Predicting the Success of Bank Telemarketing using various Classification Algorithms

Author: **Muneeb Asif**

3rd semester

Degree project: 2nd Cycle HT17, 15 hp

Subject: Independent Project I

Masters in Applied Statistics

Örebro University School of Business

Supervisor: Farrukh Javed, Senior Lecturer, Örebro University

Examiner: Nicklas Pettersson, Senior Lecturer, Örebro University

Abstract

In this thesis, well known methods of classification Support Vector Machine, Decision Trees, Random Forest and Artificial Neural Network have been performed. To reduce the dimensionality, feature selection best subset Logistic Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest approaches have been utilized. The focus was to check the prediction accuracy and performance of these classification methods after feature selection on full as well as reduced model. The output showed that reduce subset of variables which is attained through Random Forest has best accuracy with the classification method random forest. In this way, we can rely on the subset of variables obtained from Random Forest as it has almost same percentage of accuracy as compared to the full model.

Acknowledgements

First of all, I am thankful to my supervisor **Farrukh Javed** whose expertise, guidance and support made it possible for me to work on this thesis.

I would also like to extend my thanks for my examiner, **Nicklas Pettersson** for his thoughts and feedback which helped me to improve this thesis more.

I would also like to express my gratitude towards my **Parents** for their encouragement which facilitated me to complete this thesis on time.

Contents

1 Introduction

Marketing is technique of exposing the target clients to a product via suitable systems and channels. It ultimately facilitates the way to buy the product or service and even helps in determining the need of the product and persuade customers to buy it. The overall aim is to increase sales of products and services for enterprise, business and financial institutions. It also helps to maintain the reputation of the company.

Telemarketing is form of direct marketing in which salesperson approaches the customer either face to face or phone call and persuade him to buy the product. Telemarketing attains most popularity in 20th century and still gaining it. Nowadays, telephone (fixed-line or mobile) has been broadly used. It is cost effective and keeps the customers up to date. In Banking sector, marketing is the backbone to sell its product or service. Banking advertising and marketing is mostly based on an intensive knowledge of objective information about the market and the actual client needs for the bank profitable manner.

Making right decisions in organizational operations are sometimes proved a great challenge where the quality of decision really matters. Decision Support Systems (DSS) are classified as a particular class of computerized facts and figures that helps the organization or administration into their decision making actions. The concept of DSS originates from a balance which lies between the data generated by computer and the judgment of human. According to Rupnik & Kukar (2007) the objective of decision support systems is to enhance the effectiveness of the decisions. This is a great tool which can analyze the sales data and provide further predictions. The purposes which can be established from the DSS are such as, analysis, optimization, forecasting and simulation. A study by Power (2008) found that research subjects who use DSS for the decision making, come-up with more effective decisions than those who did not use it. Nowadays, DSS is contributing a meaningful role in many fields such as for medical diagnosis, business and management, investment portfolios, command and control of military units, and statistics.

DSS uses statistical data to overcome the deficiencies and helps the decision makers to take the right decision. Data mining (DM) plays vital role to support the Decision support systems which are based on the data obtained from the data mining models: rules, patterns and relationship. Data mining is the process of selecting, discovering, and modeling high volume of data to find and clarify unknown patterns. The objective of data mining in decision support systems is to suggest a tool which is easily accessible for the business users to analyze the data mining models.

A specific technology used within the DSS is Machine learning (ML) that combines data and computer applications to accurately predicting the results. The fundamental principle of machine learning is to construct the algorithms that can obtain input data and then predict the results or outputs by using the statistical analysis within satisfactory interval. ML allows the DSS to obtain the new knowledge which helps it to make right decisions.

Machine Learning can be mainly classified in 2 categories i.e. supervised learning and unsupervised learning. In supervised learning, the output of algorithm is already known and we use the input data to predict the output. The examples of supervised learning are regression and classification. In contrast, unsupervised learning we only have input data whereas no corresponding output variables are selected. The example of unsupervised learning is clustering.

Feature selection is the process of selecting the subset of relevant variables from the model. It identifies the most important attributes which help to predict the output. By using this techniques, we can reduce the curse of dimensionality, prevent model from overfitting and shorter the training time. In this way parsimonious model can be achieved with minimum number of parameter and good explanatory predictive power.

1.1 Purpose

The purpose of this study is to check the accuracy and performance of several models for classification. The full model which is used in this study consists of 16 independent variables. Feature selection approach has been used to select the best subsets of variables and then different type of classification algorithms have been utilized to check their accuracy and performance. The full model is then compared with the reduced model, obtained through feature selection, in terms of classificaion accuracy.

1.2 Outline

In the first section, introduction of decision support systems, data mining and machine learning have been defined. Section 2 is about the literature review which states the previous researches in the field of classification and machine learning. In section 3, methods of statistical techniques have been described which are utilized in this study work, section 4 explains the detail of data and variables. Section 5 is about theory and models which presents the detail of feature selection approaches, classification methods and their algorithms. Section 6 is about the results which are output of these statistical methods. In section 7, the discussion and conclusion have been discussed. Section 8 is references and section 9 is for appendix in which figures and tables are included.

2 Literature Review

This section explains the previous research work which have been already done in classification using ML techniques.

The data which is used in this study work is the data of customers of a Portuguese banking institution. The similar data set has been used in Moro et al. (2011, 2014). In Moro et al. (2011), the aim of this study was to find the model that can increase the success rate of telemarketing for the bank. The statistical techniques of data mining which have been used in their research are Support Vector Machine (SVM), Decision Tree (DT) and Naive Bayes. The performance of these models was checked through the Receiver Operator Characteristics (ROC) curve (detail of ROC curve is given in section 5). Among all these statistical techniques, SVM comes up with the most efficient results. Regarding attributes, Call duration was the most relevant feature which states that longer calls tend increase the success rate. After that month of contact, number of contacts, days since last contact, last contact result and first contact duration attributes respectively.

In Moro et al. (2014), objective of the study was to predict the success of bank telemarketing. Data set which they used in their research was consists of 150 attributes and is complete data

set of the period 2008 to 2013. They compare the 4 data mining models i.e. Logistic Regression (LR), Decision Tree, Support Vector Machine and Neural Network (NN). The best result was obtained by the neural network while decision trees discloses that probability of success in inbound calls are greater.

Statistical learning algorithms have successfully been used in many research problems for classification. For example, Qi et al. (2018) conducted a research to find out the fault diagnosis system for reciprocating compressors. Reciprocating compressors are extensively used in petroleum industry. Data was taken from oil corporation (5 years operational data) and uses the Support Vector Machine to analyze it. They come up with the results that SVM accurately predicts the 80% right classification to find the potential faults in compressor.

