Classification of Telemarketing Data using Different Classifier Algorithms

VenkateshYadav, M. SreeLatha, T. V. RajiniKanth

Abstract—Globalisation, growth of new technology usage and tough competition, made the banks to adopt new approaches to get competitive advantage to enlarge customer databases and also to generate customer satisfaction. In the present days the banks are trying to enhance customer base to meet their business targets for which they follow various approaches like Internet banking, Direct Tele Marketing, Mobile Banking, etc. Apart from banking services to customers, Banks are also selling Insurance policies to the customers through Tele Marketing and by which their business is expanding exponentially. In this paper, various Machine Learning Algorithms like Random Forest, Random Tree, Rep Tree, Naïve Baye's, J48 Decision Tree before and after refinement of data and advanced Statistical techniques were applied for effective analysis of Bank's Tele Marketing Data in order to enhance number of subscribing customer's.

Keywords: Machine Learning Algorithms, Tele Banking, Subscribing Customers, Advanced Statistical techniques, Direct Tele Marketing

I. INTRODUCTION

Telemarketing is a well-organized and professional marketing technique of generating quality business leads for your organization, and a quick technique to connect with new clients personally, arrange appointments and make deals over the phone. Direct marketing is promoting products which are having high sales in the market, Direct marketing is mostly used in Banking , Insurance industry and Retailing . It is increasing day to day.

By using Telemarketers you can get best potential clients/prospects for your company's products or services and know their needs, and requirements, which can help your company to earn more sales revenue with a telemarketing campaign. Telemarketing is classified into Inbound telemarketing and Outbound telemarketing .In Inbound Telemarketing your prompt prospects or interested clients call you to make queries about your product or services. As a marketing professional, you avail the opportunity forward by convincing the prospect, arranging an appointment, or even close the deal over the phone. Outbound Telemarketing involves your marketing specialist make calls for your company to a targeted audience or your targeted locations and a specific group of people to convince them to purchase your business products or services**.**

Revised Manuscript Received on August 05, 2019.

A.Venkatesh Yadav, , Research Scholar, Acharya Nagarjuna University: Guntur, Andhra Pradesh, India.
E-mail: venkatesh357@gmail.com
Dr. M. SreeLatha, Professor & Head, RVR & JC Engg. College, Chowdavaram, Andhra Pradesh, India.
E-mail: lathamoturi@rediff.mail.com
Dr. T. V. RajiniKanth,Professor & Dean R&D, SNIST, Hyderabad, Telangana, India.
E-mail: rajinitv@gmail.com

II LITERATURE SURVEY

LilianSing'oei et al [1] did an analysis on bank Sample marketing data and Proposes that data mining (DM) methods are the best way for Bank direct marketing Campaigns. Sergio Moro et al [2] did case study on telemarketing calls data of selling bank deposits and by using DM methods identified the success of telemarketing and compared DM classification methods Logistic Regression (LR), Neural Network (NN) and Support Vector Machine (SVM), Decision Trees (DT) and found good results from NN Method. MuneebAsif [3] in his thesis applied different classification methods, namely SVM, DT, Random Forest (RF) and Artificial Neural Network (ANN). The results showed that reduce subset of variables which is attained through Random Forest has the best accuracy with the classification method random forest. Q.R. Zhuang, Y.W. Yao et al [4] did case study on term-deposits customers data and stated that deposits are having problems from economic instability and competition in the market. By using DM SPSS model to identify customers for term-deposit and knowing the customer needs to improve bank marketing.

Charles X. Ling et al [5] stated that during the testing with DM techniques, many challenges were identified. Few of them have high differences in class distribution. Dr. Md. Rashid Farooqi et al [6] stated that banking industry identified DM is one of key area for business improvement. Many DM techniques are used in banking application like marketing term deposits, Retail banking, CRM, risk identification like fraud detection. Anthony Rahul Golden S [7] studied about digitization of Indian banking sector for better services.

