International Journal for Innovative
Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 06th Feb 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=ISSUE-02

DOI: 10.48047/IJIEMR/V12/ISSUE 02/20

Title **Classification of Customer Churn in Financial Sectors using Machine Learning Algorithms**

Volume 12, Issue 02, Pages: 127-135

Paper Authors
Pavan K, Purna N, Pavan M, Dravens T , Kanya Kumari L

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Classification of Customer Churn in Financial Sectors using Machine Learning Algorithms

[1]Pavan K, [2]Purna N, [3]Pavan M, [4]Dravens T , [5]Kanya Kumari L
[1,2,3,4] Department of Information Technology, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh.
[5]Assistant Professor, Department of Information Technology, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh.
[1]pavankothapalli47@gmail.com,[2]purnanelluri27@gmail.com
[3]pavanmatta13@gmail.com,[4]dravenstirumalasetti73@gmail.com, [5]kanyabtech@yahoo.com.

Abstract
Customer Churn is also called customer attrition. The main motivation behind predicting churn is to categorize the customers into loyal and disloyal. The loyal customers are Non-churn (NC) and the disloyal customers are Churn (C) or the customers who have a high chance of leaving the service are churns and those who regularly use the service are non-churn. Predicting the churn helps the organization to identify the customer who could leave before they do so and be proactive with communication and actions to stop that. By doing so we can prevent the loss of customers of the organization on a large scale. To predict the churn, we are using several algorithms like Logistic regression (LR), Random Forest (RF), Support Vector Machine (SVM) K-Nearest Neighbor(KNN), Decision Trees(DT), and Light Gradient Boosting Machine (LGBM). The algorithms are tuned using Hyper parameter tuning. The best results are obtained for LGBM.

Keywords: Logistic regression , Random Forest , Support Vector Machine , K-Nearest Neighbor **,** Decision Tree, and Light gradient-boosting machine.

Introduction

For any organization in the Financial sector, the main goal is to be profitable. To be profitable they must focus on gaining new customers and retaining existing customers. For gaining new customers they advertise their features, add new plans and improve the experience but for sustaining the existing user they need to have a customer-brand relationship other than the banking services. The Customer Churn is introduced for this purpose. Customer Churn is also called customer attrition. The main motivation behind predicting churn is to categorize the customers into loyal and disloyal. The loyal customers are Non-churn (NC) and the disloyal customers are Churn (C) or the customers who have a high chance of leaving the service are churns and those who regularly use the service are non-churn [1]. Predicting the churn helps the organization to identify the customer who could leave before they do so and be proactive with communication and actions to stop that [2]. By doing so we can prevent the loss of customers of the organization on a large scale. To predict the churn, we can use several algorithms like Logistic regression (LR), Random Forest (RF), light gradient-boosting machine (LightGBM), Support Vector Machine (SVM), Decision Tree, k-nearest neighbors algorithm (KNN). The algorithms are tuned using Hyperparameter tuning. Their accuracies are calculated and the algorithms with the best accuracies are chosen and ensembled to create a better-performing model. Then a model is developed which can predict whether a customer is a churn.

In the recent decades the bank, credit union, or other financial institution (including a brokerage firm) have seen a enormous growth due to the rising demand in checking and savings accounts, loan and mortgage services, wealth management, providing Credit and Debit Cards, Overdraft services. They developed a separate management system called as Customer Relationship Management for gaining new customers and for providing services. Gaining new customers is by regular promotion and innovating new ideas and adding new experiences to their services. But retaining the customers is a different process. The first and foremost step is to identify the customers who are most likely to leave. It is called churn. Customer churn is one of the major obstacles for a financial intuition. Prediction of the customer churn can benefit the organization in many ways. We use ensembled algorithms to predict the customer churn and to develop a model. As Customer retention costs less when compared to gaining new customers to a company. By analysing various studies we can say that gaining a new customer costs nearly four to twenty five times more of retaining a customer who is already a previous user of that company [21].

