Analytics in Bank Marketing

Giridhar Manoharan, Aditi Deepak Thuse

Department of Computer Science, Washington State University

December 27, 2016

Abstract

The global financial crisis in 2008 made credit on international markets more restricted for banks, turning their attention to internal clients and their deposits to gather funds. This led to a demand for knowledge about client's behavior towards bank deposits and especially their response to telemarketing campaigns.

We describe a data mining approach, to extract valuable knowledge from a Portuguese bank telemarketing campaign data. We employed several classification models like Naïve Bayes, Logistic Regression, Decision Trees, Support Vector Machines and Random forests. Of the classification models we used, Random forests showed the best results. Followed by which we conducted a data analysis technique called sensitivity analysis using the 'rminer' package in 'R', to extract useful knowledge from the data, such as the best month for contact, the influence of the last campaign result and other inferences.

1 Introduction

When business communicates directly to the customers at the time of advertising then it is called as direct marketing. Direct marketing can be done using various methods including cell phones, emails, online advertisement and so on. Well executed direct marketing campaigns will result in the increase in profit and return on investment of the business. Direct marketing is attractive to many marketers because its positive results can be measured directly.

In this paper, the focus is on targeting the customers in a direct marketing campaign of a bank. Many people receive phone calls from the bank asking whether they are interested in subscribing in a term deposit. Some people may get annoyed by these phone calls and some may be really in need of such services. It is very time-consuming and costly for making calls to every customer. Hence by using machine learning and data mining techniques we can ease out the process of direct marketing.

This study considers real time data of a retail bank from Portugal which performed direct marketing campaigns using phone calls (telemarketing). Results of each campaign are gathered together. We are using five machine learning classification models like Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree, and Random forest. Among these, the best model is chosen for prediction. We have also used the data mining techniques for inferring from the data.

The remainder of the paper is organized as follows. In the next section, we define the problem statement. Then, Section III describes models and measures. In section IV, we present our implementation analysis. Section V reports the results. In section VI we discuss the related work and finally, we conclude this paper in section VII.

2 Problem Definition

Our problem statement is to target the customers who are most likely to subscribe in term deposit by identifying the important characteristics of the customer to increase the bank's profit. By doing this we are filtering the customers so that number of making calls will get reduced. This also leads to a reduction in cost and increase in return on investment of the bank.

Following are the contributions of the paper: (i) Preprocessing of Dataset. (ii) Identifying the important characteristics which are affecting customer's enrollment. (iii) Selecting the best model for the dataset from machine learning. (iv) Sensitivity analysis by data mining techniques.

3 Models and Measures

3.1 Support Vector Machine

Support vector machine (SVM) is a discriminative classifier which uses supervised learning (labeled training data). This algorithm results in the optimal hyperplane and separate cases into different class labels. Support vector machine is used for classification, regression, and outlier detection analysis.

$$\frac{1}{2}w^2 + C\sum_{i=1}^{N}\zeta_i \ where \ y_i(w\phi(x_i)+b) \geq 1-\zeta_i \text{ and } \zeta_i \geq 0 \tag{1}$$

In above equation, where C is the cost, w is the weight vectors, b is a bias which represents parameters for handling non-separable data (inputs).

Kernel methods are algorithms for pattern analysis, which are used in support vector machine (SVM) in machine learning. In our study, we are using a linear kernel with the cost value 0.1. Following are the different types of kernels:

- Linear kernel
- Polynomial kernel
- RBF Kernel
- String kernels

3.2 Naïve Bayes

Naïve Bayes is a generative model of machine learning which is used for classification which uses the Bayes' Theorem with the assumption of independence among predictors. That is the presence of one feature does not depend on another feature given the class label. Naïve Bayes calculates the posterior probability. The equation for the naïve Bayes is as follows:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{2}$$

where learning p(x|y) is a density estimation problem, p(y) is class prior probability and p(x) is predictor prior probability.

