PREDICTION MODELLING FOR CUSTOMERS UPSELLING –APPLICATION OF VARIOUS MACHINE LEARNING ALGORITHMS

Dr A. Mansurali ***
Assistant Professor (Senior Grade)
PSG Institute of Management
PSG College of Technology
Coimbatore, Tamil Nadu
E-mail: mansurali@psgim.ac.in

V.Harish
Assistant Professor
PSG Institute of Management
PSG College of Technology
Coimbatore, Tamil Nadu
E-mail: harish@psgim.ac.in

Lokesh B
PSG Institute of Management
PSG College of Technology
Coimbatore, Tamil Nadu
E-mail:
lokeshbalasundaram@gmail.com

Abstract—**Banks with a customer base is always looking out to convert its existing customers for up selling their products. Selecting the existing customers for up selling the products is beneficial in many ways like targeting the known customers, targeting the right customers, cost of acquiring the customer is less and the behavior and profile of the customer is known. A banking dataset has been chosen for the purpose of applying various algorithms. Dataset contains 5000 data points, 12 predictor variables and one outcome variable which is whether the customer has reacted to the campaign for a personal loan or not. The success rate of the previous campaign was about 9 percent. The problem at the hand now is to identify existing clients that have higher chance to subscribe for purchasing a loan or for any new products and focus marketing effort on such clients. The objective of the study is to apply a classification approach to predict which clients are more likely to subscribe for personal loans using various models and machine learning algorithms and select the best algorithm based on the different evaluation parameters say AUC, Classification Matrix etc..,. This will increase the success ratio for up selling while at the same time reduce the cost of the campaign. The Various algorithms applied are logistic regression, CART, neural network etc.., using R. The output of the model is further used to predict the probable customers to target and also to understand their profile characteristics.**

Keywords— Bank, Predictor, Classification, Model, Marketing

1. Introduction

Banks often conduct many campaigns to upsell their products to the targeted customers. While conducting such campaigns banks collect data about the customers who take part in such campaigns. Banks use the data collected during such campaign to evaluate the success of the campaigns by measuring the number of customers who responded positively to such campaigns. Bank use these data to establish relationship between customers in future and target them by offering customized products and services since 95% of customers are influenced to a bank by the marketing campaigns of the banks [4]. As banks have huge and rich data about individual customers it becomes hard for banks to identify a suitable algorithm to better micro target existing customer or a new customer. Micro targeting increases the probability of the customer to subscribe to the service offering [7]. Hence, various machine learning techniques are employed for classifying the data set which has 5000 data points and 12 predictor variables and one outcome variable to find an algorithm which better targets the customer.

2. Methods

R programming language which is a software environment for data analysis and statistical computing and modelling is chosen to devise five different classification algorithms, to analyses the performance of each algorithm on the chosen bank marketing data set. The outcome of the analysis is to select the algorithm which can better target the customer, that is, the algorithm which have greater accuracy among the five. Out of the five chosen algorithms 2 are decision trees-based model and 3 are regression-based models. The five chosen techniques are

2.1 CART – Classification and Regression Trees

Classification and regression trees (CART), a statistical Model developed by Breiman et al (1984) [1], is a classification tool used to classify an object into two or more populations. CART methodology is outlined into three stages namely stage 1 – selecting variable and split points using splitting criterion for growing tree using recursive partition technique, stage 2 – Pruning procedure incorporating minimal cost complexity measure and finally, stage 3 – methodology to select optimal tree which yields lowest error.

2.2 LPM – Linear Probability Model

Linear probability model is a linear regression model which establishes relationship between dependent variable and independent variable(s). It varies from conventional regression model by characteristic of the independent variable since all the independent variables used to establish relationship will have values either 1 or 0 i.e., they can take one binary values. This implies that the customer can be either targeted or not targeted. LPM can be used for inference and estimation, Prediction and classification, and selection bias [3]. Here it is used for the purpose of prediction and classification, to classify the customers and predict whether customer will subscribe to personal loan or not.

