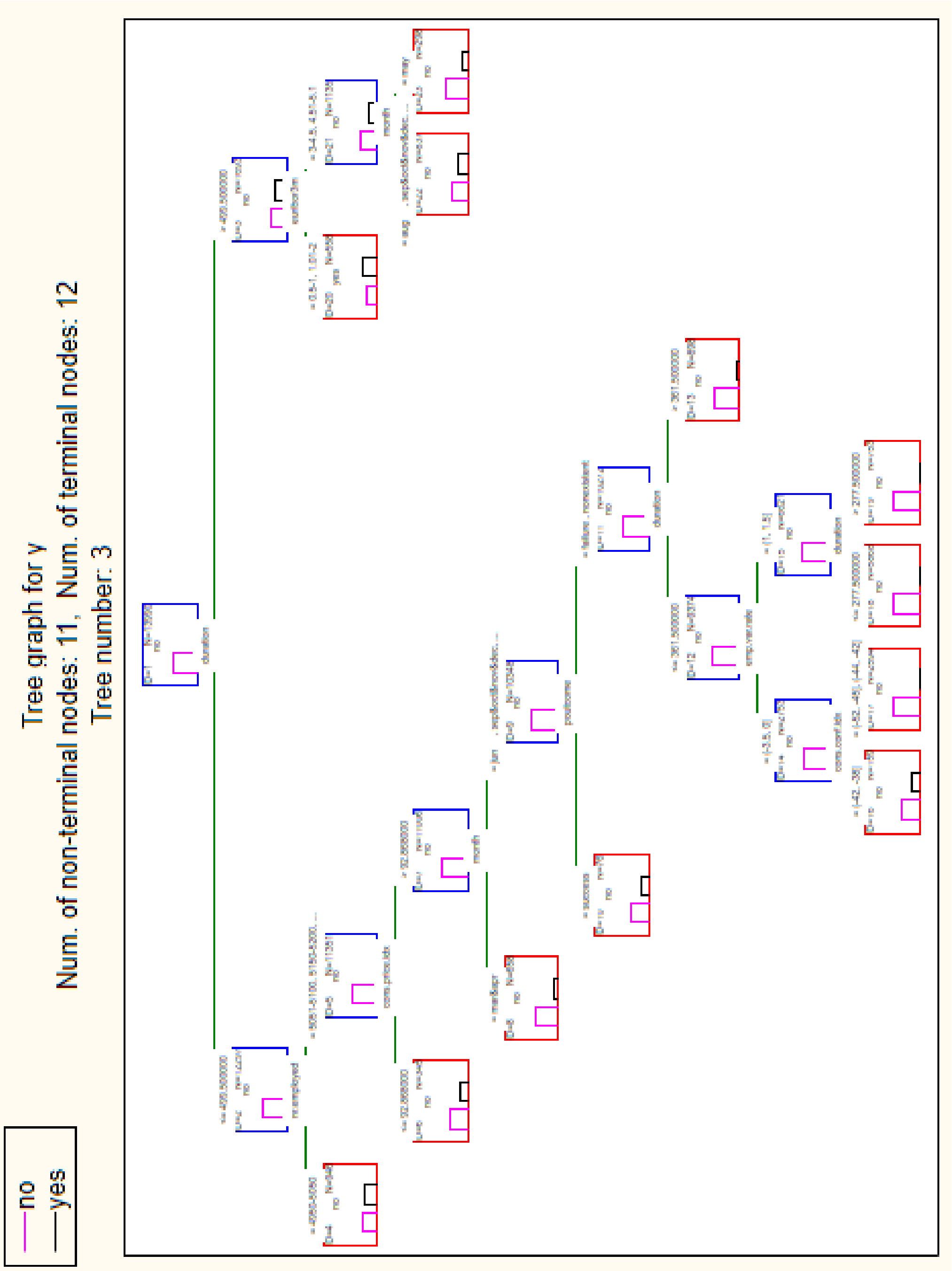
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 21. Classification matrix chart for Random Forest model training set



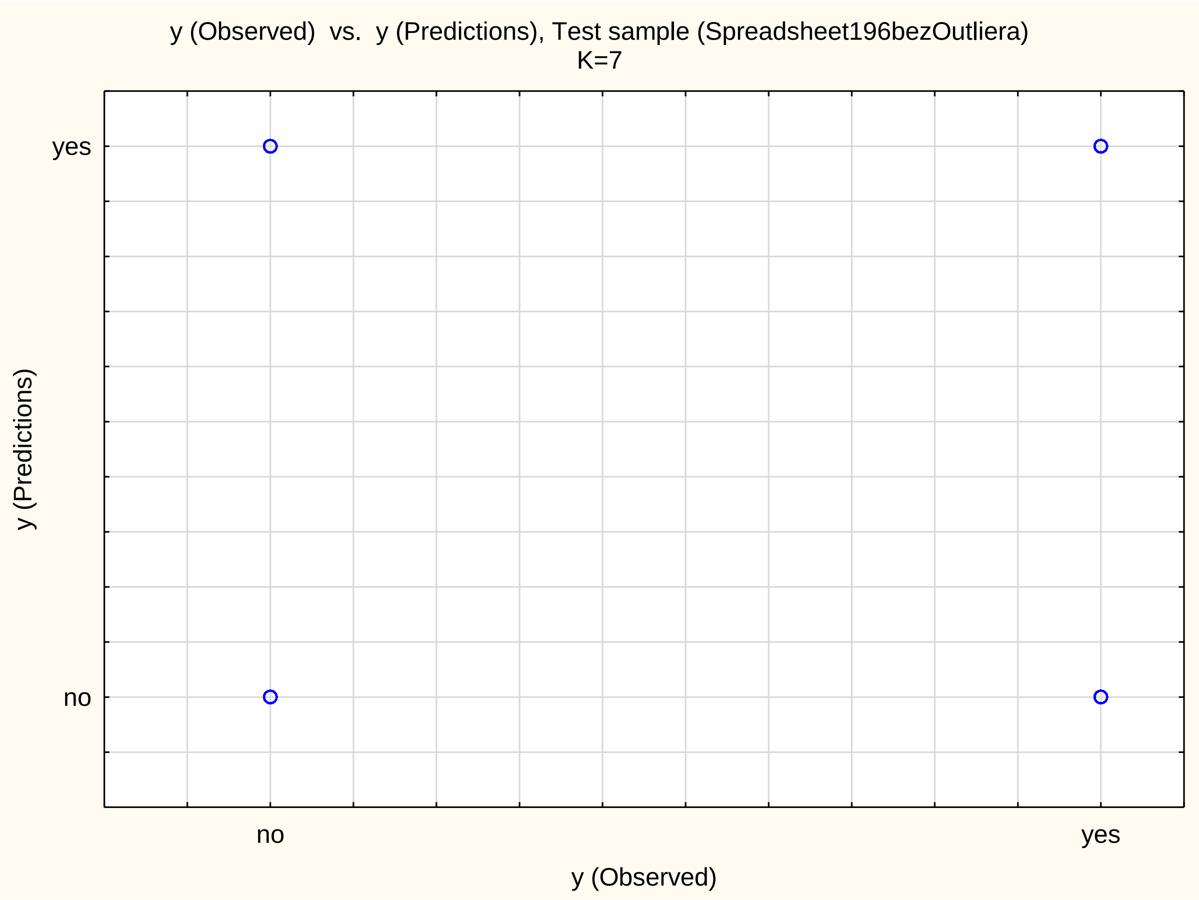