3.3 Logistic Regression

Logistic regression is the descriminative model which estimates the probability using logit function. according to this probability, the prediction is made.

$$p(y|w) = \frac{1}{1+e^{-wx}} \tag{3}$$

In above formulation, we assume that x and y are probabilistically related (parameterized by w) and Goal is to learn from the training data using Maximum Likelihood Estimation (MLE)

3.4 Decision Tree

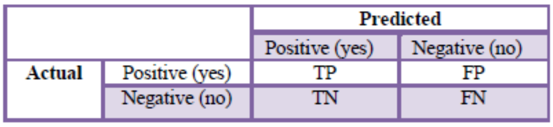
Decision tree algorithm partitions the data samples into two or more subsets so that the samples within each subset are more similar than in the previous subset. This is a recursive process; the resulting subsets are then split again, and the process repeats until the homogeneity criterion is reached or until some other stopping criterion is satisfied. As the name implies, this model recursively splits data samples into branches and construct a tree structure for improving the prediction accuracy. Each tree node is either a leaf or a decision node. All decision nodes have to split, testing the values of some functions of data attributes. Each branch of the decision node corresponds to a different outcome of the test. It is a more popular technique because the procedures are relatively straightforward to understand and explain, and the procedures address a number of data complexities, such as non-linearly and interactions, that commonly occur in real data.

3.5 Random Forests

Random forests are an ensemble learning method for classification. It operates by constructing a multitude of decision trees at training time. To classify a new object, each tree gives a classification and the classification having the most votes is chosen to be the predicted class of the new object. It uses a technique called bootstrap aggregation to overcome the problem of over-fitting to the training set that happens with decision trees. Bootstrap aggregation, also known as bagging which helps to reduce the variance of a model on unseen data.

3.6 Classification Accuracy

The performance of each classification model is evaluated using classification accuracy. It is calculated from confusion matrix which contains information about actual and predicted classifications done by a classification system. Figure 1. shows the confusion matrix for a two-class classifier.



		Predicted	
		Positive (yes)	Negative (no)
Actual	Positive (yes)	TP	FP
	Negative (no)	TN	FN

Figure 1: Confusion Matrix

True Positive (TP) is the number of correct predictions that an instance is true. True Negative (TN) is the number of correct predictions that an instance is false. False Positive (FP) is the number of incorrect predictions that an instance is true. Finally, False Negative (FN) is the number of incorrect predictions that an instance is false.

Classification accuracy is equal to the sum of TP and TN divided by the total number of cases N.

$$Accuracy = \frac{TP + TN}{N} \tag{4}$$

3.7 Receiver Operating Characteristic

The receiver operating characteristic or ROC curve is a plot of the true positive rate against the false positive rate for the different possible cut points of a diagnostic test. It illustrates the performance of a binary classifier. The true positive rate(TPR) also known as sensitivity, recall, or probability of detection, is the proportion of positives that are correctly identified.

$$TruePositiveRate, TPR = \frac{TP}{TP + FN} \tag{5}$$

The false positive rate (FPR), also known as false alarm ratio, is the ratio of the number of negative events wrongly categorized as positive and the total number of actual negative events. It is equal to $1 - Specificity$.

$$FalsePositiveRate, FPR = \frac{FP}{FP + TN} \tag{6}$$

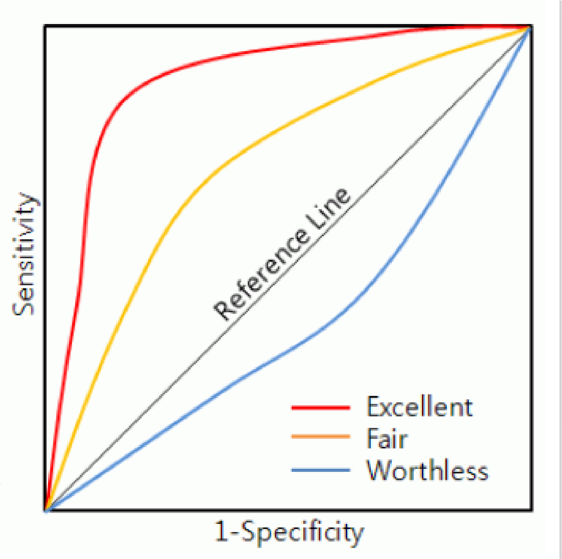


Figure 2: Receiver Operating Characteristic

Above shows the trade off between TPR (sensitivity) and FPR (1-specificity). The closer the curve is to the left side border and the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

