****

SCHOOL OF ARCHITECTURE, COMPUTING AND ENGINEERING

Department of Engineering and Computing

**Knowledge creation in banking marketing using machine learning techniques**

**Yana Mihova**

**ID No: 1520143**

A report submitted in part fulfilment of the degree of

BSc (Hons) *Computer Science*

Supervisor: Dr. Amin Karami

CN6103

Academic Year 2018-19

# Abstract

Globalization and technological advances has created an extremely competitive market. This also has an impact on the banks. In recent years, banking and direct database marketing have become an important strategy for understanding customer needs. The success rate of banking marketing depends on the achieved results and decisions. In order to make more accurate predictions, statistical tools and methods are been used.

This report examines how to use machine learning techniques to analyse and make predictions in banking marketing using existing dataset. The purpose of building the models is to predict whether the client will subscribe for a term deposit. This report presents the different stage of data analysis such as data preparation and cleaning, building the models and model testing. Finally, the results of machine learning techniques are evaluated and analysed. Although there is no significant difference in the decision tree algorithm’s accuracy, C5.0 achieved a higher percentage. Linear regression model presents the relationship between quantitative features.

**Keywords:** Machine learning, supervised machine learning techniques, Decisions tree, linear regression, R, Banking marking

# Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr Amin Karami for his encouragement, patience and help he gave me throughout the whole project process. I would like to thank the University of East London and the lecturers for giving me the knowledge and skills to complete this difficult journey.

Finally, I would like to thanks to my beloved family members for their greatest support during my studies.

Contents

[Abstract ii](#_Toc6253334)

[Acknowledgments iii](#_Toc6253335)

[Introduction and aim of the project 1](#_Toc6253336)

[Chapter 1: Background and literature review 2](#_Toc6253337)

[1.1 Machine learning 2](#_Toc6253338)

[Supervised learning 3](#_Toc6253339)

[Unsupervised 3](#_Toc6253340)

[Reinforcement learning 4](#_Toc6253341)

[1.2 Linear Regression 4](#_Toc6253342)

[Simple linear regression model 4](#_Toc6253343)

[Multiple linear regression model 5](#_Toc6253344)

[Nonlinear regression model 5](#_Toc6253345)

[1.3 Decision trees 5](#_Toc6253346)

[1.4 Machine learning techniques in banking marketing 7](#_Toc6253347)

[Chapter 2 Proposed solution and methodology 8](#_Toc6253348)

[2.1 Proposed solution 8](#_Toc6253349)

[2.2 Methodology 8](#_Toc6253350)

[Chapter 3 Legal, Ethical and Social Approval 9](#_Toc6253351)

[Chapter 4 Description of the dataset 10](#_Toc6253352)

[4.1. Dataset information 10](#_Toc6253353)

[4.2 Attribute Information 10](#_Toc6253354)

[Client data 10](#_Toc6253355)

[Campaign data attributes 11](#_Toc6253356)

[Other data attributes 11](#_Toc6253357)

[Socio-economic attributes 11](#_Toc6253358)

[Output variable (desired target) – Term Deposit 11](#_Toc6253359)

[Chapter 5 Understand banking marking dataset 12](#_Toc6253360)

[5.1 Preparing the directory 12](#_Toc6253361)

[5.2 Load the dataset 12](#_Toc6253362)

[5.3 Exploring of banking marking dataset 12](#_Toc6253363)

[5.3.1 Create the data frame 12](#_Toc6253364)

[5.3.2 Structure of the dataset 13](#_Toc6253365)

[5.3.3 Dimension of the dataset 13](#_Toc6253366)

[5.3.4 Content of the dataset 13](#_Toc6253367)

[5.3.5 Summary of the data 14](#_Toc6253368)

[Chapter 6 Data cleaning and preparation methods 16](#_Toc6253369)

[6.1 Missing values 16](#_Toc6253370)

[6.2 Outliers 16](#_Toc6253371)

[6.2.2 Outlier detection 17](#_Toc6253372)

[6.2.2.1 Gaussian distribution method (normal distribution) 17](#_Toc6253373)

[6.2.2.2 Box-and-whisker plot (boxplot) and IQR rule 18](#_Toc6253374)

[6.3 Distribution and visualization of variables 22](#_Toc6253375)

[6.4 Converting the data type 31](#_Toc6253376)

[6.5 Creating training and test dataset 31](#_Toc6253377)

[Chapter 7 Building the models 32](#_Toc6253378)

[7.1 Decision trees 32](#_Toc6253379)

[7.1.1 Types of decision trees 32](#_Toc6253380)

[7.1.2 Decision Classification trees in R 32](#_Toc6253381)

[7.1.2.1 CART (Classification and Regression Trees) using rpart package 32](#_Toc6253382)

[7.1.2.1.1 Building and training the model on the data 32](#_Toc6253383)

[7.1.2.1.2 Test and evaluate the performance of the model 34](#_Toc6253384)

[7.1.2.2 C5.0 decision trees 35](#_Toc6253385)

[7.1.2.2.1 Building and training the model on the data 35](#_Toc6253386)

[7.1.2.2.2 Test and evaluate the performance of the model 36](#_Toc6253387)

[7.1.2.3 Comparison of Decision trees algorithms 37](#_Toc6253388)

[7.1.3 Linear Regression in R 37](#_Toc6253389)

[7.1.3.1 Simple Linear Regression 38](#_Toc6253391)

[7.1.3.2 Multiple Linear Regressions 40](#_Toc6253392)

[Chapter 8 Conclusion 41](#_Toc6253393)

[8.1 Summary 41](#_Toc6253394)

[8.2 Future work 41](#_Toc6253395)

[References and Bibliography 42](#_Toc6253396)

[Appendices 50](#_Toc6253397)

[Appendix A – Initial Proposal Form 51](#_Toc6253398)

[Appendix B – Final Proposal Form 52](#_Toc6253399)

[Appendix C - Project Plan 53](#_Toc6253400)

[Project milestones 53](#_Toc6253401)

[Gantt chart 53](#_Toc6253402)

[PDP (Personal Development plan) 53](#_Toc6253403)

[Appendix D - Legal, Ethical and Social approval 55](#_Toc6253404)

[Appendix E Poster Viva/Demonstration Event 57](#_Toc6253405)

**List of Figures**

[Figure 1Traditional programming and machine learning (Ajanki, 2018) 2](#_Toc6237469)

[Figure 2 Simple linear regression model (Montgomery, 2012, p.1) 5](#_Toc6237470)

[Figure 3 Multiple linear regression model (Montgomery, 2012, p.4) 5](#_Toc6237471)

[Figure 4 Multiple linear regression model (Yan, 2009, p.3) 5](#_Toc6237472)

[Figure 5 Decision tree structure (Brid, 2018) 6](#_Toc6237473)

[Figure 6 Entropy of the class (Lantz, 2013, p.125) 6](#_Toc6237474)

[Figure 7 InfoGain (Lantz ,2013, p.126) 6](#_Toc6237475)

[Figure 8 Gini Index (Zaki, 2014 p. 487) 7](#_Toc6237476)

[Figure 9 Project’s working directory 13](#_Toc6237477)

[Figure 10 loading the data 13](#_Toc6237478)

[Figure 11 the class of object 13](#_Toc6237479)

[Figure 12 Structure of the banking marketing dataset 14](#_Toc6237480)

[Figure 13 Dimension of dataset 14](#_Toc6237481)

[Figure 14 Column names of banking marketing dataset 15](#_Toc6237482)

[Figure 15 Overview of first 6 and last 6 rows 15](#_Toc6237483)

[Figure 16 Summary of data in early stage 16](#_Toc6237484)

[Figure 17 Target percentage proportion 16](#_Toc6237485)

[Figure 18 No missing values in banking marking dataset 17](#_Toc6237486)

[Figure 19 if-else statement for missing values 17](#_Toc6237487)

[Figure 20 The continuous spectrum from normal data to outliers (Aggarwal, 2013, p. 4) 18](#_Toc6237488)

[Figure 21 Gaussian distribution of the values within the standard deviation from the mean (Seo, 2006, p.5) 18](#_Toc6237489)

[Figure 22 Outlier detection of Age using 3SD in R 19](#_Toc6237490)

[Figure 23 Box-and whisker plot (Kohl, 2015, p. 40) 20](#_Toc6237491)

[Figure 24 Boxplot () in R 20](#_Toc6237492)

[Figure 25 Boxplot visualization of Age with outliers 20](#_Toc6237493)

[Figure 26 cons.conf.idx outliers 21](#_Toc6237494)

[Figure 27 Boxplot visualization of other attributes 21](#_Toc6237495)

[Figure 28 IQR range in R 22](#_Toc6237496)

[Figure 29 ggplot function 23](#_Toc6237497)

[Figure 30 Graph of distribution between job and target data (y) 23](#_Toc6237498)

[Figure 31 Graph of distribution between age and target data (y) 24](#_Toc6237499)

[Figure 32 Graph of distribution between marital and target data (y) 24](#_Toc6237500)

[Figure 33 Graph of distribution between education and target data (y) 25](#_Toc6237501)

[Figure 34 Graph of distribution between job and contact 25](#_Toc6237502)

[Figure 35 Graph of distribution between default and target (y) 25](#_Toc6237503)

[Figure 36 Graph of distribution between loan and target (y) 26](#_Toc6237504)

[Figure 37 Graph of distribution between housing loan and target (y) 26](#_Toc6237505)

[Figure 38 Graph of distribution between previous and target (y) 27](#_Toc6237506)

[Figure 39 Graph of distribution between poutcome and target (y) 27](#_Toc6237507)

[Figure 40 Graph of distribution between campaign and target (y) 27](#_Toc6237508)

[Figure 41 Graph of distribution between months of last contact with client 28](#_Toc6237509)

[Figure 42 Graph of distribution between days of the week of last contact with client 28](#_Toc6237510)

[Figure 43 Graph of distribution between duration of the last contact with client in seconds and target(y) 29](#_Toc6237511)

[Figure 44 the number of days after the customers was last contacted from a previous marketing campaign 29](#_Toc6237512)

[Figure 45 employment variation rate Emp.var.rate 30](#_Toc6237513)

[Figure 46 consumer price index Cons.price.idx 30](#_Toc6237514)

[Figure 47 confidence price index Cons.conf.idx 31](#_Toc6237515)

[Figure 48 3 month Euro Interbank Offered Rates Euribr3m 31](#_Toc6237516)

[Figure 49 number of employees nr.employed 31](#_Toc6237517)

[Figure 50 Class argument for each column 32](#_Toc6237518)

[Figure 51 converting the numeric variable 32](#_Toc6237519)

[Figure 52 Spit the dataset into training and test 32](#_Toc6237520)

[Figure 53 Building the Rpart tree model 33](#_Toc6237521)

[Figure 54 Summary of the CART training model 34](#_Toc6237522)

[Figure 55 Results of the CART training model 34](#_Toc6237523)

[Figure 56 Visualization of the CART model 35](#_Toc6237524)

[Figure 57 Test dataset prediction of CART 35](#_Toc6237525)

[Figure 58 CART Calculation of Accuracy 35](#_Toc6237526)

[Figure 59 *CrossTable()* in CART model 36](#_Toc6237527)

[Figure 60 Cross table CART results 36](#_Toc6237528)

[Figure 61 Building the C5.0 model 36](#_Toc6237529)

[Figure 62 C5.0 information 37](#_Toc6237530)

[Figure 63 Evaluation of the C5.0 training model 37](#_Toc6237531)

[Figure 64 Test dataset prediction of C5.0 37](#_Toc6237532)

[Figure 65 C5.0 Calculation of Accuracy 37](#_Toc6237533)

[Figure 66 Cross table C5.0 results 38](#_Toc6237534)

[Figure 67 Scatterplot matrix for quantitative features 39](#_Toc6237535)

[Figure 68 Simple linear regression in R 39](#_Toc6237536)

[Figure 69 Summary results of simple linear regression 40](#_Toc6237537)

[Figure 70 Multiple linear regressions in R 41](#_Toc6237538)

[Figure 71 Summary results of multiple linear regressions 41](#_Toc6237539)

**List of Tables**

[Table 1 Normal distribution of each attribute within third standard deviation of the mean 19](#_Toc6237540)

[Table 2 IQR with inner and outer fences for each attributes 22](#_Toc6237541)

[Table 3 Comparison of Decision tree algorithms 38](#_Toc6237542)

# Introduction and aim of the project

In the last years, machine learning (ML) has grown into one of the most significant IT and Artificial intelligence (AI) branches. This is a specific sub-group of AI based on the idea that the machine can learn by identifying patterns and make predictions in various data problems with minimum human intervention. Machine learning is a data analysis method that is widely used in various business and industrial sectors. The main reason for that because ML can build predictive models to produce better predictions and achieve the desired level of accuracy, leading to better outcomes.

The aim of the project is to find how to use machine learning techniques for analysis and making predictions using existing dataset in banking marketing. To find how they can be used together in a process of converting raw data to effective decision making knowledge. Building the predictive models will help to predict whether the client will subscribe for a term deposit.

This report will describe the different stages of preparation and implementation of the predictive models, staring with a literature review on machine learning techniques; in particular, linear regression and decision trees and how machine learning techniques are used in banking marking.

Once the literature review of these techniques has been revised, a methodology will be composed on how to pursue the investigation. The methodology is needed to establish how the implementation of the models will continue.

After the establishment of the methodology, the methods of data cleaning and preparation methods on the raw data will be described and explained. The project will identify probabilities and visualize the results in order to improve the solutions and achieve desired outcomes. At the same time, a good understanding of banking marketing dataset will be provided so that the scope of the analysis can be clearly defined.

The implementation will include building the linear regression and Decision trees machine learning techniques using R. The dataset used in this project is downloaded from UCI Irvine machine learning repository ([https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing))( direct marketing campaigns of a Portuguese banking institution )At the same time, the practical development will try to test how the theory works and find better solutions.

The report will include visualizing the results, testing and evaluating the performance of the models and solving various data problems.

Then, by analysis of the performance and achieved results, and summarise them at the conclusion, the aim and objectives of the project will be fulfilled.

# Chapter 1: Background and literature review

This project examines how to use machine learning techniques to analyze and make predictions in banking marketing using existing dataset. In order to find how they are used in a process of converting raw data to effective decision making knowledge, the literature review on machine learning methods was carried out for existing research papers and books; in particular, linear regression and decision trees. By the same token, the use of machine learning techniques in the banking marketing sector is investigated based on last published articles and research papers. The main purpose of the literature review is to give a clear idea of what was written as a theory of these techniques and how they were used to make predictions in banking marking.

## Machine learning

Today, there are many research articles and books that explore machine learning as a specific area of Artificial Intelligence used to perform various tasks which are too complex with minimum human efforts. (Bonaccorso, 2017, p.7) Machine learning also is defined as teaching the machine to learn and solve specific problem. The machines learn from the data by identifying patterns. Nowadays, there is a huge variety of data that can be processed by machines. As a result, processed information can be used as decision-making knowledge and solve different data problems. (Lantz, 2013, p 8) The interaction between knowledge and information (data) is essential and influences every step of building the solution. (Usuelli, 2014) It could also be said that, machine learning is a method of designing a series of small processes to solve a problem, known as an algorithm. (FSB, 2017, p.4) Hence, these algorithms transform raw data into intelligent knowledge. Machine learning algorithms take the data, identify the patterns that could be used and implement the model to produce outputs. (Lantz, 2013, p. 9) Thus, machine learning techniques implement precise models that analyse complex and big data and produce more accurate results. Machine learning models are able to adapt independently as learning from previous already existed computations to produce reliable decisions based on more accurate results. (Sas Institute, 2019) Moreover, machine learning systems can perform more complex processes by learning from data instead following predefined instructions. (The Royal Society, 2019)

The difference between machine learning and traditional programming can be seen in the figure below.

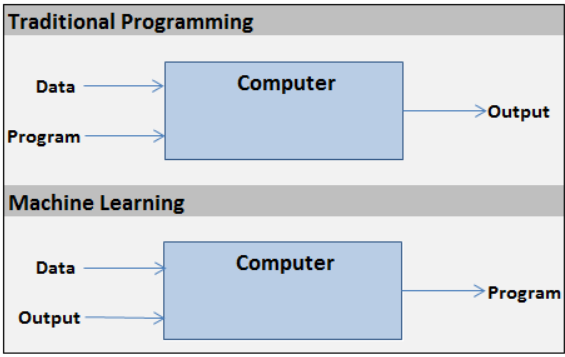


Figure 1Traditional programming and machine learning (Ajanki, 2018)

It could be also being said that machine learning as a subset of AI allows machines to learn without an explicit program with pre-programmed rules. In traditional program is needed input and program in a form of logic form and then return output. In contrast, machine learning algorithms get input and output and give a reusable program that can be applied to a new input to give one output. (Bhatia, 2018)

In addition, there are different machine learning techniques and algorithms that can be applied to deal with the same data problems, but achieve different levels of accuracy and desired results. (Usuelli, 2014) Choosing the right machine learning algorithm depends on the type of data that needs to be analysed and the specific task to be done, as well as the human involvement. For these reasons, machine learning is divided into two different groups: supervised learning algorithms used to build predictive models, and unsupervised used to construct descriptive models and one additional group, known as Reinforcement learning. (Lantz, 2013, pp.18-20)

Supervised learning

There is a specific target data, known as labelled data. The dataset is separated into training and test set. This type of algorithm learns by obtaining a set of inputs at the same time with correct outputs and comparing the actual inputs with the target outputs to detect the errors. Then the model has been modified accordingly. (Usuelli, 2014) In other words, supervised learning is not about human participation, rather the fact that the target data perform a supervisory role, which it shows the task that need to be done. The algorithm tries to optimize the function (model) to find the combination of the values of the characteristics leading to the target output. (Bonaccorso, 2017) In addition, it is called supervised because the process of learning from the training data can be seen as a teacher who supervised the process of learning. (Brownlee, 2016)

Supervised learning is mainly used in classification and regression tasks such as mortgage decisions or banking marking with different target campaigns, etc. (FSB, 2017, p.5) Classification task is when the output is a category, for example, the weather is “cold” or “hot”, while regression task produces the output as a real number such as the temperature during the weekend. Supervised learning can be presented as Y=f(X), where the x is input data and y is output data and algorithm learns with mapping function form x to y. (Brownlee, 2016)

Supervised learning is also used in a prediction of numeric data, for instance incomes, test scores or temperature in the further week, etc. In order to predict numeric values, a usual form of numeric predictions fit linear regression model to the input data. There are various supervised machine learning algorithms such as linear regression, decision tress, random forest, Support Vector Machine (SVM), logistic regression. Most of them perform both classification and regression tasks. (Lantz, 2013, p.21)

Unsupervised

Thereis no specific target data. The training actually is unlabelled without any outcomes to predict and estimate. (Usuelli, 2014) The system learns without references, correct answers and teacher. The unsupervised learning detects patterns within the data by identifying groups (clusters) based on similar characteristics. (FSB, 2017, p.5) It is usually used in retailing or customer basket analysis, etc. For instance, customers can be clustered (grouped) based on their interests and offered with items similar to desired goods or promotions. (Sas Institute, 2019) Lanz describes as a pattern discovery and used to find associations in the data. (Lantz, 2013, p.21) Unsupervised techniques are self-organizing maps, k – means clustering, Principal component analysis (PCA), Dimensionality reduction, hierarchical clustering, Kernel PCA, etc.

Reinforcement learning

This learningis applied in gaming, navigation systems, self-driving cars and robotics. By reinforcement, the algorithm detects through trials and errors which actions give the greatest rewards. (Sas Institute, 2019) Moreover, these algorithms are goal-oriented to complete a specific task using a system of rewards (feedbacks) and series of actions to maximize these rewards. (FSB, 2017, p.5)

In recent years, there has been remarkable progress in machine learning, which increased its capabilities in a number of applications. For example, it can be used in image recognition system as these that are used on social media; Voice recognition system used by Google assistant, Siri, Alexa(The Royal Society, 2019); filter email spam, prevent fraud credit activities, self-driving vehicles, target advertising and marketing, weather forecast and natural disasters estimates, produces better financial predictive analysis, etc. (Lantz, 2013, p.8) Machine learning is not able to perform everything, such as determining causation. In general, machine learning algorithm is used to identify patterns. Some of these patterns are unknown to the human. However, machine learning is increasingly used to understand and solve complex problems along with other different tools and domain experience. (FSB, 2017, p.6)

This report is considered in supervised machine learning, in particular, linear regression and decision tree. The following sections provide explanation of both techniques.