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 22. Random forest tree example



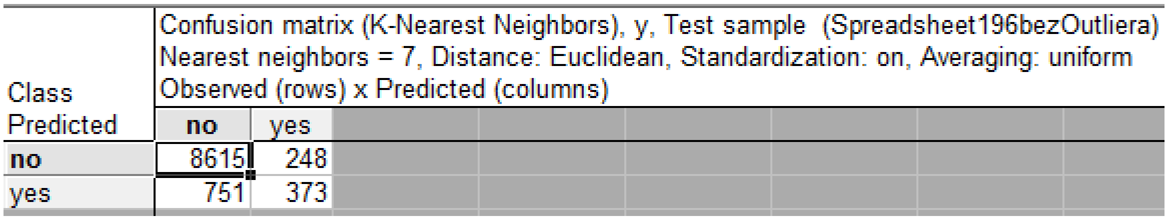
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 23. Observed versus predicted K-Nearest Neighbor



Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 24. Confusion matrix for K-Nearest neighbor model

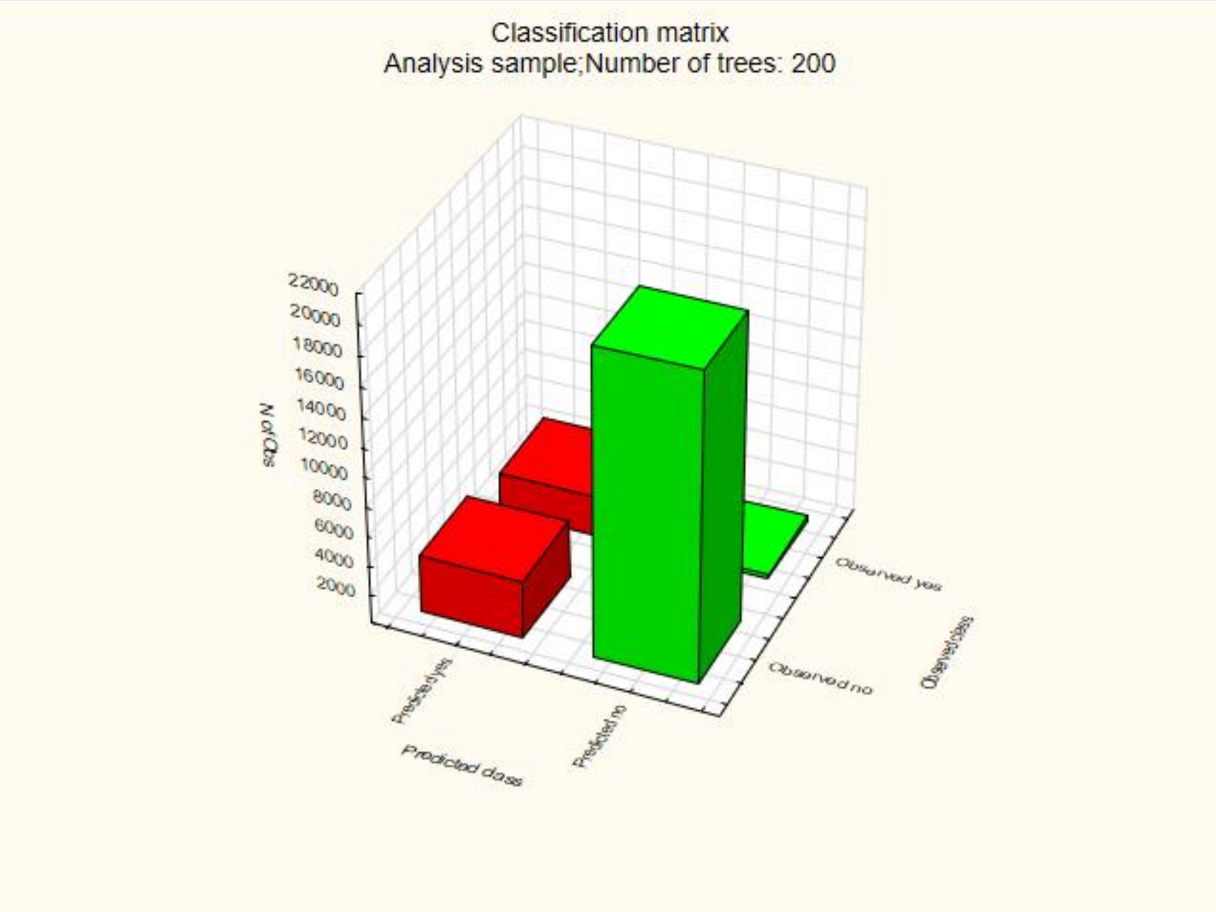


Confusion matrix (K-Nearest Neighbors), y, Test sample (Spreadsheet196bezOutliera)