Dr. Agboola A. A. [8] did case study on banking services in Nigeria. Highly used services are ATM, online banking, Telephone Banking. He stated that Telephone Banking improves customer relationship, customer trust if they resolve the banking issues. Dr. SurajitGhoshDastidar et.al [9] stated that the Indian government started campaigning that all customers should prefer using digital banking. After analysing the data transactions identifies that net banking is used mostly. V Vimala [10] did case study on customers having challenges with online banking security using online banking services and proposed that bank should give online security protection, virus scanner, Hacking alerts to customers. HosseinHassani et al [11] stated that analyzing the huge data by using DM techniques in Banking industry to improve customer needs by using best strategies in banking. MerveMitik et al [12] did case study on common

challengers for marketing on the bank product. The first step is to cluster the customers and second one is classification method for product and Marketing channel offers. Indrajani et al [13] found the algorithms that are suitable for fraud detection on online banking transactions. Israel González-Carrasco et al [14] did case study on identifying similarities between bank transaction data by using business analytics and applied intelligence methods on big data. Justice Asare-Frempong et al [15] used four classifiers methods namely, Multilayer Perceptron NN, DT (C4.5), LR and RF on bank direct marketing sample data and achieved the best results on customer response. Yogesh Sanjay Golecha [16] did case study on bank customer data set to know whether customers are interested in investing in log-term deposit or not. OlatunjiApampa [17] has studied Classification techniques used for designing bank dataset.TeemuVartiainen [18] has studied use of DM techniques on telemarketing dataset of a bank, to identify the potential customers for investment in long term deposits. M. Purnachary [19] has analyzed a case study on two Classification techniques Bayes Net, Naïve Bayes on bank marketing dataset, which is available in UCI Repository and found best results from Bayes Net classification.

III PROPOSED APPROACH

Preprocess the data set [13] for removal of noise and use imputation methods then refine the data set and saved in to user suitable format. Various Regression techniques were applied on it and developed predictive models. Initially, in the pre-processing stage Skewness, Kurtosis and Grubb's Test will be conducted on the data set to find whether the data is normally distributed or not apart from testing for the presence of Outliers etc. Anderson Darling Test was applied on data set whether it is normal Distributed or not.

SKEWNESS: It indicates distortion of data from the Normal distribution.

Skewness = (Mean – Median) / Standard Deviation - (A)

KURTOSIS: It describes the distribution and finds farthest values in either tail.

Kurtosis $= (\sum_{i=1}^{n}(x_i - \overline{x})^4/n)/(\sum_{i=1}^{n}(x_i - \overline{x})^2/n)^2 - 3$ ---- ***(B)***

GRUBB'S TEST:

It is defined as: $G = \frac{max|Y_i - \overline{Y}|}{s}$ -- - (***C)*** with $\overline{Y}$ is the sample mean and ***s*** is the standard deviation.

ANDERSON's DARLING NORMALITY TEST

It is a statistical test to find a dataset comes from a certain probability distribution or not.

Anderson-Darling statistic equation:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(X_i) + \ln(1 - F(X_{n-i+1}))]$$

(D)

Then Linear and Quadratic Regression Equations were found and displayed by equations (E), (F) and (G) respectively depending on the number of Predictor variables vs Response Variable. Here Balance attribute was considered as Response variable and others as Predictor variables. Finally the given data set was converted to the desired format and applied k-means Clustering technique for further analysis. The resultant clustered data set was classified using Decision tree (J48 Algorithm) classifier and also Random forest for effective classification and prediction. Finally the results were summarized for useful conclusions.

IV. RESULTS AND ANALYSIS

The Skewness, Kurtosis and Grub's Test values are shown in Table-1. The values of the attributes 'Age', 'Balance', 'Day', 'Duration', 'Campaign', 'Pdays', and 'Previous' are positive. It indicates that the distribution is skewed towards the right for all the attributes. Positive skewness means that the mean of the data values is bigger than the median, and the data distribution is right-skewed. The degree to which portfolio returns appear in the tails of our distribution is measured as Kurtosis. Kurtosis for normal distribution has 3, which indicates that a normal distribution does have some of its mass in its tails. A distribution with a kurtosis bigger than 3 has more returns out its tails than the normal, and one with kurtosis less than 3 has fewer returns in its tails than the normal. There are more outcomes for the attributes, namely Balance, Duration, Campaign, Pdays and Previous. Out of these attributes Campaign, Previous and Balance have higher risks as outcomes are more compare to other attributes. The Grub's test values are indicating that except the attribute Day all other attributes are having Outliers. The increasing order of the attributes having outliers is Age, Pdays, Duration, Previous, Campaign and Balance. It is observed that the Attribute Balance has highest Outliers and are replaced with mean. The attributes Month, Day, Age have no outliers and so replacement with mean is not required.