3.8 Lift Curve

One of the popular technique in direct marketing is the lift curve [1] which is used in data mining model. It is a measure of a performance of a targeting model while predicting or classifying the examples. It measures the response with respect to the population as whole measured against random choice targeting model. A classifier which predicts the potential customer who is likely to subscribe in the term deposit. Lift chart illustrates how the model is good compared to random guessing when there are n number of samples.

$$x = Yrate(t) = \frac{TP(l) + FP(l)}{P + N}, \; y = TP(t) \tag{7}$$

4 Implementation Analysis

4.1 Bank Direct Marketing Data

4.1.1 Source of the Data

This research is focused on the targeting the customers through telemarketing phone calls who has the highest possibility to subscribe into the services like term deposit.The dataset which is used in this paper is from direct marketing campaigns of Portuguese retail bank.

The Data is available in public domain and can be downloaded from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. This dataset is analyzed by S. Moro, P. Cortez and P. Rita[2].

4.1.2 Understanding of Dataset and Attribute Information

The dataset used is related to 17 campaigns that occurred between May 2008 and November 2010. Total number of records in this dataset are 45211 with 16 features and a binary output. This dataset is unbalanced, as only 5289 (11.69%) out of 45211 records are related to the success. Following are the attributes and description of dataset.

Attribute Name	Description	Type	Missing Values
Personal Information			
Age	Age of the client	Numeric	0
job	Client's occupation	Categorical	288
marital	Marital status	Categorical	0
education	Client's education level	Categorical	1857
Bank Client Information			
default	whether the client has credit	Categorical	0
balance	Bank balance in account	Numeric	0
housing	whether client has housing loan	Categorical	0
loan	whether client has personal loan	Categorical	0
Information of the last call			
contact	Type of contact communication	Categorical	13020
day	Day that last contact was made	Categorical	0
month	Month that last contact was made	Categorical	0
duration	Duration of last contact(seconds)	Numeric	0
campaign	Number of contacts performed in this campaign for this client	Numeric	0
pdays	Number of days since client was contacted in last campaign	Numeric	0
previous	Number of contacts performed before this campaign for this client	Numeric	0
poutcome	Outcome of the previous marketing campaign	Categorical	36959

Table 1: Attributes

Output Variable: there is one column in table that corresponds to target variable or the class label.

Attribute Name	Description	Type
y	whether the client has subscribed for a term deposit	Binary (yes or no)

4.2 Data Preprocessing

At the initial stage, we have collected the dataset. In table 1, we can see that there are missing values of some of the attributes. Processing the missing values is the first obstacle in modeling. These missing values must be processed before applying different models of machine learning. There are different methods for data imputation – (i)Imputation by discarding the data, (ii)random imputation of single variable, (iii)imputation of several missing variables, (iv)model-based imputation, (v)multivariate imputation.

In our study, we have used the multivariate imputation using MICE which is R package [3]. It is one of the useful packages which uses multivariate imputation via chained equations. Instead of creating single imputation that is calculating mean, it calculates the multiple imputations that take care of uncertainty in missing values.

The assumption behind this concept is that the missing data is Missing at Random(MAR) that is a probability of missing value depends on the observed value and can be predicted. When there are missing values in a continuous feature then linear regression is used and when there are missing values in a categorical feature then logistic regression is used. Since in our dataset contains missing values only in categorical features, we have used logistic regression for imputation. By doing this, an imputed dataset is obtained.

