ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

Churn Prediction in Financial Institutions Using Machine Learning

[1] Malik Mubasher Hassan, [2]Tabasum Mirza
[1] Dept. of ITE, BGSB University Rajouri (J&K)-India-185234
[2] Dept. of School Education Govt. of J&K (India)

Abstract: Customer churn prevention is one of the important components of CRM(Customer Relationship Management) in financial institutions like banks and predictive modeling of customer churn can help in preventing the churn from actually occurring thus saving banks from losses. In this research study we are presenting a comparative analysis of different popular machine learning algorithms for the challenging problem of churn prediction by cross validation on the basis of performance metrics like accuracy and kappa coefficients. Our results determined the Random Forest algorithm as the best possible classifier for prediction of customer churn in financial institutions with almost 86% accuracy in predictions.

Keywords: Churn predictions, financial institutions, and machine learning algorithms.

I. INTRODUCTION

Customers are the most important asset in the financial market and in the current scenario where a large number of financial institutions are working hard to survive in the global market there is high competition between the institutions, preventing customer attrition is one of the top challenges[1][2]. In financial institutions customers are directly linked to profits and losing a customer means loss of profit[3]. Acquisition of a new customer is a difficult task and requires much more effort and cost than retaining an existing customer and long tenure customers tend to produce more profits[4]. However churn cannot be completely avoided but if we can predict customers that may churn, efforts can be made to reverse the decision of the customer to churn by addressing issues which may lead to the decision. Improving quality of services, incentives and offers, loyalty cash, complaint redressal and other remedial strategies may help to reduce the actual churn from occurring[5][6][7][8]. Machine learning algorithms can be used to predict the customers who are at the risk of churn by analyzing customer demographics and other related data[9][10][11][12][13][14].

Although a lot of research has been done in this area and a number of techniques have been used by various researchers to predict customer churn in financial institutions there is lack of accuracy and an ideal model is yet to be formed that can work with large datasets and generate predictions with impoverished performance[15][16][17][18][19]. This paper uses various machine learning algorithms like Logistic Regression, SVM (Support Vector Machines), Random Forests, Naive Bayes Classification, CART (Classification and Regression trees) to predict customer churn in financial institutions.

2 .OBJECTIVES

a. To predict customer churn in financial institutions based on demographic and other factors using popular machine learning algorithms on random data samples.
b. To compare performance of different algorithms and determine the best possible solution among them.

3.METHODOLOGY

We have used a data set of 10000 bank customers out of which 8000 records were used for training the models and 2000 records for testing purpose .We used linear regression to check the degree of correlation of variables with customer churn. The framework for the customer churn prediction is given in figure 1.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

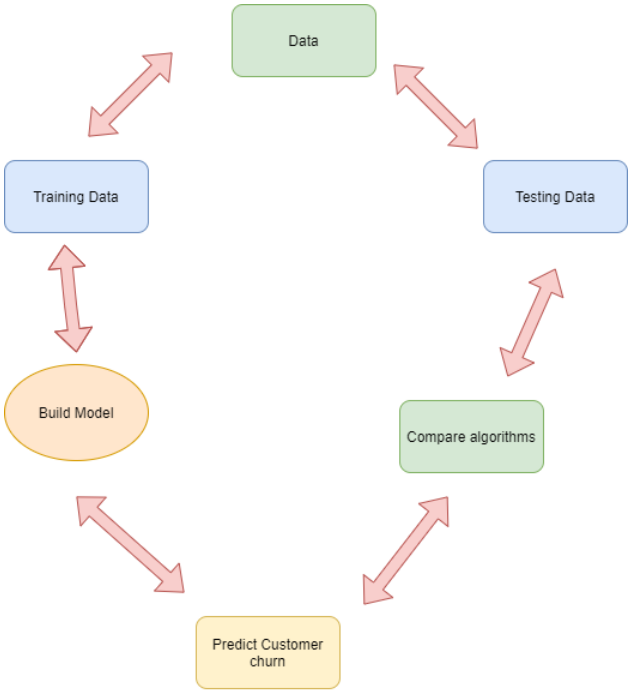


Fig 1. Research framework for customer churn prediction

We are using R-language for coding of algorithms, prediction and testing of model performance. R libraries used were e1071, CARET, naivebayes, rpart, random Forest, klaR, ggplot2.