TABLE-1: Skewness, Kutosis and Grubb's Test Values

S. No.	Attribute	Skewness Values	Kurtosis Values	Grub's Test Values
1	Age	0.6990374	3.347063	4.33
2	Balance	6.592054	91.29128	23.18
3	Day	0.09456412	1.960291	1.83
4	Duration	2.77058	15.51487	10.63
5	Campaign	4.740767	40.1265	15.18
6	Pdays	2.715269	10.947	8.30
7	Previous	5.871361	54.9364	14.44

The Fig.1 shows the graph for Skewness, Kurtosis and Grubb's Test Values of 7 Attributes Age, Balance, Day, Duration, Campaign, Pdays and Previous. It is observed from the graph that Skewness, Kurtosis and Grubb's Test Values are behaving in the same way. They have higher values for the Attributes Balance, Previous, Campaign. The outliers are also more in these attributes values.

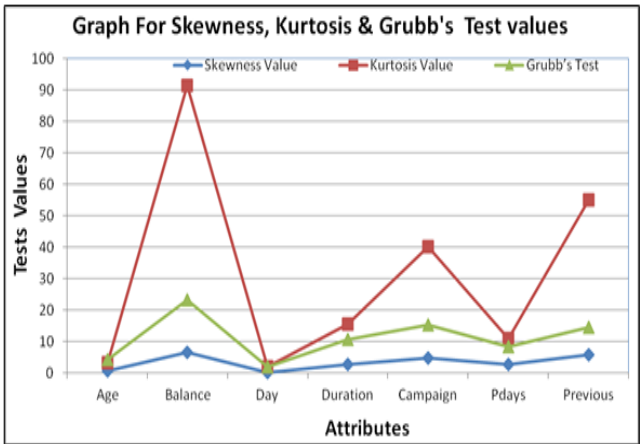


Fig.1: Graph with Attributes along X- axis and Skewness, Kurtosis and Grubb's Test values across Y -axis

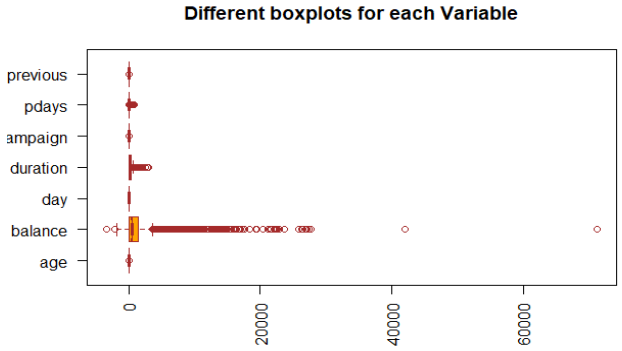


Fig.2: Outliers of the 7 attributes are shown above using Box plots

The Box Plots for the attributes are shown below **Fig.1**, **Fig.2** and **Fig.3** for outlier's detection in the data set