4.3 Feature Selection

The variable importance plot is derived from randomForest package in R. It ranks features based on their contribution to the prediction of the output variable. It has two measures namely Mean Decrease in Accuracy (MDA) and Mean Decrease in Gini (MDG). Mean decrease in accuracy measures the decrease in accuracy of the output when that variable is not considered for prediction. Mean Decrease in Accuracy (MDA) or Permutation Importance is assessed for each feature by removing the association between that feature and the target. By random permutation of the values of the feature and then measuring resulting increase in error, mean decrease in accuracy can be calculated. The influence of the correlated features is also removed. Mean decrease in gini measures decrease in node impurities when spitting on a feature. Gini Importance or Mean Decrease in Impurity (MDI) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. We made use of this plot to select the top 10 features - "duration, poutcome, month, day, pdays, age, housing, job, previous, and campaign."

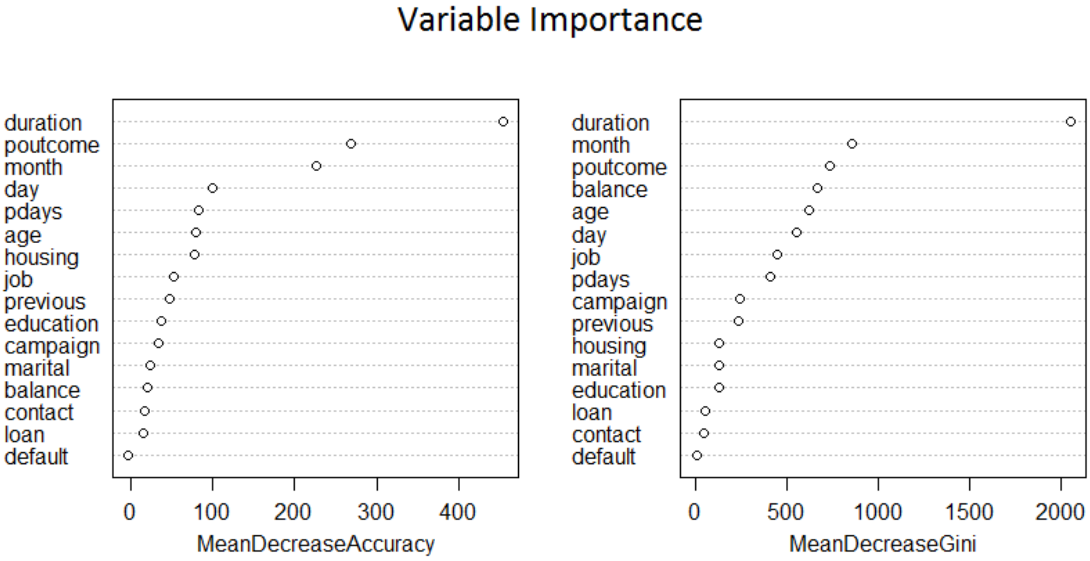


Figure 3: Variable Importance plot from random forests package in R

4.4 Machine learning Models

In this work, we are testing five binary classification machine learning models which are implemented in R tool: Support vector machine(SVM), Logistic regression(LR), naïve Bayes(NB), decision tree(DT) and Random forest. Imputed data is fed to these machine learning models for the classification task. For this, we divided this data into 80% training and 20% validation data. We made use of R function called createDataPartition under caret library for this purpose.

SVM: In support vector machine, we have used the linear kernel and tuned the hyper-parameter that is the cost (C= 0.01,0.1,1,10). This parameter tuning is done by grid search oversupplied parameter ranges. Best hyper-parameter is chosen among these and with this best hyper-parameter, training data is trained with SVM and predicted using the validation data. We got 0.1 as the best hyper-parameter after tuning. Accuracy for SVM is 91.57 %

Logistic Regression: After the prediction of validation data, it returns the probabilities in the form of p(y=1|x). Our decision boundary is 0.5. If the probability is greater than 0.5 then y=1 otherwise y=0. Accuracy for logistic regression is 90.86.

Naïve Bayes: We computed the conditional posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule. Accuracy for Naïve Bayes is 89.47. We got the lowest accuracy in this classifier.

Decision Tree: The training data was fed into the rpart function which is a decision tree library in R. Rpart function uses the ID3 decision tree greedy algorithm for the classification task.