Similarly, Gil & Johnsson (2010) did a research in medical field for diagnosing the urological dysfunctions using SVM. In this research data was collected from the 381 patients who are suffering from a number of urological dysfunctions to check the overall worth of decision support system. The fivefold cross validation has been utilized for the robustness. The outputs of this study describe that for the purpose of classification SVM attained the accuracy of 84.25% .

Nogami et al. (1996) utilized the machine learning in decision support system. In their research they introduce the air traffic management for the future which can manage the flight schedule and flow of air traffic professionally. Their system involves many decision makers and utilized it with the neural network. They require such system which can make the optimal decision in the critical situation. Their simulation studies prove that system which is based on neural network is performed more efficiently than the current air traffic system.

Another research by Cramer et al. (2017) the machine learning methods are used in time series for rainfall prediction. Data was derived from the 42 cities including climatic features.They tried Support vector regression, NN, and k nearest neighbors. After performing these methods they come up with the results that machine learning methods have predictive accuracy.

Wang & Summers (2012) used the machine learning in field of radiology. They used it for the neurological disease diagnosis images, medical image segmentation and MRI images. They come-up with the results that machine learning identifies the complex patterns. It also helps the radiologists to make right decisions. Furthermore, they suggest that development of technology in machine learning is an asset for long term in the field of radiology.

Machine learning algorithms are also used in the field of applied mathematics. For instance, Barboza et al. (2017) did a research to predict the models for developing of bankruptcy by using the SVM and random forest methods. The data was taken from the Salomon Center database & Compustat about North American firms from period 1985 to 2013 with observations of more than 10,000. After applying SVM and RF techniques they compare the results with the ordinary used methods such as discriminant analysis and logistic regression. They concluded that ML techniques are come up with 10% averagely more accurate results than usual methods.

To find the risk factors about failure of banks Le & Viviani (2017) conducted a research. In their study, a sample of 3000 US banks was analyzed by using 2 traditional statistical methods i.e. discriminant analysis and logistics regression. Then they compare these methods with the machine learning methods i.e. SVM, ANN and k-nearest neighbors. The results of this study illustrate that ANN and k-neighbors method gives the accurate predictions as compared to the

traditional methods.

3 Method

This section describes the method and statistical techniques which have been utilized in this study work.

In the first step, feature selection approach has been used to extract the relevant variables from the data. For the purpose of best feature subset the LASSO, best-subset logistic regression and random forest have been utilized. After this procedure of feature selection, the process of classification has been executed by using the machine learning techniques Support vector machines, Decision Tress, Random forest and Artificial neural network on the full model as well as on the feature subsets obtained from the first step. The statistical software which is used for the analysis purpose in this study is R-package.

4 Data

This section illustrates the data and information of all the variables.

Data set which is utilized for this research has been taken from University of California, Irvine (UCI) machine learning repository website(http://archive.ics.uci.edu/ml) which is openly available for the public for research purpose. This data set reflects the real information which is related with direct marketing campaigns of a Portuguese retail bank, from period of May 2008 to November 2010, in a total of 45,211 phone contacts (observation) and 17 attributes including response variable. The marketing campaigns were based on phone calls. In many of the cases, more than one contact to the same customer was essential in order to know that if the product (bank term deposit) will be subscribed ('yes') or not subscribed ('no'). The details of 17 attributes are following.

Variable	Description
age	numeric, age of client
job	categorical, type of job (admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services)
marital	categorical, marital status (married, divorced, single. Here "divorced" states the both divorced or widowed)
education	categorical (unknown, secondary, primary and tertiary)
default	binary, customer credit is in default (yes,no)
balance	numeric, average yearly balance (in euros)
housing	binary, status of housing loan (yes,no)
loan	binary, clients personal loan (yes,no)
contact	categorical, contact communication type (unknown, telephone, cellular)
day	numeric, the last contact day of the month range (1-31)
month	categorical, last contact month of the year
duration	numeric, last contact duration (in seconds)
campaign	numeric, number of contacts performed during this campaign
pdays	numeric, number of days that passed by after the client was last contacted from a previous campaign
previous	numeric, number of contacts which are made before this campaign
poutcome	categorical, result or outcome of the previous marketing campaign (unknown, other, failure, success)
y	binary, (desired target) output variable whether client subscribed a term deposit or not

5 Theory and Models

This section explains the theory and concepts of statistical approaches which have been used in this study. The algorithm of these statistical techniques are also discussed in this section.

5.1 Feature selection and Approaches

Feature selection is the procedure of narrowing down a subset of variables, attributes or features which can be used within the predictive modeling process. It plays vital role in the machine learning, specifically if the problem is higher dimensional. As high dimension data may consist of irrelevent variables therefore it is important to do variable selection, as pointed out in Friedman et al. (2001). Its objective is to find the small subset instead. Izetta et al. (2017) mentioned that feature selection method is better than the method typical weights averaging which are mostly used in the implemented of all the recent Machine Learning libraries. According to Trambaiolli et al. (2017) the precision of accuracy in the data can be attained only by using the appropriate features.

In the problem of feature subset selection, a learning algorithm is tackled with the issue of choosing some subset of features where upon to emphasis on its concern, whereas remaining can be skipped. Feature selection is a crucial task in machine learning. The process of selecting subsets of appropriate variables or features from the set of data can significantly decrease the chances of data overfitting as well as cost of the computational classification algorithms. The algorithm of feature subset selection examine for the subset which is the best one, utilizing the induction algorithm itself that is also a subset of that function by assessing the feature subsets.

The feature subset with the most important assessment is picked as the concluding set on which the induction algorithm is performed. The resulting classifier is then assessed on the test set which is independent that was not utilized throughout the search before (Kohavi & John (1997)).

To recognize the best feature subset, the algorithm should requires to inspect all of the possible feature subsets which can be in feature space. The technique which is most well-known for feature selection is to choose the features with the most consequence and importance to the target variable. Feature selection gives the results more efficient as compared to include all the features in the model.

Methods of feature selections reduces the complexity of the models (Izetta et al. (2017)). According to Kohavi & John (1997) the same level accuracy in feature subset can be achieved by using the different subsets therefore it is not compulsory that the optimal feature subset will be the unique one. For example, if two variables are fully correlated with one another, in this case one attribute can be exchanged for the other one. For the purpose of obtaining the best feature subset which reflects the highest possible accuracy, according to definition is that which algorithm chooses as an optimal attribute subset. Approaches for feature selection which have been used in this study are following.

5.1.1 Best-Subset Logistic Regression

Best-subset logistic regression algorithm works like backward stepwise regression. Stepwise regression is the method of fitting the regression in which at each step the subtraction or addition of variables take place. There are types of stepwise regression i.e. forward stepwise, backward stepwise and hybrid stepwise. In forward stepwise regression, it starts from null model. Null model is the model which consists of only the intercept but no explanatory variables. Then it fits the linear regression and add the variable which causes to lower the residuals sum of squares. After that, it attempts with two variables. This process continues until all the variable combinations satisfied the condition. At the end it come-up with only those variables which are significant. Backward stepwise regression is the inverse of forward selection. It starts from the full model and remove the variable with largest p-value. The procedure continues until only those variables which are significant remain in the model. Mixed selection is the mixture of both methods. At each step, it checks the p-value of the variable and decides, whether variable should be include or exclude in the model.