To overcome such problems, companies need to use highly accurate methods to identify customer churn so that they can avoid the risk of future churning [22]. Just by minimizing the percentage of churn by 5%, the profits of a company could get increased up to 75% [23]. We can see that retention of customers has more impact on companies' profit than gaining new customers. The customers who churn the company are divided into two categories. They are Voluntary Churners and Non-voluntary Churners. Identifying non-voluntary churners are easy when compared to voluntary churners. Because non-voluntary churners are those churned by the company itself due to random reasons like misuse of the service or not paying to the service in the right time. Voluntary churners are difficult to identify as the customers consciously terminate the service from the company. Voluntary churning can be categorised into two, Incidental churning and Deliberate churning. Incidental churning is a churning where the services is terminates due to changes in situations where the customer could not continue with the service. For example, the financial situation of the customers might not allow them to continue with the service and results in churning. The geographical location of the customer can also become a reason for incidental churning if the company might not be able to provide the service in that particular location.

Customer churn can be managed by two approaches. These two approaches are reactive and proactive. When the company receives a request from the customer to cancel the service then this approach is said to be a reactive approach. By this, the company offers an incentive to the customer to continue the services provided by them. The other approach that companies adapt to support the customers is the proactive approach. Using the proactive approach, most of the companies predict the customer churn by applying various accurate machine learning algorithms. Here, the company strives to identify the customers who are thinking of churning before they stop the services from the company. Through this, companies come up with special incentives to keep the customers from churning away from the services of the company [24].

Literature Survey

Implementation of Machine learning has become one of the important components of an organization because of the fact that it gives the enterprises, view of trends in customer's behaviour and helps to develop the business operational patterns, as well as supports the development of new products. Many of today's top leading companies make use of the machine learning method as their central part of the operations. Machine learning has taken a significant role in the competitive differentiation of different organizations.

By analyzing various studies we can say that gaining a new customer costs nearly four to twenty-five times more than retaining a customer who is already a previous user of that company. Just by minimizing the percentage of churn by 5%, the profits of a company could get increased up to 75%. The results of experiments showed that the proposed systems for churn prediction perform with an accuracy of 82% using the Random Forest algorithm[1].

This paper proposes an accurate way to predict customer churn using LSTM model and the data is preprocessed using SMOTE technique. Thus, the system is more useful for 5 organizations to find the customers with more chances to become churn. The results of experiments showed that the proposed systems for churn prediction performs with an accuracy of 74% and which is much better than the system without SMOTE technique [2]. The experimental results show that the accuracy rate of the model has reached 80%, and the AUC value is more than 78%. The model can be used to predict whether the user may be lost in the future, reserve enough time for user retention activities, and provide a lot of valuable information to help marketing personnel to formulate a feasible user retention scheme, which has a wide range of industry application prospects [16]. From this experiment, it could be inferred the Random Forest model works best for this particular use case with a prediction accuracy of 91% on the testing data before grid search [17]. The experimentation was conducted on the churn modeling dataset from Kaggle. The results are compared to find an appropriate model with higher precision and predictability. As a result, the use of the Random Forest model after oversampling is better compared to other models in terms of accuracy [18]. The summary of Literature is Represented in Table1.

Table 1. Literature Survey

Reference No	Title	Method	Result
[1]	Comparison of Machine learning algorithms on Predicting Churn within Music streaming service	Random Forest Algorithm.	Achieved 82% Accuracy
[2]	Customer churn prediction in telecom using machine learning in big data platform	XG Boost, Random Forest, Decision Tree.	Achieved 86% Accuracy Using XG Boost
[3]	A Survey on Churn Analysis in Various Business Domains	XG Boost, Random Forest, Decision Tree, KNN, SVM, LGBM.	Telecommunications, Marketing, Human Resources, Insurance, Commerce, Music streaming service
[4]	Prediction of Type 2 Diabetes using Machine Learning Classification Methods.	Logistic Regression , KNN, SVM, Naive Bayes ,Decision Tree, Random Forest.	Achieved 86% Accuracy Using Random Forest.
[5]	An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach.	Decision Tree, Logistic Regression, SVM.	Senti Churn model proved its efficiency with 88% Accuracy.
[6]	Methods for churn prediction in the pre paid mobile telecommunication industry.	Neural Networks, SVM, Naïve Bayes.	Achieved 85.44% Accuracy Using SVM.
[7]	Estimating customer churn	Logistic Regression	Achieved 82% Accuracy