## Linear Regression

Many researchers are interested in the relationship between one variable and other variables. For instance, how the house prices depend on the size of the house. The relationship between these two variables falls within the scope of regression analysis. Accordingly, it is a statistical method that explores such a relationship. (Yan, 2009, p.2) In addition to that, other researchers describe regression analysis as a statistical technique that investigate and make a model of relationship between variables. This is most used statistical technique and applied in almost all areas such as engineering, economics, marketing, banking, etc. (Montgomery, 2012, p.1) Regression analysis refers also to the term linear regression. (Yan, 2009, p.3) Linear regression is defined as a supervised machine learning approach useful for predicting quantitative outcome. (James, 2017, p.59) Furthermore, the aim of this technique is to build a mathematical model for presenting and explaining the existing relationship between the variables. (Seber, 2003, p.3)These variables are called dependent variables (known as predicted variables) typically presented by *y* and independent variable, known as predictors) presented by *x*. (Yan, 2009, p.2)

In short, linear regression is a statistical model used to explain and present the relationships between independent and dependent variables. There are three types of linear regression depending on the number of independent variables used. (Yan, 2009, p.2) The models assume that the dependant (predicted) variables are continuous. Linear regression widely used in various engineering sectors, marketing, finance and banking, etc. (Lantz, 2013, pp. 159 -161),

### Simple linear regression model

This model is straightforward approach to present and predict the relationship between two variables. (James, 2013, p.61) One of them is single predictor *x* and another variable is the dependant *y*. The simple linear regression can be represented by the mathematical form:



Figure 2 Simple linear regression model (Montgomery, 2012, p.1)

Where *y is* the predicted variable, β0is *y* intercept, β1 is the slope of regression line also known as regression coefficient, *x* is the predictor and ε is a component that present the random error. (Yan, 2009, p.10) The error is assumed to have zero mean or uncorrelated unknown variance. (Montgomery, 2012, p.1) The method to estimate the value of β0 and β1 is called Least Squares. This method tries to minimize the sum of the squared distance between actual value y and actual value *x*. (Montgomery, 2012, p.2)

### Multiple linear regression model

This is a model where the independent variable is more than one. (Hastie, 2001, p.50) The mathematical form of this model is:



Figure 3 Multiple linear regression model (Montgomery, 2012, p.4)

### Nonlinear regression model

This model (also known as growth model) assumes that dependent and independent variables are in nonlinear relationship. Can be presented by:

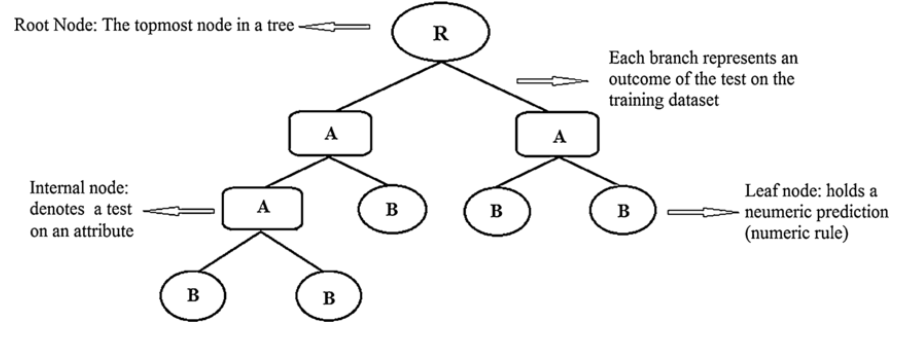


Figure 4 Multiple linear regression model (Yan, 2009, p.3)

This model is more complicated in terms of variable selections, model selections, outlier detection and estimation of the model parameters. (Yan, 2009, p.3)

## Decision trees

According to Lantz (2013, p.120), a decision tree is supervised machine learning techniques that build the model in a tree structure form. The decision tree model contains a series of logical solutions that can be presented in flowchart structure. Furthermore, the decision tree algorithm divides the dataset into two or more subgroups, so the sample in each subgroup is more consistent than in the previous subgroup. This process is called a recursive process and is repeated until homogeneity of the criterion is achieved. (Elsalamony, 2014, p.14) A graphical representation of the structure of decision tree is shown in Figure 5.

Figure 5 Decision tree structure (Brid, 2018)

At the top of the structure is the root node. Branches represent the rules of the decisions, which decisions are indicated by internal nodes, and finally determinate by leaf nodes (terminal nodes) that contain the outcome of the decisions (categorical or continues values), also indicating the class labels. (Lantz, 2013) Each branch of the decision tree responds to a different outcome. (Elsalamony, 2014, p.14)

There are various performances of decision trees, but one of the most famous is the C5.0 algorithm developed by Ross Quinlan as an improved version of his previous C4.5, which improved ID3 (Iterative Dichotomiser ) algorithm and use splitting algorithms, including entropy based on information gain. (Lantz, 2013, p.124) This technique works by splitting the data based on the attribute with highest information gain. Each new subset is split again based on different criteria until the process cannot be continuing further. At the end, the low-level splits are reviewed and these that do not have significant contribution to the result of the model are removed or pruned. (Elsalamony, 2014, p.14) The pruning process actually reduces the size of the decision trees to obtain more meaningful and accurate results. In order to avoid the growth of large decision trees, early stop, also known as pre-pruning, is applied. This process avoids unnecessary splitting. In this manner, post-pruning can be involved in growing a large tree then by pruning criteria based on the errors of the node’s rates to reduce the size. Hence, this approach is more effective because it is difficult to determine the depth of the decision tree without growing. (Lantz, 2013, p.128) In additional, C5.0 is quite beneficial when some values are missing and a large number of inputs occur as problems and increase the accuracy of the results as a measure the purity, using entropy which is based on an information gain. (Elsalamony, 2014, p.16) The entropy can be specified by the mathematical formula:

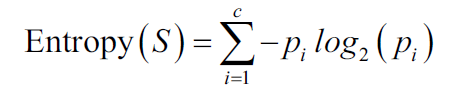


Figure 6 Entropy of the class (Lantz, 2013, p.125)

In this formula, the *S* is different class label and *p* presents the proportions of the values in the class *i.*

The information gain is calculated as difference between the entropy of class before splitting and partition after the split (S2).



Figure 7 InfoGain (Lantz ,2013, p.126)

Another many used algorithm is CART (Classification and Regression Tree) by Breiman, Friednam and C.J. Stone AND Olsehn. (Lantz, 2013, p.187) CART use Gini method as an attribute selection to compute the impurity of the sample. CART model is represented as a binary tree. The R implementation of CART algorithm is known as RPART (Recursive Partitioning and Regression Trees).The *rpart* algorithm divides the dataset recursively, which means that the subgroups that result from the division are further split up to a predetermined termination criterion. (Brownlee, 2016) At each stage, the division is made on the basis of the independent variable that leads to the lower value of Gini and higher possible reduction in heterogeneity of the predicted (dependent) variables. (Zaki, 2014, p. 487) Can be presented by:

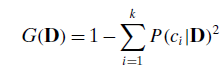


Figure 8 Gini Index (Zaki, 2014 p. 487)

Decision trees can perform both classification and regression problems. (Zaki, 2014, p.485) Thus, there are two types of decision trees: classification and regression. Classification tree is used to predict qualitative output rather than regression for quantitative. It could also be said that, classification tree used for categorical predicted outcome such as yes and no, and regression tress used for continuous (real number) such as a price of houses (James, 2017, pp. 303 -316)

This method is widely used in the banking sector because of their accuracy and the ability to express a statistical model in a simple language. (Lantz, 2013, p.128) The research on this problem is presented in the next section.

## Machine learning techniques in banking marketing

Among the different financial organisations, banks are the main component and most active participant in the financial industry. (Nataraja, 2018, p.1) Performance of the banking has an important role in financial stability at all. In additional, globalization and technological advances has created an extremely competitive market. This affects the banks as well. (Bakar, 2009, p.176) In recent years, banking and direct database marketing have become an important strategy for understanding the customer needs. (Apampa, 2016, p.85)Furthermore, marketing activities seek new tactics to promote the goods and services to the customers. (Grzonka, 2016, p.36)Hence, there is a need of decision makers in order to make more accurate decisions. In order to make better prediction statistical tools and methods are used. Choosing the right methods is still a valid issue. (Bakar, 2009, p.176)Machine learning techniques are widely used in banking marking because of their ability to recognize patterns and calculate relationship in the dataset. (Choudhry, 2018, p.4)

There are various machine learning techniques used in existing dataset in banking marking for creating effective decision making knowledge. Bakar, N., Tahir, I., (2009, p.176) used multiple linear regression and artificial neural network to Malaysian banks data. From their study, it can be concluded that multiple linear regression can be used as a simple tool for predicting and exploring the relationship between the dependent variable and independent variables. This technique explains in simple manner the important variables of the bank performance and the effect of the contributing factors. By the same token, Elsalamony (2014, p.12) compared four techniques: back propagation of neural network (MLPNN), Logistic regression analysis (LR), naïve Bayes classifier (TAN), and the decision tree model (C5.0) on the dataset direct marketing campaigns of a Portuguese banking institution to classify the bank deposit subscription. The experiment has shown that C5.0 has better performance than others and the most important attribute is duration. However, on the same dataset Grzonka, D., Suchacka, G., Brorowik,B,(2016, p.46) applied four classification methods: decision trees, bagging, boosting and random forests. They achieved best results using random forest. Also, best results using random forest, Apampa(2016, p.100) has achieved after applying classification algorithms such as Logistic Regression, Naïve Bayes ,Decision Tree and Random Forest ensemble. However, in this research, attributes as duration, contact, month and housing significant contribute to the deposit description. Hence, the best results can be obtained by finding the right techniques and attributes.

In this project is used the dataset from a direct marketing campaign of a Portuguese banking institution, used in the above surveys. This report will discuss the use of linear regression and a decision tree as predictive models to convert raw data into decision-making knowledge, in which case, to predict whether the client will subscribe for a term deposit.

# Chapter 2 Proposed solution and methodology

## 2.1 Proposed solution

The aim of the project is to find how to use machine learning techniques for analysis and making predictions using existing dataset in banking marketing. To find how they can be used together in a process of converting raw data to effective decision making knowledge.

The dataset used for this project relates to direct marketing campaigns of a Portuguese banking institution. The purpose of building the models is to predict whether the client will subscribe for a term deposit. In order to accomplish the project problem, the proposed solution can be described in following steps:

**Step.1 identifying the problem, collecting, describing the dataset**

The dataset is downloaded from UCI Irvine machine learning repository ([https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)). This step includes a detailed description of the dataset and identifying the scope of analysis.

**Step.2 Exploring, data cleaning and preparation**

This includes understanding and applying the data cleaning techniques for missing data and attributes that do not contribute significantly to the results. The preparation of the data includes splitting the dataset into a training and test dataset.

**Step.3 Building the models**

Building models such as linear regression and Decision trees and training the model on the data, in this project, simple and multiple linear regression, and C5.0 and CART decision tree.

**Step.4 Test and evaluate the performance of the models**

This stage of the project analyses the archived results and performance of the modules.

**Facilities**

R and R programming environment (R studio)

## 2.2 Methodology

In order to find how to use machine learning techniques for analysis and making predictions using existing dataset in banking marketing are used two machine learning – Decision trees and Simple and Multiple Linear Regression, in particular – CART based on Gini method of attribute selection and C5.0 – Entropy. Data cleaning and preparation methods are applied such as statistical methods to check for missing values and outliers, converting the datatype, visualization of variables using *ggplot*, method of dividing the dataset into training and test data, confusion matrix to evaluate the accuracy of the models and Scatterplot matrix for visualization of the correlation between the quantitative features. The analysis is performed by R language and environment. R is free software for statistical computing. R language is a function based. To take advantages of R, R packages are installed and reloaded each time by libraries with the same name. (Kohl, 2015, p.10) In this report, all R libraries and packages are always mentioned in the corresponding sections of the data preparation, building and evaluation of the model result.

# Chapter 3 Legal, Ethical and Social Approval

In this project is used collected data, so legal, ethical and social standards are considered. The dataset is downloaded from UCI Irvine machine learning repository.

The used dataset not includes any personal information such as names, address, phone number, bank accounts.

# Chapter 4 Description of the dataset

## 4.1. Dataset information

The dataset used in this project has been downloaded from UCI Irvine machine learning repository ([https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)). The dataset relates to direct marketing campaigns of a Portuguese banking institution. The campaigns are based on phone calls to the clients. Typically, it requires more than one contact with the same customer to determine if the bank term deposits will be (or not) subscribed. (UCI, 2014) The purpose of building the models is to predict whether the client will subscribe for a term deposit and will focus on marking efforts on these clients. On the other hand, by analysing the marking campaign that the bank has made most recently and the identification of the patterns will help to draw conclusions in order to improve further marketing campaigns. (Bachmann, 2019)

The files with data within the bank-additional folder are:

1. **bank-additional-full.csv**- includes 41188 examples and 20 inputs in an order by date
2. **bank-additional.csv** – includes 4119 examples (10%) in a random selection from 1and 20 inputs
3. **bank-additional-names.txt** – overall description of the dataset

The files with data within the bank folder are:

1. **bank-full.csv** – includes all examples and 17 inputs in an order by date (older version)
2. **bank.csv –** includes 10% of all examples and 17 inputs in a random selection from 3 (older version)

In this project is used bank-additional-ful.csv file. This dataset contains 41188 records with 21 observations (attributes) per record. Each record has 20 explanatory observation about the client and 1 observation of a term deposit (subscribed or not).

## 4.2 Attribute Information

### Client data

* **Age** – Age of the client – (numeric)
* **Job** – Client’s type of job – (categorical: admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown )
* **Marital** – marital status – (categorical: divorced, married, single, unknown, note: divorced means divorced or widowed)
* **Education** – education level – (categorical: basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)
* **Default -** if the client has credit in default – (categorical: no, yes, unknown)
* **Housing –** if the client has housing loan – (categorical: no, yes, unknown)
* **Loan –** if the client has a personal loan **–** (categorical; no, yes, unknown)

### Campaign data attributes

* **Contact –** communication contact type – (categorical: cellular, telephone)
* **Month –** month of last contact with client **–** (categorical: January – December)
* **Day –** day of the week of last contact with client – (categorical: Monday –Friday)
* **Duration –** duration of the last contact with client, in seconds – (numeric). This input is for benchmark purpose only and not reliable for predictive model because highly effect the output target (if the duration is 0 the target (y) is no)

### Other data attributes

* **Campaign –** Number of contacts with the client during this campaign including last contact – (numeric)
* **Pdays –** Numberof days from last contact with client from the previous campaign**-** (numeric; 999 means customer was not contacted previously)
* **Previous** – Number of the contact before this campaign with this client –(numeric)
* **Poutcome**- Outcome of the previous marketing campaign –(categorical: failure, nonexistent , success

### Socio-economic attributes

* **Emp.var.rate -** Quarterly indicator of employment variation rate – (numeric)
* **Cons.price.idx –** Monthly indicator of consumer price index - (numeric)
* **Cons.conf.idx -** Monthly indicator of confidence price index - (numeric)
* **Euribor3m -** Daily indicator of euribor 3 month rate –(numeric)
* **Nr.employed -** Quarterly indicator of number of employees – (numeric)

Output variable (desired target) – Term Deposit **-** subscribe for a term deposit –(binary : yes or no)

# Chapter 5 Understand banking marking dataset

## 5.1 Preparing the directory

Before importing the data, the working directory in R environment should be set up. Two functions are used: *getwd()* to check the working directory in the window *Console* and *setwd()* to change the path of the directory that includes the folder with the stored dataset. (Willems, 2018) In this project, the dataset is stored in the same folder as the R file. The figure below presents the working directory.



Figure 9 Project’s working directory

## 5.2 Load the dataset

The banking marking dataset used for this project is stored in *CSV* format (Comma Separated Values). In order to load this data into R, the function *read.csv()* is used. It is depicted in figure 10, where the *header* argument is set to *true* to interpret the first row of the file as the attribute names and separated by a semicolon using the *sep* argument. The argument *stringsAsFactors* allows determining whether strings are converted as factor variables or not. The default value of the *stringsAsFactors* is *true*. (Lantz, 2013) In this case, the default value is accepted in order to present categorical data. The factors contain a predefined set of values sorted in alphabetical order, knows as *levels.* (Engel, 2019)



Figure 10 loading the data

## 5.3 Exploring of banking marking dataset

The next step after loading data into R studio is to examine the dataset to understand all attributes and values and to use it more effectively when building the predictive models. Better understanding of the data helps to better implement machine learning techniques to solve a data problem.

### 5.3.1 Create the data frame

When the dataset is loaded into R environment, a *BankMarkeringData* object is created in a data frame structure. Data frame is a special list of elements of any type, but they are of equal length. (Engel, 2019) The main advantage of this data structure is to store the data in a form of a spreadsheet, where each column represents an attribute and each row contains values for that column. This is an extremely useful data structure when the dataset has many attributes in different types. (Bali, 2016, p.21) The *class ()* function is used to check the class (type of element) of the object, as shown in the following figure. (Engel, 2019)



Figure 11 the class of object

### 5.3.2 Structure of the dataset

Each column is represented in a vector where all values have the same data type such as integers, factors, characters, etc. The *str()* function inspects the structure of the data frame and provides more details about the data in each column. The statement *obs*. tells how many observations or examples the data include. (Lantz, 2013, p.43) Figure 12 depicts the structure of the banking marketing dataset. It expects 41188 observations with 21 variables, which are a combination of integer, numeric and factor data types. The factors contain a predefined set of values sorted in alphabetical order, knows as *levels*.

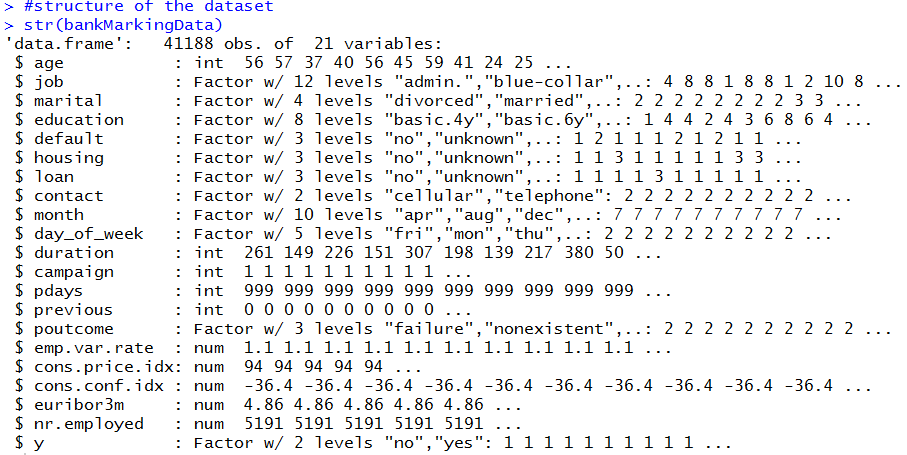


Figure 12 Structure of the banking marketing dataset

### 5.3.3 Dimension of the dataset

The size of the dataset, known as a dimension of the object, can be verified with *dim()* which returns a vector. The first element is the number of rows and the second is the number of columns. Figure 13 shows the dimension of the dataset with 41188 rows and 21 columns. (Lantz, 2013)



Figure 13 Dimension of dataset

### 5.3.4 Content of the dataset

In order to understand in more details the banking marking dataset, R provides function *colnames()* to observe the columns. As shown in the figure below, this function presents in a comprehensible way the names of each column that are included in this dataset. In addition, this would help to build models such as giving good knowledge of all attributes and choosing those who make a significant contribution to the results. (Engel, 2019)

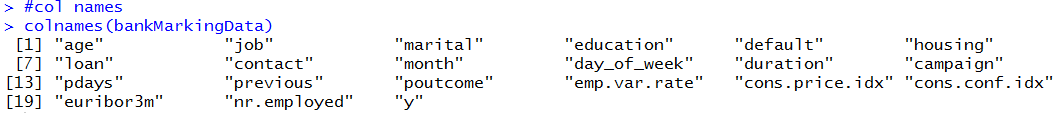


Figure 14 Column names of banking marketing dataset

The *head ()* and *tail()* functions are used to examine the content of the data by displaying the first 6 and last 6 rows. They provide a quick overview of the data as shown in figure 15. Another significant benefit of using these functions is to check if the file is read correctly with all attributes and values. (Matloff, 2011)

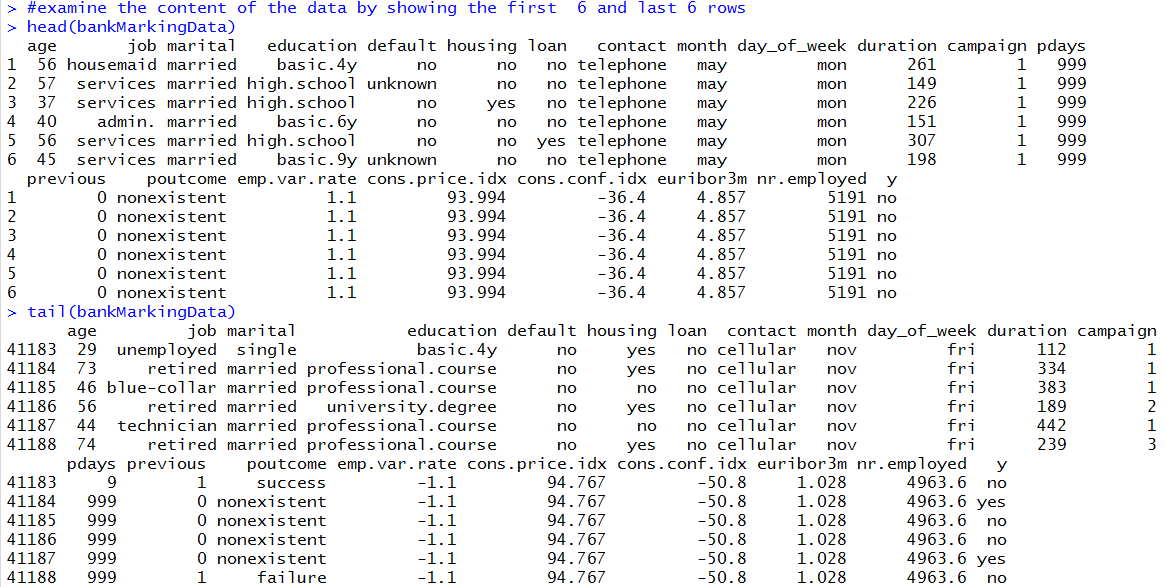


Figure 15 Overview of first 6 and last 6 rows

### 5.3.5 Summary of the data

The command *summary()* reports the summary of descriptive statistics for any numeric column and count statistics on each categorical column. For numeric variables *summary()* gives the range of *min* and *max* value, *quartiles*, *median* and *mean* and for the categorical variable gives the count number of values for each category. For instance, the age of the customers is represented in a range of min (17) and max value (98) and the job is divided in different categories as admin job has the highest number. Also, the numeric variables have several quartiles. The first quartile, known as lower quartile, is a value that cut off the first 25% of data when sorted in ascending order. The second quartile, called as median is the value that cut off the first 50%. The upper quartile is the third one which cut off 75 % of the data. They tell how much of the data is below a certain value. (Yau, 2019)

*Summary ()* is a most useful tool in R commands for investigating the data because provides a full overview of all attributes and helps to avoid various data issues as missing values, invalid values, outliers and too wide or narrow data ranges. (Zumel, 2016, p.4)

As shown in figure 16, *summary ()* command is used to observe the banking marking dataset in an early stage. It provides information about 21 observations with their values. As seen below some of categorical attributes are *unknown* or *other* (not equal to NAs). These values can be accepted as class labels. However, they can be treated as missing values and removed by data cleaning techniques. In this project, they are assumed as a class label. There are two attributes (*emp.var.rate* and *cons.conf.idx*) with negative values by default. The attribute *pdays* has value of 999 that means customer has not contacted previously.