Best subset selection is an important phase of regression modeling. A regression in which response variable is categorical for example (success, failure) and explanatory variables can be a mixture of both continuous as well as categorical variables. There can be many scenarios in which dependent variable has binary outputs. In this case the logistic regression is used. As dependent variable is categorical so we use the best subset approach using logistic regression to find the best subset.

Lawless & Singhal (1978) suggested a strategy for effectively screening the non-normal regression models, in this manner giving the principle of best subsets non-linear regression. Now a days these methods have already built-in in latest software or can be utilize as a software packages. For example, if we have a pair (Y,x) where Y is response variable with binary output coded as 0 or 1 and $x' = (x_0, x_1, x_2, ..., x_p)$ is the vector of p+1 covariates which are supposed to be measured without any error. The logistic regression model illustrates that

$$Pr(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$$

where

$$\pi(\mathbf{x}) = e^{g(x)}/(1 + e^{g(x)})$$

and

$$g((\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Then the function of likelihood data will be

$$L(\beta) = \left[\sum_{i=1}^{n} y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)\right]$$

where $\pi_i = pi(x_i)$

Most of the statistical packages use the reweighted least square method to get the maximum likelihood estimator of β. it can be expressed as

$$\widehat{\beta} = (\mathbf{X'VX})^{-1}\mathbf{X'Vz}$$

5.1.2 LASSO (Least Absolute Shrinkage and Selection Operator)

The main objective of lasso is to find and identify the feature subset whose coefficients are non-zero Zhang et al. (2017). Lasso is very powerful technique in regression and prevent the model for overfitting by simplify the function. One of the main benefit of using this technique is that it can be used for the purpose of best subset selection. In this way, it skipped the extraneous variables from the model and comes up with only that variables which are model relevant. It is simple technique which is based on regression and works with significance of p-value for the selection of best subset Bardsley et al. (2015). It becomes popular for the regression and classification problems with various independent variables.

According to Friedman et al. (2001) the estimates of lasso are

$$\widehat{\beta}_{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} (yi - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

Lasso can also be written in the equivalent Lagrangian form which is

$$\widehat{\beta}_{lasso} = \underset{\beta}{argmin}\{\frac{1}{2} \sum_{i=1}^{N} (yi - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

Lasso does the sort of somehow continuous subset selection due to the constraints nature which assembles t sufficiently small that pushes parameters associated with unnecessary variables to zero.

The variables which cause for the overfitting of model can be removed using lasso even though accuracy of the model remains the same. Lasso stabilizes its work just by addition if penalty term into the function of likelihood. Lasso can be performed in R by using the glmnet package.

5.1.3 Random Forest for variable selection

Random Forest is non-parametric statistical technique and depends on decision trees. It utilizes the idea of aggregation since they need to require few conditions on the model that is produced using the observed data (Genuer et al. (2015)). Random forest algorithm is efficient and work for both regression as well as classification problems. RF is set of the classification trees which is also called decision tree. It comprises with leaf nodes and other explanatory variables which are exist on the other intermediate node. By gathering the leaf nodes and explanatory variables that refers to be class variables which is also called the decision variables or predictor variables (Jaiswal & Samikannu (2017)).

The principle on which RF works is to join numerous binary decision trees on the base of various samples generated by bootstrap which can be gathered from the learning sample or by selecting the each node randomly as a subset of the independent variables Genuer et al. (2010). These are some properties of random forest according to Jaiswal & Samikannu (2017).

5.2 Support Vector Machine (SVM)

Support vector machines has been effectively utilized in numerous applications over the recent years. The great speculation capacity of SVM is that, it separate the two classes for which these are generally suitable. Among all the strategies of classification, the SVM has been broadly known.

Usually, the performance time of a SVM model for classification takes longer and relates directly to the quantity of support vectors which can be sometimes challenging for some ongoing applications. For the purpose of classifying the data point, a SVM calculates the dot product for the given test point with each of support vector either in the feature space or in the space of input after the conversion through a function i.e. kernel. In this manner, the time for the execution rises with the rising of number of support vectors (Panja & Pal (2017)).

Algorithm of SVM

According to Friedman et al. (2001) the classifier of SVM is following

$$\underset{\beta,\beta_0}{argmin} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \forall_i,$$

where C is the cost parameter, The lagrange (primal) function is

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i,$$

we will minimize w.r.t β, β_0 and ξ_i and set these derivatives to zero which gives

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i,$$

$$0 = \sum_{i=1}^{N}\alpha_i y_i$$

$$\alpha_i = C - \mu_i, \forall_i$$

The positivity constraints $\alpha_i,\ y_i,\ \xi_i \geq 0\ \forall_i$

5.3 Decision Trees (DTs)

Decision Trees play s significant role in the field of data mining as they are really fast to construct as compared to the other data mining methods. They can easily handle the data even if it comprises of mixture of numeric and categorical predictor variables (Friedman et al. (2001)). The algorithm of DTs is to split the data-set to accomplish a homogenous classification for the dependent variable. At every part, objective of algorithm goes for diminishing the entropy of the dependent variable in the subsequent datasets by selecting the ideal part from various explanatory variables.

The key benefit of this technique is that it is low-cost in terms of computation as well as no assumption is required for the distribution of the variable. Beucher et al. (2017) proves DT as robust for the redundant variables and for the missing data as well. In the processing of decision trees, various DTs are usually joint to get a best predictive output by collaborative the approach called bagging, which produces the various bootstrap samples from the data and in this way generates a classifier from every sample of the bootstrap. After that, the predictions for the classifiers are then consolidated by the voting.

According to Beucher et al. (2017) a DT is a stream chart like a progressive tree pattern which is made out of three essential components. First is decision nodes which are relates to the variables. Second is branches which relates to distinctive possible value of the variable. The third module is called leaves which consists of objects that commonly have a place with a similar class or that are fundamentally the same. Such illustration of pattern enables us to encourage decision rules which can be utilized to classify the new examples.

Algorithm of DTs

According to Friedman et al. (2001) the algorithm of single decision tree is

$$\mathcal{I}_l^2(T) = \sum_{t=1}^{J-1} \hat{\iota}_t^2 I(v(t) = l)$$

Is the measure of relevance for each predictor variable X_l whereas, the sum of $J-1$ is the inner node of the tree. On the every single node t, each of the input variables $X_{v(t)}$ is used to split the portion that is linked with that specific node into two sub portions, and then response values to be fitted in every single constant.