Five machine learning algorithms were trained and evaluated on random samples of 10 independent variables to predict the customer churn. The dependent variable Exited is strongly correlated with these factor variables and can take two values 'Yes' or 'No'. The algorithms used include logistic regression, Random Forest, SVM (Support Vector Machine), CART (Classification and Regression Trees) and Naïve Bayes algorithm. The variables are depicted in the table 1.

The prediction results were compared using a confusion matrix for false negatives and false positives. Each algorithm is then evaluated on certain performance metrics and the one with the highest Kappa coefficient and accuracy is selected as the best fit model for churn prediction.

The predictive capability of the given algorithms is evaluated with the Kappa coefficient as shown below:

$$(K) = \frac{P^o - P^{\mathrm{h}}}{1 - P^{\mathrm{h}}}$$

Where:
P_o= Observed agreement among raters
P_h = Hypothetical probability of chance agreement
The strength of the kappa coefficients varies in the following manner: 0.01-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 substantial agreement; 0.81-1.00 almost perfect agreement[20][21][22][23].

Table 1 Variables used for churn prediction

Dependent Variable	**Value**
Exited	0-No,1-Yes
Independent Variables	**Value**
CreditScore	Can take valid numeric value
Geography	Can take valid Character Value
Gender	Male/Female
Age	Can take valid numeric value
Tenure	Can take valid numeric value
Balance	Can take valid numeric value
NumOfProducts	Can take valid numeric value
HasCrCard	0-No,1-Yes
IsActiveMember	0-No,1-Yes
EstimatedSalary	Can take valid numeric value

4. RESULTS

Linear Regression Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2637	-0.6540	-0.4482	-0.2626	3.0261

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

Table 2 Linear regression coefficients.

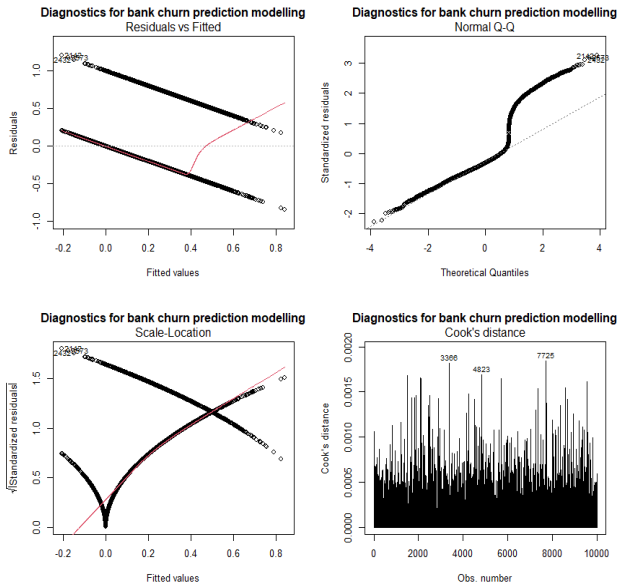
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.304e+00	2.753e-01	-12.001	< 2e-16 ***
CreditScore	-8.421e-04	3.151e-04	-2.673	0.00752 **
Geography Germany	7.859e-01	7.635e-02	10.293	< 2e-16 ***
Geography Spain	6.129e-02	7.862e-02	0.780	0.43560
Gender Male	- 5.379e-01	6.130e-02	-8.774	< 2e-16 ***
Age	7.518e-02	2.922e-03	25.730	< 2e-16 ***
Tenure	-1.958e-02	1.053e-02	-1.860	0.06293
Balance	2.461e-06	5.773e-07	4.262	2.02e-05 ***
NumOfProducts	-1.032e-01	5.358e-02	-1.927	0.05401
HasCrCard	-8.017e-02	6.656e-02	-1.205	0.22835
IsActiveMember	-1.100e+00	6.505e-02	-16.912	< 2e-16 ***
EstimatedSalary	2.124e-07	5.313e-07	0.400	0.68928

Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8050.6 on 7999 degrees of freedom
Residual deviance: 6784.0 on 7988 degrees of freedom
AIC: 6808

Fig 2. Linear regression diagnostics



ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

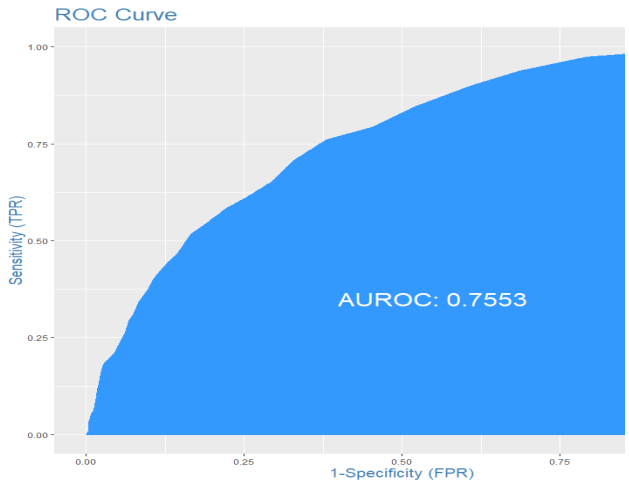


Fig 3. ROC curve for linear regression

Confusion Matrix

	Customer will Churn	**Customer will Not Churn**
Customer will Churn	True positive (__)	False negative (__)
Customer will Not Churn	False positive (__)	True negative (__)

Confusion matrix for SVM

p	0	1
0	7447	1587
1	516	450

Confusion matrix for Naive Bayes

p	0	1
0	7709	1437
1	254	600

Confusion matrix for CART

P	0	1
0	7781	182
1	1225	812

Confusion matrix for Random forest

P	0	1	Class. error
0	**7660**	**303**	0.03805099
1	**1077**	**960**	0.52871870

Accuracy

Accuracy is the proportion of the total number of predictions that were correct and can be calculated from the following equation:

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Fp+Tn+Fn}$$

Where, Tp= True Positives
Tn= True Negatives
Fp= False Positives
Fn= False Negatives[24][25][26].

Table 3 Comparison of algorithms on the basis of accuracy.

Accuracy

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
CART	0.7972028	0.8256186	0.8359169	0.8349662	0.8451161	0.8611389	0
GLM	0.7967968	0.8060000	0.8094044	0.8101014	0.8146111	0.8228228	0
RF	0.8361638	0.8558919	**0.8614309**	0.8610009	0.8660000	0.8761239	0
SVM	0.7960000	0.7960000	0.7962038	0.7963001	0.7967968	0.7970000	0
NB	0.7912088	0.8210000	0.8255869	0.8240691	0.8319580	0.8418418	0

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

Table 4 Comparison of algorithms on the basis of Kappa coefficient

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CART	0.2760874	0.3694224	0.3987965	0.3974947	0.4296595	0.4936183	0
GLM	0.1770839	0.2083629	0.2280815	0.2307033	0.2499799	0.2938740	0
RF	0.3445448	0.4430011	**0.4619756**	0.4647483	0.4917275	0.5372789	0
SVM	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0
NB	0.2218114	0.3258721	0.3577370	0.3500061	0.3780324	0.4175400	0