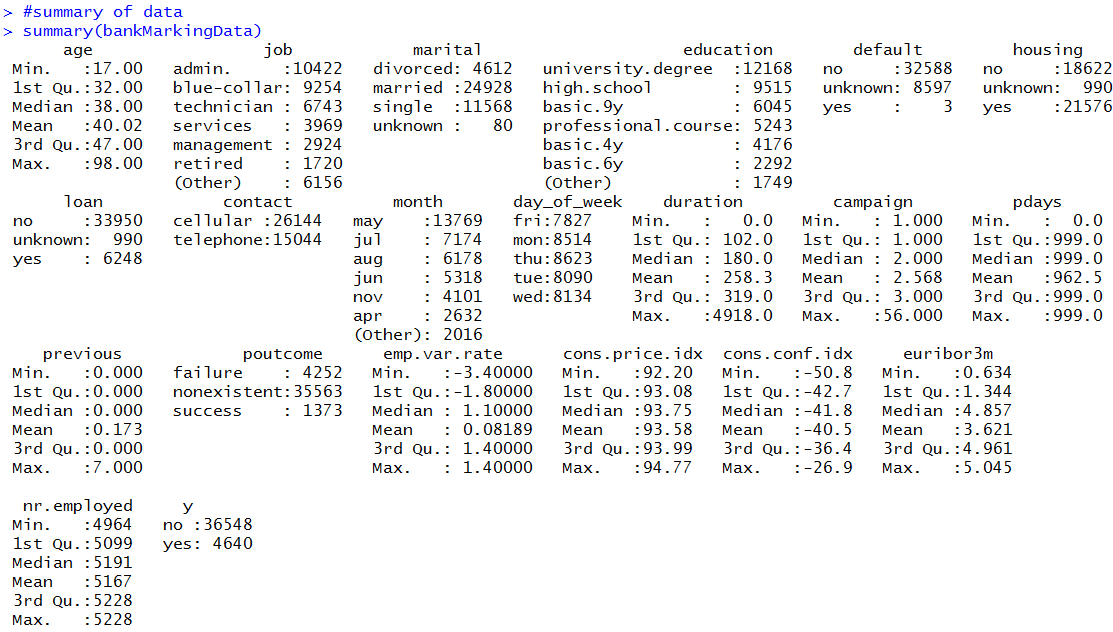


Figure 16 Summary of data in early stage

As seen in figure above, the target variable with attribute name *y* represents whether the client subscribed for a term deposit. More customers have not subscribed for a term deposit. The percentage proportion of *yes* and *no* valuesis calculated as follows. *Table()* function of R set of commands is used to produce contingency table, known as a *cross tabulation*. The table represents the frequency distribution of categorical variables in a matrix format. The table outputs provide the list of categories and a count of the number of values in each category. R has ability to calculate the percentage proportion directly by *prop.table()* command multiply by 100 and rounded by *round()* function to the first digit after the decimal point. (Lantz, 2013, pp. 56-57) The target percentage is depicted in the following figure. As shown, the percentage of customers who have subscribed to a term deposit is 11.3%, which is lower than percentage of customers who are not subscribed.

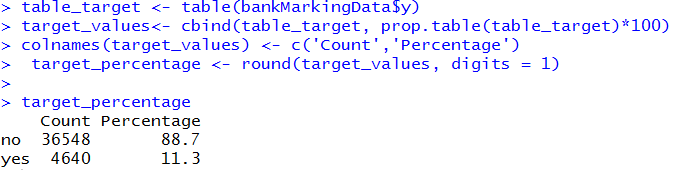


Figure 17 Target percentage proportion

# Chapter 6 Data cleaning and preparation methods

Data cleaning and preparation includes cleaning the data form missing values, outliers and attributes that do not contribute significantly to the results, converting the categorical (factor) variables into machine-readable format and visualizing results, and dividing the dataset into training and test set. (Lants, 2013)

## 6.1 Missing values

The missing values can be a significant problem in analysing the data and affecting the accuracy of the achieved results. In order to prevent that, missing values should be removed at an early stage. (Lantz, 2013, p.280) There are various approaches to check the missing values in a dataset. In this project are used two of them by using R command *is.na()* and a deep check by *if-else* statement.

The command *is.na()* return true if values are missing, then the sum of them is computed by *sum ()* and returned to each column by *sapply().* As seen in figure below there are no missing values in banking marking dataset.

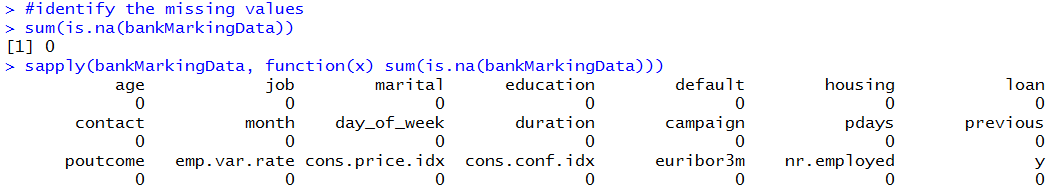


Figure 18 No missing values in banking marking dataset

A deep check of the missing values is shown in the figure 19 using the *if-else* statement. The R function *which()* return the position of element when it meets the specified condition. (Kabacoff, 2017) In particular, it returns the message *“No missing values are found”* when the position of the element is *True* in a logical vector.

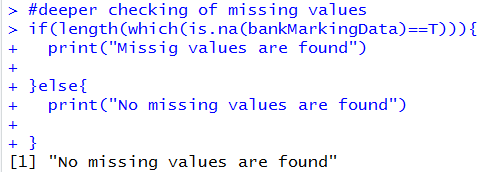


Figure 19 if-else statement for missing values

## 6.2 Outliers

Outliers are extreme values that are distant from majority of other observations. (Brownlee, 2013) These are data points that differ significantly from other data. The outlier is also indicated as deviant, discordant, abnormality or anomaly in statistical literature. All data points are placed on continuous spectrum from normal data up to noise and at the end to anomalies as shown in figure 20. The outliers are distinguished as weak and strong. It depends on how they deviate from the rest of the data. (Aggarwal, 2013, pp.1-5)

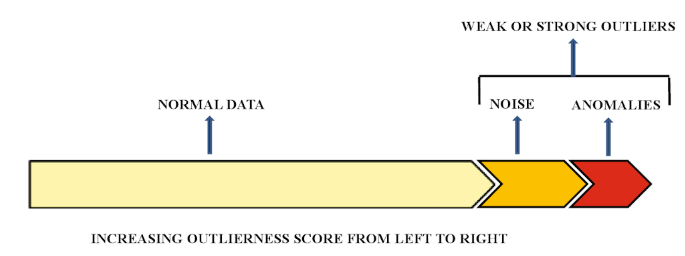


Figure 20 The continuous spectrum from normal data to outliers (Aggarwal, 2013, p. 4)

These values can distort and mislead the training process and lead to longer training time, poor model performance and less accurate results. (Brownlee, 2013) Hence, the outlier detection is an important stage of data analysis. (Seo, 2006) The process of identifying outliers as also refers as outlier modelling, outlier mining, anomaly and novelty detection. There many outlier detection methods such as statistical methods and extreme values analysis, proximity methods based on clustering methods (such as k-means algorithm), projection methods based on higher-dimensional methods (such as PCS, SOM or Sammon’s mapping). (Brownlee, 2013)

### 6.2.2 Outlier detection

The outlier detection methods used in this project are the statistical methods based on extreme value analysis that focuses on univariate methods, visualized by box and whisker plot (boxplot) and a distribution (Gaussian) method.

#### 6.2.2.1 Gaussian distribution method (normal distribution)

This method assumes that during each measurement the values follow the normal distribution with an equal number of measurements below and above the mean value. It is illustrated by a bell-shaped curve. (ScienceDirect, 2018)There are three standard deviations from mean values used to identify the outliers in the Gaussian distribution method: 1 SD (68%), 2 SD (95%), 3 SD (99.7%). In order to identify the outliers in a large sample of data can be used 3SD. The observations which are outside of the range of these deviations might be considered as outliers. (Brownlee, 2018)

The following figure shows the Gaussian distribution (normal distribution) of the values within 1, 2 and 3 standard deviation from the mean value that about 68%, 95%, 99.7%, respectively.

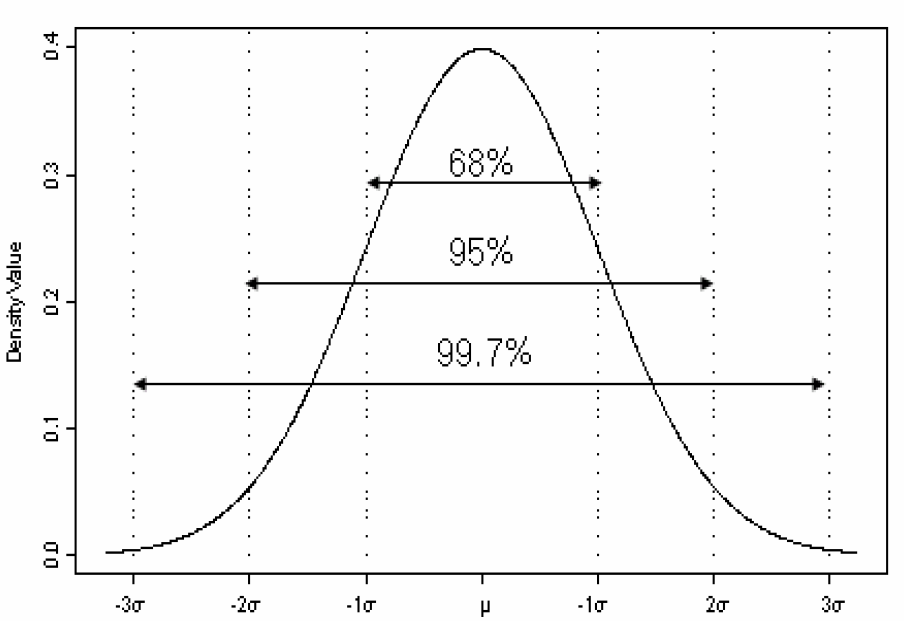


Figure 21 Gaussian distribution of the values within the standard deviation from the mean (Seo, 2006, p.5)

In this project, the mean and standard deviation of the each attribute is calculated and then outliers are identified as more or less than 3SD from the mean value, as shown in figure below.

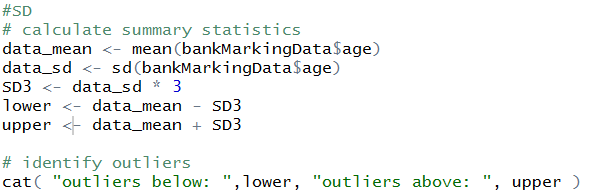


Figure 22 Outlier detection of Age using 3SD in R

Detailed results about outlier detection using Gaussian distribution method (normal distribution) of each attribute within third standard deviation of the mean are given in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attributes** | **Mean** | **SD** | **3SD** | **Lower** | **Upper** |
| Age | 40.02 | 10.43 | 31.26 | 8.76 | 71.29 |
| Campaign  Previous | 2.568  0.173 | 2.77  0.495 | 8.31  1.48 | -5.74  -1.31 | 10.88  1.66 |
| Poutcome  Emp.var.rate  Cons.conf.idx Cons.price.idx  Euribor3m  Nr.employed  Marital  Job  Education  Default  Housing  Loan  Contact  Month  Day-of-week  Duration  Pdays | 1.93  0.082  -40.5  93.58  3.62  5167  2.17  4.72  4.75  1.21  2.07  1.33  1.37  5.23  3.00  258.3  962.5 | 0.36  1.57  4.63  0.58  1.73  72.25  0.61  3.59  2.14  0.41  0.96  0.72  0.48  2.32  1.39  259.28  186.91 | 1.48  4.71  13.88  1.74  5.20  216.75  1.83  10.78  6.41  1.22  2.96  2.17  1.44  6.96  4.19  777.84  560.73 | -1.31  -4.63  -54.39  91.84  -1.58  4950.28  0.35  -6.06  -1.66  -0.01  -0.88  -0.84  -0.07  -1.73  -1.19  -519.54  401.77 | 11.66  4.79  -26.62  95.31  8.82  5383.79  3.99  15.51  11.16  2.43  5.03  3.50  2.81  12.19  7.19  1036.14  1523.23 |

Table 1 Normal distribution of each attribute within third standard deviation of the mean

As seen in the table above, it can be assumed that several attributes such as *age, camping, cons.conf.idx* have outliers. The attribute *pdays* has a number 999. It means the clients have not been previously contacted. Thus, they cannot be considered as outliers. However, not all data can be treated by a Gaussian distribution. The following sector provides explanation of other outlier candidate detection methods – boxplot and IQR rule. They are efficient statistic techniques for summarising data that cannot be treated by a Gaussian distribution. (Brownlee, 2018) They are very useful methods ever since do not make assumptions about distribution, nor do not depend on the mean value or standard deviation. (Seo, 2006)

#### 6.2.2.2 Box-and-whisker plot (boxplot) and IQR rule

One of most important and at the same time simplest graphical display in descriptive statistics is the box-and-whisker plot (boxplot), also known as Tukey’s method. The boxplot provides well summarised information of median, lower (Q1) and upper (Q3) quartiles, the inter-quartile range (IQR) is the calculated interval between Q1 and Q3. The outliers are individual plotted points that out of the range. (Seo, 2006) The box-and-whisker plot is illustrated in the following figure.

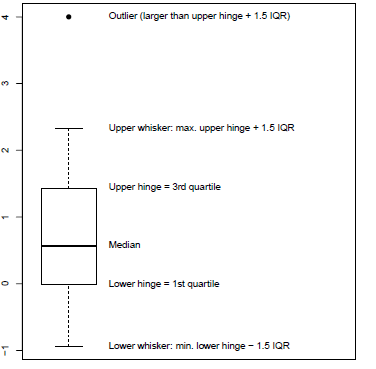


Figure 23 Box-and whisker plot (Kohl, 2015, p. 40)

The proposed method above can be referenced to Interquartile Range method, known as IQR range rule. It’s an efficient statistic technique for summarising data that cannot be treated by a Gaussian distribution. This method defined Q1-(1.5\*IQR) and Q3+ (1.5\*IQR) as *inner fences*, Q1-(3\*IQR) and Q3+ (3\*IQR) as *outer fences.* (Seo, 2006) The outlier can be found by determining where the observation is related to inner and outer fences. If one observation is more extreme than one of outer fences, then it is an outlier and in particular is called a *strong* or *probable outlier.* If one data value is between the respective internal (inner) and external (outer) fences, then this value is a suspicious outlier, known as a *weak* or *possible outlier*. (Aggarwal, 2013, p. 4)

The method of *boxplot* is performed and illustrated in the following figures, where the attribute age is used. In order to obtain the boxplot of the variable, the *boxplot ()* is used. In addition, several parameters are specified such as *main* to add the title of the figure and *xlab* (the horizontal axis). The function *out()* and *print()* find and print the outliers, respectively.(Lantz, 2013, p.50)

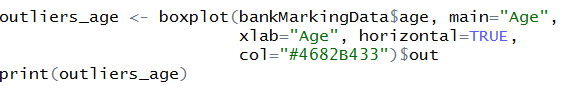


Figure 24 Boxplot () in R

As seen in figure 25, the boxplot method detects the outliers over 70.

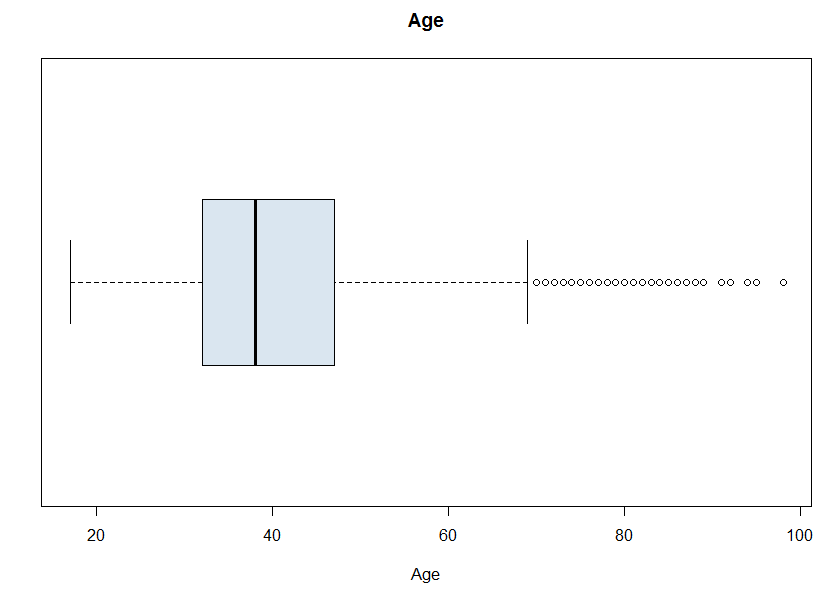
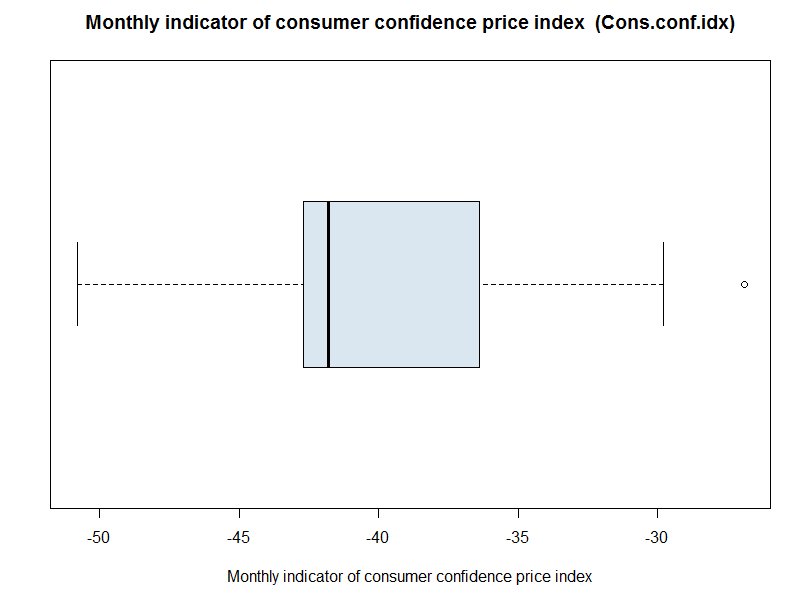


Figure 25 Boxplot visualization of Age with outliers

The boxplot method of *cons.conf.idx* detects outlier values after -30. A sample of printed outliers is shown in figure below.