This simple approach is then generalized to the additive tree extensions it is basically averaged on the trees

$$\mathcal{I}_l^2 = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_l^2(T_m)$$

Because of stabilizing effect of average, the above equation goes to be more consistent as compared to the precious one. Both equations refer to the squared relevance so for the calculation of actual relevance, we thus require their corresponding square roots.

5.4 Random Forest for Classification

Random forest can also be used for the purpose of classification. It is one of the most broadly used machine learning algorithm for classification. Either the response variable is continuous or categorical, it works in both cases. According to Friedman et al. (2001) random forests starts to become stable at around 200 trees, whereas at 1000 trees the boosting of this still keeps on improving. If trees are much smaller or there is a presence of shrinkage then process of boosting starts to reduce. The important function of the RF is the utilization of out of bag (OOB) samples. For each value $z_i = (x_i, y_i)$, in which term zi not appears that makes the RF predictor by averaging only those trees which are consistent to the bootstrap samples. The OOB error estimate is then nearly identical of that which is getting by N fold cross validation. In contrast to the other non linear estimators, it is possible to fit the RF in one sequence with the cross validation being completed. The training can be finished if the OOB error stabilizes itself.

Algorithm of RF

The algorithm of RF is considered as best in term of accuracy. Even the data is too large or includes thousands of input variables, though the efficiency does not decrease and at the same time prevent to be overfitting as well and there is no need of data pruning in it. It can be used for methods i.e. selection of best subset as well as imputation of the missing values and in both cases it performs very fine and efficient. The forest which is produced as an output is also proficient for adding the data for future.

In RF we have a learning set which is $L = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ which should contain the observations of independent random vector (X,Y), where X is a vector of explanatory variables i.e. $X = (X^1, ..., X^P)$ and $X \epsilon \Re^p$ and Y is the class label if the in the case of classification.

5.5 Artificial Neural Network (ANN)

The primary preferences of ANNs are their capacity to deal with datasets containing large number of observations. It can also manage to estimate the nonlinear relationship. According to Beucher et al. (2017) artificial neural networks are directed machine learning approaches regulate the relationship between the set which is known for the training points and distinctive natural attributes with intention of classification for the new one. It prevent the model from overfitting and also is not affect by the outliers.

Artificial Neural Network resembles the human brain in learning over the data storage and training. It is made and prepared over a particular input information training pattern. Throughout the procedure of learning, the results of NN is then matched to the target value and in this way, algorithm has been accomplished to reduce the error as minimum between the two values. This process of reducing error is utilized through MSE (Karouni et al. (2011)).

Algorithm of ANN

According to Friedman et al. (2001) the parameters of neural networks which are not known are called weights. In this way, we try to find the values for that unknown parameters that makes the model fit well for the training data. we can denote the complete set of weights (unknown parameters) by θ For regression purpose, sum of squared error has been used for the measure of fit

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2$$

For the purpose of classification, cross-entropy (deviance) or squared error can be utilized

$$R(\theta) = -\sum_{k=1}^{K} \sum_{i=1}^{N} y_{ik} \log f_k(x_i)$$

The corresponding classifier is

$$G(x) = argmax_k f_k(x)$$

In the hidden units, the neural network is accurately the linear logistic regression model with cross entropy error function and softmax activation function. The method which is used for the estimation of parameters in neural network is maximum likelihood.

5.6 Confusion Matrix

Confusion matrix is the table which expresses the actual and predicted number of classification. This table is used to define the performance of the classification. There are four terms which are used in this table i.e. True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). TP are cases which are actually 'true' and the test has also predicted it as the 'true' value, whereas TN are those numbers which are actually 'false' and the test has also predicted them as 'false'. On the other hand, FP are those cases which are predicted by the test as 'true' but in actual they are 'false'. FP is also known as the type-I error. While FN are those numbers which are predicted by the test as 'false' however in actual they are 'true'. FN is also known as the type-II error. Following table provides the layout of the confusion matrix.

		Predicted
	TN	FN
Actual	FP	TP

Table 1: Confusion Matrix

Accuracy tells that how accurate the specific test performs the classification. Accuracy of a test can be find by following formula.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

5.7 Receiver Operating Characteristic (ROC)

The Receiver operating characteristics (ROC) curve is the plot of true positive rate (TPR) against the false positive rate (FPR). TPR is also called the sensitivity of the test which is the probability of positives that are correctly identified whereas the FPR is the probability of false rejecting. Recently, these have been broadly utilized machine learning to gauge the classifier's performance. According to Zhang et al. (2015) ROC curve is a valuable tool for assessing paired classifiers which are based on their performance. One of known characteristics of ROC curve is their ability of discriminating the true state of subjects. The Area under curve of ROC is referred as AUC which is measure of classifier performance. Analysis of ROC is the best tool to choose the optimal model. For the ROC curve, the TPR is plotted at the x-axis while the FPR is plotted at the y-axis.

5.8 Cross Validation

Cross validation is a way to assess the performance of a prognostic model. In general the k-fold cross-validation is used, which is a method of partitioning the original data into k subsamples of equal size. At that point, single sub sample is examined as the validation of data for the purpose of testing the model while the rest of the $k-1$ sub samples are then utilized as training data. This procedure is then repetitive k times and every k sub sample is then utilized precisely one as per validation data.

There are various sorts of cross validation, for example, k-fold cross validation, leave one out cross validation, $k \times 2$ cross validation and so on. The type of cross validation which is used in this study is the k-fold cross validation where k is set to be 10. The purpose of cross validation is to check the performance of model that how accurate the predictive model is. The procedure

of k-fold cross validation can be understood more clearly with the following figure.

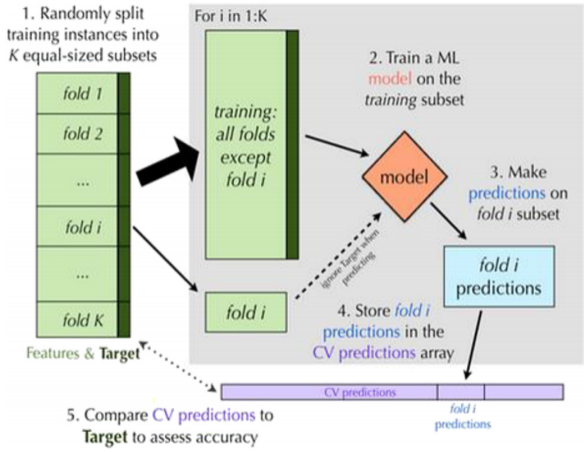

Figure 1: Flow chart diagram of k-fold cross validation (Rohani et al. (2018))

6 Results and Analysis

This section consists of outputs and tables of features selection and different classification approaches which have been discussed in the section 5.

6.1 Feature Selection

As discussed in the section 5, there are 3 techniques which have been used for the purpose of finding the best subset. In this regard, this section shows the outputs of best-subset logistic regression, LASSO and random forest. Below, we provide a brief detail about the chosen of feature selection methods.