Data is partitioned at each node of the tree based on the feature having the maximum information gain value at that node. Validation accuracy of this classification model is 90.58

Random Forest: The training set was supplied to the randomForest function in R. This function has two hyper-parameters - the number of attributes randomly sampled as candidates at each split of the tree node given by the 'mtry' function parameter and the number of trees to grow given by 'ntree' function parameter. We made use of the default value for 'mtry' which is square root of the number of attributes in the input. For 'ntree,' we tried with three different values and the validation accuracies obtained are show in Table 2. From the table, we can say that as we increase the number of trees in random forests, the accuracy on the validation set is increased by a little amount. We selected 'ntree' value to be 1500 for comparing this model with other models which is discussed in the next section.

Number of trees	Validaton accuracy (in %)
50	91.71
500	91.85
1500	91.89

Table 2: Random forest accuracy variation by varying number of trees grown

Models	Accuracy
Random Forest	91.89
SVM	91.57
Logistic Regression	90.86
Decision Tree	90.58
Naive Bayes	89.47

Table 3: Accuracy Table

5 Results and Discussion

5.1 Model Comparison

We used Receiver Operating Characteristic (ROC) curves to provide a good comparison between the various machine learning models that we have used. The 'rminer' package provides a function called 'mgraph,' which was used to plot ROC curves of different models in the Figure 4. From the figure, it is evident that the performance of random forests is better than the other models followed by SVM. This is because random forests is an ensemble learning model which uses multiple classifiers for prediction whereas other models have just one single classifier. The bagging technique used in random forests helps reduce the variance on unseen data thus leading to higher prediction accuracy.

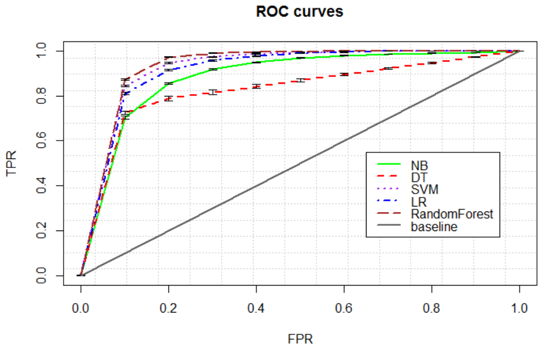


Figure 4: Receiver Operating Characteristic - Model Comparison

From figure 5, it is evident that lift for random forest is greater than all the model following by SVM and logistic regression.Hence, we can say that random forest presents the best predictive result among SVM, Logistic Regression, Naïve Bayes and decision tree.

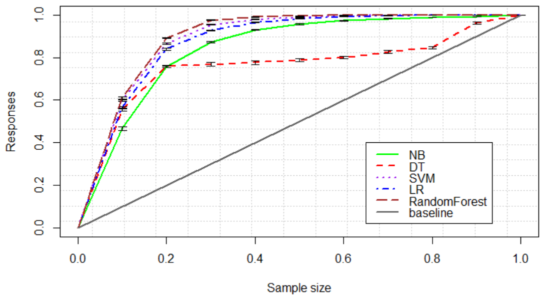


Figure 5: Lift Curve - Model Comparison

5.2 Inferences

As mentioned in variable importance, data driven models like Random forests and Decision trees are easy to understand by humans and can tell which attributes are more important in terms of their contribution to the prediction of the class variable. But in complex data driven models like SVM, it is difficult to understand which attributes are influencing more on the model for prediction of class variable. Hence we adopted a procedure called sensitivity analysis to derive inferences from the model. Sensitivity analysis measures how a model is influenced by each of its input attributes. This helps to quantify the contribution of a given attribute for the model.
We selected 4 important attributes - "Last call duration, outcome of previous marketing campaign, number of days passed after previous campaign, and month in which last call was made." By plotting Variable Effect Characteristic (VEC) curve which is a plot between the average influence of a given attribute (in x-axis) on the model probability of success (y-axis), we can get more input influence details and detect patterns in the data.