Nearest neighbors = 7, Distance: Euclidean, Standardization: on, Averaging: uniform
Observed (rows) x Predicted (columns)

Class Predicted	no	yes
no	8615	248
yes	751	373

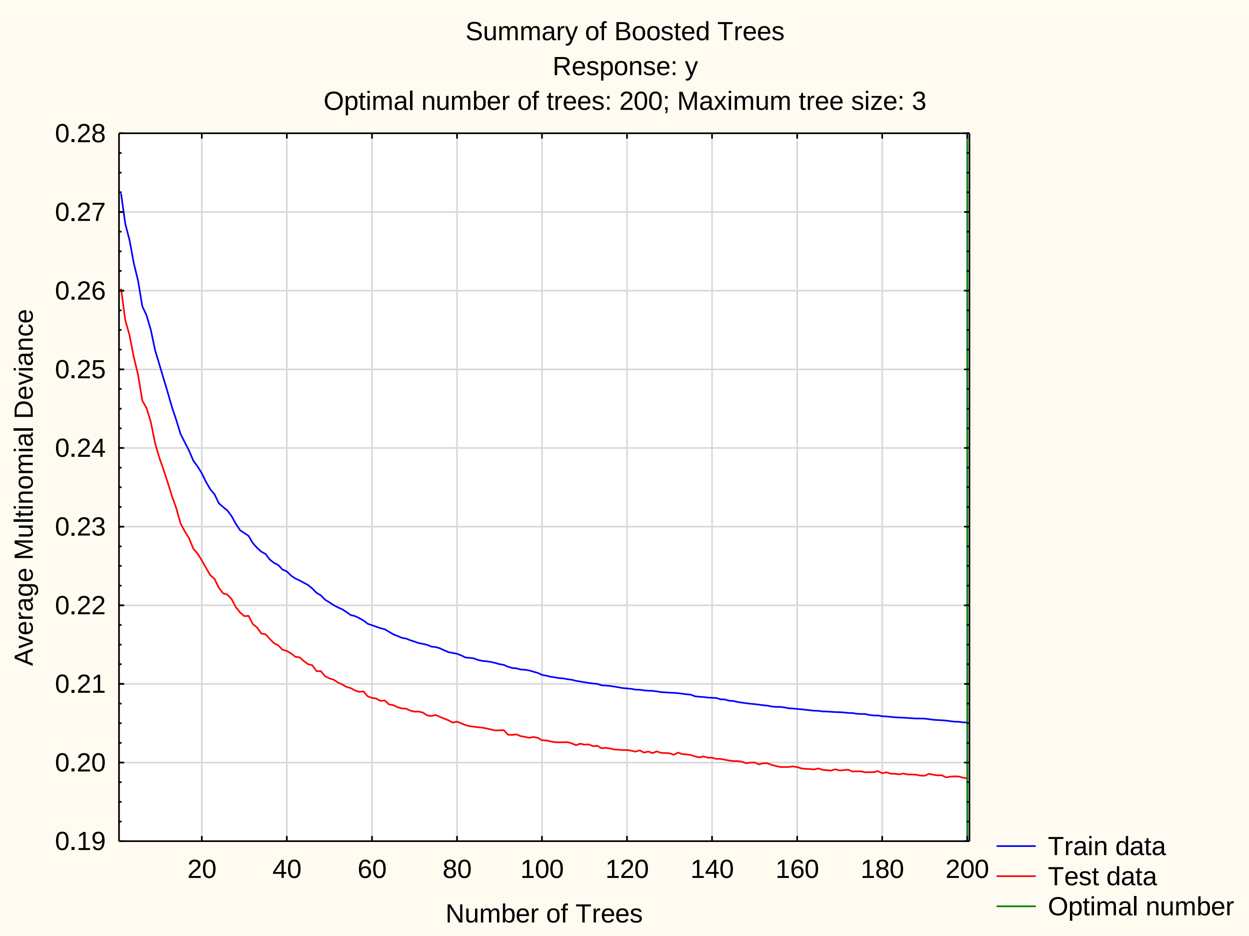