6.1.1 Best-Subset Logistic Regression

Since the dependent variable is categorical in the data so logistic regression is used for the purpose of selecting the subset of variables. This process of selecting the best subset resembles stepwise logistic regression in which it first consider the full model, then at each step it exclude the variable which is insignificant. At the end it comes up with the subset of those variables which are significant. The variables which have been selected as important by this statistical technique are 10 out of 16 and marked as cross as shown in Table 2.

6.1.2 Random forest

Random forest is also used in this study for the aim of selecting best subset. Figure 2 presents the variable importance chart which illustrates the importance of each variable in the data according to this statistical technique. In this chart, threshold value 0.03 has been chosen for

selecting the important variables. There are 10 variables which reach this threshold value. According to this graph, *duration* is the most important variable as compare to the other variables. Then *age*, *balance*, *day*, *month*, *job*, *pdays*, *poutcome* are also chosen as significant variables. In contrast, *marital*, *default*, *loan*, *housing*, *contact*, *previous* are the least important and not selected in the model. Following is the variable importance plot which represents the importance of each attribute in the data and this is obtained after applying the random forest technique.

In Figure 3, the ROC curve has been plotted which compares the full model with the subset of variables from random forest. There are two lines in this plot. Red line represents the ROC curve for full model whereas the blue line represents the feature subset selected by random forest. By comparing these two lines we can see that full model and the reduced model gives almost the same accuracy.

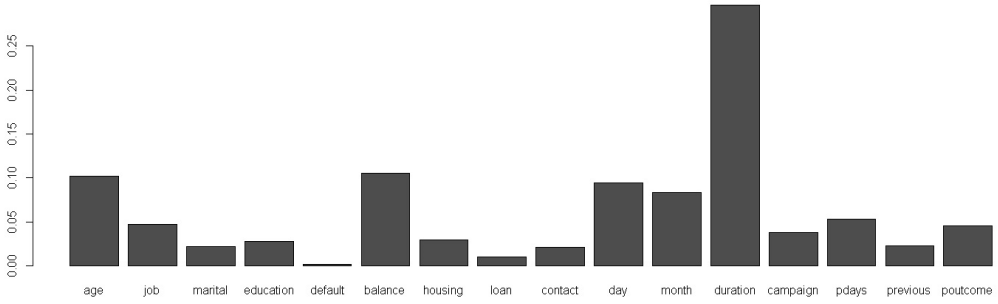

Figure 2: Variable Importance Plot

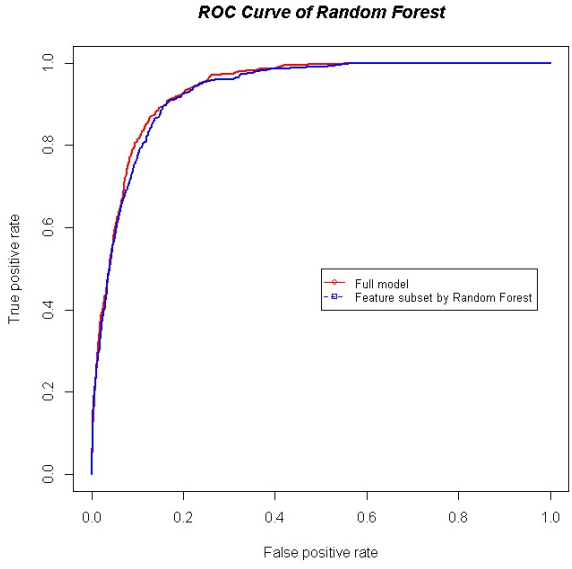

Figure 3: ROC curve for Random Forest

6.1.3 LASSO

LASSO penalizes the zero value to the coefficients of that variable which are irrelevant. At the end, it comes up only with the subset of variables which are relevant. After performing the LASSO, it selects 10 variables out of 16 as best variables such as *marital*, *education*, *balance*, *housing*, *loan*, *contact*, *duration*, *campaign*, *pdays*, *poutcome*. The reduced model selected by LASSO and logistic regression are almost the same instead of one variable. Whereas, *education*, *balance*, *duration*, *campaign* are variables which are selected as significant by all three feature selection approaches. Table 2 shows the variables which have been selected after executing these approaches.

Table 2: Variables Selected by Feature Selection

Variables	**Best-Subset LR**	**LASSO**	**Random Forest**
age			×
job			×
marital	×	×	
education	×	×	×
default			
balance	×	×	×
housing	×	×	
loan	×	×	
contact	×	×	
day			×
month			×
duration	×	×	×
campaign	×	×	×
pdays	×		×
previous		×	
poutcome	×	×	×

6.2 Classification

In this section, now we proceed to presenting the results for the classification algorithms used in this study. The evaluation of these models are based on over all accuracy and area under curve (AUC). The processing time of these approaches is also mentioned in each subsection.

6.2.1 Support Vector Machines

We start with the SVM based classification. Table 3 demonstrates the SVM for classification, in which results of three best subsets which have chosen by using the feature selection techniques are given in terms of their accuracy and AUC. SVM is analyzed on the data set using 10-fold cross validation. The time for the processing of SVM, as compare to all other classification methods which have been used in this thesis, is greater. It took approximately 1 hour to execute one subset. As there are 4 models so it took almost 4 hours to finish the task.

Table 3: SVM for Classification output

Feature Subset method	**Accuracy** (in percentage)	**Area under curve (AUC)**
Full Model	89.63	0.7726
Best-subset LR	89.39	0.7520
LASSO	89.31	0.7535
Ranfom Forest	89.71	0.7720

Below we present confusion matrix in Table 4 for classification obtained through SVM. Confusion matrix shows the performance of the classification that how correctly classified and misclassified the specific test made. The accuracy of the test can be measured by taking sum of TP and TN and divide it on the total number of observations.

			Predicted (%)	
			No	**Yes**
Actual (%)	**Logistic Regression**	**No**	86	8.40
		Yes	2.30	3.39
	LASSO	**No**	86.17	8.39
		Yes	2.30	3.14
	Random Forest	**No**	86.09	8.08
		Yes	2.21	3.62

Table 4: Confusion matrix for classification obtained through SVM

6.2.2 Decision Trees

We analyzed the decision tree approach on the data with 10-fold cross validation. Table 5 illustrates the outputs obtained after analysing the Decision tree technique, which represents the accuracy in second column and AUC in third column. Decision Tree has been analysed with less computational effort, and this can be seen as an advantage of this approach.

Table 5: DTs for Classification output

Feature Subset method	**Accuracy** (in percentage)	**Area under curve (AUC)**
Full Model	89.63	0.7854
Best-subset LR	89.45	0.7617
LASSO	89.50	0.7625
Ranfom Forest	89.87	0.7835

Table 6 is the confusion matrix for classification obtained through Decision Trees. This confusion matrix represents the actual vs predicted number of classification.

