National College of Ireland

Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques

MSc Research Project
Data Analytics

Yogesh Sanjay Golecha
x16137272

School of Computing
National College of Ireland

Supervisor: Lisa Murphy

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Yogesh Sanjay Golecha
Student ID:	x16137272
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Lisa Murphy
Submission Due Date:	11/12/2017
Project Title:	Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques
Word Count:	6160

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques

Yogesh Sanjay Golecha
x16137272
MSc Research Project in Data Analytics

11th December 2017

Abstract

This paper proposes to develop a predictive model to analyse the behavior of the customer in the banks, whether they will be applying for long-term deposit in the banks or not. The dataset is from UCI machine learning repository for the Portuguese Banking Institution for the direct marketing Campaigns. The predictive model has been developed using various machine learning techniques like Adaptive Boosting, Support Vector Machines, Logistic Regression and Decision Trees. The aim for the research is to develop a predictive model that can help the banks in acquiring more knowledge of the customers behavior for making long-term deposits in the bank and to get detailed idea about what factors contribute in achieving higher predictions. This analysis can help banks in maintaining their customers and to avoid financial risks in the banks. The model developed uses only some basic attributes of the customers that can be easily accessed and gathered by the banks. The predictive model is trained and tested for each machine learning algorithm and finally compared based on accuracy obtained by each model. From the accuracy obtained for all the four algorithms, it has been observed that Adaptive Boosting(Adaboost) has the highest accuracy and high ability to predict the behavior of the customer for applying long-term deposit in banks.

1 Introduction

1.1 Project Background and Motivation

A sector that play a very significant part in the Commercial and Economic backdrop of any country is the banking sector. Data Mining technique can play a key role in providing different methods to analyse data and to find useful patterns and to extract knowledge in this sector (Vajiramedhin and Suebsing; 2014). Data mining helps in the extraction of useful information from the data (Turban et al.; 2011). According to (Venkatesh and Jacob; 2016), machine learning has more capability to gather information from the data, which results in more frequent use of data mining methods in the banking sector. Due to a large amount of data gathered in banks, data warehouses are required to store these data. Analyzing and identifying patterns from such data can be useful for Banks

to identify trends and acquire knowledge from these data. By the acquired knowledge from these data, organizations can more clearly understand their customers and improve the services they provide. Such an understanding of data can help organizations to gain success and improve the decision support system. As stated by (Raorane and Kulkarni; 2011), customers behavior must be understood by any organization to improve their business.
(Moro et al.; 2013) says, Analyzing the banks information and understanding the regular patterns can help banks to give better administrations to their clients. From the Bank Telemarketing information, different examples can be dissected, and learning can be extricated to give better consumer loyalty and to make significant strides towards Mining valuable data from the information. As stated by (Keller and Kotler; 2015), to enhance any business, advertising efforts assumes an essential part in drawing in the clients to the administrations gave by the associations.

1.2 Research Question

How can Predictive Analysis Using Multiple Machine Learning Techniques support decision making process in the banking sector to improvise the model to predict weather a customer will apply for long-term deposit in bank and in improving the business of the bank?

2 Literature Review

2.1 Introduction

According to (Suebsing and Vajiramedhin; 2013), Many organizations before offering the services to their customers, analyse the data from the previous customers and takes decisions to avoid any failure for the campaigns. Predicting such bank data of the customers can help in finding the hidden patterns a help in success for such marketing campaigns. According to (Moro et al.; 2013), due to the worldwide budgetary crisis, the credits for banks are limited, and the focus for banks is to accumulate funds from their customers. So, accumulating such data and providing services according to that can be of immense help to gain successful marketing campaign. Issues such as Behavior, Psychology, Mindset and Motivation needs to be considered to analyze and improve the marketing ability for the organizations (Raorane and Kulkarni; 2011). Managing and Maintain the data of the customers can help in getting patterns and trends to analyze to generate new strategies to attract new customers (Fayyad et al.; 1996). Machine Learning has a better ability to capture meaningful patterns from the data, also applications of the data mining methods in the sector of banks is increasing enormously (Venkatesh and Jacob; 2016). Classifications can be performed using machine learning algorithms, which can be used to segregate the data into distinct categories (Radhakrishnan et al.; 2013).

2.2 Related Work

According to the analysis performed by (Moro et al.; 2014), they used data mining for analyzing the direct marketing dataset for the Portuguese bank using CRISP-DM methodology. The goal for their study was to develop a predictive model for improving the effectiveness of the direct marketing campaigns by reducing the contacts performed using