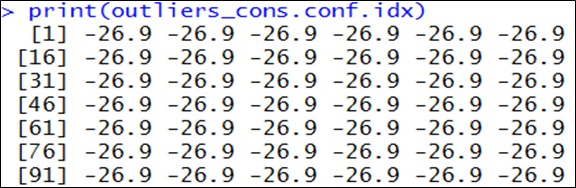


Figure 26 cons.conf.idx outliers

By the same token, other attributes for outliers were checked. As shown in figure 27 , the boxplot preview of *nr.employed, eurbor3m, emp.var.rate, cons.price.idx* do not detect any outliers that could affect the accuracy of the results.

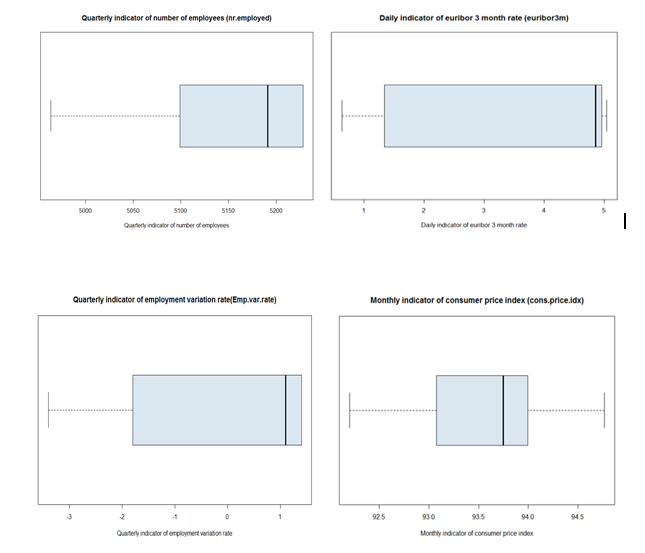


Figure 27 Boxplot visualization of other attributes

The following figure represents the IQR with inner and outer fences.

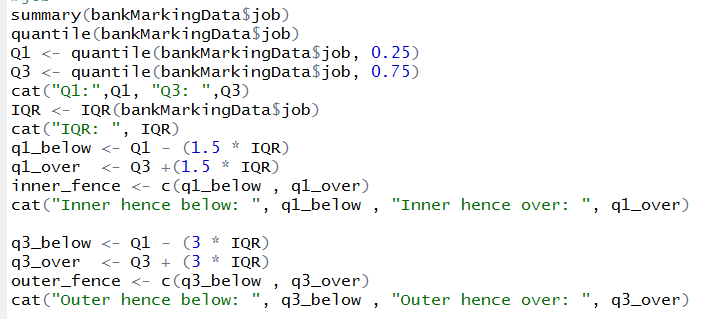


Figure 28 IQR range in R

Detailed results about outlier detection using Interquartile Range (IQR) method of each attribute are given in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Attributes** | **Median** | **Q1** | **Q3** | **IQR** | **Inner fences** | **Outer fences** |
| Age | 38 | 32 | 47 | 15 | 9.5 69.5 | -13 92 |
| Job | 3.00 | 1 | 8 | 7 | -9.5 18.5 | -20 29 |
| Marital | 2.00 | 2 | 3 | 1 | 0.5 4.5 | -1 6 |
| Education | 4.00 | 3 | 7 | 4 | -3 13 | -9 19 |
| Default | 1.00 | 1 | 1 | 0 | 1 1 | 1 1 |
| Housing | 3.00 | 1 | 3 | 2 | -2 6 | -5 9 |
| Loan | 1.00 | 1 | 1 | 0 | 1 1 | 1 1 |
| Contact | 1.00 | 1 | 2 | 1 | -0.5 3.5 | -2 5 |
| Month | 5.00 | 4 | 7 | 3 | -0.5 11.5 | -5 16 |
| Day\_of\_week | 3.00 | 2 | 4 | 2 | -1 7 | -4 10 |
| Duration | 180.0 | 102 | 319 | 217 | -223.5 644.5 | -549 970 |
| Campaign | 2.00 | 1 | 3 | 2 | -2 6 | -5 9 |
| Pdays | 999 | 999 | 999 | 0 | 999 999 | 999 999 |
| Previous | 0 | 0 | 0 | 0 | 0 0 | 0 0 |
| Poutcome | 2.00 | 2.00 | 2.00 | 0 | 2 2 | 2 2 |
| Emp.var.rate | 1.10 | -1.80 | 1.40 | 3.2 | -6.6 6.2 | -11.4 11 |
| Cons.price.idx | 93.75 | 93.08 | 93.99 | 0.92 | 91.69 95.37 | 90.318 96.75 |
| Cons.conf.idx | -41.8 | -42.7 | -36.4 | 6.3 | -52.15 -26.95 | -61.6 -17.5 |
| Euribor3m | 4.86 | 1.34 | 4.96 | 3.62 | -4.08 10.39 | -9.51 15.81 |
| Nr.employed | 5191 | 5099 | 5228 | 129 | 4905.6 5421.6 | 4712.1 5615.1 |

Table 2 IQR with inner and outer fences for each attributes

As seen in the table above, it can be assumed that several attributes such as *age, campaing, cons.conf.idx, default, loan, previous, poutcome*. They can be considered as outliers because they are more extreme than the observation of the outer fences. In addition, by comparing the median value with the quartiles, it is shown whether the data is skewed. Furthermore, the results of outlier detection may depend on the methods or distribution of the data. However, need to be check if there is any improvement of analysis before removing them. (Seo, 2006) In this case, they do have not significant to the result.

## 6.3 Distribution and visualization of variables

The use of graphics for data examination is known as *visualization*. (Zumel, 2016, p.9) Visualization of variables is useful in term of detecting data problems. At the same time, data visualization provides knowledge about the distribution of variables. (Lants, 2016, p.49) Among the many visualization systems in R, *ggplot2* is one of the most versatile and used for declarative graphing. The main advantage of *ggplot2* is the application of graphics grammar and provides a consistent system of defining and building the graphs. To take advantage of ggplot2, the package ggplot2 should be installed once and reloaded each time by *library(ggplot2).* *ggplot2* starts a plot with *ggplot()* function and creates a coordinate system. The first argument of *ggplot()* is the dataset used in the graph, in this project the data is *bankMarketingData*. It creates an empty graph to which, one or more layers of different types can be added by *geom* functions. The function *geom\_bar()* create a layer of bars to a plot. The function *geom()* has a *mapping* argument. This determines how variables in the dataset are mapped to visual characteristics. The argument *aes()* specifies the variables associated to x and y axes. The algorithm that uses *stat\_count()* to calculate new values for the graph is known as a *stat* (short term of statistical transformation). The default value of *stat* argument is *count,* mapped to the y- axis. In order to change the appearance of the legend and axis labels is used the function *theme().*(Grolemund, 2017) The practical application of *ggplot()* function is depicted in figure 29.

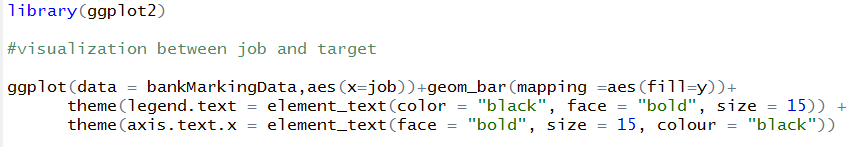


Figure 29 ggplot function

The function calculates and visualizes the distribution between two variables – job of the customers and target data (yes and no), shown in figure 30. The admin and technician are the job descriptions with the highest number of clients that subscribed a term deposit.



Figure 30 Graph of distribution between job and target data (y)

The visualization in the graph below shows the distributions between the age attribute and target data. As seen in figure 31, clients in the age range between 25 and 60 are subscribed the term deposit. Another, significant factor in this distribution is customers under 25 and 60 over are also subscribed the deposit. Accordingly, the all values in age attribute are used in building the predictive models in order to predict whether the client will subscribe for a term deposit and will focus on marking efforts on these clients.

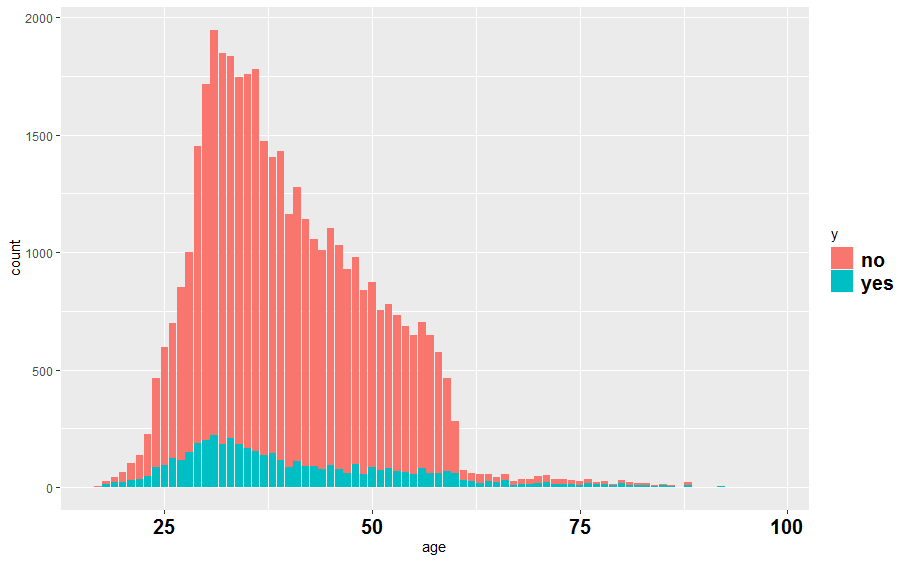


Figure 31 Graph of distribution between age and target data (y)

The figure below provides the graphical representation between marital statuses of the customers and target data. As shown in the graphic, married customers are a marital status with the highest subscriptions in a term deposit. Also, the observations in the job category *unknown* do not provide subscription for a term deposit.

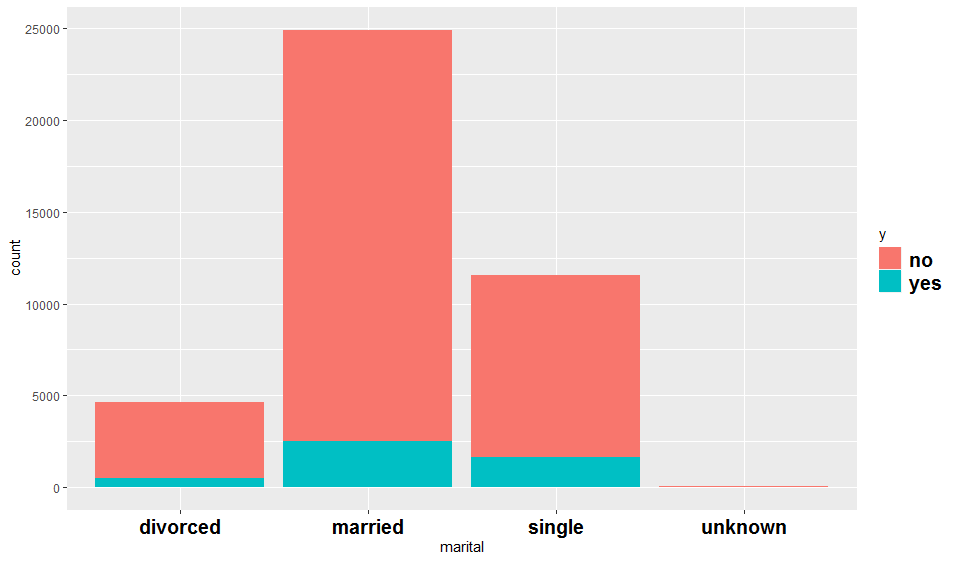


Figure 32 Graph of distribution between marital and target data (y)

As shown in figure 33, the customers who are highly educated have subscribed a term deposit. All illiterate people have not subscribed a term deposit.

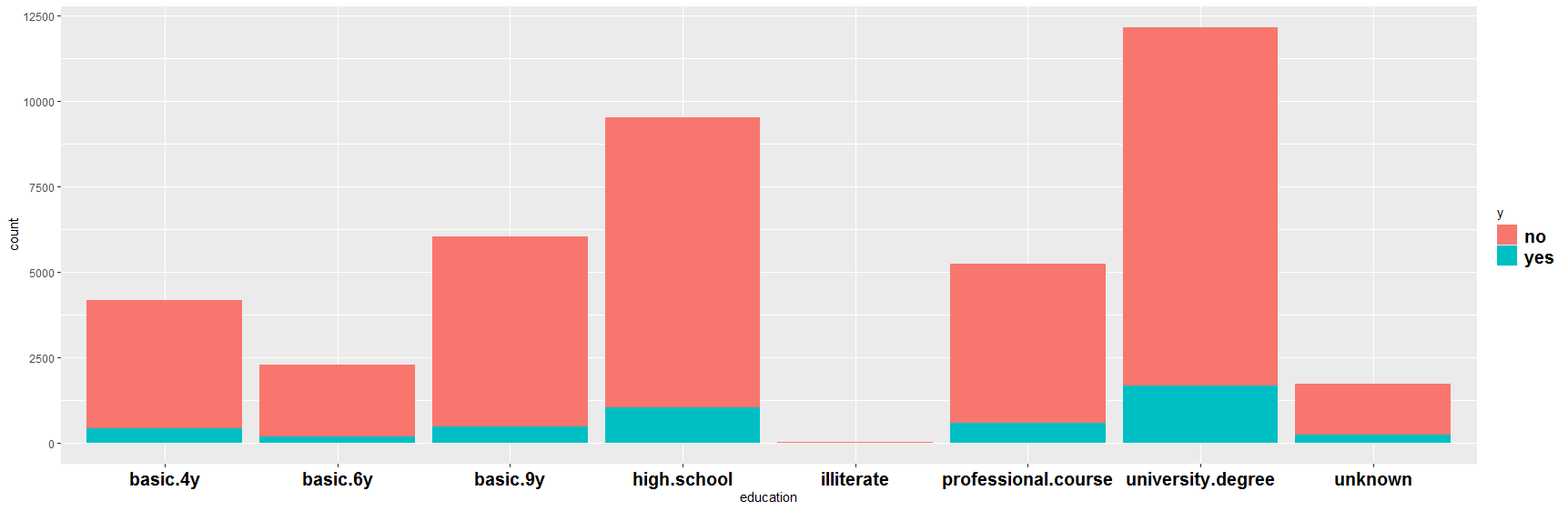


Figure 33 Graph of distribution between education and target data (y)

More customers are contacted by cellular phone than a home telephone. The graphical distribution between job and contact attribute is depicted in figure 34.

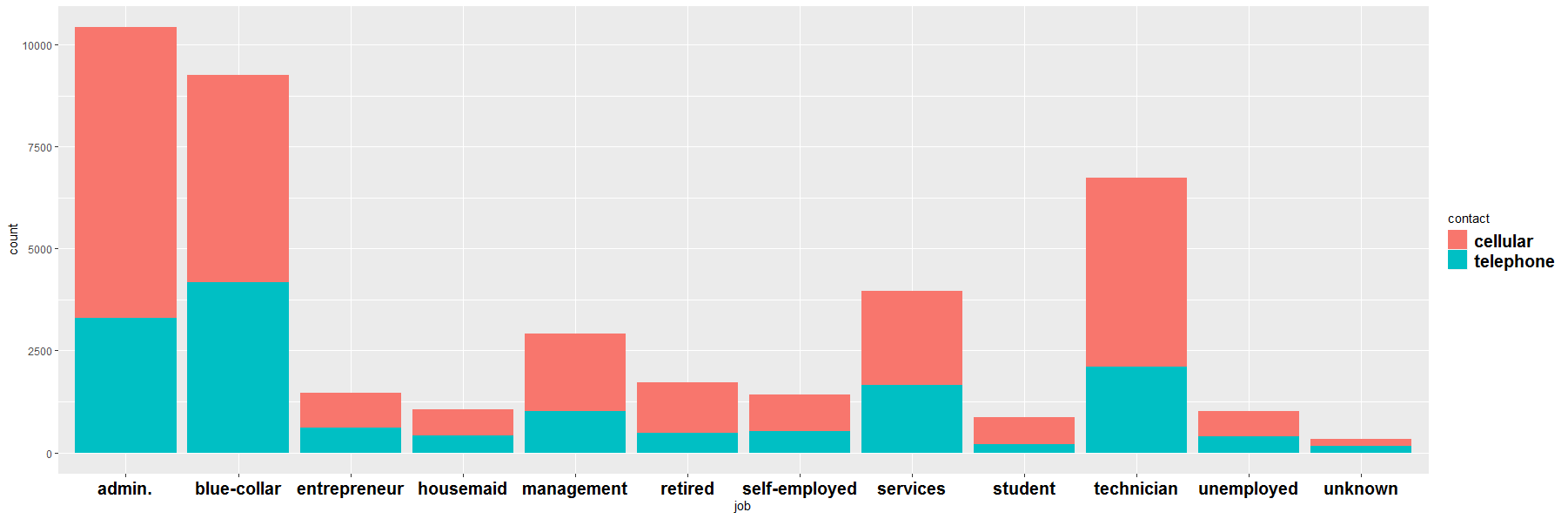


Figure 34 Graph of distribution between job and contact

The *default* attribute provides information if the client has a credit or not. Figure 35 shows the graphical visualization between customers who have a credit and which ones do not. As seen in the graphic, customers with credit has a small number than the number of customers with no credit. Also, the *unknown* category has a significant number of customers. Thus, the distribution between default and target is shown in the figure below. A small number of unknown customers have subscribed a term deposit.

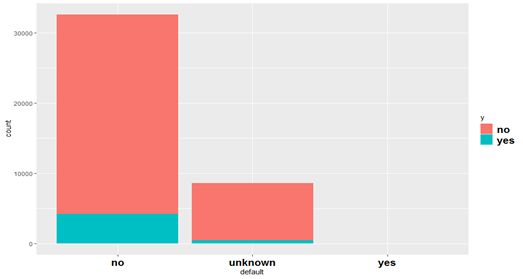
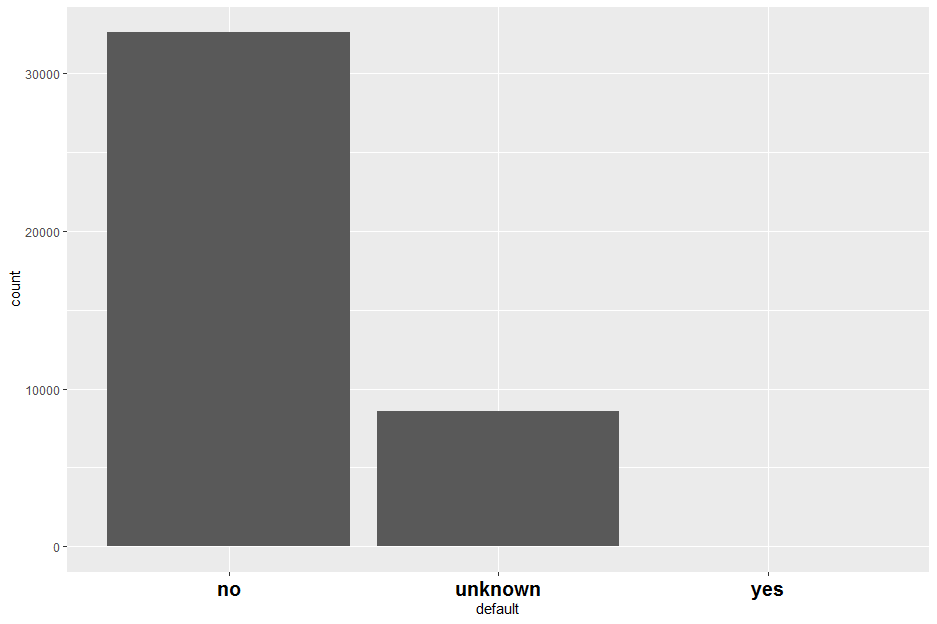


Figure 35 Graph of distribution between default and target (y)

By the same token, graphical visualization of distribution between customers with personal loan and target is depicted in figure 36. More customers without a personal loan have subscribed a term deposit.