			Predicted (%)	
			No	**Yes**
Actual (%)	**Logistic Regression**	**No**	85.60	8.20
		Yes	2.35	3.85
	LASSO	**No**	85.96	8.15
		Yes	2.17	3.54
	Random Forest	**No**	85.92	7.81
		Yes	2.32	3.77

Table 6: Confusion matrix for classification obtained through Decision Tree

6.2.3 Random Forest

Table 7 shows the results generated by Random forest on three models in terms of accuracy and AUC. The processing time for random forest took almost 40 minutes as refers to single model and there are 4 models in total including full model. It's processing time depends on the number of trees which have been used while implementing this method. The number of trees used in this approach are 500.

Table 7: Random Forest for Classification output

Feature Subset method	**Accuracy** (in percentage)	**Area under curve (AUC)**
Full Model	90.63	0.9373
Best-subset LR	89.84	0.8725
LASSO	89.66	0.8682
Ranfom Forest	90.56	0.9289

Similarly, Table 8 is the confusion matrix for classification obtained through Random Forest which represents the percentage of classification on the reduced subsets.

			Predicted (%)	
			No	**Yes**
Actual (%)	**Logistic Regression**	**No**	85.90	7.40
		Yes	2.10	4.60
	LASSO	**No**	85.74	7.78
		Yes	2.38	3.92
	Random Forest	**No**	85.60	6.75
		Yes	2.69	4.96

Table 8: Confusion matrix for classification obtained through Random Forest

6.2.4 Artifical Neural Network

Finally we summarized the classification results obtained from the Neural Network approach. Table 9 presents the results in terms of accuracy and AUC. The processing time which ANN took is roughly 50 minutes and it depends on the number of hidden layers which have been used

during this analysis. In this technique, the data has been divided into two parts i.e. training set includes 80% data and the remaining 20% of data is in test set. The number of hidden layers for all the models which have been utilized in this statistical approach is 3. As we tried the number of hidden layers from 1 to 5, and the most accurate results were come up by using 3 hidden layers. As there were 4 models in total so it took around 4 hours for implementing this technique. Below are the output of artificial neural network table and plots.

Table 9: Artificial Neural Network for Classification output

Feature Subset method	**Accuracy** (in percentage)	**Area under curve (AUC)**
Full Model	89.03	0.8930
Best-subset LR	89.70	0.8869
LASSO	89.68	0.8872
Ranfom Forest	88.89	0.8536

Following is the plot of neural network for the full model. This figure consists of input layer, hidden layer and output layer. In the hidden layer number of nodes are 3. At output layer, there is dependent variable while explanatory variables are located at input layer.

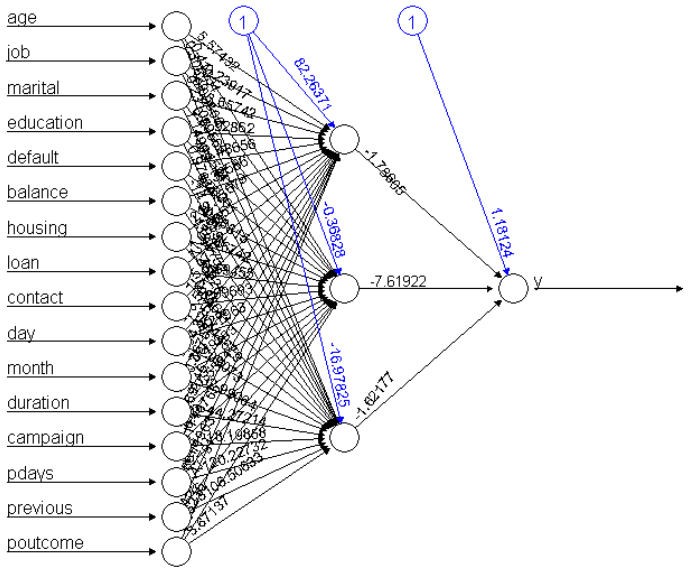

Figure 4: Artificial Neural Network plot for full model

Below is the neural network plot which is performed on the subset obtained through LR subset. In input layer, there are 10 explanatory variables as LR selects 10 variables to its feature subset. There are 3 nodes located at hidden layer while at output layer there is dependent variable.