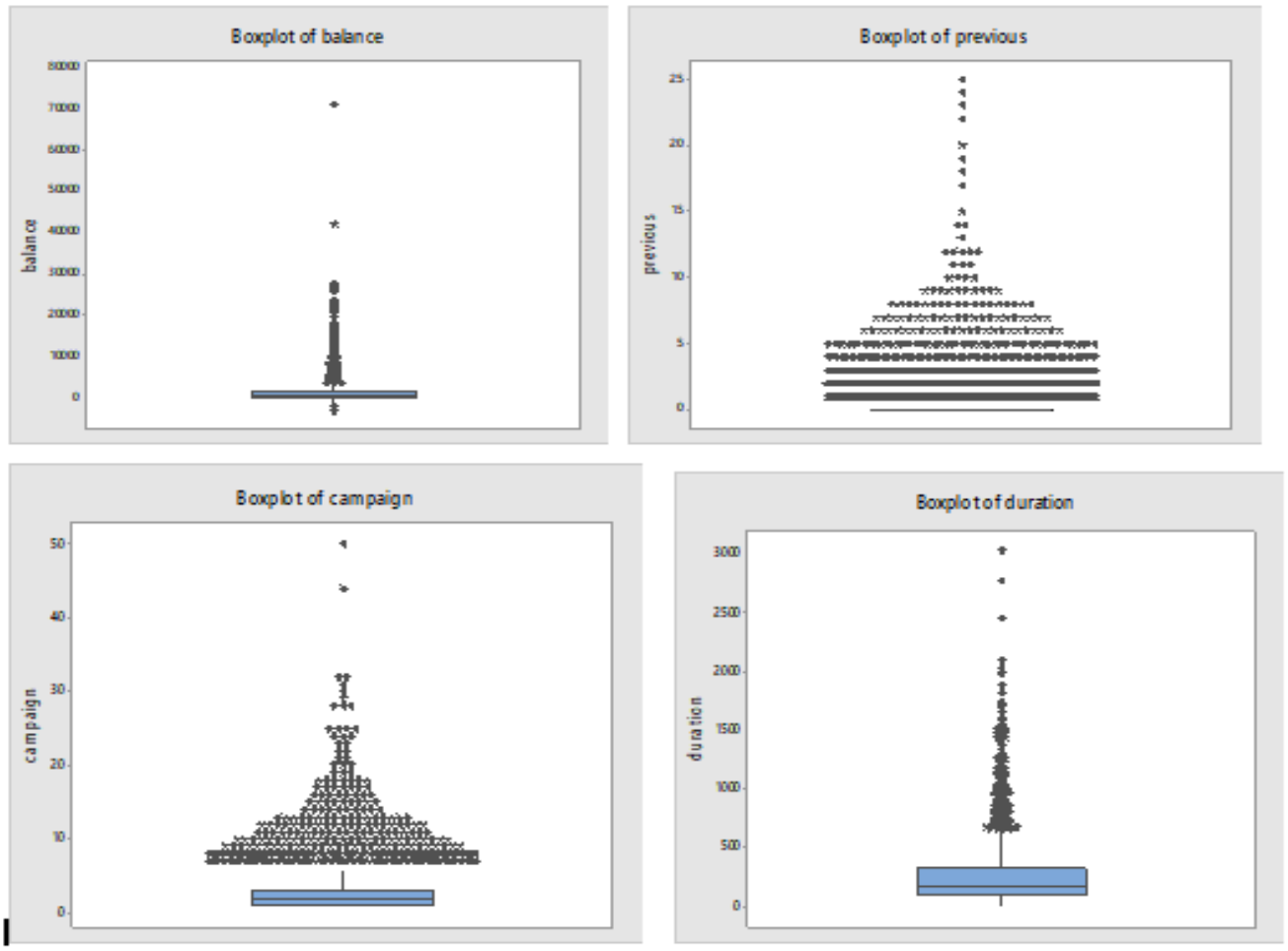


Fig.3: Outliers of the 4 attributes Balance, Previous, Campaign and Duration are shown above using Box plots