phone. In the model developed by (Moro et al.; 2014), lift analysis plays an important part for the decision-making process. (Moro et al.; 2014) used 29 features for analyzing the data. They used AUC and ALIFT to analyze the algorithms. The techniques applied were SVM, Decision Tree and Neural Networks. According to their research, attribute Call Duration contributed the most as compared to other attributes. So, in our proposed model Call Duration has been included for the analysis. Under the analysis using different classification algorithms, they found SVM to provide the best AUC with 0.938.
According to the study by (Moro et al.; 2011), where they used CRISP-DM methodology for then bank telemarketing data to compare Nave Bayes, Decision Tree and Support Vector Machine(SVM). The results obtained by (Moro et al.; 2011) were, that SVM gave higher predictive accuracy as compared to Nave Bayes and Decision Trees.
According to the analysis performed by (Apampa; 2016), where he analyzed that the results obtained for the balanced dataset using 17 attributes, were more accurate as compared to the original unbalanced dataset. In the model, they compared Decision Tree, Nave Bayes and Logistic Regression, where he found that the Decision Trees performed more better with AUC and CA value of 76.60 percent. The main aim of (Apampa; 2016), was to understand whether balanced or unbalanced data would give him higher accuracy. The results were the increase in AUC (0.939) when the response variables were balanced.
According to (Elsalamony; 2014), he analyzed the bank marketing dataset using classification accuracy, specificity and sensitivity using 17 features. They used four different data mining techniques: MLPNN (Multilayer Perception Neural Network), TAN (Tree Augmented Nave-Baye), LR (Logistic Regression) and C5.0. In the research performed by them, they found that C5.0 performed better as compared to the other three data mining models.
According to the study performed by (Nachev; 2015) for the bank telemarketing dataset, using 70 percent training data and 30 percent testing data using four different data mining models: Logistic Regression, Neural Networks, Nave Bayes and Quadratic Discriminant Analysis. Taking accuracy for the comparative analysis, the results were that the Neural Networks performed good as compared to other models except of its poorly saturated data. Also, QDA returned better characteristics when measured using AUC.
To identify different patterns in data, machine learning proves to be efficient (Bishop; 2006).
Prediction can be improved if the data that is being analyzed is of excellent quality and accurate viz, no missing values, noise and duplicates (Phillips; 2013). The main objective for the analysis is to identify the customers that have more probability to apply for the long-term deposit in the banks. We use different machine learning techniques to identify weather a customer will be interested in making term-deposits in the banks. Four machine learning techniques have been implemented by using tools and programming languages to predict whether a customer will invest in bank or not. The Techniques are:

- Adaptive Boosting, due to its speed of boosting and accuracy (Bühlmann and Hothorn; 2007)
- SVM, due to its classifications using hyperplanes (Cortes and Vapnik; 1995)
- Logistic Regression, due to its flexibility to allow arbitrary values (Nelder and Baker; 1972)
- Decision Trees, due to its easy features and quick analyzing abilities (Loh; 2011)

All the Classification algorithms mentioned above, resulted in acquiring better results when compared using the Accuracy. (Domingos; 2012) says, data mining is an important task for classification. The results were obtained by different authors for the classification algorithms were optimized for the bank telemarketing dataset using unique features. The performance evaluation was based on the classification error, sensitivity, specificity and accuracy. But the most common metric used by all the authors was the Accuracy.So, my research focus on accuracy for the algorithms. Also, picking the algorithms used before which performed best with the models previously researched and adding some new predictive algorithms and comparing them using R will result in new findings and accuracy for different machine learning techniques. Each technique is described in the implementation section.

3 Methodology

Our research project focuses of predicting the customer response for applying long-term deposit in the banks. To build a flow for our system, the model must be defined. For an efficient execution of the process, the flow must be implemented properly. After the study of the different modelling techniques, we have decided to use Cross Industry Standard for Data Mining (CRISP-DM) for our system as it suits best for our research. (Chapman et al.; 2000) describes CRISP-DM as popular methodology for deriving success of any Data Mining Project. It is a Six-stage process which can be used to define any business strategy for a successful implementation of the project. (Chapman et al.; 2000) states, for a success implementation and modeling of the large data, the CRISP-DM (Cross-Industry Standard Process for Data Mining) model would be an effective methodology. It is a six-stage process which is important for decision making in the organizations. As there is a concept of business understanding in Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques which is among the most significant part of the project CRISP DM is the most suitable methodology for the Project as compared to SEMMA and KDD. Below is the model for the CRISP-DM and the modified version of CRISP-DM that is designed according to the flow of our research project. According to the project there have been an inclusion of distinct phases in the methodology that is describe in the following section. The Original CRISP-DM and the Modified CRISP-DM is shown below:

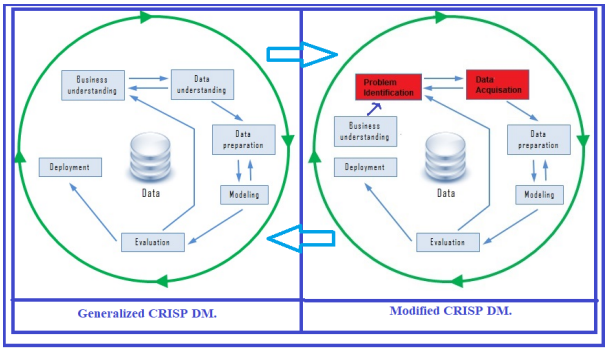


Figure 1: CRISP-DM

3.1 Problem Identification

This predictive model also keeps a focus on the marketing and investment returns campaigns so in accordance to the Business Understanding these features should also be identified. Identifying the problem related to the research is the basic step for the successful implementation of the project. The main goal for the research is to predict the correct response variable (yes or no), and to prevent banks from spending time and money on the customers, who would be less probabilistic to apply for the term deposits in the banks. This project uses basic attributes of the customers like age, marital status gender, duration of call, etc. to implement a predictive model which analyzes these attributes and predicts whether a customer is liable to deposit in bank or not.