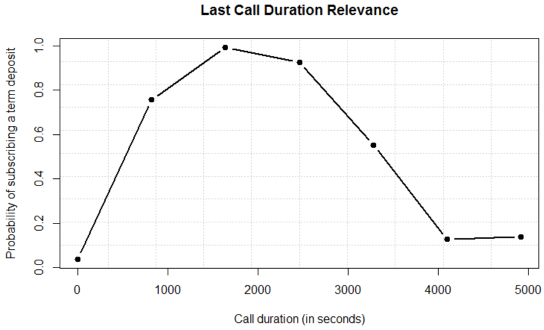


Figure 6: VEC Curve - Last call duration

Figure 6 shows that average influence of last call duration on the probability of a customer subscribing to term deposit. It can be seen from the figure that the probability of success is high if the call duration is more than 1000 seconds. This makes sense, since a successful sell requires a deeper dialog to describe the product and may be create empathy with the client. However as seen in the figure, after a certain threshold of 3000 seconds, the probability of success starts to decrease suggesting the client is only trying to be sympathetic but does not want to buy the product.

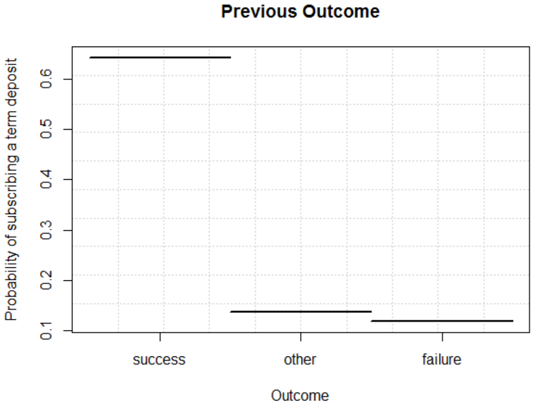


Figure 7: VEC Curve - Outcome of previous campaign

Figure 7 shows that average influence of the outcome of previous campaign on the probability of a customer subscribing to term deposit in the current campaign. The plot shows that, if the previous outcome is successful then probability of success in this campaign is also high. From this we can infer that those customers who subscribed to the term deposit previously were happy with the services offered by the bank and are willing to subscribe to term deposit again.

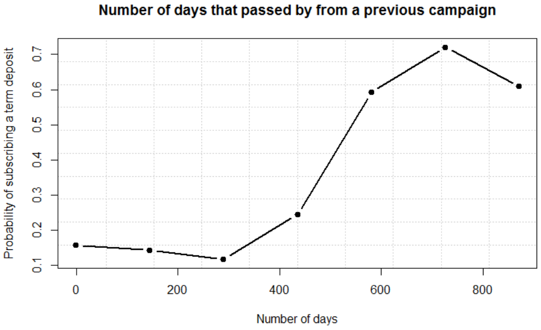


Figure 8: VEC Curve - Number of days passed after previous campaign

Figure 8 shows the average influence of the number of days passed by after previous marketing campaign held with that customer on the model probability of success. The plot shows that if the number of days passed after previous marketing campaign is more than 600 (approximately 20 months), then the probability of a customer subscribing to a term deposit is high. Normally customers prefer high return on investment for their term deposits and that will normally be in the range of 12 to 18 months. Those customers who are willing to subscribe might have subscribed during earlier term and based on the benefits and experience, they may opt for subscription in the current campaign. One other inference we can make is that those who have subscribed after getting the returns might have suggested their friends or family members regarding the benefits of the bank's deposit service and this might have also caused an increase in probability. Another inference is that people might have watched the performance of the bank for a certain period (like 600 days in this case) and after ascertaining the credibility of the bank customers may want to subscribe to the service. That is, customers expect the performance of the bank over time to be stable and the money they invest would be safe.