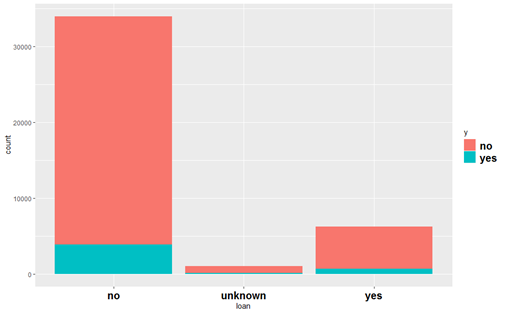
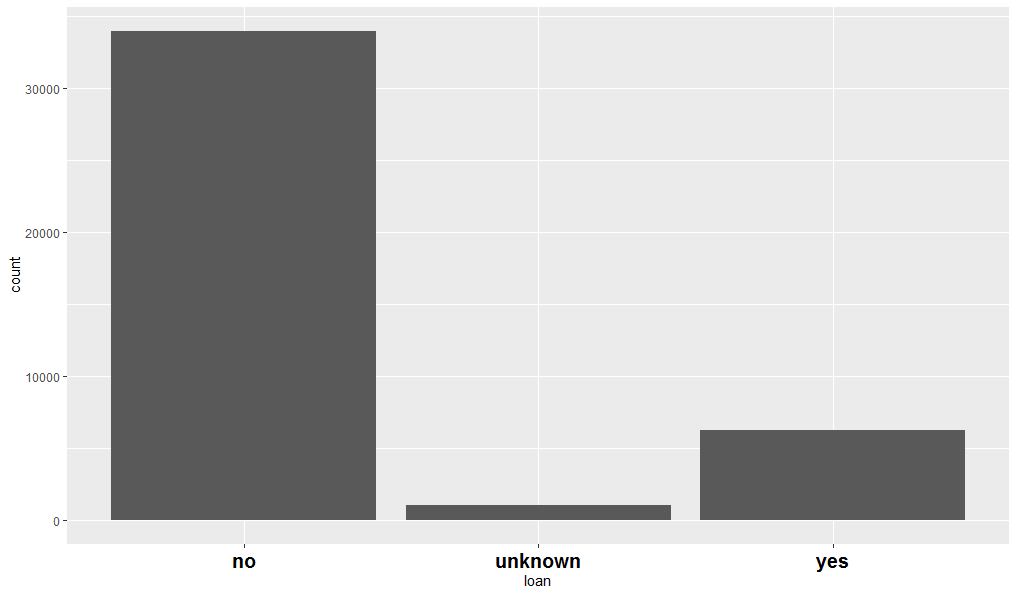


Figure 36 Graph of distribution between loan and target (y)

The number of customers with a housing loan who subscribed a term deposit is highest than other categories such as clients without a housing loan and clients without information about this type of loan.

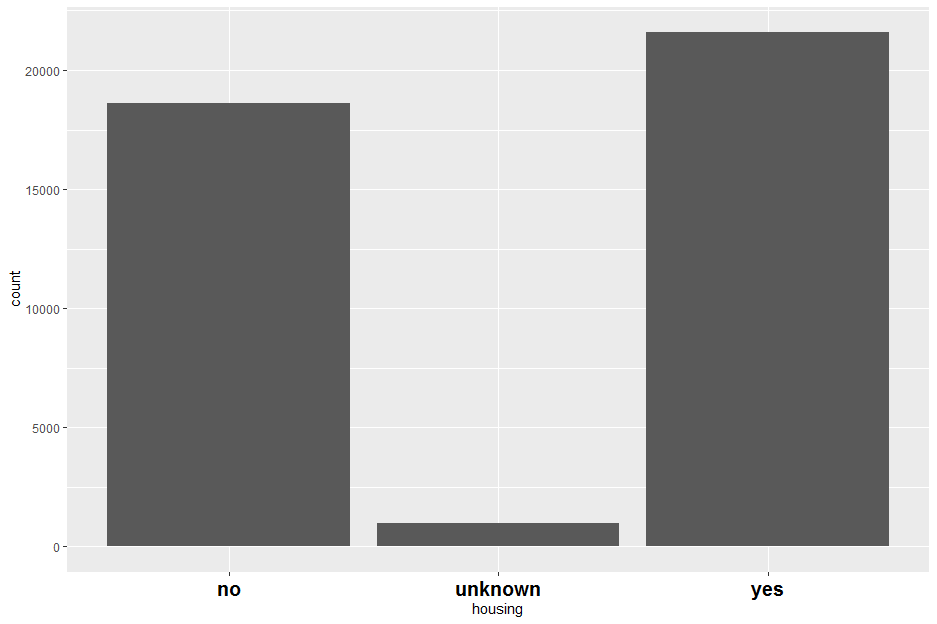
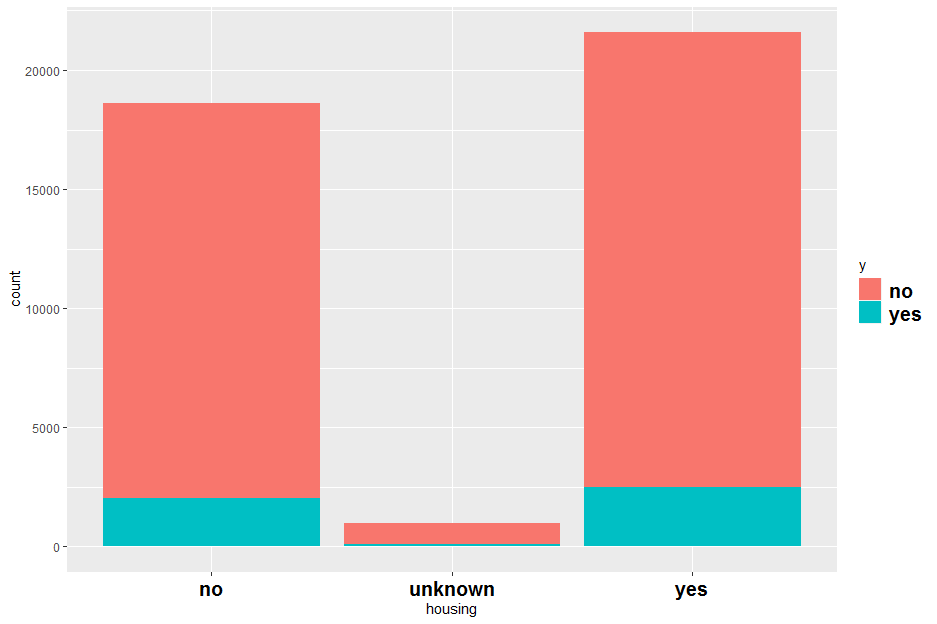
 

Figure 37 Graph of distribution between housing loan and target (y)

The previous number of contacts that made before this campaign and for this client has less effect on the term subscription then clients without previous number of performed contacts. The graphical distribution between previous number of contacts and target data is represented in the following figure.

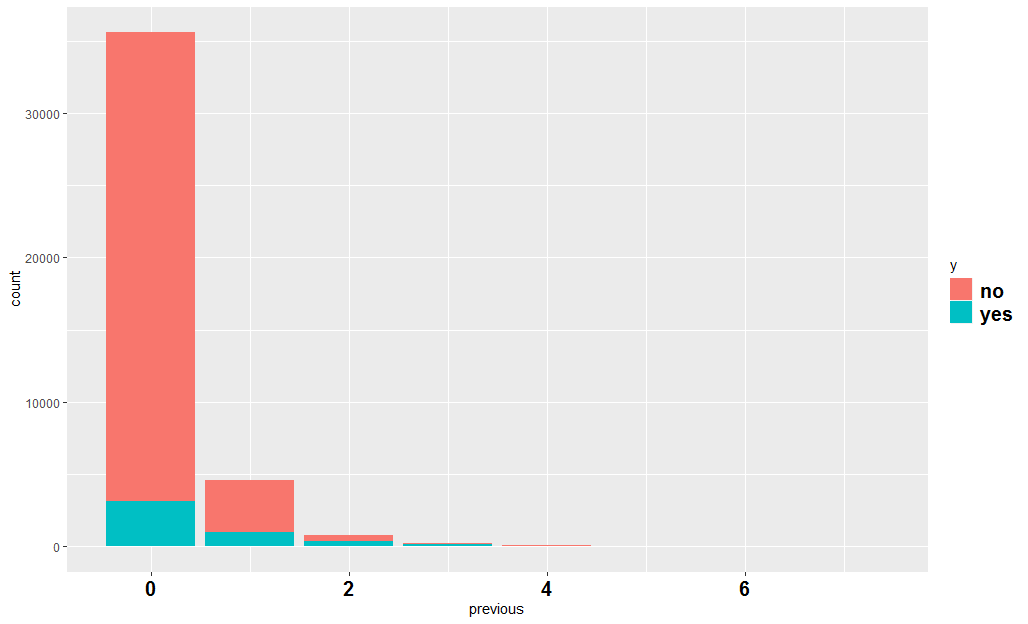


Figure 38 Graph of distribution between previous and target (y)

Among the customers who have subscribed a term deposit, the outcome of previous marketing campaigns with *nonexistent* category has the highest number of subscriptions. The graph distribution between *poutcome* and target is shown in figure 39.

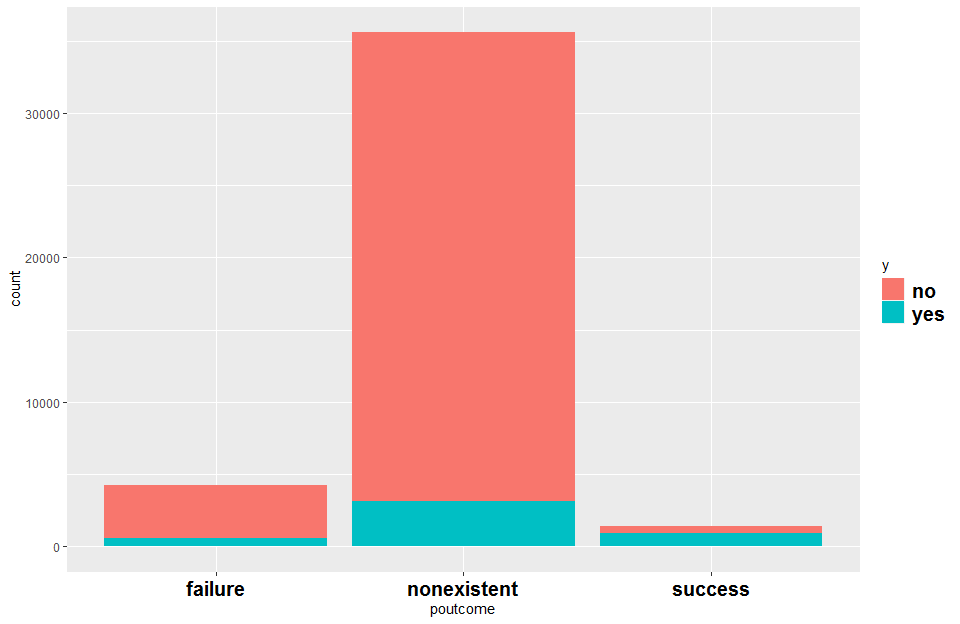
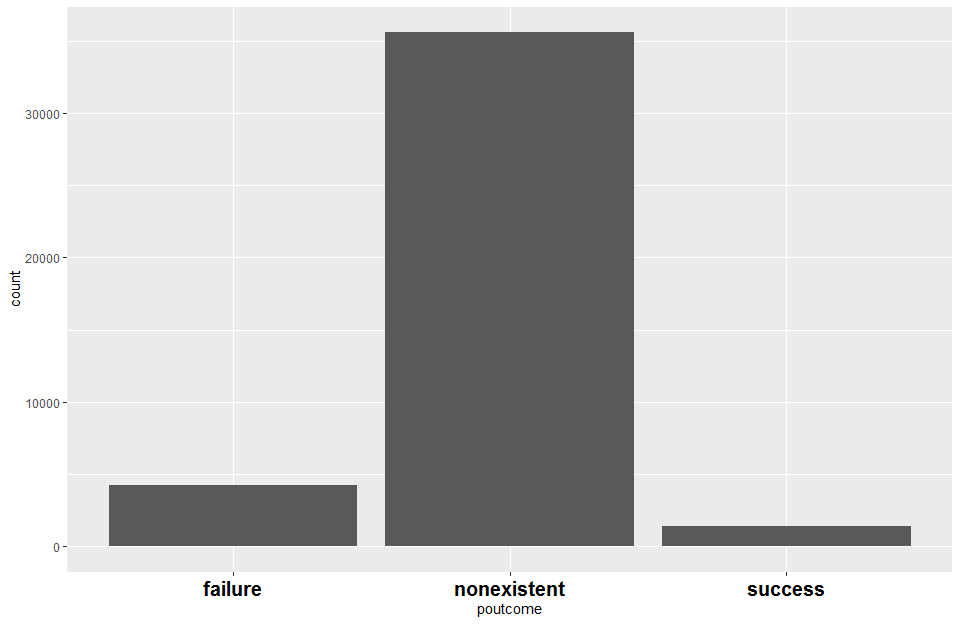


Figure 39 Graph of distribution between poutcome and target (y)

The graphical of the number of contacts with the client during this campaign including last contact is present in the figure below. The contact with the customers has less effect on the term subscription.

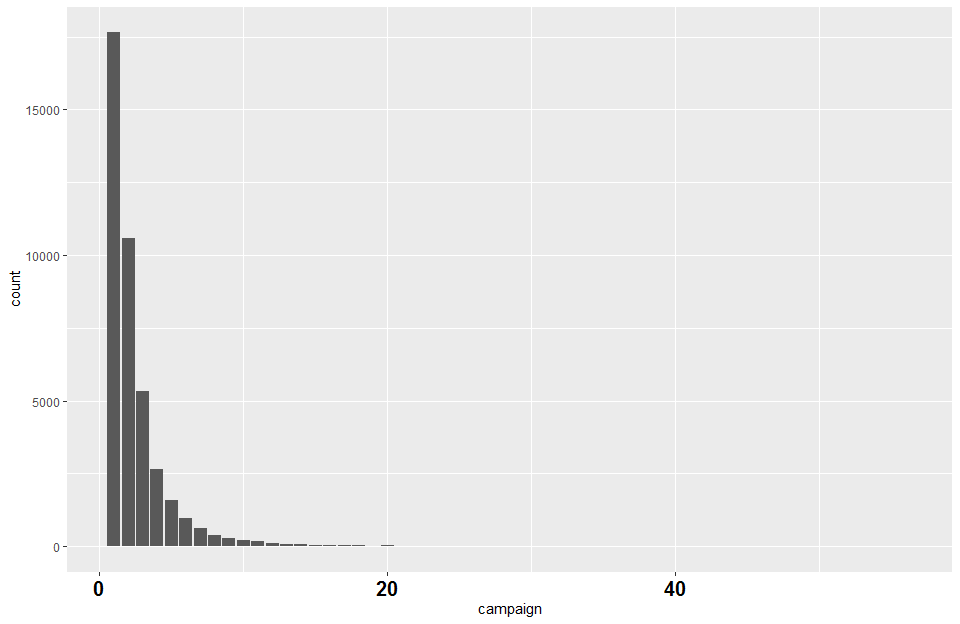
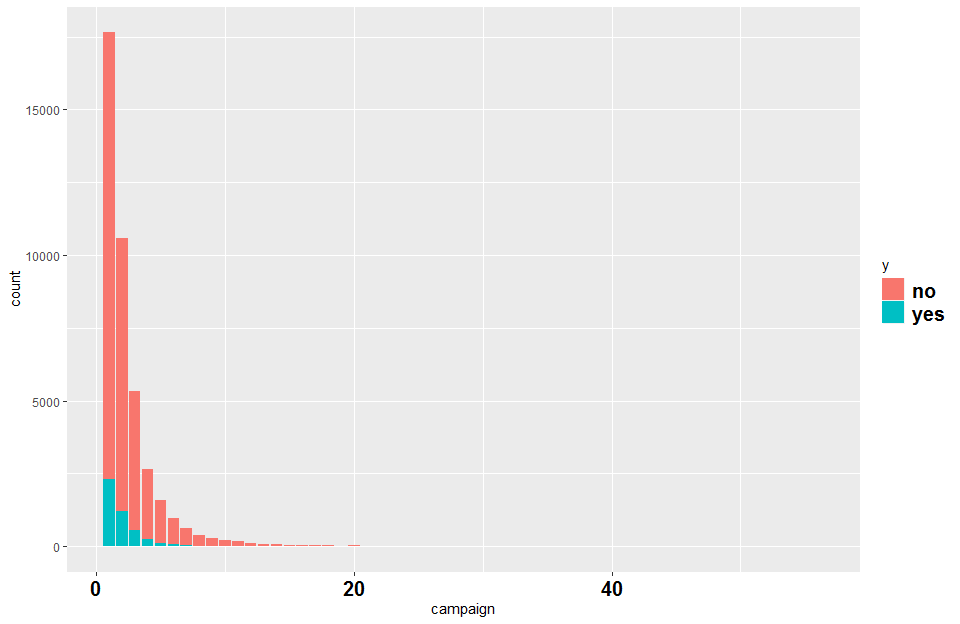
 

Figure 40 Graph of distribution between campaign and target (y)

Among the months of the year, May is the month with the highest number of last contact with client.

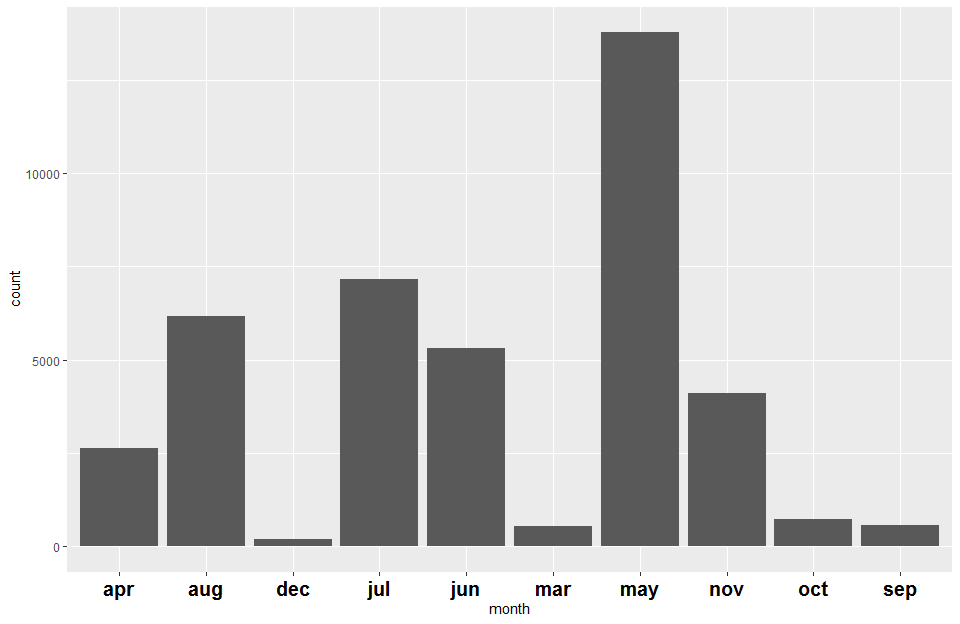


Figure 41 Graph of distribution between months of last contact with client

In the same manner, in figure 41 the days of the weeks of last contact with client are presented. Thursday is the day when most customers were last contacted for.

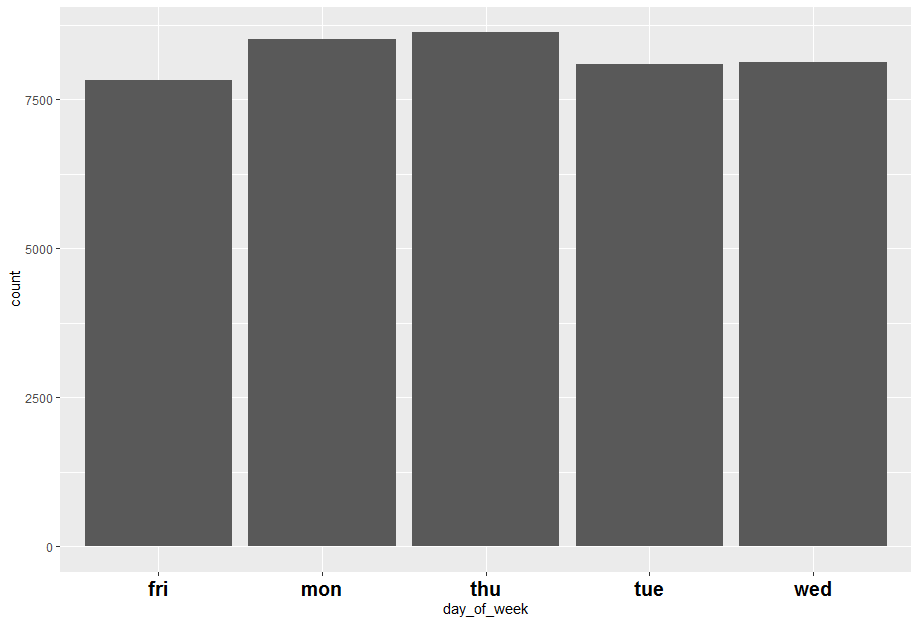


Figure 42 Graph of distribution between days of the week of last contact with client

The duration of the last contact with client in seconds is shown in the figure below. This attribute is for benchmark purpose only. It cannot be a reliable factor for the predictive models because strongly influences the output target value. If the duration is 0 the target (y) is no, respectively. However, the campaign is based on phone calls to the clients, so this attribute is included in the model building.