3.2 Data Acquisition

Another phase that has been added and plays a significant part in the project is Data acquiring. The data that is acquired should consists of all the pre-requisite attributes based upon while then predictive model cane be formulated. For analysis, various dataset was considered, but the dataset we decided to use was the Portuguese Bank-additional dataset from the UCI machine learning repository due to its demographic set of variables and the large set of information provided as compared to other small data sets.
The Link for the Dataset is: https://archive.ics.uci.edu/ml/datasets/bank+marketing

3.3 Data Preparation

This is a very important stage in the process of data mining, as using incorrect data for analysis can lead to the inaccurate results and errors in the output. So, the process of data preparation needs to be considered important as the data that is fed for the implementation and modelling depends on the data accuracy and the structure of the data. Quality data preparation is the most important task for that data analysis process (Pyle; 1999). For the process of data preparation for our research Excel and RapidMiner is considered to execute the process. For the process of data preparation following tasks were performed:

- The dataset that we used consisted of the Missing values. The missing values were removed from the data to ensure the quality of the data. Also, the unknown values were removed to ensure the accuracy is correct.

- Data set contained various unknown categories in different columns. So, for the purpose of making a accurate dataset, it has been filtered and the unknown values were removed.

- The categorical attributes were encoded (nominal to numeric conversion) for the ease of modelling. Ex. Marital Status (1= Married, 2= Divorced, 3= Single).

- Feature Discovery and Data Elimination:

  The process of selecting relevant attributes for the model, is the main task in the data mining process, since it eliminates the irrelevant features and makes the model more accurate and it enhances its performance. (Pyle; 1999) states while automatic selection of features can be useful, the best way to perform feature selection

is to manually select the variables based on the knowledge of the domain or problem. So, a clear understanding for the attributes is the key feature in the process of feature discovery. Here we are going to select the attributes manually that we find best suits out research objective. The objective for our research is to get the customer response to apply with the bank based on the features that can be easily collected and analyzed. Also, some attributes were selected by using iterative attribute selection in rapid miner by testing different sets of attributes together to find which combination gives the best predictive model for our analysis. By using RapidMiner, editing the Read CSV operator multiple times for different attributes, the best group of the attributes was shortlisted, and our data was finalized. Identification of the attributes that can contribute more towards getting accuracy is the main purpose of Feature Discovery. The original dataset consisted of 21 columns out of which we have considered 10 columns based on the correlation with each other and the attributes that best suit out proposed idea. The attributes that we considered for out model are:

Attribute	Type	Description
Age	Numeric	Age of the Customer
Job	Categorical	Type of Job
Marital	Categorical	Marital Status of the Customer
Education	Categorical	Educational Level
Default	Categorical	Has Credit in Default or Not
Housing	Categorical	Housing Loan or not
Loan	Categorical	Personal Loan or Not
Duration	Numeric	Time spent by the Customer on The Call
Euribor	Numeric	Euribor 3 Months Rate
y	Categorical	Label Variable

Figure 2: Data Attributes and Their Description

- Outlier Detection: (Maddala and Lahiri; 1992) says, Outlier detection is the test to detect data entries that are different from the observations is important for the process of data preparation. The input data is used to check whether it contains any outliers or not. For example, if we consider age for detecting the outliers, it should be in the range from 1-99. It cant be 120 or something more. This is the main purpose of Outlier Detection. We have performed Outlier detection in RapidMiner, and from the graph, we can see that the data does not have any outliers.

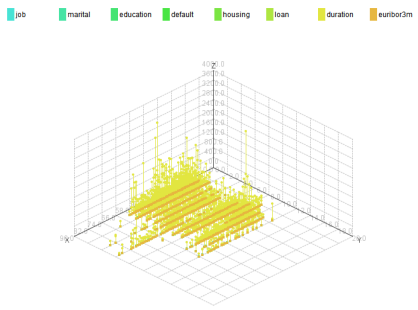


Figure 3: Outlier Detection

3.4 Prediction Modelling and Evaluation

Data that we have acquired through the Data Preparation stage is fed into modelling and evaluation. We have applied four different machine learning techniques for tested in R. The techniques are SVM, Decision Tree, Adaptive Boosting and Logistic Regression. We have applied k-fold cross validation to improve the accuracy for our model. For each model, we have tested the results and found the accuracy for each of the model for the label y which states weather the customer, based on the attributes selected for analysis, will apply for long-term deposit in bank or not. The output is in the terms of Yes or No. Each model gives the prediction table and the accuracy for each technique. The below diagram shows the flow for our model. The steps are described in detail in the Implementation section.

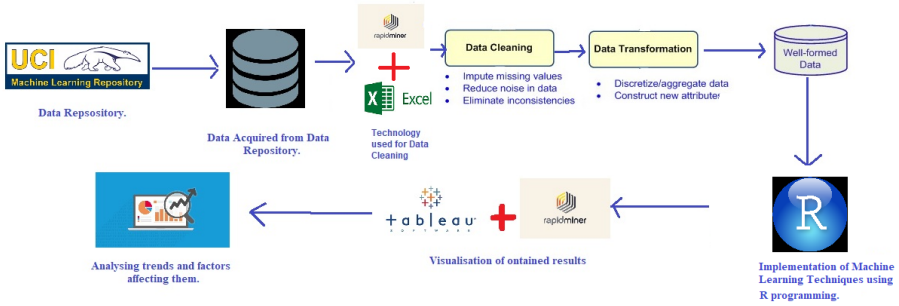


Figure 4: Process Flow

Note: The process of implementation for all the four algorithms(SVM, Decision Tree, Logistic Regression and Adaptive Boosting) was performed both in R as well as RapidMiner. But same results with a very low differences were obtained. As, the focus of the project is on analyzing the customers for long-term deposits in the banks, and the technology comparison is not the focus of the research, so the overall paper illustrates the implementation phase in one technology that is R.