2.3 Logistic Regression

Logistic Regression is used to predict binary outcome of the dependent variable based on given set of independent variables. Instead using outcome variable to predict it uses function of dependent variable to predict. Here it is used to predict the outcome of the campaign to upsell the personal loan to existing customers. The basic equation of logistic regression function is

$$\mathrm{Ln}\ (p\ (y=1)/1-p)(y=1) = Z$$

$$Z = Beta0 + Beta1\ (x1) + Beta2\ (x2) + \ldots\ Beta\ n\ (Xn)$$

Here

Z is the target variable (outcome variable)

$Beta0 + Beta1\ (x1) + Beta2\ (x2) + \ldots\ Beta\ n\ (Xn)$ is the predictor.

2.4 LDA – Linear Discriminant analysis

LDA is a Classification technique used for classifying observations into predefined classes[9]. The model creates set of predictor functions based on training dataset called discriminant functions. The linear function representing the model is

$$y = a + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

where parameters $a_1,a_2,..,a_n$ are determined in such a way that discrimination between the groups is maximum.

2.5 Random forest

Random forest technique was introduced by Breiman (2001) [2]. It is a powerful ensemble machine learning algorithm which creates multiple decision trees and every observation is fed into each decision tree. Finally, it combines the output generated by each decision tree and most common output in each observation is used as final output.

3. Dataset Description

The data set consist of 5000 entries with 12 predictor variable and 1 outcome variable. For the purpose of training only 70% of data is used and the rest 30% is used to test accuracy of the algorithm in use. The description the data set is shown below in the table 01

ATTRIBUTE	TYPE	DESCRIPTION
Age..in.years.	Integer	Age of the customer
Experience..in.years.	Integer	Work Experience of the customer
Income..in.K.month	Integer	Monthly Income of costumer in thousands
ZIP.Code	Numeric /Discrete	Postal zip code
Family.members	Integer	Number of members in the customer's family
CCAvg	Float	
Education	Categorical	Educational status of the customer
Mortgage	Integer	Mortgage amount
Securities.Account	Binary	Does customer has Securities Account? Yes = 1, No= 0
CD.Account	Binary	Does customer has CD account? Yes = 1, No= 0
Online	Binary	Does customer has online banking account? Yes = 1, No= 0
CreditCard	Binary	Does customer holds a credit card? Yes = 1, No= 0

Table 01 – Description of Dataset

4. Exploratory Data Analysis

In this section a brief exploratory analysis is performed on the data to get hold about the data that is to be analyzed. The table 02 shown below shows the values of Mean, Standard deviation (SD), Median, Mean Absolute Deviation (MAD), Minimum value (Min), Maximum value (Max), Range and skewness for 11 predictor variables except zip code since its of nominal type. The average age of the participants in the campaign is 45.33 years with an average income of 73.77 K per month.

ATTRIBUTE	MEAN	SD	MEDIAN	RANGE
Age..in.years.	45.3384	11.4631	45	44
Experience..in.years.	20.1046	11.4679	20	46
Income..in.K.month.	73.7742	46.0337	64	216
Family.members	2.397132	1.1450	2	3
CCAvg	1.937938	1.7476	1.5	10
Education	1.881	0.839	2	2
Mortgage	56.4988	101.71	0	635
Securities.Account	0.1044	0.3058	0	1
CD.Account	0.0604	0.2382	0	1
Online	0.5968	0.4905	1	1
CreditCard	0.294	0.4556	0	1

Table 02- Data description table

Figure 01 is the plot of correlation diagram to understand the multicollinearity among all the predictor variables. In the plot its is seen that there exist multicollinearity between variables Experience..in.years. and Age..in.years. Thus, this multicollinearity is removed before training the algorithms.

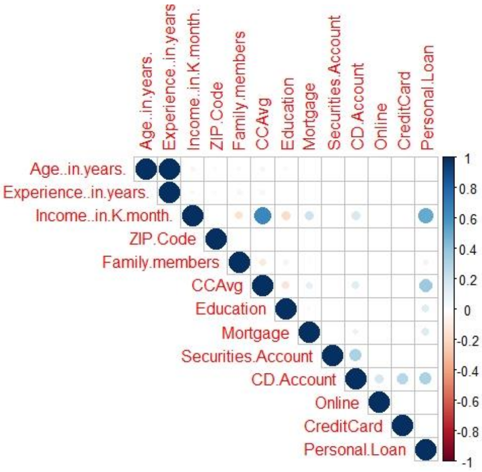

Figure 01 – Correlation plot

5. Algorithm Evaluation Criteria

All the algorithms are evaluated based on various error metrics namely

a. Sensitivity – The proportion of actual positive classes that are correctly identified.