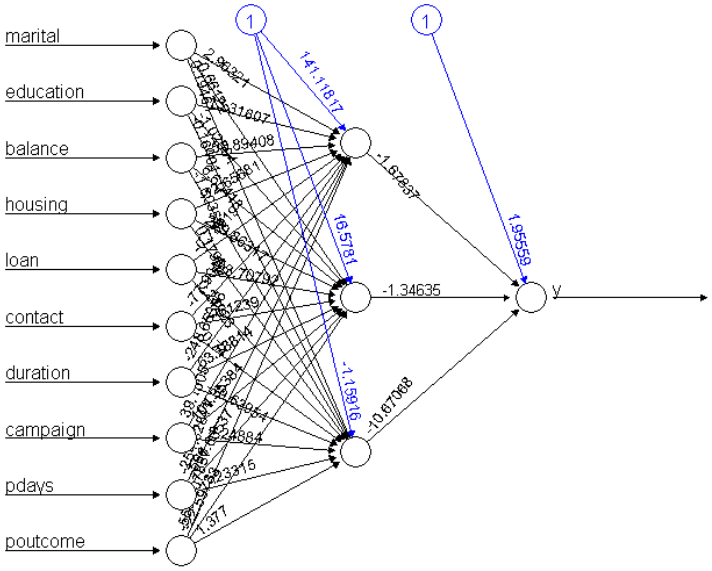

Figure 5: Artificial Neural Network plot for feature subset from LR

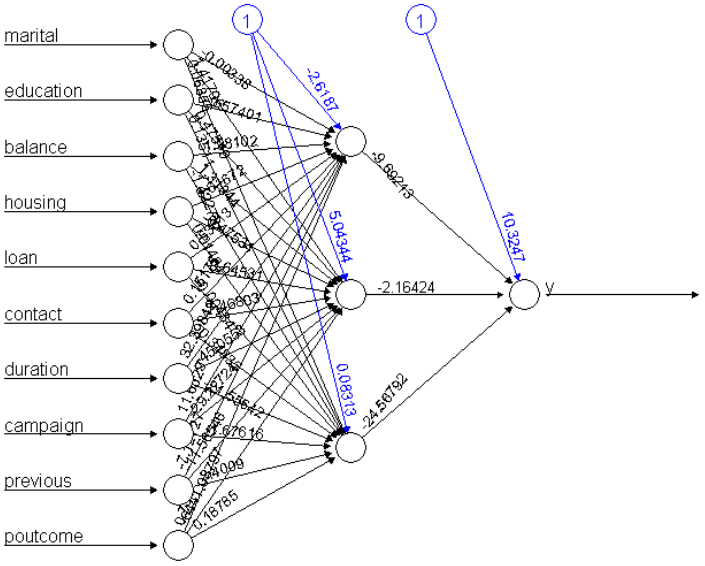

Figure 6: Artificial Neural Network plot for feature subset from LASSO

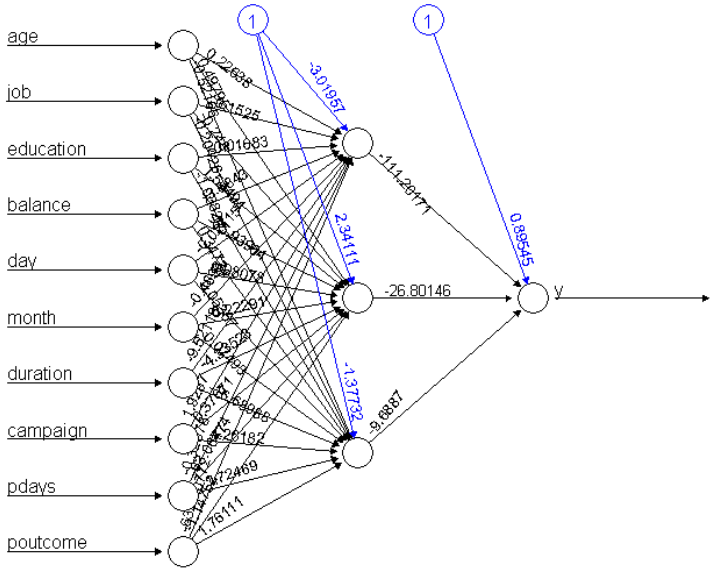

Figure 7: Artificial Neural Network plot for feature subset from Random Forest

7 Discussion and Conclusions

This section consists of discussion of the output of statistical techniques that are used for classification and conclusion.

Figure 2 illustrates variable importance chart generated by random forest for the feature subset selection. As from the figure, the most relevant variable according to the random forest is *duration* which means that longer the duration of call the more success chances are. Beside this variable the important variables are balance and age.

Table 2 is the table of comparison of best subsets generated by using the feature selection approaches. The table presents 3 subsets as output result of statistical attribute selection techniques. This table shows that subset attained from the LASSO and logistic regression have almost same variables except the variables *pdays* and *previous*. Regrading subset from the random forest, this subset is bit different from both other subsets because chosen variables are different but overall there is not so much difference between all of them.

Table 3 is the output of classification of SVM. In this table there is full model as well as best chosen models by the statistical approaches is presented in terms of accuracy and AUC. The accuracy of the model shows that how accurate prediction can the model predict and area under curve illustrates the performance of that specific model. Table 2 describes the full model in which all 16 variables are included, shows 89.63% of accuracy with 0.7726 area under the curve, for subset selected by logistic regression gives the SVM classification result with accuracy of 89.39% with 0.7520 AUC. In case of model from LASSO, 89.31% accuracy was achieved with 0.7535 area under curve of ROC. The subset obtained from random forest technique provided 89.71% accuracy with 0.7720 AUC. By comparing these results within SVM classification, after full model the best results are attained from subset generated by random forest.

Table 4 is the confusion matrix for classification obtained through SVM. It consists of 3 confusion matrices that are LR, LASSO, RF. As from the table in first confusion matrix for the LR, 86% are true negative whereas true positive are 3.39%. If we add TP and TN, we will come-up with the accuracy of the test i.e. 89.39%. Similarly, false positive and false negatives represent the misclassification from the test. In case of LR through SVM, the misclassification is 10.70%. Similarly, In the confusion matrix on the subset from LASSO, the accuracy is 89.31% and the misclassification is 10.69%. In case of random forests confusion matrix, 89.71% accuracy has been achieved which is the best accuracy as compared to the accuracy of the reduced subsets of LR and LASSO.

Table 5 is output results for classification obtained from performing the another classification approach of Decision trees, which have been performed on full model as well as the 3 best models which are selected earlier. On the full model it attains the accuracy of 89.63% with 0.7854 are under the curve. In case of feature subset obtained from logistic regression, the accuracy of 89.45% has been attained with the performance of 0.7617. The model obtained from LASSO has accuracy of 89.50% with performance of 0.7625. Regarding random forest model, the accuracy of 89.87% has been achieved with area under curve of 0.7835. By comparing the results within then classification technique of DTs, next to the full model, the best results of DTs achieved on the subset of variables are those selected by the random forest.

Table 6 illustrates the confusion matrix obtained through the classification technique of de-

cision trees. In this table the first confusion matrix refers to feature subset obtained from LR, in which 89.45% accuracy has been achieved while 10.55% is the misclassification. Afterwards, there is confusion matrix for the LASSO subset where 89.50% has been correctly classified and 10.32% are misclassified. In confusion matrix obtained from random forest subset, 89.69% are correctly classified whereas 10.13% are misclassified. The accuracy of Random Forests subset is higher as compared to the other feature subsets.

Table 7 is the table of output results generated by random forest for classification on all 4 best models including the full model with all the variables. In case of full model, the accuracy is 90.63% with 0.9373 AUC value. This is the most accurate value is achieved on the full model, if we compare the results of all the classification techniques. The subset attained from LR has accuracy of 89.84% with the 0.8725 AUC value. In case of subset chosen by LASSO, the best-subset has the 89.66% value of accuracy with 0.8682 value of AUC. Regarding the best subset selected by random forest the random forest classification achieved the accuracy of 90.56% with the value of AUC 0.9289. By comparing the results, random forest for classification performed best on all 4 models as compare to the other techniques for classification.

Table 8 demonstrates the confusion matrix obtained through the classification method of Random Forest. In this table the first confusion matrix is for the LR subset, where RF ensured 90.50% correctly classified. The percentage of correctly classification of Random Forest on subsets of LASSO and RF are 89.66% and 90.56% respectively. Similar in other confusion matrices, the accuracy of Random Forest classification is higher as compared to the other classification methods.

Table 9 demonstrates the output results for classification of artificial neural network technique, which represents the 89.03% of accuracy on full model with 0.8930 AUC value. In case of best subset selected by LR, the accuracy of 89.70% has been achieved with 0.8869 value of area under the curve. LASSO subset got the accuracy of 89.68% with the value of AUC 0.8872. The model obtained from the random forest is 88.89% accurate and has 0.8536 value of AUC. In this table 5. It can been be clearly seen that best accurate results are achieved by the subset of variable elected by random forest.

Figure 3 is the ROC curve for random forest, the full model is compared with the best subset selected by random forest. As this is the most accurate feature subset as well as classification technique, so the difference between the performances of full model and selected subset can be seen in this ROC curve which is rarely small and very close to each other. In this way, we do not need the full model and thus go for the feature subset selection instead.