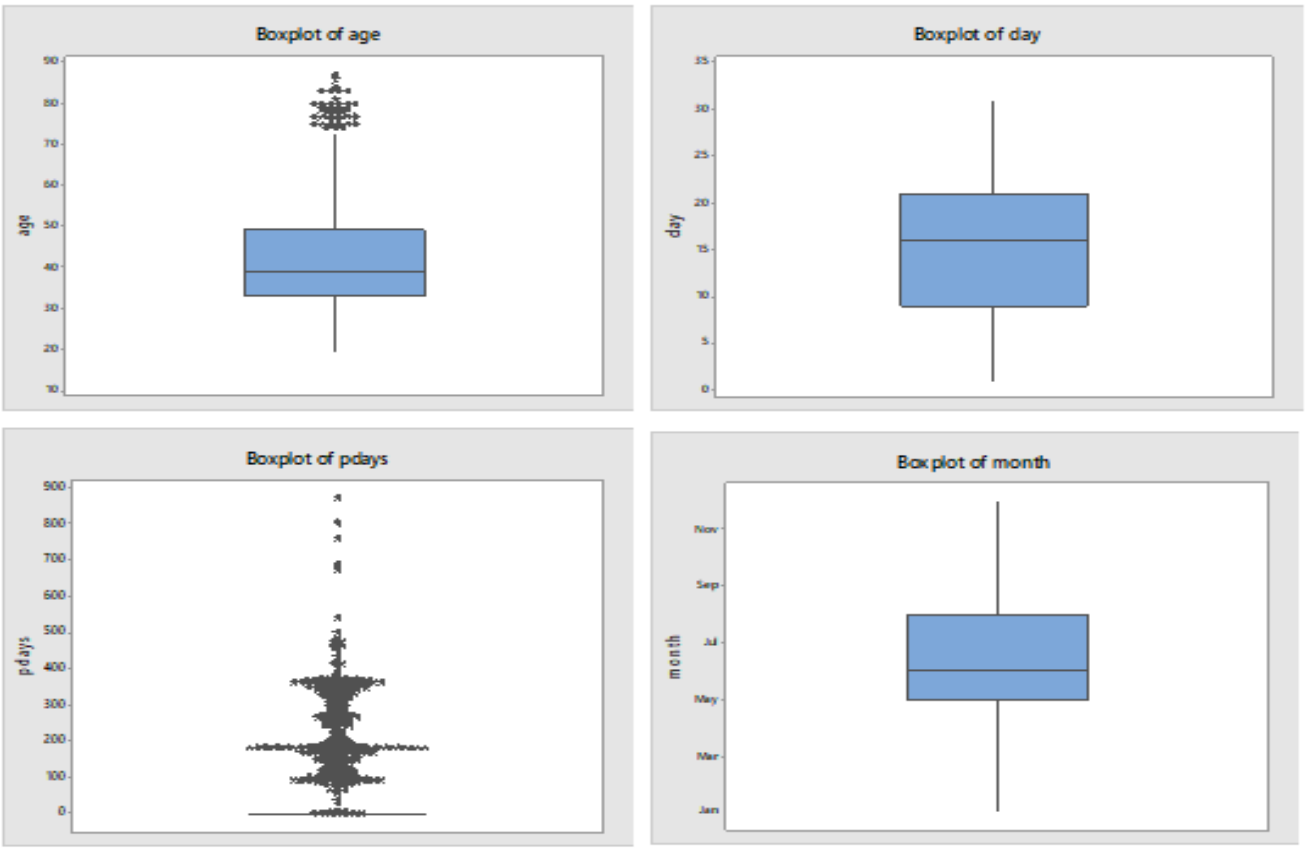


Fig.4: Outliers of the 4 attributes Age, Day, Pdays and Month are shown above using Box plots

The linear regression equation for the attributes Age and Balance is given by equation ***(E)***

Balance = 532.2 + 21.25 Age --- (E)

The polynomial regression (i.e. Quadratic Regression) equation for the attributes Age and Balance is given by equation ***(F)***

Balance = 1325 - 16.43 Age + 0.4199 Age ^ 2 -- (F)

The regression equation for the 7 attributes is

Balance = - 162228 + 19.32 Age – 1.94 Day + 3.735 Month – 0.180 Duration

-8.5 Campaign + 0.143 Pdays + 53.9 Previous --- (G)

The refined data set after replacement of outliers by mean is subjected to

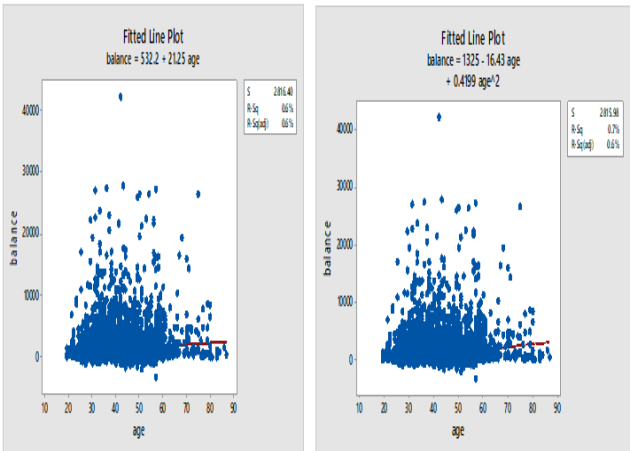


Fig.5: Graphs of Linear and Quadratic Regressions of Balance vs Age

The following Table-2 shows the 5 clusters of the refined data set formed after the application of k- Means clustering algorithms having the attribute fields Age, Marital , Job , Default , Education, Housing, Loan , Balance, Day, Contact, Duration, Month, Pdays, Campaign, Poutcome, Previous and y. The detailed analysis of 5 clusters is given below.

Cluster 0: The Age is third highest among the clusters with Self-employed category having marital status as Divorced and education as tertiary and no credit for default attribute and have lowest balance amount across clusters and has neither housing nor personal loan. The customer was contacted 9 days before through mobile communication in the month of February with average call duration of 266 seconds with an average of two times for the present campaign and an average of 0.5 times in the last campaign. The number of days that passed after last marketing campaign was 27 days with an outcome was unknown and finally customer has not subscribed term deposit.

Cluster 1: The Age is the lowest among the clusters with technician category having married with secondary education and no credit for default attribute and has third highest balance amount across clusters and has housing loan but not personal loan. The customer last contacted was 14 days before through mobile communication in the month of May with average call duration of 259 seconds with an average of two and half times for the present campaign and an average of 1 time in the last campaign. The number of

days that passed after last marketing campaign was 86 days with an outcome was unknown and finally customer has not subscribed term deposit.