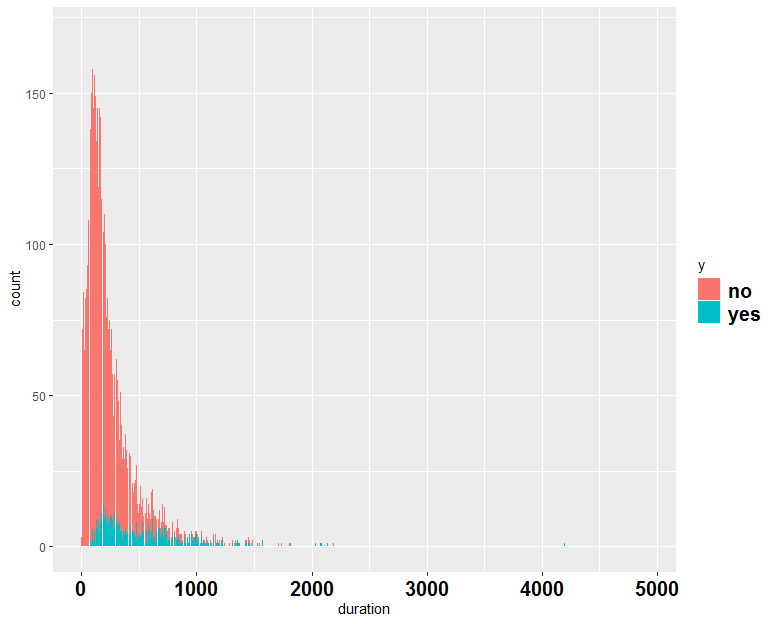
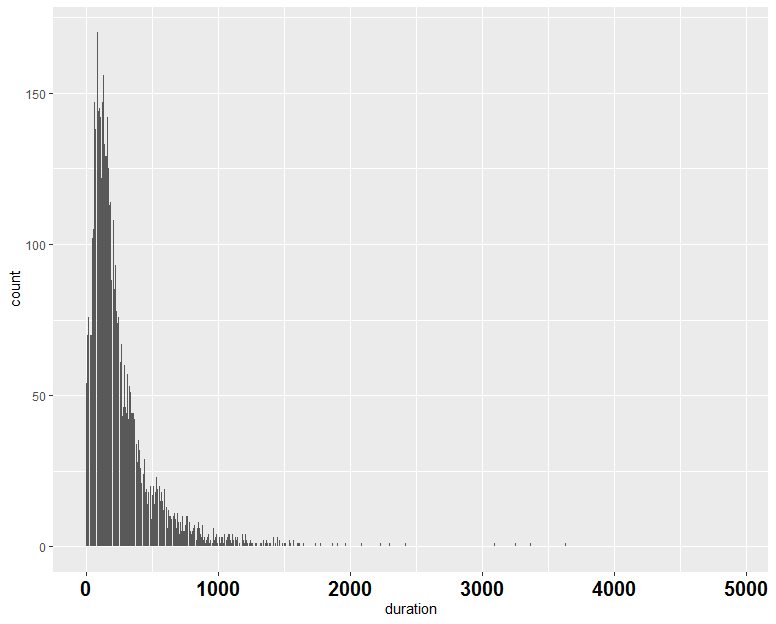


Figure 43 Graph of distribution between duration of the last contact with client in seconds and target(y)

The number of days after the customers were last contacted from a previous marketing campaign is graphically represented in figure 44. The number 999 means clients have not been previously contacted.

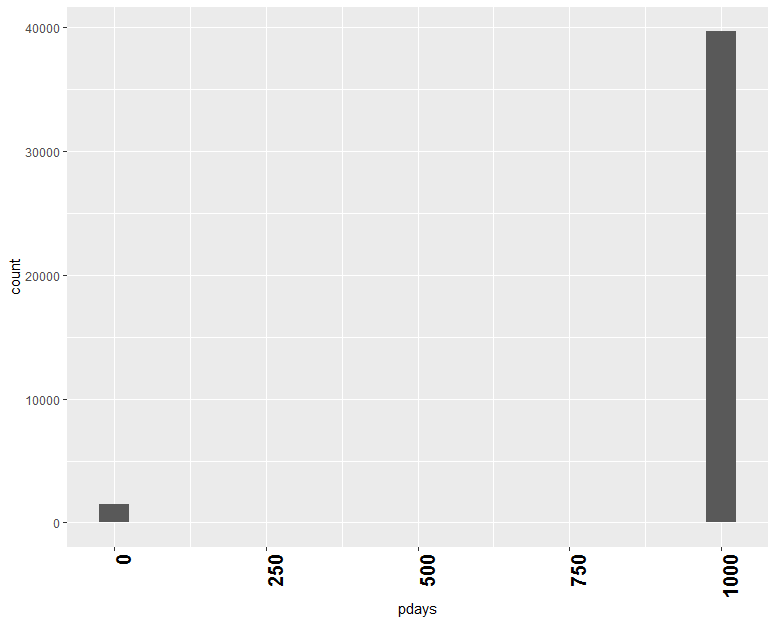


Figure 44 the number of days after the customers was last contacted from a previous marketing campaign

Quarterly indicator of employment variation rate **Emp.var.rate** is represented in figure 45. The employment rates are a measure of the use of available working resources (available for work). They are computed as the ratio of employed to the population in the working age. This is economic indicator and related to the contact date (quarterly collection of data) (OECD, 2018)

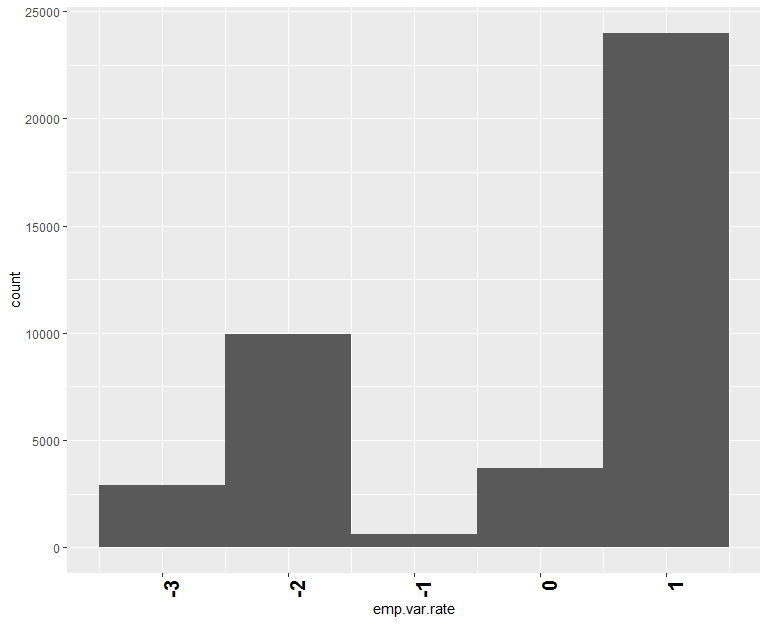


Figure 45 employment variation rate Emp.var.rate

Monthly indicator of consumer price index **Cons.price.idx** is presented in the following figure. This economic indicator measures the changes in the rate of inflation and the purchasing power of the currency. The consumer price index gives an overview of the current prices of the goods in the basket and services compared to prices in the same period in the previous year in order to shown the impact of inflation on purchasing power. (BD dictionary, 2019)

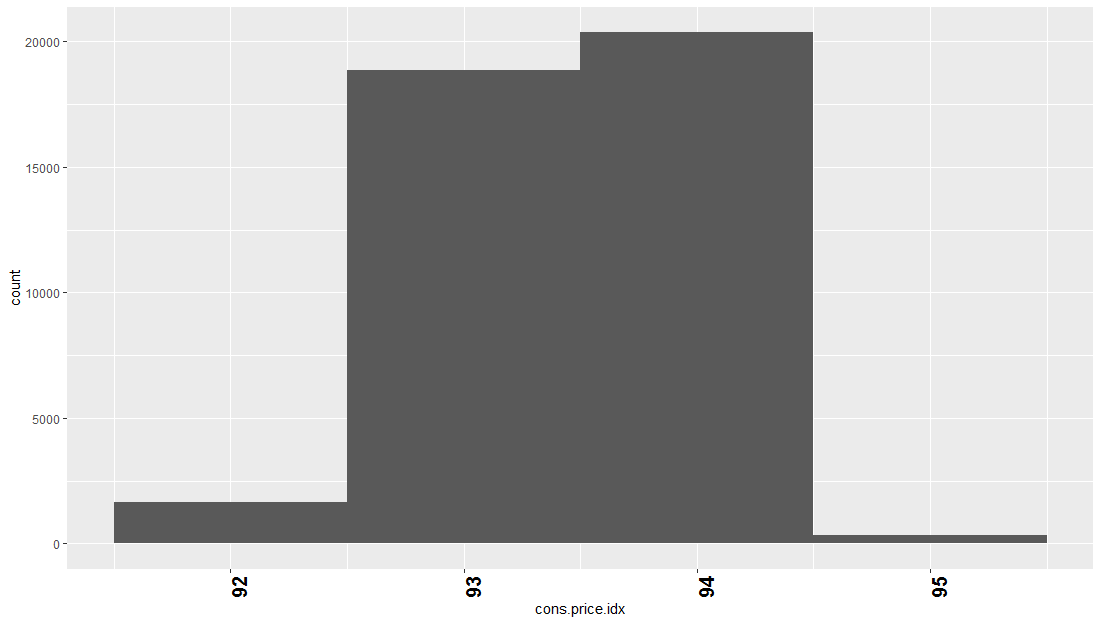


Figure 46 consumer price index Cons.price.idx

Monthly indicator of confidence price index **Cons.conf.idx** is depicted in figure 47. This indicator gives an indication of the future development of household consumption and savings based on responses to their expected financial situation and their saving potential. (OECD Data, 2019)

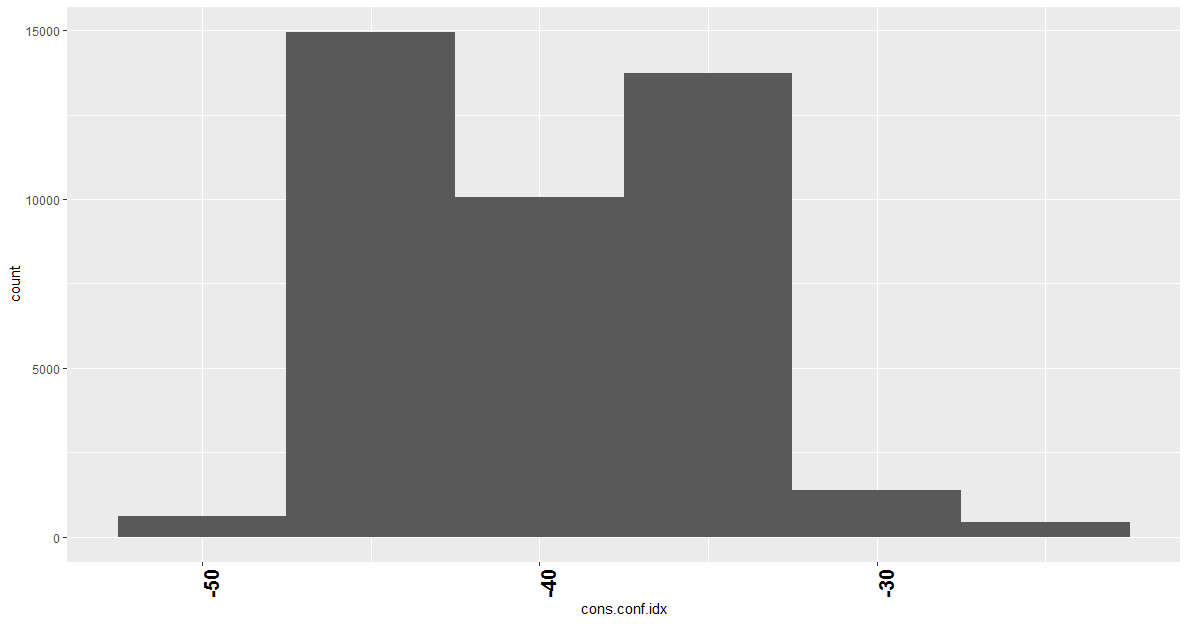


Figure 47 confidence price index Cons.conf.idx

Euribr3m is a daily indicator of 3 month Euro Interbank Offered Rates based on average interest rates. The three-month Euribor rate is the interest rate at which a collection of European banks lending euro-denominated funds in which the loans have a maturity of three months. (Global rates, 2018)

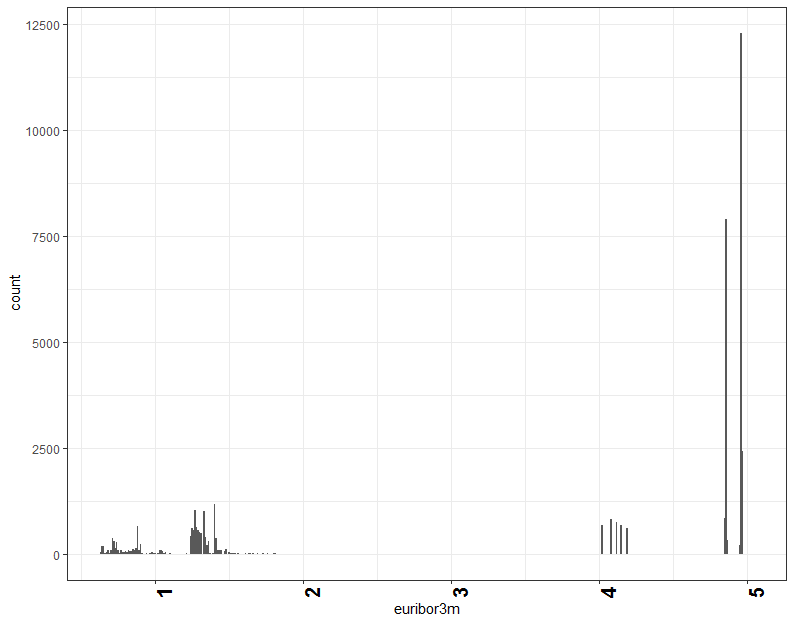


Figure 48 3 month Euro Interbank Offered Rates Euribr3m

The quarterly indicator of the number of employees **nr.employed** is graphically represented in figure 49.

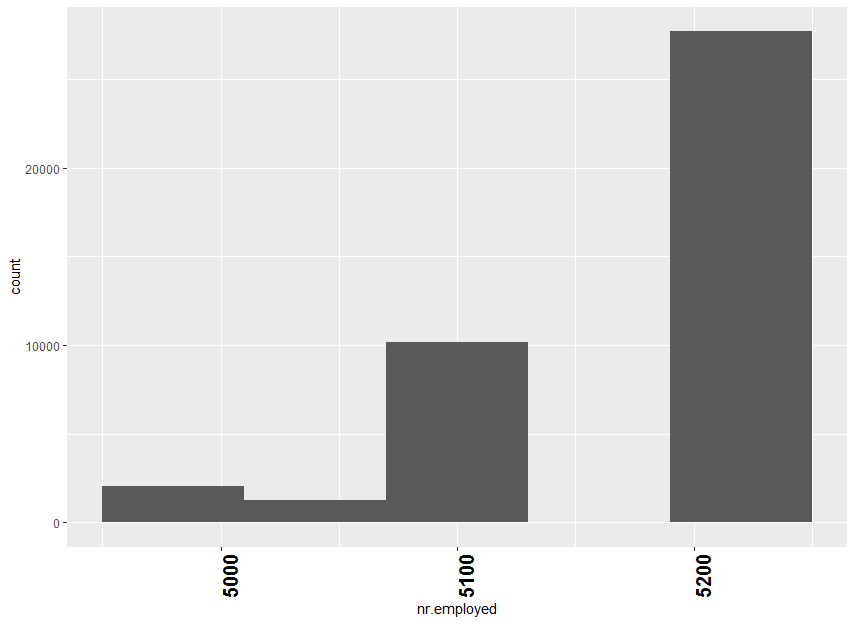


Figure 49 number of employees nr.employed

## 6.4 Converting the data type

Converting of the variable, known as coercion, is part of data cleaning and preparation techniques. In order to review the class of each variable, *sapply()* with a class argument for each column in banking marking dataset is used before converting. (Jonge, 2013, p.18)

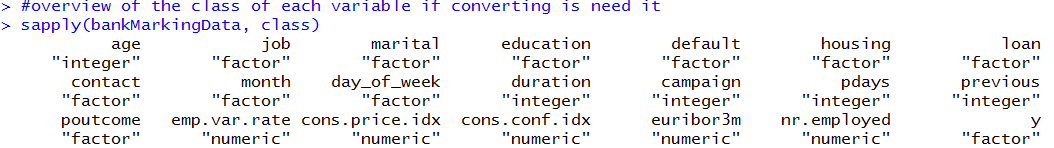


Figure 50 Class argument for each column

In order to convert in numeric variable is used function *as.numeric().*

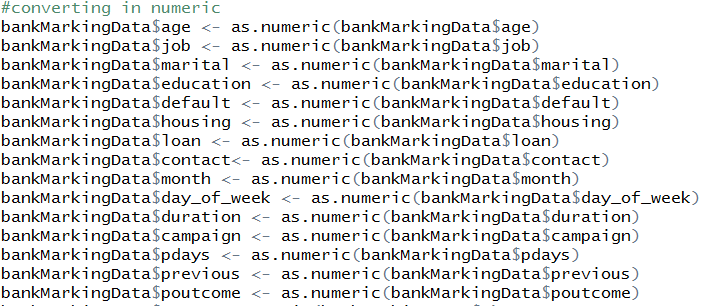


Figure 51 converting the numeric variable

## 6.5 Creating training and test dataset

The preparation of the data includes splitting the dataset into a training and test dataset. The training dataset is used to train or fit a model. However, the training dataset cannot provide reliable estimates of the accuracy of the model’s results on new data. This is in contrast with the idea of building the model to make predictions on new data. For this reason, the dataset is divided into two subsets: a training dataset to build the model and test dataset to evaluate the model’s performance on unseen (new) data. (Brownlee, 2018, p. 5)

As shown in figure 52, banking marketing dataset is divided into two portions. The size of the data is 70: 30 ratio in the training and test dataset, respectively. In order to generate random numbers in a predetermined sequence, the function *set.seed()* is used. In this way, producing identical results are ensured even if the analysis is repeated. In this way, all results are reproducible. (Lantz, 2013, p. 132) The function *sample()* is used to take a sample in a specific size. *nrow()* returns the number of rows in the data frame. (RDocumentation, 2017)

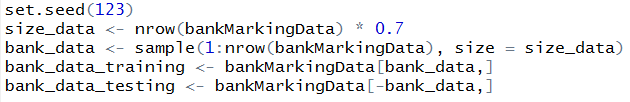


Figure 52 Spit the dataset into training and test

# Chapter 7 Building the models

## 7.1 Decision trees

### 7.1.1 Types of decision trees

Decision trees are seen as one of most used supervised predictive learning techniques. (Brownlee, 2016) Decision tree model can be applied in both classification and regression task. Thus, there are two types of decision trees: classification and regression. Classification tree is used to predict qualitative output rather than regression for quantitative. It could also be said that, classification tree used for categorical predicted variables (outcome) such as *yes* and *no*, and regression tress used for continuous (real number) such as a price of houses (James, 2017, pp. 303 -316)

In this project is used decision classification tree because the purpose of building the models is to predict whether the client will subscribe (yes or no) for a term deposit (variable y).

### 7.1.2 Decision Classification trees in R

R provides various performances of decision trees using different algorithms and packages. In this project are built CART (Classification and Regression Trees) using *rpart* package based on Gini as methods of attribute selection, and C5.0 using package C50 and Entropy method.

### 7.1.2.1 CART (Classification and Regression Trees) using rpart package

CART use Gini method as an attribute selection to compute the impurity of the sample.