3.5 Implementation

This stage consists of implementing the Machine Learning Techniques for that data that we have selected for our model. For the Implementation, we have used four Machine learning techniques and each technique is tested in R to check which tool and technique helps in the maximum contribution and more accuracy for our model. Thereafter, the best model was evaluated by comparing the accuracy for all algorithms. All the algorithms are described in detail.

1. Basic Functions used in R:

    - read.csv ()- to read the data from the directory
    - summary ()- to show the data summary or information
    - predict ()- to predict the model object
    - prediction ()- extract the prediction values based on the predict () which returns a data frame
    - performance ()- to perform predictor evaluations for the model and derive its accuracy and other statistics
    - dim ()- to get the dimensional statistics for the test or train data
    - confusionMatrix ()- to derive the matrix and other statistical information (accuracy, sensitivity, specificity, kappa, etc.) for the model
    - set.seed ()- this function is used to set the random sequence for the starting number, to obtain the same results each time the code is executed

3.5.1 Logistic Regression

Logistic Regression is one of the most popular classification method in data mining. This method is being used when the target variable is in the binary form (Ex. Yes/no, good/bad, etc.). The equation of Logistic regression for fitting the curve is y=f(x) where is a categorical variable. The use of this model is to predict y for the given set of predictors x. In our model, the y is yes/no. The predictors can be mix-both continuous and categorical. To avoid the problem of class imbalance and to improve the performance for our model, k-fold validation of 10 folds has been applied to the data. We will discuss the results for both as follows:

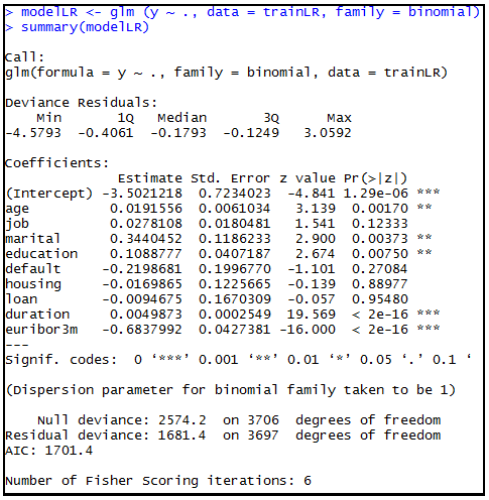
We have used RStudio to perform logistic regression. Steps performed, and results obtained are as follows:

- Performing basic operations:

  Basic operations include installing libraries, reading the dataset, checking the summary for the data and normalizing the data to avoid inconsistent data inputs. We derived head() and summary() for the data we used. Also, normalized data was checked before applying the algorithms onto it.

- Applying K-fold cross validation to the data:
  The data must be cross validated to estimate the performance as the imbalances class may affect the accuracy and performance of the model. So, the data has been applied 10 folds cross validation to enhance the performance for the model.

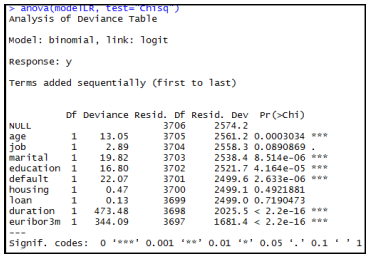
- Defining the model:
  To perform logistic regression in R, model must be applied. glm () is flexible for the predictor variables to be linear or binomial (Olsson; 2002). As, out predictor variable is binomial, we use the family=binomial. The output for the model summary is as follows:



```
> modelLR <- glm (y ~ ., data = trainLR, family = binomial)
> summary(modelLR)

Call:
glm(formula = y ~ ., family = binomial, data = trainLR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5793  -0.4061  -0.1793  -0.1249   3.0592

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.5021218  0.7234023  -4.841 1.29e-06 ***
age          0.0191556  0.0061034   3.139  0.00170 **
job          0.0278108  0.0180481   1.541  0.12333
marital      0.3440452  0.1186233   2.900  0.00373 **
education    0.1088777  0.0407187   2.674  0.00750 **
default     -0.2198681  0.1996770  -1.101  0.27084
housing     -0.0169865  0.1225665  -0.139  0.88977
loan        -0.0094675  0.1670309  -0.057  0.95480
duration     0.0049873  0.0002549  19.569  < 2e-16 ***
euribor3m   -0.6837992  0.0427381 -16.000  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2574.2  on 3706  degrees of freedom
Residual deviance: 1681.4  on 3697  degrees of freedom
AIC: 1701.4

Number of Fisher Scoring iterations: 6
```

Figure 5: Model for Logistic Regression

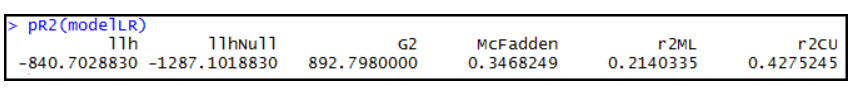
- Analyzing table of variance:
  To analyze the table for variance anova () has being used. The results obtained are as follows:



```
> anova(modelLR, test='Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       3706     2574.2
age        1    13.05      3705     2561.2 0.0003034 ***
job        1     2.89      3704     2558.3 0.0890869 .
marital    1    19.82      3703     2538.4 8.514e-06 ***
education  1    16.80      3702     2521.7 4.164e-05 ***
default    1    22.07      3701     2499.6 2.633e-06 ***
housing    1     0.47      3700     2499.1 0.4921881
loan       1     0.13      3699     2499.0 0.7190473
duration   1   473.48      3698     2025.5 < 2.2e-16 ***
euribor3m  1   344.09      3697     1681.4 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: Output for anova()

- Computing different pseudo-R Measures:
  Various pseudo-R2 measures have been equated for the model using the pR2() function. The generated output is given below:



```
> pR2(modelLR)
         llh      llhNull           G2     McFadden         r2ML         r2CU
-840.7028830 -1287.1018830   892.7980000     0.3468249    0.2140335    0.4275245
```

Figure 7: Output for pR2()

- Predictive ability of the Model:
  The predictive ability of the model is derived using predict(), prediction() and

performance() function. The accuracy curve is generated using plot() for the model. The accuracy derived for the logistic regression is 89.47 percent.
The curve obtained is as follows:

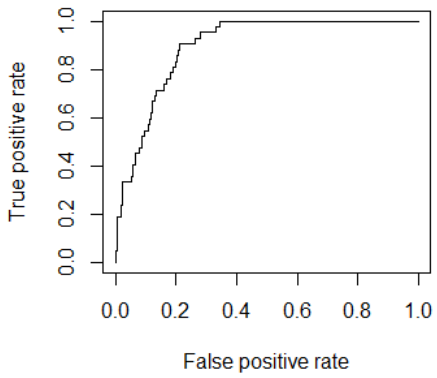


Figure 8: Accuracy Curve

3.5.2 Adaptive Boosting

Adaptive boosting is one of the popular machine learning algorithm, that is used for boosting the model. The main advantage of using this algorithm is it increases the competitive accuracy and quick learning capacity of the process (Ling and Li; 1998) (Friedman et al.; 2000). According to (Efron and Gong; 1983), this method increases the speed for prediction for any models.
We used RStudio to perform Adaptive boosting. Libraries used: gbm and cvAUC.

- The gbm (Generalized Boosting Models) is used to model the algorithm by using distribution= adaboost
- cvAUC () is used to derive the cross-validated accuracy for the model. We divided the data into train and test. Then we calculated the accuracy for the model. Below is the screenshot of the code for Adaptive Boosting.

The process for implementing Adaptive Boosting is as follows:

- Basic Operations:
  Reading the data, Installing the libraries, shuffling the data and using set.seed() to derive same results for every execution are the basic operations that were performed for the process of adaptive boosting.
- Applying k-fold validation to the model:
  The k-fold cross validation was applies using folds=10, so as to divide the data into training and testing sets.
- Defining the model and collecting its statistics:
  To fit any boosting model like adaboost, gradient boost, etc., gbm(Generalized Boosting Model) is a popular function as it takes less time to execute and give

better results (Ridgeway; 2007). The following is the output for using gbm () to apply adaboost model.