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 25. Classification matrix chart for Boosted Trees model



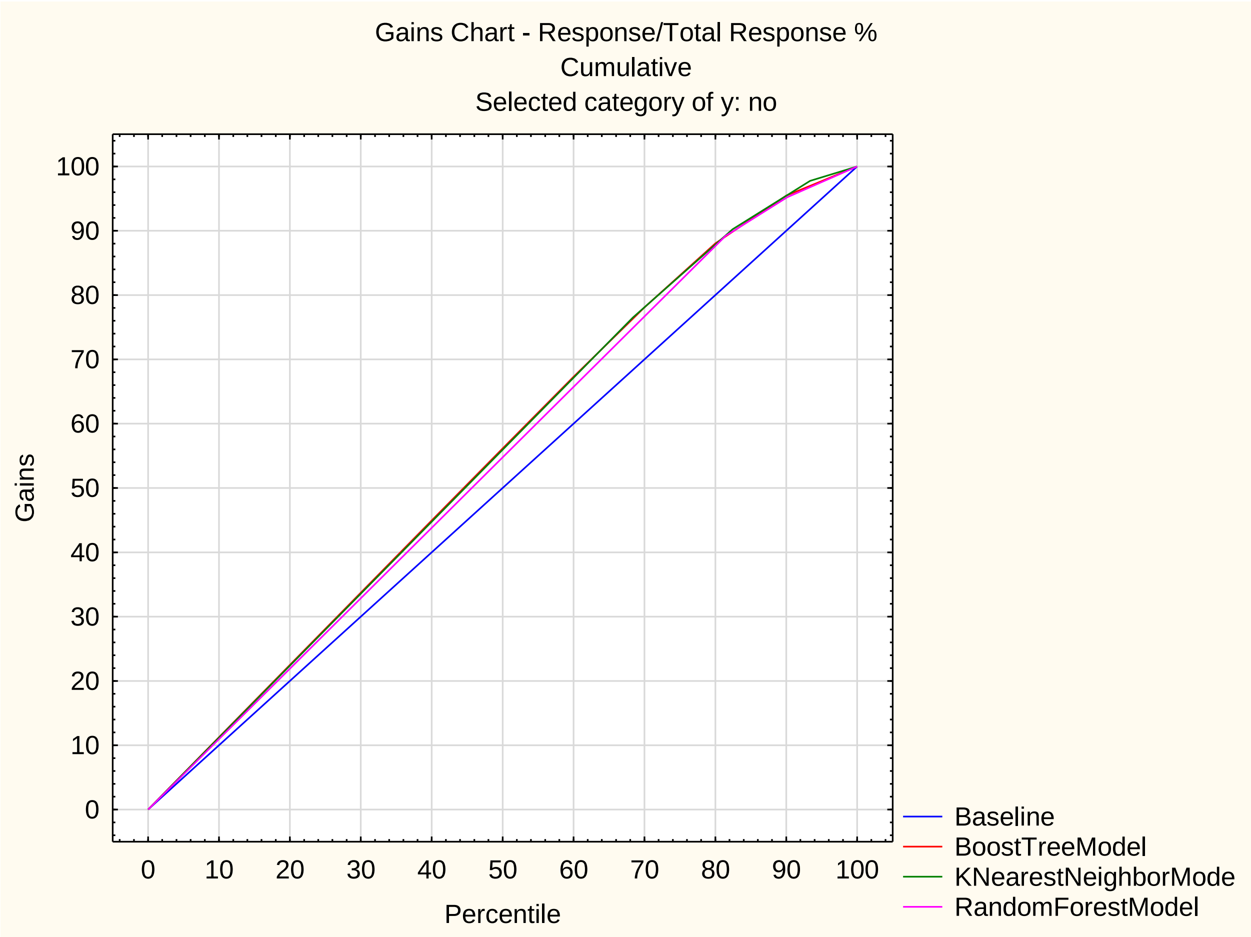
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 26. Summary of Boosted trees model, train and test data