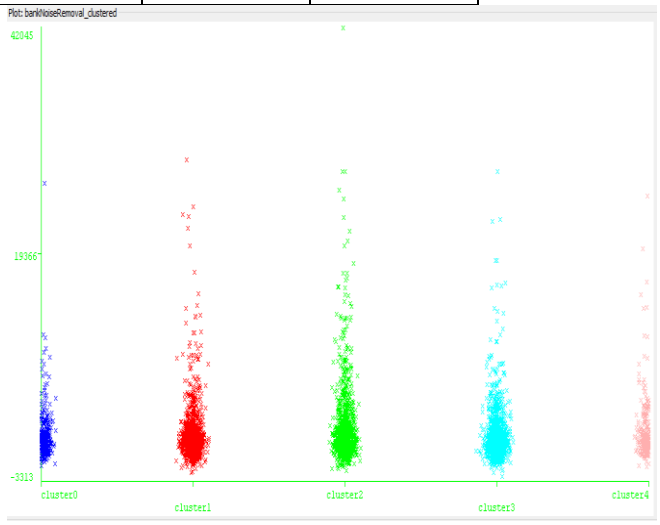
Cluster 2: The Age has the second highest among the clusters with Management category having married with tertiary education and no credit for default attribute and has second highest balance amount across clusters and has neither housing loan nor personal loan. The customer last contacted was 19 days before through mobile communication in the month of August with average call duration of 239 seconds with an average of 3.3 times for the present campaign and an average of 0.3 times in the last campaign. The number of days that passed after last marketing campaign was 19 days with an outcome was unknown and finally customer has not subscribed term deposit.

Cluster 3: It has the highest cluster size. The Age has the last but one among the clusters with blue-collared category having married with secondary education and no credit for default attribute and has last but one balance amount across clusters and has housing loan but not personal loan. The customer last contacted was 18 days before through unknown channel communication in the month of May with average call duration of 261 seconds with an average of 3.0 times for the present campaign and an average of 0.05 times in the last campaign. The number of days that passed after last marketing campaign was 4 days with an outcome was unknown and finally customer has not subscribed term deposit.

TABLE-2: Clustered Data sets after k-means clustering is applied on refined data

Attribute	Full Data(4521)	Cluster 0 (339) 7%	Cluster 1 (1259) 28%	Cluster 2 (1328) 29%	Cluster 3 (1340) 30%	Cluster 4 (255) 6%
age	41.1701	42.0265	38.6585	42.2831	40.203	51.7176
job	management	self-employed	technician	management	blue-collar	retired
marital	married	divorced	married	married	married	married
education	secondary	tertiary	secondary	tertiary	secondary	secondary
default	no	no	no	no	no	no
balance	1407.2263	1196.8732	1208.2923	1757.0542	1202.1784	1924.7176
housing	yes	no	yes	no	yes	no
loan	no	no	no	no	no	no
contact	cellular	cellular	cellular	cellular	unknown	cellular
day	15.9153	9.6962	13.7847	18.7123	16.7776	15.6039
month	may	feb	may	aug	may	oct
duration	263.3504	265.8702	258.9936	238.6536	261.0082	422.4353
campaign	2.7832	2.3658	2.4639	3.2312	2.9418	1.749
pdays	39.5828	27.7286	85.8801	18.6431	3.7955	123.8706
previous	0.5373	0.5339	0.996	0.3276	0.0485	1.9373
poutcome	unknown	unknown	unknown	unknown	unknown	failure
y	no	no	no	no	no	yes

Cluster 4: It is the lowest cluster size. It has highest Age across the clusters with retired category having married with secondary education and no credit for default attribute. It has the highest balance amount across all clusters with neither housing loan nor personal loan. The customer last contacted was 16 days before through mobile channel communication in the month of October with average call duration of 423 seconds with an average of 1.8 times for the present campaign and an average of 2 times in the last campaign. The number of days that passed after last marketing campaign was 124 days with an outcome was Failure and finally customer has subscribed term deposit.