Fig. 4

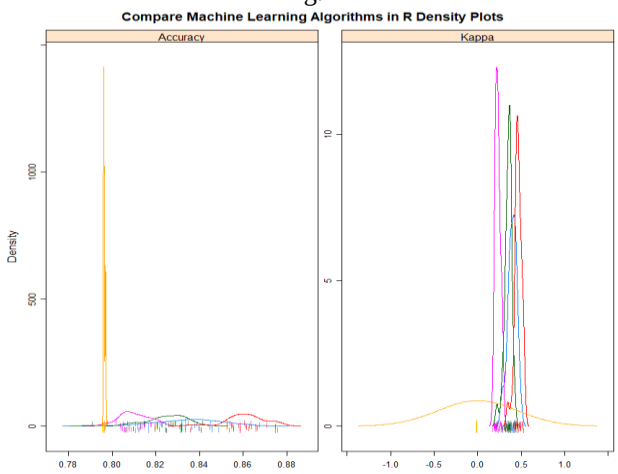


Fig. 5

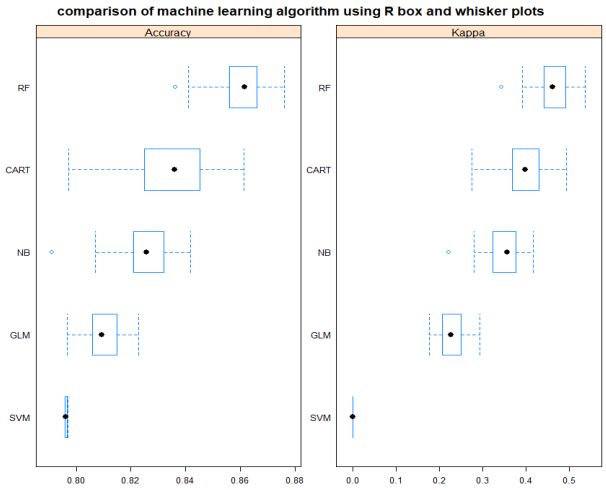


Fig. 6

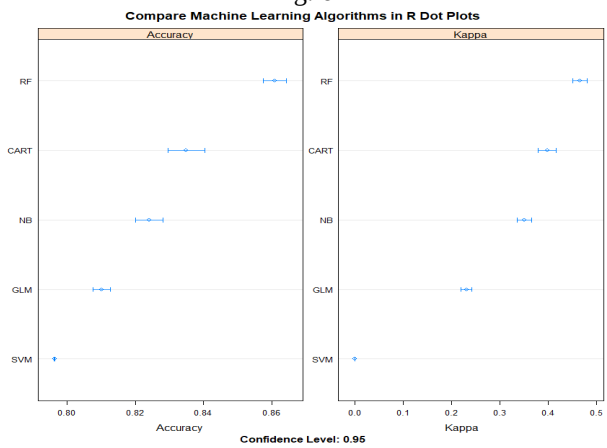
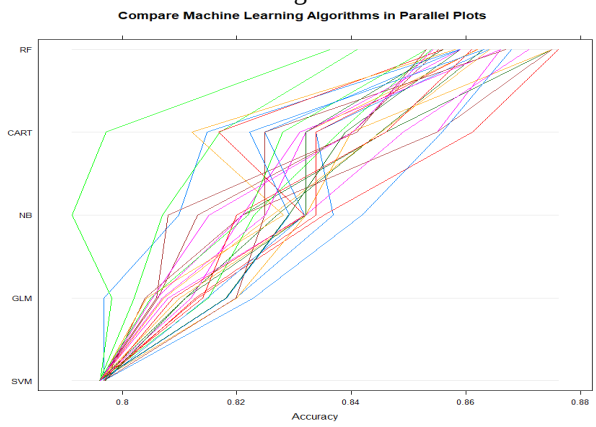


Fig. 7



DISCUSSION

When we compared churn prediction results of various algorithms like Logistic regression, SVM(Support Vector Machines),Random Forest, Naive Bayes algorithm and CART(Classification and Regression Trees) it was found that Random Forest algorithm (0.8614)produced most accurate results followed by CART. The Kappa coefficient was also highest for the Random forest algorithm followed by CART (0.4619). Thus Random Forest was established as the most efficient supervised machine learning algorithm for predicting customer churn

in financial institutions on our data set. In future more advanced techniques and algorithms can be explored to solve the problem with different datasets.