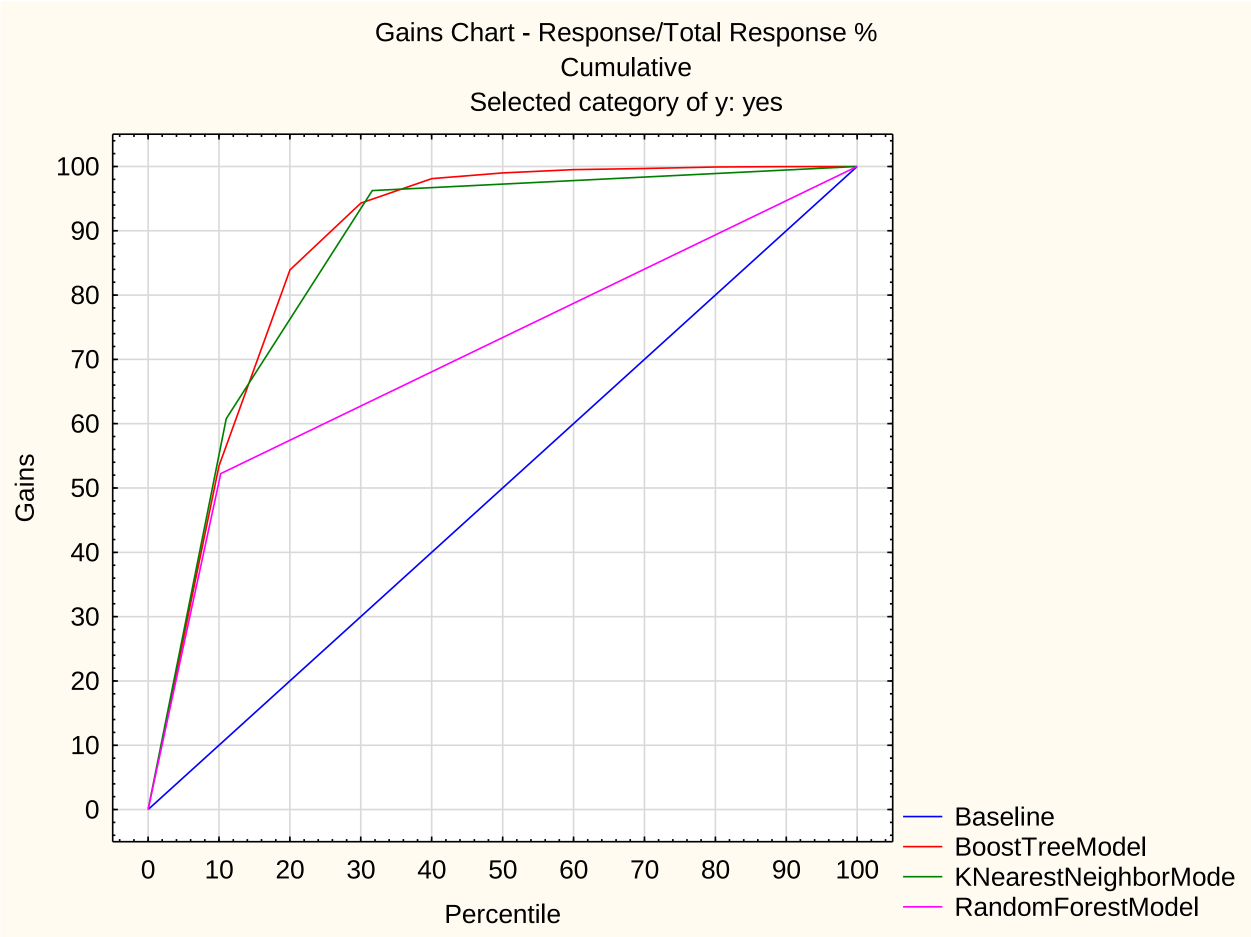
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 27. Gains chart for all three models for category “no”



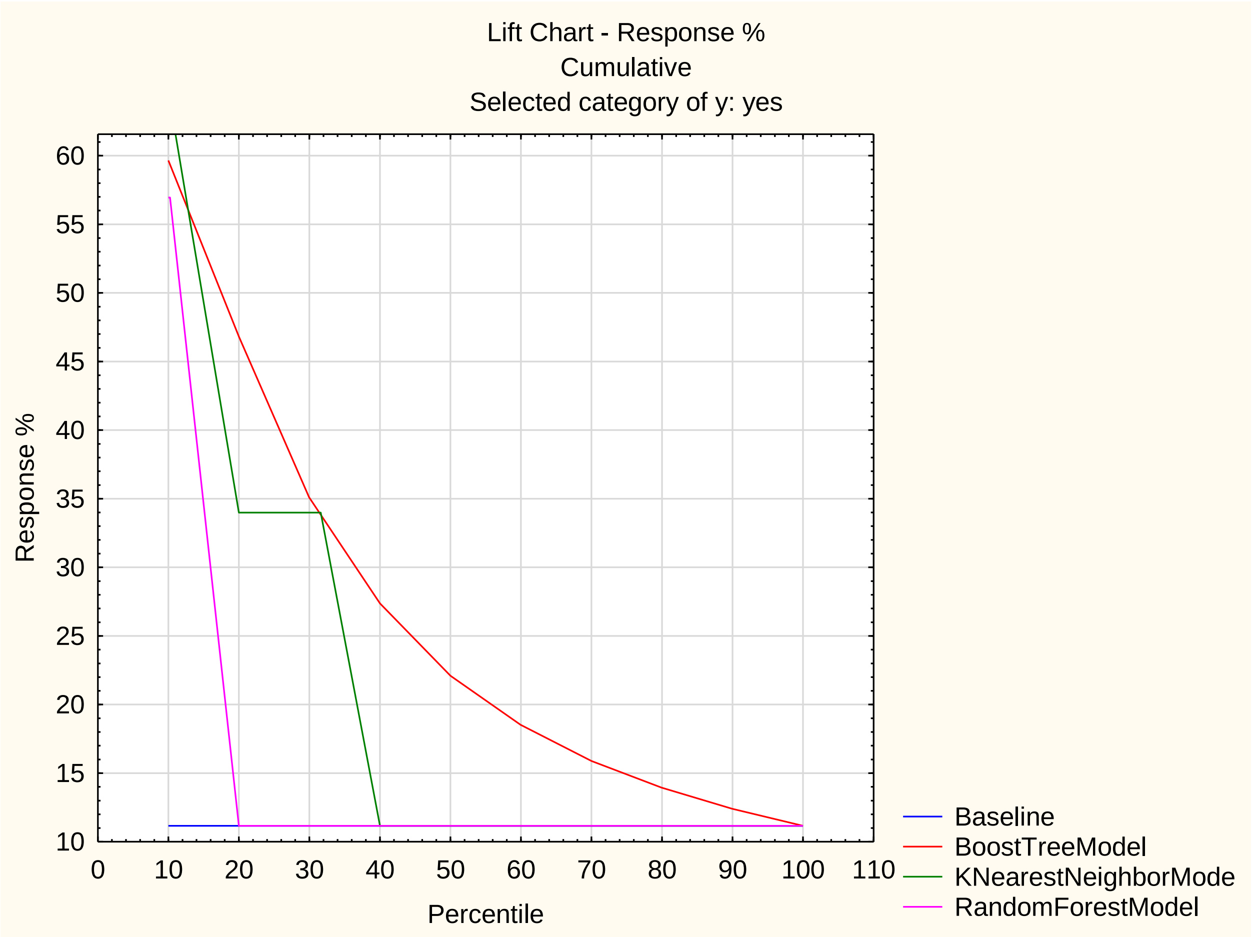