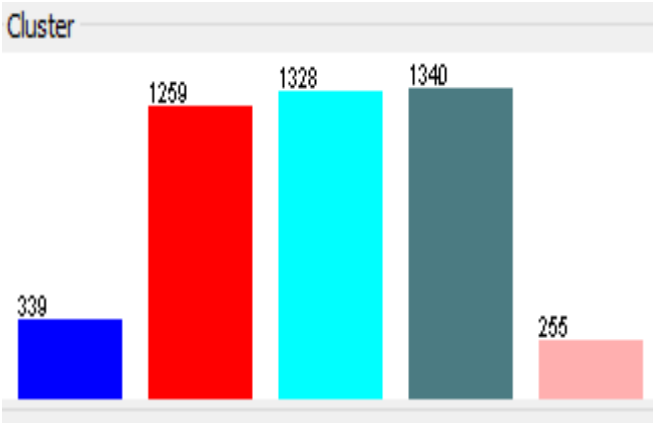


Fig.6: The graph for clusters vs Balance and Clusters Sizes

The Fig.6 shows Clusters vs Balance attribute and sizes of the clusters and it is observed that cluster-3 is the largest followed by cluster-2 and the smallest cluster is cluster-4.

TABLE-3: Performance comparisons between Classifiers

		Without Refinement		After Refinement			
S.No.	Parameters	J48	J48	Random forest	Random Tree	REP tree	Naïve Baye's
1	Number of Instances	4521	4521	4521	4521	4521	4521
2	Number of Attributes	18	18	18	18	18	18
3	Time taken to build the model in Sec	0.16	0.11	0.2	0.09	0.17	0.11
4	Time taken to test the model in Sec	0	0.02	0	0	0	0
4	Correctly Classified Instances	4204 (92.9883%)	4216 (93.3%)	4129 (91.3294%)	4521 (100%)	4225(93.4528 %)	3982 (88.0779%)
5	Incorrectly Classified Instances	317 (7.0117 %)	305 (6.7%)	392 (8.6706 %)	0 (0%)	296(6.5472 %)	539 (11.9221%)
6	Kappa statistic	0.5705	0.5976	0.4593	1	0.6177	0.469
7	Mean absolute error(MAE)	0.1182	0.1156	0.1273	0	0.1062	0.1458
8	***Root mean squared error (RMSE)***	0.2431	0.2404	0.2563	0	0.2304	0.3143
9	Relative absolute error %	57.934	56.6563	162.3908	0	52.0416%	71.4734
10	Root relative squared error %	76.1388	75.2946	80.2761	0	72.163 %	98.4351
13	Accuracy	0.93	0.932	0.913	1	0.935	0.88
14	Precision	0.99	0.99	0.983	1	0.987	0.9175
15	Recall	0.934	0.939	0.934	1	0.945	0.922

The formulae for performance metrics are shown below by the equations (H), (I) and (J)

Accuracy = (TP + TN) / (TP + TN + FP + FN) ---- (H)

Precision = TP / (TP + FP) --- (I)

Recall = TP / (TP + FN) --- (J)

It is observed from the Table-3 that the classifier accuracy was more for the refined clustered data set than without Refined Non-clustered data set. The performance of the Random Tree classifier is high when compared the performance of all other classifiers namely J48, Random Forest, Naïve Baye's, REP Tree and J48 classifiers (without refined). The classifier J48 performance is enhanced if the data set is refined which is shown in the above Table – 3 in terms of Number of Correctly classified instances, time to build the model, kappa statistic, Accuracy, Recall, Precision. So out of all the classifiers Random Tree has highest Performance.

V. CONCLUSIONS

The overall conclusions across all clusters are Aged people having secondary Education and retired having good balance amount and without loans when contacted through mobile with long call duration for explaining and clarifying their doubts will definitely prefer term deposit than other customers during October month to leisurely plan for

their future returns. The present or past campaign should take place minimum of 2 times with a huge gap of days between the previous and present campaigns. The divorced, Married with house loan and people at management level positions do not prefer term deposits in most of the cases. The classifier Random Tree has highest performance than all other classifiers namely J48, Random Forest, Naïve Baye's. Another interesting observation found is that performances of the classifiers are increasing when applied on refined clustered dataset than without Refined Non-Clustered data set. The performance of the classifier J48 decision tree on refined clustered data set is high than Non refined and Non clustered data set.