b. Specificity – The proportion of actual negative classes which are correctly identified

c. Accuracy – The proportion of total predictions that were correct.

d. Area under the ROC Curve (AUC) – The ROC (Receiver Operating Curve) is the plot between True positive rate and False positive rate. Area under the curve is found to bring down ROC to single number.

e. Gini Coefficient – Gini is the ratio between ROC and the diagonal line & the area above the triangle and its calculated using formula

Gini = 2*AUC - 1

f. K-S – Kolmogorov-Smirnov chart measures the performance of classification models, more accurately K-S is the degree of separation between positive and negative distributions.

6. Experimental Analysis

6.1 CART

The dataset is split into training and testing data sets. The model is built upon training dataset and used 12 attributes to train the classifier and built the decision tree, shown in figure 02. To explain the decision tree as shown in figure 02, node 1 at which variable Income..in.K.month. is chosen for splitting has 90% of observations are class 0 i.e. 90% of the observations are having income less than 114K per month. Similarly at the respective nodes different variables are used to split and classify the data. When the same dataset is pruned to minimal cost complexity factor, the root node error is 0.094 which is null error. When the predicting data is fed into the validated algorithm to get a classification matrix or confusion matrix which evaluates the algorithm and its shown in confusion matrix in table 03. Confusion matrix is a table which measures the classifier performance[11]. Finally all the algorithm evaluation metrics, as mentioned in section 5 of this paper, have been found to compare it with other algorithm.

	Expected result		
Prediction		No	Yes
	No	1349	6
	Yes	29	116
Classification Error	0.094		

Table 03 – CART Confusion matrix

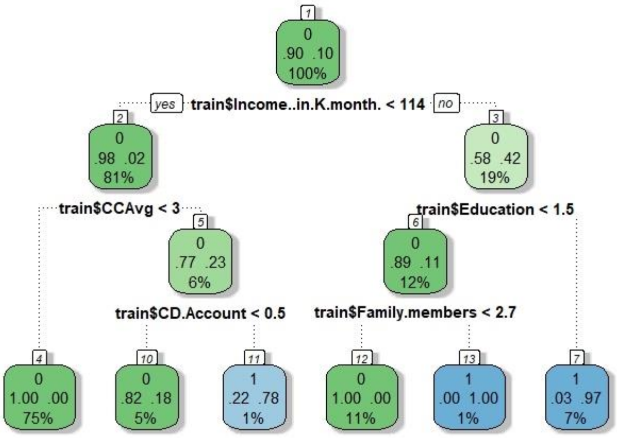

Figure 02 – Decision tree

Inference – From the CART model it can be inferred that there are high chances for customers who have an average income of more than 114K per month and with an education level of 2 or more to respond to the campaign. Thus, subscribing to personal loan. Since they are more educated and have more than average income it is most likely they will respond to the campaign.

6.2 Linear Probability Model

When the linear probability model have been executed in the training dataset it has been found that variables Income..in.K.month., Family.members, CCAvg, Education, Mortgage, Securities.Account, CD.Account, Online, CreditCard are found to be significant and the R squared value 1.924% and Adjusted R squared error is 1.896%. All significant variables are retained and the LPM Model is executed again, the model yielded R squared value of 38.1% and 37.94%. When the model is applied for prediction dataset and a confusion matrix is created as show in table 03, the classification error is 0.0758. Finally, the model is evaluated for the metrics shown in section 5 of this paper to compare it with other algorithms.

Inference – If the bank focuses on customer having mortgage, securities.account, cd account online banking and a credit card there is a high probability for the customer to respond positively to the campaign. Hence if banks target such existing customers the success rate of campaign will be high.