```
> model_GB <- gbm(formula = y ~ .,distribution = "adaboost",data = train_GB,n.trees = 100,interaction.depth = 5,
+             shrinkage = 0.3,bag.fraction = 0.5,train.fraction = 1.0,n.cores = NULL)
> print(model_GB)
gbm(formula = y ~ ., distribution = "adaboost", data = train_GB,
    n.trees = 100, interaction.depth = 5, shrinkage = 0.3, bag.fraction = 0.5,
    train.fraction = 1, n.cores = NULL)
A gradient boosted model with adaboost loss function.
100 iterations were performed.
There were 9 predictors of which 9 had non-zero influence.
```

Figure 9: Model for Adaptive Boosting

- Deriving the accuracy for the model:
  The accuracy that we derived using Adaptive Boosting is 93.37 percent. Adaptive Boosting is good in deriving better predictions for the long-term deposit customers, and we can say that this model is good to apply for such a dataset as this is giving higher accuracy as compared to other models.

3.5.3 Decision Tree

According to (Grzonka et al.; 2016), Decision trees works by separating data to groups. They are very easy to understand and to derive decisions from them. It works by analyzing the nodes by traversing the data until it reaches to its desired decisions. Decision Trees works by sub-dividing a large set of data into smaller subsets and then applying a simple set of decision rules. With the division of the decision tree, them members of the tree becomes more like each other. The main advantage of the decision trees is that it separated a large heterogeneous group into smaller homogeneous groups by considering the label variable citepradhakrishnan2013application.

R was used to perform Decision Trees and to derive the plot and accuracy for the model. Two libraries were used are explained as follows:
1) rpart ()- for recursive partitioning of classification trees
2) caret ()- for deriving the confusion Matrix.
The steps for implementing Decision tree is as follows:

- Installing rpart() and caret() libraries.
- Reading the data.
- Shuffling the data.
- Applying k-fold cross validation.
- Plotting the tree: The tree that has been derived is shown below:

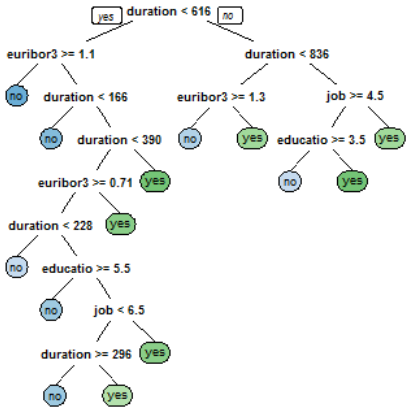


Figure 10: Decision Tree

- The Confusion Matrix and the Decision tree statistics like accuracy, kappa, sensitivity, specificity, etc. was derived. The accuracy that we got using Decision tree in R is 91.02 percent. So, after deriving the accuracy we can say that this model gives us the accuracy which can be competitive to compare with the other algorithms.

3.5.4 Support Vector Machine(SVM)

According to (Suykens and Vandewalle; 1999), SVM is used for the classification problems and proved to give better results for the predictive analysis. It uses Hyperplanes to find the distance between the attributes. The result is the hyperplane that gives the largest minimum distance for the examples.

In SVM, three libraries are used:
1) PerformanceAnalytics- for analyzing the performance for attributes
2) tabplot- to plot the charts
3) kernlab- to define different SVM kernels (ex. Polydot, rbfdot, etc.) for machine learning
Each library is evaluated and tested with the data to find which suits the best for our model. Also, correlation charts and table plots are derived to estimate the structure and interdependency between the variables. Data is divided into training and testing. Three kernels are tested (viz. polydot, rbfdot and tanhdot) to find which kernel fits the best for our model. The process is as follows:

- Basic Operations Performed: Basic operations are installing the libraries (tabplot,kernlab,e1071 and performanceAnalytics),using set.seed as 10, normalizing the data and deriving the summary for the data, so as to check the data consistency.

- Applying K-fold validation: To ensure the accuracy and to avoid the problem of class imbalance, k-fold cross validation has been applied to the dataset, where folds=10.

- Deriving tableplot and correlation matrix: The plots derived for the dataset are as follows:

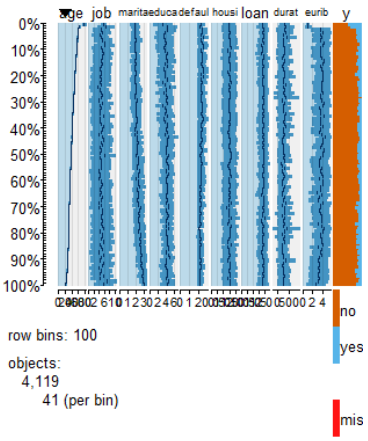


Figure 11: Table Plot for SVM

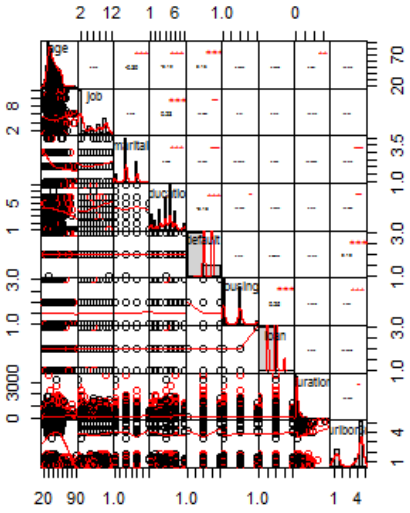


Figure 12: Correlation Matrix for SVM

- Model for SVM using Polynomial Kernel:
  The following is the model for the kernel polydot-

```
> model_SVM<-ksvm(y ~ .,data=train_SVM, kernel="polydot")
 Setting default kernel parameters
> model_SVM
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Polynomial kernel function.
 Hyperparameters : degree =  1  scale =  1  offset =  1

Number of Support Vectors : 796

Objective Function Value : -791.7107
Training error : 0.103318
```

Figure 13: Model for SVM using Kernel polydot

Result Obtained:

```
                Accuracy : 0.8908
                  95% CI : (0.8566, 0.9192)
    No Information Rate : 0.8811
    P-Value [Acc > NIR] : 0.302
```

Figure 14: Output for SVM using Kernel polydot

- Model for SVM using Hyperbolic Kernel:
  The following is the model for the kernel tanhdot-

```
> Hyperbolic<-ksvm(y ~ .,data=train_SVM, kernel="tanhdot")
 Setting default kernel parameters
> Hyperbolic
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Hyperbolic Tangent kernel function.
 Hyperparameters : scale =  1  offset =  1

Number of Support Vectors : 609

Objective Function Value : -16035.89
Training error : 0.163744
```

Figure 15: Model for SVM using Kernel tanhdot

Result Obtained:

```
                Accuracy : 0.8252
                  95% CI : (0.7851, 0.8607)
    No Information Rate : 0.8811
    P-Value [Acc > NIR] : 0.9996
```

Figure 16: Output for SVM using Kernel tanhdot

- Model for SVM using Radial Kernel:
  The following is the model for the kernel rbfdot-

```
> Radial<-ksvm(y ~ .,data=train_SVM, kernel="rbfdot")
> Radial
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0852294879759811

Number of Support Vectors : 838

Objective Function Value : -718.9705
Training error : 0.085514
```

Figure 17: Model for SVM using Kernel rbfdot

Result Obtained:

```
               Accuracy : 0.9029
                 95% CI : (0.8701, 0.9297)
    No Information Rate : 0.8811
    P-Value [Acc > NIR] : 0.09535
```

Figure 18: Output for SVM using Kernel rbfdot

- Comparing output for SVM using different Kernels: Three different kernels were tested with SVM. The Polynomial Kernel(polydot). The Hyperbolic Kernel(tanhdot) and the Radial Kerenel(rbfdot). Accuracy derived using different kernels is 1) polydot- 89.08, 2)tanhdot- 82.52 , 3)rbfdot- 90.29. So, we select the best accuracy that we derived using kernel rbfhdot. So, we can state that for our SVM model the Radial kernel worked the best.

4 Evaluation

In the stage of Evaluation, the decision is made weather which model fits the best for our data to predict which Customer is more liable to apply for the Term-Deposit in the Bank. Evaluation of the Models is based on the performance of the models in terms of their Predictive Accuracy. Accuracy for each model is tested in R and visualized in R. Model with higher accuracy is selected for the Evaluation and Decision-Making Process. The Results are analyzed based on the number of predictions correctly classified by the algorithms.

Below are the table showing accuracy obtained using R:

Data Mining Techniques	Accuracy
Adaboost	93.37%
Decision Tree	91.02%
Support Vector Machines	90.29%
Logistic Regression	90.22%

Figure 19: Comparing Accuracy

From the model we designed, the Adaboost gave the highest Accuracy of 93.37 percent, which was followed Decision Tree u with 91.02 percent of accuracy and so on. Also, by analyzing the Weight by Gain Ratio for our model, it has been derived that the highest contributing factors for our dataset is Duration which is followed by the Euribor Rate. The chart illustrating the result for the Weight by Gain Ratio was derived using Rapid-Miner is as follows:

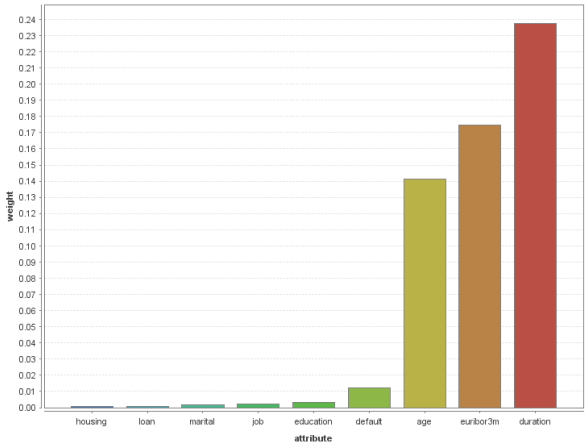


Figure 20: Weight by Information Gain Ratio Chart

5 Visualization

To analyze and understand the data that we have researched, some Visualizations were performed to understand the data, to get a brief overview of the dataset that we have chosen for the research. The visualizations were performed in Tableau, as it is known to be a very popular and powerful tool to perform visualizations. The Visualizations are as follows:

- Result for Time Spent on Calls based on Age and Job Role:
  People who have an account in bank spend some time on calls with the banks to understand the offers given by the banks to their customers. However, the details for the time the customers are on the calls with them are recorded in the banks database. So, analyzing the time by considering the age of the customer and the job role has given us some interesting information and understandings for the age groups and the job role of the people who are more likely to apply for the Term-Deposit in Bank. The Visualization is given below:

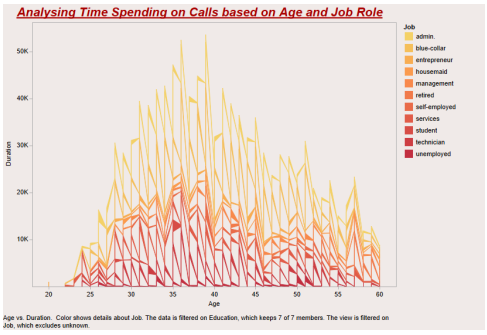


Figure 21: Duration Of Calls Based on Age and Job

Results: The results for the above visualization states that the people in the age group of 35-45 are more likely to invest in the banks. Also, it shows that the people

with the jobs roles like administrators and blue-collar jobs are more likely to invest as compared to other job roles. So, banks who tends to attract more customers, should focus of the two criteria of age group and job profiles, so that they can get a successful marketing campaign and get more customers.

- Marital Status of the people applying for Term-Deposit:

  As we know that banks always have the information about their clients. Based on such an information like the marital status of the customer, it can be of immense help to understand that which category (married, single or divorced) of people are more likely to invest in the banks.

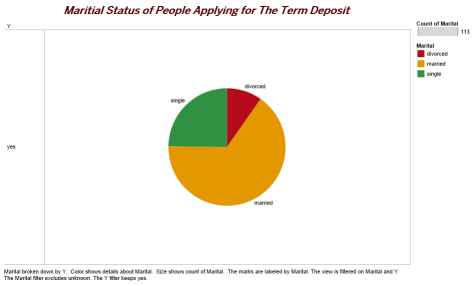


Figure 22: Marital Status for People Applying In Banks

Results: From the visualization it can be analyzed that the people who are married are more probabilistic in applying for the term-deposits in the banks. So, it can be said that the banks should keep on managing with the customers who are married, and try to focus and provide better services to the other two categories of the customer, so as to attract them for making the Term-Deposits in the banks.

- Dependency between Euribor Rate and the People applying for Deposits in Banks The Euribor (Euro Interbank Offered Rate), which is published by the European Money Markets Institute, are the daily interest rates. The reason to perform analysis on this attribute was to find relation between the Euribor rate and the people applying for the term deposits in the banks.

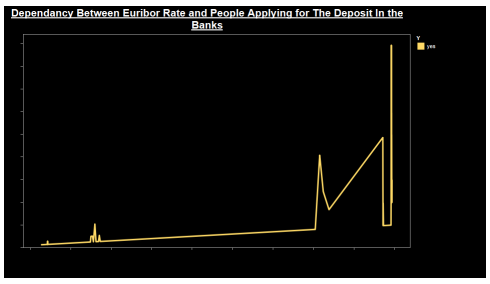


Figure 23: Marital Status for People Applying In Banks

Results: From the above visualizations, the increase in the interest rates makes more customers attracted towards investing in the banks. So, banks should always

analyze the Euribor rates before offering services so as to get a better customer response.

- Educational Status of people applying for Term-Deposit in Banks

  Education plays a vital role in the behavior and decision-making ability for any person. Analyzing such information can result in understanding the basic structure of the customers behaviors towards applying for Term-Deposits in Banks and educational qualifications. The visualization for the same is as follows:

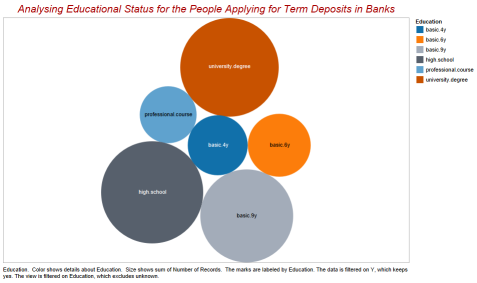


Figure 24: Educational Status for People Applying In Banks

Results: From the above visualization, it can be derived that the people holding a University Certificate or High School Certificate are probability to apply in banks as compare to the people holding other qualifications.

6 Conclusion and Future Work

The main aim of the research was to analyze the Customer behavior for making a Term deposit in banks, by considering some basic attributes that can be collected and analyzed easily. The research was evaluated using four different machine learning techniques (Adaptive Boosting, SVM, Decision Trees and Logistic Regression). Also, all these techniques were tested using R. The best results were obtained for performing Adaptive Boosting with the highest predictive accuracy of 93.37 percent. Also, the most contributing variable for the prediction was Duration (Duration of time spent by the customer on the calls) which was followed by the Euribor Rate. The implemented system performs a predictive analysis for the customers who are more responsive to apply for the term-deposits in the bank. The model can help banks in focusing and providing services to the customers who have more probability to apply for term-deposits in the bank. In Future, the model can be improved by considering a large dataset as compared to our model. Also, some ensemble models must be used to get more accuracy.

References

Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction, *Journal of International Technology and Information Management* **25**(4): 6.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, springer.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* pp. 477–505.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Cortes, C. and Vapnik, V. (1995). Support vector machine, *Machine learning* **20**(3): 273–297.

Domingos, P. (2012). A few useful things to know about machine learning, *Communications of the ACM* **55**(10): 78–87.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *The American Statistician* **37**(1): 36–48.

Elsalamony, H. A. (2014). Bank direct marketing analysis of data mining techniques, *International Journal of Computer Applications* **85**(7).

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*, Vol. 21, AAAI press Menlo Park.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), *The annals of statistics* **28**(2): 337–407.

Grzonka, D., Suchacka, G. and Borowik, B. (2016). Application of selected supervised classification methods to bank marketing campaign, *Information Systems in Management* **5**(1): 36–48.

Keller, K. L. and Kotler, P. T. (2015). *Framework for Marketing Management*, Pearson.

Ling, C. X. and Li, C. (1998). Data mining for direct marketing: Problems and solutions., *KDD*, Vol. 98, pp. 73–79.

Loh, W.-Y. (2011). Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(1): 14–23.

Maddala, G. S. and Lahiri, K. (1992). *Introduction to econometrics*, Vol. 2, Macmillan New York.

Moro, S., Cortez, P. and Laureano, R. (2013). A data mining approach for bank telemarketing using the rminer package and r tool.

Moro, S., Cortez, P. and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing, *Decision Support Systems* **62**: 22–31.

Moro, S., Laureano, R. and Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology, *Proceedings of European Simulation and Modelling Conference-ESM'2011*, Eurosis, pp. 117–121.

Nachev, A. (2015). Application of data mining techniques for direct marketing, *Computational Models for Business and Engineering Domains* .

Nelder, J. A. and Baker, R. J. (1972). *Generalized linear models*, Wiley Online Library.

Olsson, U. (2002). Generalized linear models, *An applied approach. Studentlitteratur, Lund* **18**.

Phillips, R. (2013). Optimizing prices for consumer credit, *Journal of Revenue and Pricing Management* **12**(4): 360–377.

Pyle, D. (1999). *Data preparation for data mining*, Vol. 1, morgan kaufmann.

Radhakrishnan, B., Shineraj, G. and Anver Muhammed, K. (2013). Application of data mining in marketing, *IJCSN International Journal of Computer Science and Network, ISSN (Online)* pp. 2277–5420.

Raorane, A. and Kulkarni, R. (2011). Data mining techniques: A source for consumer behavior analysis, *arXiv preprint arXiv:1109.1202* .

Ridgeway, G. (2007). Generalized boosted models: A guide to the gbm package, *Update* **1**(1): 2007.

Suebsing, A. and Vajiramedhin, C. (2013). Accuracy rate of predictive models in credit screening, *Applied Mathematical Sciences* **7**(112): 5591–5597.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers, *Neural processing letters* **9**(3): 293–300.

Turban, E., Sharda, R. and Delen, D. (2011). *Decision support and business intelligence systems*, Pearson Education India.

Vajiramedhin, C. and Suebsing, A. (2014). Feature selection with data balancing for prediction of bank telemarketing, *Applied Mathematical Sciences* **8**(114): 5667–5672.

Venkatesh, A. and Jacob, S. G. (2016). Prediction of credit-card defaulters: A comparative study on performance of classifiers, *International Journal of Computer Applications* **145**(7).