##### 7.1.2.1.1 Building and training the model on the data

To take advantage of *rpart*, the package *rpart* should be installed once and reloaded each time by library (*rpart*) as shown in figure 53. To build and grow the tree model is used *rpart*() function where target data y, training dataset and method *class* for classification tree are included. (Kabacoff, 2017)

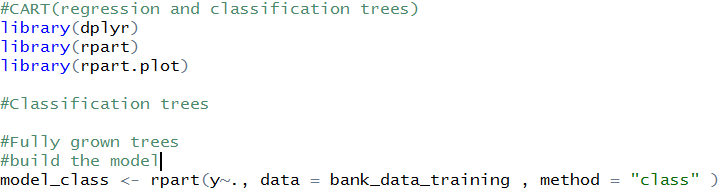


Figure 53 Building the Rpart tree model

In order to examine the results are used *summar*y(), as shown in the figure below. This function provides information about the number of observations, important variables, primary and surrogate splits. (Kabacoff, 2017)

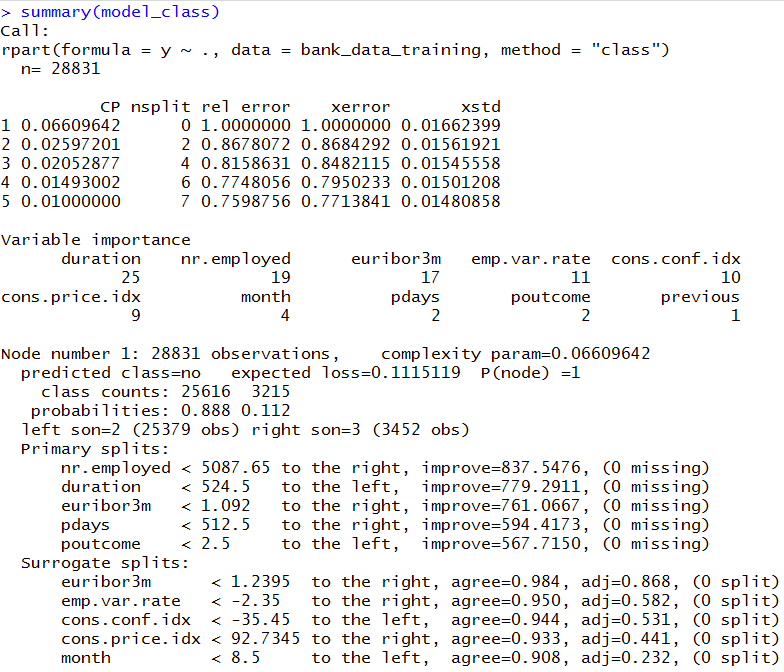


Figure 54 Summary of the CART training model

The *print*() function provides a clear view of all terminal nodes with the number and percentage of *yes* and *no* for each of node as shown in figure 55.

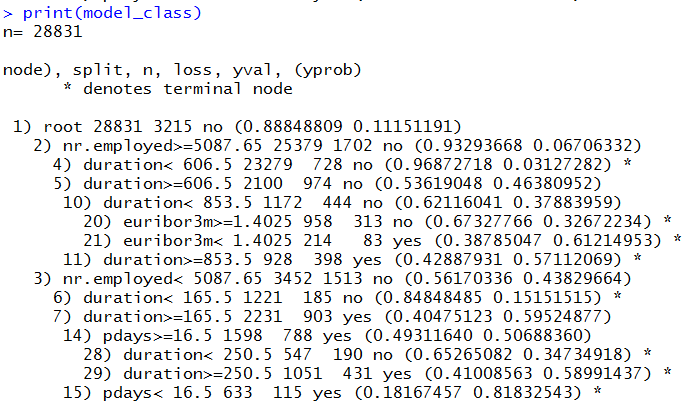


Figure 55 Results of the CART training model

In order to visualize the tree flowchart structure of the model is applied the *fancyRpartPlot ()* function from the package *rattle*. Each node presents is a predictor variable that help to determine if the bank term deposits will be (or not) subscribed. In the *rpart* trees, the left branch is with true condition. (Kabacoff, 2017)

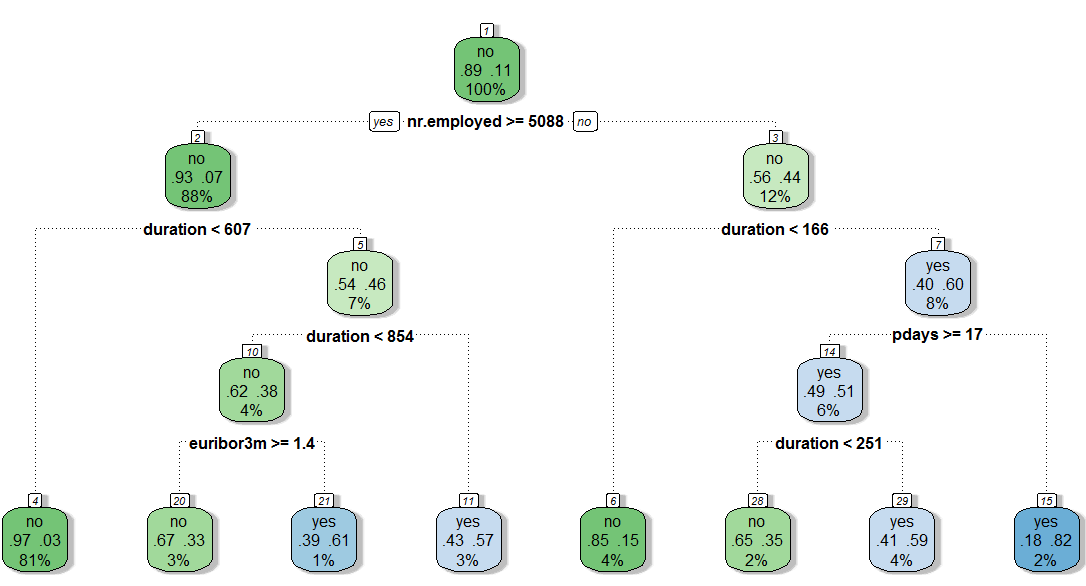


Figure 56 Visualization of the CART model

The results from the figure above can be interpreted as 89% of the customers at the top node are employed and 11 % are not. Then, the tree is divided according to call duration. 93% of the clients were called for less than 607 seconds and 7% for more time. 97% of the customers who were called in less time did not subscribe the term deposit and 3% of them subscribed. The interpretation of the results continues with the customers who have been on call for more than 607 seconds duration. 57% of them have been on call with duration more than 854 seconds and 43% for less time. Similarly, 67% of the customers who have been on call in less than 854 seconds did not subscribe and 33% did. Whereas, 43 % of clients who been held on the call for more than 854 seconds subscribed and 57 % of them not.

##### 7.1.2.1.2 Test and evaluate the performance of the model

In order to apply the test dataset is used the *predict()* function as shown in the figure 57.



Figure 57 Test dataset prediction of CART

Then, the accuracy of the model is calculated as shown in the figure below using confusion matrix.

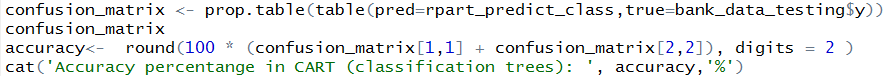




Figure 58 CART Calculation of Accuracy

A confusion matrix is a technique that presents and visualizes the performance of the classification model. The table gives information about true positive (TP) predicted values, false positive (FP) incorrect predicted values, false negative (FN) incorrect predicted negative values and true negative (TN) correct predicted negative values. (Brownlee, 2018) The *CrossTable*() function form the package *gmodels* is used as shown in figure 59.

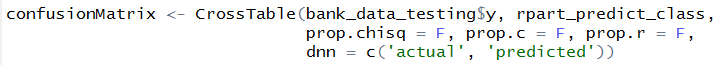


Figure 59 *CrossTable()* in CART model

The results are depicted in figure 60. Out of 12357 observations, the model predicted that 10505 have not subscribed and 781 have made correctly.

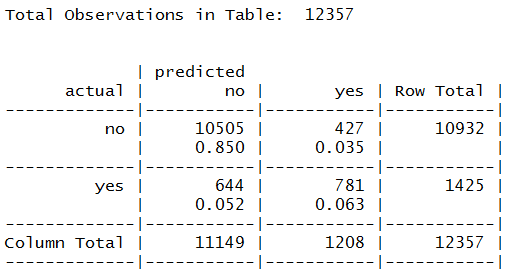


Figure 60 Cross table CART results

### 7.1.2.2 C5.0 decision trees

C5.0 is an improved version of his previous C4.5, which improved ID3 (Iterative Dichotomiser ) algorithm and use splitting algorithms, including entropy based on information gain and compute the homogeneity of a sample.. (Lantz, 2013, p.124)

##### 7.1.2.2.1 Building and training the model on the data

In order to build the C5.0, a package C50 is installed. The build of the model includes most used commands, as depicted in figure 61. The 21st column in *bank*\_*marketing*\_*training* is the class variable, so it is excluded as an independent variable. (Lantz, 2013, p.134)

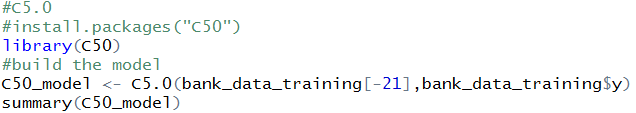


Figure 61 Building the C5.0 model

As shown in the figure 62, *print()* function provides basic information about the tree such as generated call function, the number of predictors (20), the samples (28831) used for the tree growing and the tree size (60) , indicating that the tree has 60 deep decisions. (Lantz, 2013, p.135)

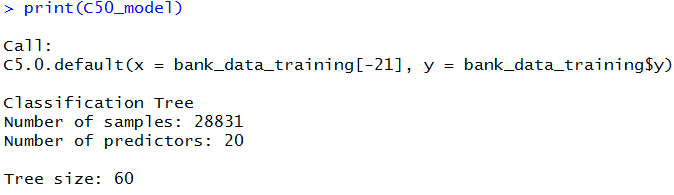


Figure 62 C5.0 information

The *summary()* provides more details about the training model. The results are depicted in the following figure. The Error statement notes that the model classified all, expect 2200 out of 28831 cases for an error rate of 7.6%. 986 actual *no* were misclassified as yes (false positive) and 1214 yes as no (false negatives). Also, the most used attributes are *duration*, *poutcome* and nr.*employes.*

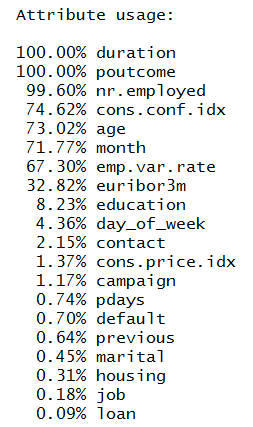
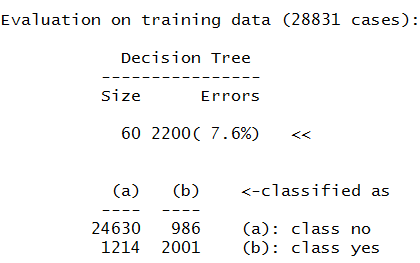


Figure 63 Evaluation of the C5.0 training model

##### 7.1.2.2.2 Test and evaluate the performance of the model

In order to apply the test dataset is used the *predict()* function as shown in the figure 64.



Figure 64 Test dataset prediction of C5.0

Then, the accuracy of the model is calculated as shown in the figure below using confusion matrix.

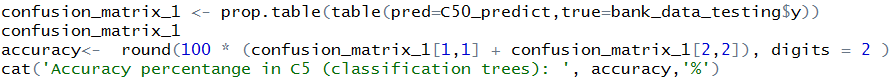




Figure 65 C5.0 Calculation of Accuracy

The results of *CrossTable()* function are depicted in figure 66. Out of 12357 observations, the model predicted that 10496 have not subscribed and 799 have made correctly.

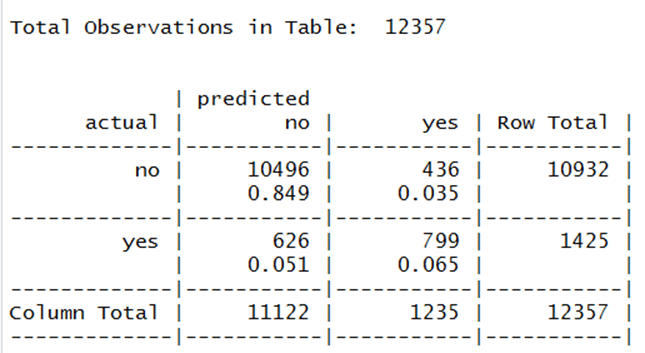


Figure 66 Cross table C5.0 results

### 7.1.2.3 Comparison of Decision trees algorithms

Detailed results about the tree algorithms are given in the comparison table below. Although there is no significant difference in the accuracy, C5.0 achieved a higher percentage.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Attribute selection method** | **Spilt** | **Accuracy** |
| CART | Gini | Binary | **91.33%** |
| C5.0 | Entropy | Multi – way | **91.44%** |

Table 3 Comparison of Decision tree algorithms

Some of the advantages of decision tree models are easy to interpret results without statistical knowledge, a clear visualization of the results in a tree structure with most important attributes. The disadvantages are easy to over fit and under fit and small changes in the training data leads to big changes in the decision results. (Lantz, 2013, p. 124)

### 7.1.3 Linear Regression in R

Linear regression is a statistical model used to explain and present the relationships between independent and dependent variables. (Yan, 2009, p.2) If is a single independent variable is called simple linear regression, otherwise multiple linear regression. (Lantz, 2013, p. 161) In contrast, linear regression cannot predict probability. In case it is used for binary modeling outcome, the result may not me provided within yes or no, 0 and 1, etc. For this reason, other techniques such as logistic regression can be applied to obtain a probability result. (Le, 2018)

The following figure shows the correlation (relationship) between two variables. The range that indicates a perfectly relationship is between -1 and +1. A very strong correlation is for values above 0.9. In order to create s scatterplot matrix is used the *pairs.panels()* function in the package *psych*. Above side is the correlation matrix and below side provides a visualization data. The diagonal includes the attributes presented by histograms. If the correlation ellipse is more stretched, it means that the correlation is stronger. (Lantz, 2013, pp. 161-178)

There are three very strong corrections between: *emp.var.rate* and *euribor3m* (0.97), *euribor3m* and *nr*.*employed* (0.95) , *emp.var.rate* and *nr.employed* (0.91). In this project, these attributes are used to build a liner regression model.

#### 

Figure 67 Scatterplot matrix for quantitative features

#### 7.1.3.1 Simple Linear Regression

To build a liner model is used *lm()* function shown is figure 68.

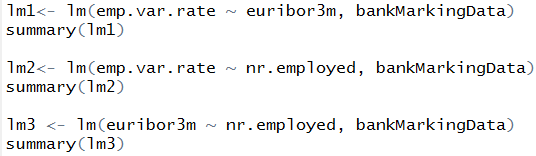
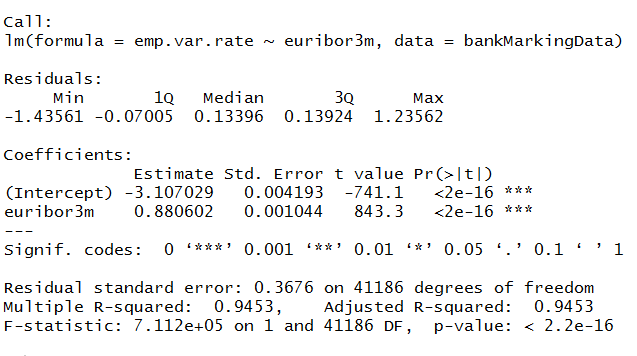
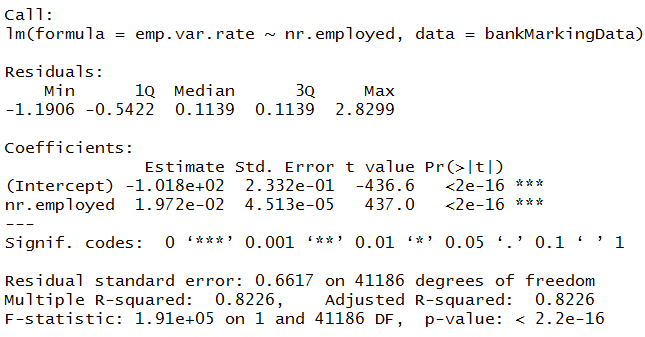


Figure 68 Simple linear regression in R

To evaluate the results, the *summary()* is used as seen in figure 69. There are few important components: *Call* – shows the used function to compute the model, *Residuals-* distribution of the residuals, the median should be near to zero, *Coefficients* –includes estimate, standard error, p-value, t- value and provides how statistically significant is the relationship. The higher t-statistic and lower p-values(≤ 0.05 to reject a null hypothesis) are more significant. The star symbol visualizes the level of significance listed by *Signif.codes*.( three starts – more significant variable). The model accuracy can be check by Residual Standard Error (RSE)(standard deviation of residual errors, better to be close to 0), Multiple R-squared (coefficient of determination, better to be close 1.0) – measure how well the model fit the data, Adjusted R-squared – degrees of freedom and it’s good to be a higher value. F-statistic provides the overall importance of the model. (Lantz, 2013, pp. 181-182) As seen in figure below, the model fits best in *emp.var.rate* and *euribor3m*, where employment variation rate and Euro Interbank Offered Rates have the strong linear relationships.





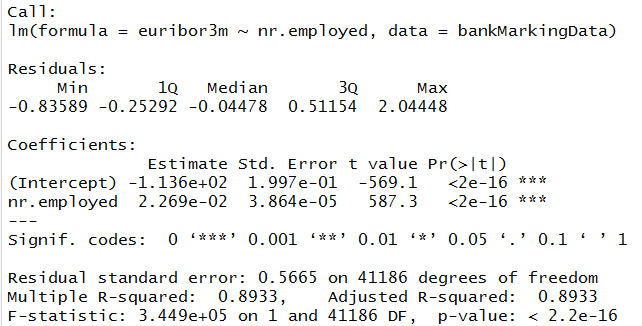


Figure 69 Summary results of simple linear regression

#### 7.1.3.2 Multiple Linear Regressions

In order to build a multiple linear model are used *emp.var.rate* , *euribor3m* and *nr.employed*. As seen in figure 70, *lm()* function is applied to build the model where *nr.employed* is the dependent variable.



Figure 70 Multiple linear regressions in R

The results are depicted in the figure below. F-statistic’s p-value is <2.2e-16. It is highly meaningful and means that at least one predictor is very significantly relevant to the predicted variable. Thereis a more significant association between the *euribor3m* and *nr.employed*, because the t-value statistic is significantly different than 0. Multiple R-squared and Adjusted R-squared are close to 1.Hence, the model explains much of the variation in the output. RSE is 20.31 corresponding to 0.45% error rate.

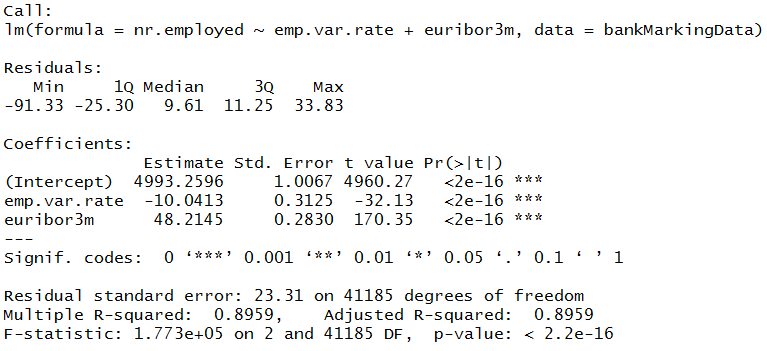


Figure 71 Summary results of multiple linear regressions

The advantages of linear regression are adaptation to model any numerical data and provide an estimation of the size and strength of the relationship between variables. The disadvantage are requires statistical knowledge to evaluate the results and compatible only with numerical data. (Lantz, 2013, p. 168)

# Chapter 8 Conclusion

## 8.1 Summary

Machine learning techniques are widely used data analysis methods in various business and industrial sectors. The main reason for that because ML can build predictive models to produce better predictions and achieve the desired level of accuracy, leading to better outcomes. Building the models is an easy and straightforward process. The main challenges in data analysis are data preparation and cleaning, the selection of appropriate models and attributes used in their implementation. The aim of the project is to find how to use machine learning techniques for analysis and making predictions using existing dataset in banking marketing. To find how they can be used together in a process of converting raw data to effective decision making knowledge. Thus, in this project are used decision tree algorithms to predict whether the client will subscribe for a term deposit. Although there is no significant difference in the result accuracy, C5.0 achieved a higher percentage. At the same time, the linear model is applied to represent the relationship between quantitative attributes and how significantly relevant they are. The simple linear model fits best in *emp.var.rate* and euribor3m, where employment variation rate and Euro Interbank Offered Rates have the strong linear relationships, while there is a more significant association between the euribor3m and *nr.employed* in the multiple regressions.

## 8.2 Future work

There are few aspects in data analysis and machine learning techniques are planned as a future work to expand and improve this research such as exploring the different outliers’ detection methods. The outliers detection is an important and interesting area of data analysis process so further investigation is considered. Improving the model’s performance using boosting and bagging ensemble methods as well as exploring and building advanced techniques such as random forest and logistic regression are also part of future work.