	Expected result		
Prediction		No	Yes

	No	1331	4
	Yes	100	40
Classification Error	0.0758		

Table 04 – LPM Confusion matrix

6.3 Logistic regression

Logistic regression model has been executed on all the 12 predictor variables and its found that only Income..in.K.month., Family.members, CCAvg, Education, Securities.Account, CD.Account, Online, and CreditCard variables to be significant. Then the logistic regression model is executed only for significant variables for 7 fisher iterations. On finding McFadden R squared [8], a pseudo R squared value which is like R squared in regression model, value for this model it comes out to 70.08%. This model is used to evaluate the prediction on validity dataset and confusion matrix is built as shown in table 05. Finally, the model is evaluated for the metrics shown in section 5 of this paper to compare it with other algorithms.

	Expected result		
Prediction		No	Yes
	No	1320	51
	Yes	15	89
Classification Error	0.04255		

Table 05 – Logistic Regression Confusion Matrix

Inference – Though the model has less error it becomes difficult if the classes are well separated also if the variables are reduced it becomes unstable [5], hence it becomes difficult to predict which customer will better respond to the campaign though it has 7 significant variables.

6.4 LDA – Linear Discriminant analysis

In LDA, this paper tries to find variables that are significant in separation, first it is proved at least on variable is significant with the help of MANOVA. The MANOVA result shows that atleast one or more variables are significant. The wilks value of 0.61401 is the extent of discrimination that is possible in the data. Fisher discriminant is developed which helps us to discriminate who will subscribe and not subscribe the personal loan. On calculating p-value it is found that the variables Age..in.years., Zip Code, Experience..in.years., Securities.Account, Online, CreditCard are significant discriminators. With this LDA model is built with two sets of output one is group means and other coefficient of linear discriminant functions. When a density plot is created as shown in figure 03, it can be seen that more than 60% overlap shown by wilks lambda. Thus, the developed model is evaluated to prediction dataset the classification error is found to be 0.0646. Finally, the model is evaluated for the metrics shown in section 5 of this paper to compare it with other algorithms

Inference – It can be inferred that the variables Age..in.years., Zip.Code, Experience..in.years., Securities.Account, Online, CreditCard are the important factors in deciding whether the customer will respond or not. If the customer has securities.account, online banking and credit card there is high probability that the customer will respond positively to the campaign. But applying LDA to the data

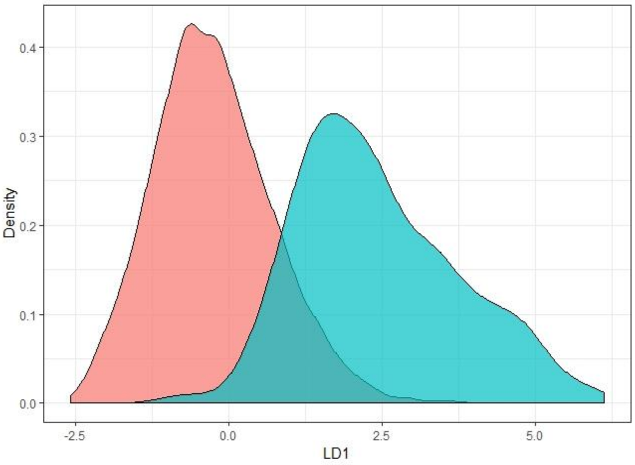

Figure 03 – Density plot

	Expected result		
Prediction		No	Yes
	No	1319	37
	Yes	54	90
Classification Error	0.0646		

Table 06 – LDA Confusion Matrix

6.5 RF - Random Forest

Random Forest model has been initially ran, with 501 trees to make the model more robust and the forests provides results of higher accuracy, on all the variables. Then all the 12 predictor variables are taken to be important variables and number of variables for splitting at each node referred to as mtry parameter is found to get optimum number of variables. It is found that mtry 6 with an accuracy of 98.5% is optimum. With this mtry maximum number of trees and nodes is found and the model is fitted. The fitted model is validated against the validation dataset the classification error of 0.01027 is found form the confusion matrix in the table 03.

	Expected result		
Prediction		No	Yes
	No	1334	14
	Yes	1	126
Classification Error	0.01027		

Table 07 – Random Forest Confusion matrix

Inference – Based on the Random forest algorithm the success lies in the number of trees used to model. Thus, higher the trees higher the accuracy. But when the data becomes huge it becomes more complex to construct a tree and less intuitive. Even if the results are accurate it may not depict the true picture of the data. Hence targeting customer with random forest is complex and time consuming.