REFERENCES

1. LilianSing'oei and Jiayang Wang, "**Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing**", IJCSI International Journal of Computer Science Issues, Pg.No's:198-203, Vol. 10, Issue 2, No 2, March 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, www.IJCSI.org.
2. Sergio Moro, Paulo Cortez, Paulo Rita, "**A Data-Driven Approach to Predict the Success of Bank Telemarketing",** Preprint submitted to Elsevier 19 February 2014.
3. MuneebAsif, "**Predicting the Success of Bank Telemarketing using various Classification Algorithms"**
4. Q.R. Zhuang, Y.W. Yao, and O. Liu, "**Application of Data Mining in Term Deposit Marketing**" Proceedings of the International Multi Conference of Engineers and Computer Scientists 2018 Vol II IMECS 2018, March 14-16, 2018, Hong Kong, ISBN: 978-988-14048-8-6 ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online).
5. Charles X. Ling and Chenghui Li, "**Data Mining for Direct Marketing: Problems and Solutions"** Pg. No:73- 79, Copyright ©1998, American Association for Artificial Intelligence (www.aaai.org), KDD-98.
6. Dr. Md. Rashid Farooqi, NaiyarIqbal, "**Effectiveness of Data mining in Banking Industry: An empirical study",** Pg No:827-830, Volume 8, No. 5, May-June 2017 International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697**.**
7. Anthony Rahul Golden S, "**An Overview of Digitization in Indian Banking Sector**", Indo-Iranian Journal of Scientific Research (IIJSR) Volume 1, Issue 1, Pages 209-212, October-December 2017.
8. Dr. Agboola A. A., "**Electronic Payment Systems and Tele-banking Services in Nigeria",** Journal of Internet Bank Commerce, http://www.icommercecentral.com/, ISSN: 1204-5357.
9. Dr. SurajitGhoshDastidar, Dr. Rajib Kumar Das, (2018) "**Customers' motivation to adopt digital banking: A case study of HDFC Bank in Kolkata", 11th** International Conference on science, Technology and Management (ICSTM-18), 21st JAN 2018, Pg. No: 110-116, ISBN: 978-93-86171-94-8.
10. V Vimala, "**An Evaluative Study on Internet Banking Security among Selected Indian Bank Customers"** Amity Journal of Management Research, 1(1), (63–79), ©2016 ADMAA.
11. HosseinHassani, Xu Huang and Emmanuel Silva, "**Digitalization and Big Data Mining in Banking"**,Big data and cognitive computing, Pg No: 1-13, **2018**, 2, 18, MDPI, DOI: 10.3390/bdcc2030018, www.mdpi.com/journal/bdcc.
12. MerveMitik ; OzanKorkmaz ; Pinar Karagoz ; Ismail HakkiToroslu ; FerhatYucel, "**Data Mining Based Product Marketing Technique for Banking Products",** 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), DOI: 10.1109/ICDMW.2016.0085,ISSN: 2375-9259.
13. Bank Marketing Data Set https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (UCI Repository Data.
14. Indrajani, HarjantoPrabowo, Meyliana, "**Learning Fraud Detection from Big Data in Online Banking Transactions: A Systematic Literature Review"**,Journal of Telecommunication, Electronic and Computer Engineering, Pg. No: 127-131, Vol. 8 No. 3, ISSN: 2180 – 1843 e-ISSN: 2289-8131.
15. Israel González-Carrasco, Jose Luis Jiménez-Márquez, Jose Luis López-Cuadrado, Belén Ruiz-Mezcua "**Automatic detection of relationships between banking operations using machine learning",** Information Sciences 485 (2019) 319–346, https://doi.org/ 10.1016/j.ins.2019.02.030, ©2019 Elsevier Inc. www.elsevier.com/locate/ins.
16. Justice Asare-Frempong, Manoj Jayabalan, "**Predicting Customer Response to Bank Direct Telemarketing Campaign**", Pg.No:1-4, 2017 IEEE The International Conference on Engineering Technologies and Technopreneurship (ICE2T 2017).
17. Yogesh Sanjay Golecha, "**Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques"**,National College of Ireland.
18. OlatunjiApampa, "**Evaluation of Classification and Ensemble Algorithms for BankCustomer Marketing Response Prediction",** Journal of International Technology and Information Management, Pg. No: 85-100, Volume 25, Number 4 2016© International Information Management Association, Inc. 2016, ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy
19. TeemuVartiainen, "**Telemarketing Data Analysis and Predictive Modelling"** Tampere University Of Technology.
20. M. Purnachary, B. Srinivasa S P Kumar, HumeraShaziya, "**Performance Analysis of Baye's Classification Algorithms in WEKA Tool using Bank Marketing"**,International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Pg. No: 128-133, Vol. 5, Issue 2, February 2018. ISSN (Online) 2394-2320