Figure 4 is the plot of artificial neural network of the full model which consist of input layer, hidden layer, output layer and nodes. As there are 16 variables so corresponding to this there are 16 nodes and this layer is called the input layer. In center of the plot, there is hidden layer which consist of 3 nodes. The output variable is y that last part of figure is called the output layer and has only one node. Figure 5 is the ANN plot of the reduced model given by logistic regression. This plot has 10 nodes in input layers as there are only 10 variables in the model. The hidden layer in the center comprises of 3 nodes and output layer has only one neuron. Similarly for the Figure 6 and Figure 7 which are ANN plots of subsets of variable selected by LASSO and RF respectively. In the input layer there are 10 variables so there are 10 nodes, in hidden layer there are 3 nodes and in output layer consists of 1 node as there is one dependent variable.

Table A.1 in the Appendix section is the complete table and illustrates the overall accuracy and performance of all the models in aspect with classification. In this table it is clearly revealed that all the classification results performed the most accurate classification on the subset of variables chosen by random forest. Regarding classification, the random forest for classification also gives the best accuracy results as compared to the other classification methods.

Summary

This research demonstrate the different data mining methods which is a great tool in the decision making. For this study work, real data was taken from the data base of University of California, Irvine website which is open source. In general there are 2 steps involved in this thesis work. In first step the process of feature selection has been done by using 3 statistical approaches i.e. best-subset Logistic Regression, LASSO, Random Forest. In the second step, the best subset of variables which are selected by these methods are go through for the classification. For classification purpose there are 4 computational algorithms for classification have been used that are Support vector Machines, Decision Trees, and Random Forest for classification, Artificial neural network. The aim was to check whether these classification methods give same number of accuracy and performance by using the feature selection approach. The results indicated that we do not need to go for the full model as reduced subset of 10 variables can provide almost the same accuracy. Regarding feature subset, the best subset of variables is chosen by the random forest and regarding classification, also random forest comes up with the most accurate results with 90.56% accuracy on the subset selected by random forest. As accuracy of reduced model is almost the same so one can rely on the subset of variables selected by random forest instead of full set of variables. Area under curve of ROC shows the performance of 0.929 for this selected subset. Moreover, RF reveals that the most impacting attribute is duration, afterwards there are balance and age respectively.

References

Barboza, F., Kimura, H. & Altman, E. (2017), 'Machine learning models and bankruptcy prediction', *Expert Systems with Applications* **83**, 405–417.

Bardsley, W. E., Vetrova, V. & Liu, S. (2015), 'Toward creating simpler hydrological models: A lasso subset selection approach', *Environmental Modelling & Software* **72**, 33–43.

Beucher, A., Møller, A. & Greve, M. (2017), 'Artificial neural networks and decision tree classification for predicting soil drainage classes in denmark', *Geoderma* .

Cramer, S., Kampouridis, M., Freitas, A. A. & Alexandridis, A. K. (2017), 'An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives', *Expert Systems with Applications* **85**, 169–181.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2010), 'Variable selection using random forests', *Pattern Recognition Letters* **31**(14), 2225–2236.

Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2015), 'Vsurf: An r package for variable selection using random forests.', *R Journal* **7**(2).

Gil, D. & Johnsson, M. (2010), 'Using support vector machines in diagnoses of urological dysfunctions', *Expert Systems with Applications* **37**(6), 4713–4718.

Izetta, J., Verdes, P. F. & Granitto, P. M. (2017), 'Improved multiclass feature selection via list combination', *Expert Systems with Applications* **88**, 205–216.

Jaiswal, J. K. & Samikannu, R. (2017), Application of random forest algorithm on feature subset selection and classification and regression, *in* 'Computing and Communication Technologies (WCCCT), 2017 World Congress on', IEEE, pp. 65–68.

Karouni, A., Daya, B. & Bahlak, S. (2011), 'Offline signature recognition using neural networks approach', *Procedia Computer Science* **3**, 155–161.

Kohavi, R. & John, G. H. (1997), 'Wrappers for feature subset selection', *Artificial intelligence* **97**(1-2), 273–324.

Lawless, J. & Singhal, K. (1978), 'Efficient screening of nonnormal regression models', *Biometrics* pp. 318–327.

Le, H. H. & Viviani, J.-L. (2017), 'Predicting bank failure: An improvement by implementing machine learning approach on classical financial ratios', *Research in International Business and Finance* .

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

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

Nogami, J., Nakasuka, S. & Tanabe, T. (1996), 'Real-time decision support for air traffic management, utilizing machine learning', *Control Engineering Practice* **4**(8), 1129–1141.

Panja, R. & Pal, N. R. (2017), 'Ms-svm: Minimally spanned support vector machine', *Applied Soft Computing* .

Power, D. J. (2008), Decision support systems: a historical overview, *in* 'Handbook on Decision Support Systems 1', Springer, pp. 121–140.

Qi, G., Zhu, Z., Erqinhu, K., Chen, Y., Chai, Y. & Sun, J. (2018), 'Fault-diagnosis for reciprocating compressors using big data and machine learning', *Simulation Modelling Practice and Theory* **80**, 104–127.

Rohani, A., Taki, M. & Abdollahpour, M. (2018), 'A novel soft computing model (gaussian process regression with k-fold cross validation) for daily and monthly solar radiation forecasting (part: I)', *Renewable Energy* **115**, 411–422.

Rupnik, R. & Kukar, M. (2007), 'Decision support system to support decision processes with data mining', *Journal of information and organizational sciences* **31**(1), 217–232.

Trambaiolli, L. R., Biazoli, C. E., Balardin, J. B., Hoexter, M. Q. & Sato, J. R. (2017), 'The relevance of feature selection methods to the classification of obsessive-compulsive disorder based on volumetric measures', *Journal of affective disorders* **222**, 49–56.

Wang, S. & Summers, R. M. (2012), 'Machine learning and radiology', *Medical image analysis* **16**(5), 933–951.

Zhang, X., Li, X., Feng, Y. & Liu, Z. (2015), 'The use of roc and auc in the validation of objective image fusion evaluation metrics', *Signal processing* **115**, 38–48.

Zhang, Z., Tian, Y., Bai, L., Xiahou, J. & Hancock, E. (2017), 'High-order covariate interacted lasso for feature selection', *Pattern Recognition Letters* **87**, 139–146.

Appendix

Table A.1: Overall Accuracy and AUC of all models

	SVM		**DT**		**RF**		**ANN**	
	Acc %	**AUC**	**Acc%**	**AUC**	**Acc%**	**AUC**	**Acc%**	**AUC**
Full model	89.63	0.7726	89.63	0.7854	**90.63**	**0.9373**	89.03	0.8930
Best-subset LR	89.39	0.7520	89.45	0.7617	89.84	0.8725	89.70	0.8869
LASSO	89.31	0.7535	89.50	0.7625	89.66	0.8682	89.68	0.8872
Random Forest	89.71	0.7720	89.68	0.7835	**90.56**	**0.9289**	88.89	0.8536