7. Results

Finally, a consolidated table, table-04, is drawn to compare all the algorithms against metrics mentioned in section 5 of the paper. The figures are marked in red and green to highlight the best and worst figures respectively. It can be seen that CART model performs good in Specificity, AUC and K-S, while random forest for Sensitivity and accuracy and LPM for Gini Coefficient. Overall, CART algorithm is the best model since it has highest values for metrics AUC and K-S among the five.

Metric	**CART**	**Rf**	**LPM**	**Logistic regression**	**LDA**
Sensitivity	98	99.99	93.01	98.87	96.06
Specificity	96	90	90.90	63.57	70.86
Accuracy %	97.67	98.98	92.95	95.52	93.54
AUC	0.9865	0.9496	0.9667	0.9496	0.9511
Gini Coefficient	0.8799	0.0724	1.0337	0.8275	0.8078
KS	0.9276	0.8992	0.8547	0.8992	0.7837

Table 04 – metrics table

7. Conclusion

This paper has done an extensive mining of data from the selected bank dataset to identify the customers who will respond favorably to the campaign rolled out by banks. The main objective is to attract the customers towards the personal loan campaign which ultimately give the banks a profit in terms of interest. Exploratory data analysis has been done to describe the data in general and give a glimpse of the nature of the data at hand. CART, LPM, Random forest, Logistic Regression and LDA all have been developed on the training data set containing 70% of the data from the master data set and validity of the model has been evaluated on validity dataset containing 30% of the remaining data from the master dataset considering the customer will respond as 1 or not as 0. With five different algorithms to find the best algorithm that better micro targets the customer. From the study it is identified that "CART" Algorithm to best suited among the five algorithms. Classification and Regression Tree is the algorithm chosen to perform the modelling as it has its own advantages like pruning, easy for interpretation, feature engineering on its own, cross validation parameters. All validation measures score more than 87% for CART algorithm with the key metrics showing best result, AUC is 0.9865, Gini is 0.8799 and K-S is 0.9276, among the five. If analyzed with CART model the probability maximizing the profit of the company will be more since a greater number of customers can be targeted to positively respond to the campaign. And lastly the model deployment in terms of insights using business rules, monitoring mechanism and the scope for improvement is also been traced.

References

1. Breiman L. et al. (1984) Classification and Regression Trees, Wadsworth, CA.
2. Breiman, L. (2001). Random forests. Machine Learning 45 5–32
3. Chatla, Suneel and Shmueli, Galit, Linear Probability Models (LPM) and Big Data: The Good, the Bad, and the Ugly (October 11, 2016). Indian School of Business Research Paper Series.
4. Faramarzpour et al. (2015), "The Effect of Marketing of Bank Services on Customer's Preference of Private Banks: Case Study of Mellat and Tejarat Banks in Khorasan Razavi Province", International Journal of Management, Accounting and Economics, Vol. 2, No. 3, March, 2015, ISSN 2383-2126.
5. Park, H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, *43*(2), 154-164.
6. Metcalf, A. L., Angle, J. W., Phelan, C. N., Muth, B. A., & Finley, J. C. (2019). More "bank" for the buck: Microtargeting and normative appeals to increase social marketing efficiency. *Social Marketing Quarterly*, *25*(1), 26-39.
7. Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression. Consulted page at September 10th http://www.unt.edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011.pdf, 29, 2825-2830.
8. Tharwat, Alaa & Gaber, Tarek & Ibrahim, Abdelhameed & Hassanien, Aboul Ella. (2017). Linear discriminant analysis: A detailed tutorial. Ai Communications. 30. 169-190,. 10.3233/AIC-170729.
9. M. Vedanayaki, "A Study of Data Mining and Social Network Analysis", Indian Journal of Science and Technology, Vol 7(S7), 185–187, November 2014.
10. Ivo Düntsch, "Confusion Matrices and Rough Set Data Analysis", IOP Conf. Series: Journal of Physics: Conf. Series **1229** (2019) 012055.