# References and Bibliography

Aggarwal, C., (2013) *Outlier Analysis,* USA: New York, Springer Science & Business Media

Ajanki, A., (2018) Difference between machine learning and software engineering [Online] Available at <https://www.futurice.com/blog/differences-between-machine-learning-and-software-engineering/> (Accessed: 22 January 2019)

Akwei, J., (2017) *Bank Marketing Data Classification* [Online] Available at <http://www.rpubs.com/johnakwei/330635> (Accessed: 01 February 2019)

ANALYTICS VIDHYA CONTENT TEAM, (2016) *A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)* [Online] Available at <https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/?fbclid=IwAR1PwPkaU1Ypf7tbOLSLFLTL-NBppFZwnLsOPl09-513ljya-g-3AFpQCU4> (Accessed: 10 April 2019)

Apampa, O., “Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction”, *Journal of International Technology and Information Management Volume 25-Issue 4 Article 6*, Tilburg University, January 2016, pp. 85-100

Bachmann, J., (2019) *Bank Marketing Dataset, Predicting Term Deposit Subscriptions* [Online] Available at <https://www.kaggle.com/janiobachmann/bank-marketing-dataset/home> (Accessed: 01 February 2019)

Bakar, N., Tahir, I., “Applying Multiple Linear Regression and Neural Network to Predict Bank Performance”, *International business research* *Vol.2, No. 4*, October 2009, pp. 176-183

Bali, R., Sarkar, D., (2016) *R Machine Learning By Example ,* 1st ed., UK: Birmingham, Packt Publishing Ltd

BD dictionary, (2019) *Consumer price index (CPI)* [Online] Available at <http://www.businessdictionary.com/definition/consumer-price-index-CPI.html> (Accessed: 07 March 2019)

Bhan, S., (2019)*Decision-Tree-Using-R* [Online] Available at <https://github.com/SanyTiger/Decision-Tree-Using-R/blob/master/bank.r> (Accessed: 05 February 2019)

Bhatia, R., (2018) *How Do Machine Learning Algorithms Differ From Traditional Algorithms?* [Online] Available at <https://www.analyticsindiamag.com/how-do-machine-learning-algorithms-differ-from-traditional-algorithms/> (Accessed: 22 January 2019)

Bista Solutions, (2019) *Statistical Outliers: Detection and Treatment* [Online] Available at <https://www.bistasolutions.com/resources/blogs/statistical-outliers-detection-treatment/?fbclid=IwAR2_7DPfTBtz1debUujUSLQ2aAg1Tg_5rgqHIaor87C7rW5qk7ACrXgYtI8> (Accessed: 1 April 2019)

Bonaccorso,G., (2017)*Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning*,2nd ed., UK: Birmingham, Packt Publishing Ltd

Brid, R., (2018) *Introduction to Decision Trees* [Online] Available at <https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb?fbclid=IwAR3cvAtradTTRiIvbWq5trIJjAqnFb050bnMX_LnwFo_M9cVK5UuVinYdp8> (Accessed: 27 February 2019)

Brownlee, (2016) *Classification And Regression Trees for Machine Learning* [Online] Available at <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/> (Accessed: 09 April 2019)

Brownlee, J. (2013) *How to Identify Outliers in your Data* [Online] Available at <https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/?fbclid=IwAR0mTraTqpJENa0EePMnTELUQ3ZnMw-JvUMAin1XbLKkNuJwEMomdQCZxZ8> (Accessed: 13 March 2019)

Brownlee, J., (2013) *How to Prepare Data For Machine Learning* [Online] Available at <https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/> (Accessed:05 February 2019)

Brownlee, J., (2014) *Linear Regression in R* [Online] Available at <https://machinelearningmastery.com/linear-regression-in-r/?fbclid=IwAR1EqqOJSilC7vxQLczZQmCRU1-YWsEK6bmwAFwy_yAKgOWBINb0Vsvr1F8> (Accessed: 05 February 2019)

Brownlee, J., (2016) Supervised *and Unsupervised Machine Learning Algorithms [*Online] Available at<https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/> (Accessed: 22 January 2019)

Brownlee, J., (2016*) What is a Confusion Matrix in Machine Learning* [Online] Available at <https://machinelearningmastery.com/confusion-matrix-machine-learning/> (Accessed: 10 April 2019)

Brownlee, J., (2016) *Your First Machine Learning Project in R Step-By-Step (tutorial and template for future projects)* [Online] Available at <https://machinelearningmastery.com/machine-learning-in-r-step-by-step/?fbclid=IwAR0UHR3lRm5OMiey5rMgiJ99zifEh9UByIhWxOaFqb15H_d3o4QfLCGpSJs> (Accessed: 05 February 2019)

Brownlee, J., (2018) *How to Use Statistics to Identify Outliers in Data* [Online] Available at <https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/?fbclid=IwAR3tGVR6ARCj-oOcp3DQ_EA7h2WAjpMJXyIUb_SmYuSKEG81hsEy06oZTXQ> (Accessed: 01 April 2019)

Brownlee,J.,(2018) *Machine Learning Mastery with R Mini-Course From Developer To R Machine Learning Practitioner in 14 Days*, 1st ed., Jason Brownlee

Campos, F., (2017) *A Decision Tree model for Bank Marketing Analysis* [Online] Available at <https://rpubs.com/fabiorocha5150/decisiontreemodel?fbclid=IwAR23TCDaBPGzCFVGm7Pf44BQkDdwzHhEIUL-oDut8imL1dT3wIvPdXAcOK0> (Accessed: 01 February 2019)

Carroll, J., (2018) *Beyond Spreadsheets with R*, Manning Publications

Chen, J., (2017) *which customers are more likely to respond to bank’s marketing* campaigns? [Online] Available at <https://medium.com/@jameschen_78678/which-customers-are-more-likely-to-respond-to-banks-marketing-campaigns-3f00c512268d> (Accessed: 01 February 2019)

Choudhry,M., “*Machine Learning in Banking: How To Transform both Balance Sheet Management and Customer Services Provision*”, The European Financial Review, April - May 2018, pp. 2-8

Cinaroglu, S., “Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Countries Health Expenditures”, *International Journal of Computer Applications (0975 – 8887) Volume 138 – No*.1, Hacettepe University, March 2016, pp.1-41

Conway, D., White, J., (2012) *Machine Learning for Hackers*, USA: Sebastopol, O’Reilly Media, Inc.

Data Carpentry, (2018) *Aggregating and analysing data with dplyr* [Online] Available at <https://datacarpentry.org/R-genomics/04-dplyr.html> (Accessed: 13 February 2019)

Elsalamony,H., “Bank Direct Marketing Analysis of Data Mining Techniques”, *International Journal of Computer Applications (0975 – 8887)* *Volume 85 – No 7,* Egypt *,*January 2014,pp.12-22

Engel, C., (2019) Introducing to R [Online] Available at <https://cengel.github.io/R-intro> / (Accessed: 08 February 2019)

Financial Stability Board (FSB), (2017) *Artificial intelligence and machine learning in financial services, Market developments and financial stability implications,* Switzerland: Basel, Financial Stability Board.

Global rates, (2018) *3 month Euribor interest rate* [Online] Available at <https://www.global-rates.com/interest-rates/euribor/euribor-interest-3-months.aspx> (Accessed: 07 March 2019)

Gorjanc, G., (2007) *Working with Unknown Values*, University of Ljubljana, Slovenia

Grolemund, G., Wickham, H., (2017) *R for Data Science,* O’Reilly

Grzonka, D., Suchacka, G., Brorowik,B., “APPLICATION OF SELECTED SUPERVISED CLASSIFICATION METHODS TO BANK MARKETING CAMPAIGN” *Information Systems in Management Vol. 5 (1) 36−48,* 2016, pp.36-48

Gupta, P., (2017) *Decision Trees in Machine Learning* [Online] Available at <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052?fbclid=IwAR30Kie1IbBs3AgFdNxI3VXhmySM3Bhz2bYuKwryd1SNHglZU6tyyspBnNI> (Accessed: 05 February 2019)

Gyorgy, (2016) *Data Science for Business: Homework Assignment* [Online] Available at <http://rstudio-pubs-static.s3.amazonaws.com/157695_7fe1c26be1b54b01a520a9b73ae99c85.html?fbclid=IwAR29CUlpsxC7XX6LfH5yn7b0_DsdrtDyo2exPCnARjXbnmwfe_IJLU12s18> (Accessed: 13 February 2019)

Hariharan, A., (2018) *How to Use Machine Learning to Predict the Quality of Wines* [Online] Available at <https://medium.freecodecamp.org/using-machine-learning-to-predict-the-quality-of-wines-9e2e13d7480d> (Accessed: 05 February 2019)

Hastie,T., Tibshirani,r., Friedman,J., (2001)*The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* USA: New York,Springer Science & Business Media

James,G., Witten, D., Hastie, T., Tibshirani, R., (2017) An Introduction to Statistical Leaning with Applications in R, USA: New York,Springer Science & Business Media

Jonge, E., Loo, M., (2013) *An introduction to data cleaning with R*, Statistics Netherlands

Kabacoff, (2017) *Tree-Based Models* [Online] Available at <https://www.statmethods.net/advstats/cart.html?fbclid=IwAR0wa4Vc_OzTWdDKBtb_DsD-43x7Fi5FQnhOtcD5De9sSZdRM0_rAhq1neQ> (Accessed: 09 April 2019)

Kabacoff, R., (2017) *Missing Data* [Online] Available at <https://www.statmethods.net/input/missingdata.html> (Accessed: 15 February 2019)

Kaggle Inc,(2019) *Bank Marketing UCI, The data is related with direct marketing campaigns of a Portuguese banking institution* [Online] Available at <https://www.kaggle.com/c/bank-marketing-uci> (Accessed: 01 February 2019)

Karthik, (2017) *Bank Marketing* [Online] Available at <http://rstudio-pubs-static.s3.amazonaws.com/275389_f1634d6a36354763b3b7ca42a2083415.html> (Accessed: 05 February 2019)

Kasabara, (2018) Statistical Machine Learning Essentials CART Model: Decision Tree Essentials [Online] Available at <http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/> (Accessed: 11 April 2019)

Keles, A., Keles, A., “IBMMS DECISION SUPPORT TOOL FOR MANAGEMENT OF BANK TELEMARKETING CAMPAIGNS”, *International Journal of Database Management Systems (IJDMS) Vol.7, No.5,* Turkey, October 2015,pp. 1-15

Kohl, M., (2015) *Introduction to statistical data analysis with R*, 1ST ed., BookBoon The eBook company

Kumar. V., (2019) *Bank-Marketing-Data-Analysis* [Online] Available at <https://github.com/vinay2k2/Bank-Marketing-Data-Analysis/blob/master/Bank_Marketing.R?fbclid=IwAR2IZrydpKlRTN4DBpYK55VXGBOQoqhY7LSQrt_u3325T2DiEbZFcqu3edU> (Accessed: 11 April 2019)

Lantz,B., (2013),*Machine Learning with R* , UK: Birmingham, Packt Publishing Ltd,1st ed., ISBN 978-1-78216-214-8

Laureano, H., (2018) *Bank Marketing Dataset: An overview of classification algorithms,* Statistics@Kaust

Le, J., (2018) *Logistic Regression in R Tutorial* [Online] Available at <https://www.datacamp.com/community/tutorials/logistic-regression-R?fbclid=IwAR3D5tjMdCd8EQaG4LiV8nHNSgCHjixaNDLNPUyT4hRD3KlpWojet_Bg0NY> (Accessed: 13 April 2019)

Le,J. (2018) *Decision Trees in R* [Online] Available at <https://www.datacamp.com/community/tutorials/decision-trees-R?fbclid=IwAR26IbxRb4qIrPf56bTHfxzQW_OcEN7AZqjId-DEdln3Veywj0086bjxm6w> (Accessed: 10 April 2019)

Learn by Marketing, (2019) *Decision Trees in R* [Online] Available at <http://www.learnbymarketing.com/tutorials/rpart-decision-trees-in-r/?fbclid=IwAR2NpkgvPlVRi7qrVH18NWK-AWvVXC7QjDnUKKZKCzoumzr_4cM7fO2No_o> (Accessed: 10 April 2019)

Lee, S., (2018) *Lecture notes – Artificial Intelligence*, London UK, University of East London

Lube, S., Roux, N., (2018) A Step-by-Step R Tutorial: Introduction into R application and programming, 2nd ed., BookBoon – the eBook company

Machine Leaning Plus, (2018) *Complete Introduction to Linear Regression in R* [Online] Available at <https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/> (Accessed: 05 February 2019)

Malik, F., (2018) *How Good Is My Predictive Model — Regression Analysis* [Online] Available at <https://medium.com/fintechexplained/part-3-regression-analysis-bcfe15a12866?fbclid=IwAR12DrZEKoznBgqMrAJweJounAAG9-ILm8VHaysPiacHjrC5pa6ppk2QmFg> (Accessed: 05 February 2019)

Marshall, E., Maria, S., *Multiple linear regression in R*, University of Sheffield

Matloff, N., (2011) *THE ART OF R PROGRAMMING,* USA: San Francisco, No Starch Press, Inc.

Mazlina, N., Bakar, A., Tahir, I., Applying Multiple Linear Regression and Neural Network to Predict Bank Performance, *International Business Research*, October 2009, Vol. 2, No. 4, pp.176- 183

Montgomery, D., Peck, E., Geoffrey Vining, G., (2012) *Introduction to Linear Regression Analysis*, 5th ed., USA; John Wiley & Sons, Inc.

Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Nataraja, N., Chilale, N., Ganesh,L, “FINANCIAL PERFORMANCE OF PRIVATE COMMERCIAL BANKS IN INDIA: MULTIPLE REGRESSION ANALYSIS”, *Academy of Accounting and Financial Studies Journal* *Volume 22, Issue 2,*2018,pp.1-12

OECD Data, (2019) *Consumer confidence index (CCI)* [Online] Available at <https://data.oecd.org/leadind/consumer-confidence-index-cci.htm> (Accessed: 07 March 2019)

OECD, (2018) *Employment rate* [Online] Available at <https://data.oecd.org/emp/employment-rate.htm> (Accessed: 06 March 2019)

Pandey, P., (2018) *A Guide to Machine Learning in R for Beginners: Decision Trees* [Online] Available at <https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb> (Accessed: 11 April 2019)

Pavan, (2018) *bank marketing analysis* [Online] Available at <https://rpubs.com/pavan721/bank_marketing?fbclid=IwAR2e1MIEqomxKPwWKyu-A-lDv4coANU9wkF1c1Iv0LngSNB7MLR7zgrw5XI> (Accessed: 01 February 2019)

Prabhakaran, S., (2017) *Linear Regression* [Online] Available at <http://r-statistics.co/Linear-Regression.html>

Quick, J.,(2009) *R Tutorial Series: Summary and Descriptive Statistics* [Online] Available at <https://www.r-bloggers.com/r-tutorial-series-summary-and-descriptive-statistics/> (Accessed: 12 February 2019)

Raman, S., (2017) JUST ENOUGH R Learn data analysis with R in a day, 1st ed., Sivakumaran Raman

Ratner, B.,(2017) *Statistical and Machine-Learning Data Mining,*USA, 3RD ed., CRC Press Taylor & Francis Group

Ratogi, R., Shim, K., *A Decision Tree Classifier that Integrates Building and Pruning*, USA-Korea

RDocumentation,(2017) *nrow* [Online] Available at <https://www.rdocumentation.org/packages/base/versions/3.5.3/topics/nrow> (Accessed: 03 April 2019)

RDocumentation,(2017) *sample* [Online] Available at <https://www.rdocumentation.org/packages/base/versions/3.5.3/topics/sample> (Accessed: 03 April 2019)

Reddy, A.,(2016) *Building Machine Learning* Models [Online] Available at https://rpubs.com /arjunreddyt/190610 (Accessed: 01 February 2019)

Sas Institute, (2019) *Machine Learning, what it is and why it matters* [Online] Available at <https://www.sas.com/en_gb/insights/analytics/machine-learning.html?fbclid=IwAR37oFVSsunJBtn4UuTTo92rtNqv3wKT0NODkCLnpoEyHn2drL03K7NEhBY> (Accessed: 21 January 2019)

Saxena, R., (2017) *DECISION TREE CLASSIFIER IMPLEMENTATION IN R* [Online] Available at <http://dataaspirant.com/2017/02/03/decision-tree-classifier-implementation-in-r/> (Accessed: 11 April 2019)

ScienceDirect, (2019) *Gaussian Distribution* Online] Available at <https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gaussian-distribution> (Accessed: 01 April 2019)

Seber,G., Lee, A., (2003) *Linear Regression Analysis,*2nd ed., Canada; John Wiley & Sons, Inc.

Seo, S., (2006) *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets* [Online] Available at <http://d-scholarship.pitt.edu/7948/1/Seo.pdf> (Accessed: 15 March 2019)

Sharma, N., (2018) *Ways to detect and Remove the Outliers* [Online] Available at <https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba?fbclid=IwAR137iw_0RApUNrR0dmrNCjersgML65Afjl96giDAd3E3zRa4chnff7Gw_s> (Accessed: 1 April 2019)

STHDA, (2018) *Multiple Linear Regression in R* [Online] Available at <http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/> (Accessed: 13 April 2019)

STHDA, (2018), *Simple Linear Regression in R* [Online] Available at <http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/> (Accessed: 13 April 2019)

Sullivan, L., LaMorte, W., (2016) *InterQuartile Range (IQR)* [Online] Available at <http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html> (Accessed: 1 April 2019)

Sunil, R., (2017) *Essentials of Machine Learning Algorithms (with Python and R Codes)* [Online] Available at <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?fbclid=IwAR3f02oL29naMAES6y06YX8iX9-Pdop_Mtbx-YmxBIvxwrCs6lQfWUTWhbY> (Accessed: 05 February 2019)

Tang, H., (2014) *A Comparison of Two Modeling Techniques in Customer Targeting For Bank Telemarketing*, Department of Mathematics and Statistic, Georgia State Univerisity

Taylor, C., (2018) *What Is the Interquartile Range Rule?* [Online] Available at <https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244> (Accessed: 1 April 2019)

Taylor,C., (2017) *How Are Outliers Determined in Statistics?* [Online] Available at <https://www.thoughtco.com/what-is-an-outlier-3126227> (Accessed: 13 March 2019)

Thanaki,J., (2018)*Machine Learning Solutions*, UK: Birmingham, Packt Publishing Ltd

The Royal Society, (2019) *Machine learning* [Online] Available at <https://royalsociety.org/topics-policy/projects/machine-learning/?fbclid=IwAR0cKIleDb3WYWXcxO-OhW5DlOMyzbZx8Y09xdZO718ShkTpcBwPoWjyBmE> (Accessed: 21 January 2019)

Tommoy, D., (2014) R for Data Science, Packt Publishing Ltd, ISBN 9781784392659

UCI Irvine machine learning repository, (2014) *Bank Marketing Data Set* [Online] Available at [https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) (Accessed: 12 October 2018)

Upadhyay,D., Bhavsar, C., Multiple Regression: A Multivariate Technique to Study Bank Sector, *International Journal of Science, Engineering and Technology*, 2015, Volume 3 Issue 6, pp. 20-22

Usuelli, M., (2014) *R Machine Learning Essentials,* UK: Birmingham, Packt Publishing Ltd, 1st ed., ISBN 978-1-78398-774-0

Wesley, (2013) *A Brief Tour of the Trees and Forests* [Online] Available at <https://www.r-bloggers.com/a-brief-tour-of-the-trees-and-forests/> (Accessed: 10 April 2019)

Willems, K., (2018) Reading and Importing Excel Files into R [Online] Available at <https://www.datacamp.com/community/tutorials/r-tutorial-read-excel-into-r#step2> (Accessed: 08 February 2019)

Willems, K., (2018) This R Data Import Tutorial Is Everything You Need [Online] Available at <https://www.datacamp.com/community/tutorials/r-data-import-tutorial> (Accessed: 08 February 2019)

Yamahata, H., (2019) *Bank Marketing, Bank Marketing, UCI Dataset* [Online] Available at <https://www.kaggle.com/henriqueyamahata/bank-marketing/home?fbclid=IwAR1XABxmjniLQVrZc_WGS9wN83GHQ95Kmb1JQqLw6Z_oIiFfDRF_7xH63eg> (Accessed: 01 February 2019)

Yan, X., Su, X., (2009) *Linear Regression Analysis: Theory and Computing*, 1st ed., Singapore; Hackensack, N.J.: World Scientific Pub. Co.

Yau, C., (2019) *Quartile* [Online] Available at <http://www.r-tutor.com/elementary-statistics/numerical-measures/quartile> (Accessed: 12 February 2019)

Zaki, M., Meira,W., (2014) Data mining and analysis: Fundamental Concepts and Algorithms, USA: New York, University of Cambridge

Zhang, C.,(2016)*Machine Learning on Bank Marketing Data* [Online] Available at <https://nycdatascience.com/blog/student-works/machine-learning/machine-learning-retail-bank-marketing-data/?fbclid=IwAR360rd_Q4ETzaZcC5hhkaw0rGEwNzqAr2KR9GGdN7kRtVmGftGsba0RoRY> (Accessed: 24 January 2019)

Zumel, N., Mount, N., (2016) *Exploring Data Science,* USA, Manning Publications, ISBN 9781617294181

Zumel, N., Mount, N., (2016) *Practical Data Science with R*, 2nd ed., USA, Manning Publications, ISBN 9781617294181