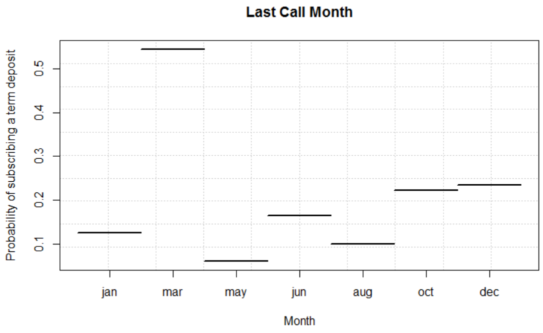


Figure 9: VEC Curve - Last call month

Figure 9 shows the average influence of the Last call month attribute on the model probability of success. It can be seen from the plot that the probability of success is very high in the month of march. One possible reason for this behavior can be said as the result of the pay hike or pay bonus offered at most of the companies in the month of march, encouraging customers to subscribe to a term deposit. This inference can be right or can be wrong. So in order to prove this inference we need more data to be collected in this respect. So this can be thought of as a cyclic process where the inferences drives us to collect more data to fill in the gaps of our knowledge about the data.

6 Related Work

Sėrgio Moro, Paulo Cortez, and Paulo Rita[4] has introduced the marketing techniques using data-driven approach for the success of bank telemarketing. In this paper, they have used machine learning techniques like logistic regression, decision trees, neural networks and support vector machine.

Ali Keles and Ayturk Keles [5] developed an intelligent bank market management system(IBMMS) for the managers who want to manage efficient marketing campaigns. It is developed using data mining and the inference engine.

Chakarin Vajiramedhin and Anirut Suebsing[6] proposed an approach for feature selection with data balancing for prediction of bank marketing. In this paper, a predictive rate is enhanced using correlation based feature relation based feature and data balancing. They have evaluated this approach using TP rate and ROC rate.

There are other marketing domains where machine learning is applying to analyze consumer's trend[7][8]. Marketing may be in many forms like email marketing, telemarketing, and digital marketing.

7 Conclusion

In the banking industry, optimizing the targeting the customers is an important issue. Our objective is successfully resolved by applying machine learning and data mining techniques. Thus by applying these techniques, banks can get a better return on investment. A number of phone calls can be greatly reduced as we are only targeting those customers who are most likely to subscribe for term deposit. Customer's trend on such subscriptions can also be analyzed.

From the evaluation metrics like accuracy, ROC curve and Lift curve, we can say that among the five classifiers random forest obtained the best predictive result. Performing sensitivity analysis on the dataset helps to find the patterns and extract useful knowledge from the data. The inferences derived can aid bank managers to better understand customer behavior and take effective decisions for the business to thrive.

In future, other machine learning techniques like neural networks can be used. An extension of this work can be done by collecting more information about the customer based on the inferences made from the current data set, thus building a robust model for targeting the subset of customers who are more likely to end up in subscribing to the services offered by the bank.

References

[1] ROC Curve, Lift Chart and Calibration Plot by Miha Vuk, Tomaz Curk

[2] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[3] https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

[4] Sérgio Moro, Paulo Cortez b , Paulo Rita - A data-driven approach to predict the success of bank telemarketing

[5] Ali Keles and Ayturk Keles- Ibmms decision support tool for management of bank telemarketing campaigns (International Journal of Database Management Systems (IJDMS) Vol.7, No.5, October 2015)

[6] Chakarin Vajiramedhin,Anirut Suebsing - Feature Selection with Data Balancing for Prediction of Bank Telemarketing

[7] http://www.skyword.com/contentstandard/marketing/how-will-machine-learning-change-the-state-of-digital-marketing/

[8] https://blogs.adobe.com/digitalmarketing/tag/machine-learning/

[9] Aptéa, C. and Weiss, S. (1997). Data mining with decision trees and decision rules, In "Future Generation Computer Systems", Vol. 13, No.2-3, pp. 197–210.

[10] Cortes, C. and Vapnik, V. (1995). Support Vector Networks, In "Machine Learning", Vol. 20, No.3, pp. 273–297.

[11] Ling, X. and Li, C., (1998). Data Mining for Direct Marketing: Problems and Solutions. In "Proceedings of the 4th KDD conference", AAAI Press, pp. 73–79.

[12] Cortez, P. and Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. In "Information Sciences", Vol. 225, pp. 1–17.