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 28. Gains chart for all three models for category “yes”



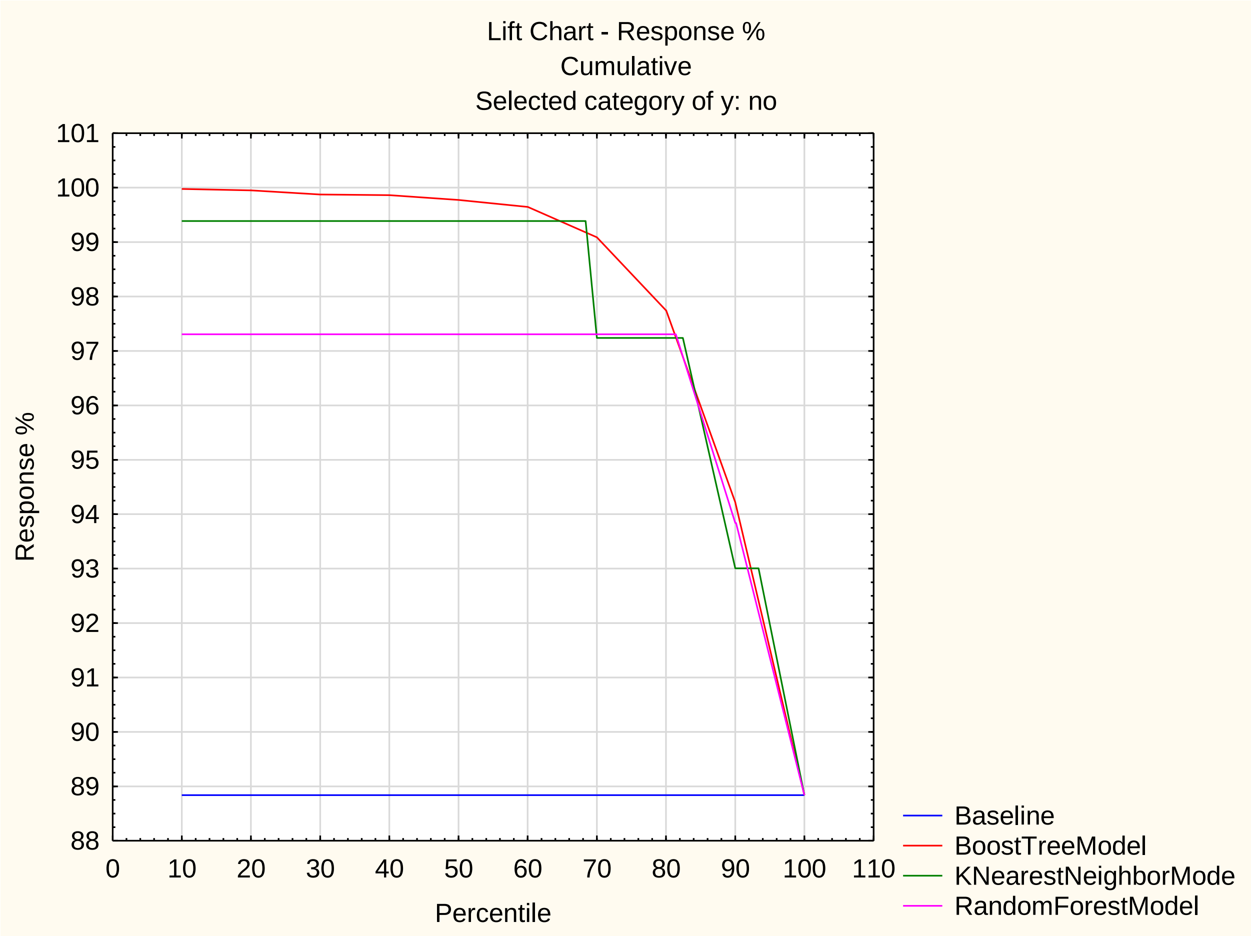
Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 29. Lift chart for all three models for category “yes”



Source: Data from (Moro et al., 2014) STATISTICA output, own processing

Appendix 30. Lift chart for all three models for category “no”



Source: Data from (Moro et al., 2014) STATISTICA output, own processing