REFERENCES

[1] I. J. Chen and K. Popovich, "Understanding customer relationship management (CRM): People, process and technology," Bus. Process Manag. J., vol. 9, no. 5, pp. 672–688, 2003, doi: 10.1108/14637150310496758.
[2] A. Ghavami, "MA S T ER ' S TH E SI S The Impact of CRM The Impact of CRM," 2006.
[3] M. Tabasum, "IMPROVING CUSTOMER RELATIONSHIP MANAGEMENT (CRM)," no. 9, pp. 70–75, 2018.
[4] M. M. Hassan and T. Mirza, "Churn Prediction in Banking Sector using Bayesian Neural Networks," vol. 6, no. 12, pp. 3343–3346, 2018.
[5] D. Wadikar, "Customer churn prediction," Masters Diss. Technol. Univ. Dublin., 2020, doi: 10.21427/kpsz-x829.
[6] N. Derby, "Reducing Customer Attrition with Machine Learning for Financial Institutions," pp. 1769–2018, 2018.
[7] A. Bilal Zoric, "Predicting Customer Churn in Banking Industry using Neural Networks," Interdiscip. Descr. Complex Syst., vol. 14, no. 2, pp. 116–124, 2016, doi: 10.7906/indecs.14.2.1.
[8] F. Li, J. Lei, Y. Tian, S. Punyapatthanakul, and Y. J. Wang, "Model selection strategy for customer attrition risk prediction in retail banking," Conf. Res. Pract. Inf. Technol. Ser., vol. 121, pp. 119–124, 2010.
[9] B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on SVM model," Procedia Comput. Sci., vol. 31, pp. 423–430, 2014, doi: 10.1016/j.procs.2014.05.286.
[10] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simul. Model. Pract. Theory, vol. 55, no. 10, pp. 1–9, 2015, doi: 10.1016/j.simpat.2015.03.003.
[11] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," J. Mark. Res., vol. 43, no. 2, pp. 204–211, 2006, doi: 10.1509/jmkr.43.2.204.
[12] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 273–281, 2018, doi: 10.14569/IJACSA.2018.090238.
[13] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Syst. Appl., vol. 36, no. 3 PART 1, pp. 5445–5449, 2009, doi: 10.1016/j.eswa.2008.06.121.
[14] C.-F. Tsai and Y.-H. Lu, "Data Mining Techniques in Customer Churn Prediction," Recent Patents Comput. Sci., vol. 3, no. 1, pp. 28–32, 2010, doi: 10.2174/1874479611003010028.
[15] K. A. Amuda and A. B. Adeyemo, "Customers Churn Prediction in Financial Institution Using Artificial Neural Network," 2019, [Online]. Available: http://arxiv.org/abs/1912.11346.
[16] S. H. Iranmanesh, M. Hamid, M. Bastan, G. Hamed Shakouri, and M. M. Nasiri, "Customer churn prediction using artificial neural network: An analytical CRM application," Proc. Int. Conf. Ind. Eng. Oper. Manag., no. July, pp. 2214–2226, 2019.
[17] H. Sayed, M. A. Abdel-Fattah, and S. Kholief, "Predicting potential banking customer churn using Apache Spark ML and MLlib packages: A comparative study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 11, pp. 674–677, 2018, doi: 10.14569/ijacsa.2018.091196.
[18] C. F. Tsai and Y. H. Lu, "Customer churn prediction by hybrid neural networks," Expert Syst. Appl., vol. 36, no. 10, pp. 12547–12553, 2009, doi: 10.1016/j.eswa.2009.05.032.
[19] A. T. Jahromi, "Predicting customer churn in telecommunications service providers," p. 88, 2009, [Online]. Available: http://epubl.ltu.se/1653-0187/2009/052/.
[20] A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," J. Basic Appl. Sci., vol. 13, pp. 459–465, 2017, doi: 10.6000/1927-5129.2017.13.76.
[21] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," Int. J. Comput. Trends Technol., vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.
[22] A. J. Viera and J. M. Garrett, "Understanding Interobserver Agreement : The Kappa Statistic," no. May, pp. 360–363, 2005.
[23] J. González Alonso and M. Pazmiño Santacruz, "Cálculo e interpretación del Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos posibles escalas tipo Likert," Rev.

Publicando, vol. 2, no. 2, pp. 62–7, 2015.
[24] K. Boggs, Liam, "Performance Measures for Machine Learning 1," pp. 1–32, 2017.
[25] K. J. Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients," 2015, [Online]. Available: http://arxiv.org/abs/1504.04646.
[26] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," Bioinformatics, vol. 16, no. 5, pp. 412–424, 2000, doi: 10.1093/bioinformatics/16.5.412.