	under competing risks.	, KNN, SVM.	Using Logistic Regression .
[8]	Predicting Customer Loyalty in Banking Sector with Mixed Ensemble Model and Hybrid Model.	XGB, Light GBM, Neural Networks.	Achieved 88% Accuracy Using Light GBM.
[9]	Effect Improved for High-Dimensional and Unbalanced Data Anomaly Detection Model Based on KNN SMOTE-LSTM	KNN, SMOTE,LSTM.	Achieved 85% Accuracy Using LSTM.
[10]	Machine-learning techniques for customer retention: A14 comparative study.	Logistic Regression , KNN, SVM, Naive Bayes, Decision Tree, Random Forest, ADA BOOST.	Achieved 88% Accuracy Using Random Forest.
[11]	Applying data mining to customer churn prediction in an Internet Service Provider	KNN, Random Forest, Logistic Regression.	Achieved 89% Accuracy Using Random Forest.
[12]	Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron(MLP)& quot.	Logistic Regression , KNN, SVM.	Achieved 80% Accuracy Using KNN.
[13]	Designing of customer and employee churn prediction model based on data mining method and neural predictor	KNN, Random Forest, Logistic Regression, Decision Tree, ANN.	Achieved 82% Accuracy Using Decision Tree .
[14]	Activation Functions and Training Algorithms for Deep Neural Network.	XGB, DNN, ANN.	Achieved 89% Accuracy Using DNN .
[15]	Bank customer retention prediction and customer ranking based on deep neural networks & quot.	Random Forest, Logistic Regression, Decision Tree, ANN.	Achieved 83% Accuracy Using ANN.
[17]	Analysis and prediction of bank user chum based on ensemble	LGBM,CAT BOOST, Random Forest.	Achieved 81% Accuracy Using Random

	learning algorithm		Forest.
[18]	Machine Learning Based Telecom-Customer Churn Prediction.	Radge Classifier, Random Forest, XGB	Achieved 79% Accuracy Using Random Forest.
[19]	Machine Learning Based Customer Churn Prediction In Banking.	KNN, SVM, Decision Tree, Random Forest.	Achieved 82% Accuracy Using Random Forest.
[20]	Bank Customer Churn Prediction Based on Support Vector Machine	SVM, C4.5, Logistic Regression.	Achieved 69% Accuracy SVM.

Dataset

The data sets used in the research papers that are analysed for churn prediction models are taken from the website Kaggle[25]. Kaggle is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve data science challenges. The data collected to work for research, is mainly sample data that is in a form of CSV file that consists 10000 unique customers. The variables used are

(1) Customer_id - unused variable.
(2) Credit_score - used as input.
(3) Country - used as input.
(4) Gender - used as input.
(5) Age - used as input.
(6) Tenure - used as input.
(7) Balance - used as input.
(8) Products number - used as input.
(9) Credit_card - used as input.
(10) Active_member - used as input.
(11) Estimated_salary - used as input.
(12) Churn - used as the target. 1 if the client has left the bank during some period or 0 if he/she has not.

Proposed Methodology

The proposed Methodology contains seven phases : Data collection, Exploratory data analysis, Feature Engineering , Train and Test data split, Modelling, Classification. The diagrammatic representation is represented in the following Figure 1. Supervised learning is part of machine learning which is most popular among classification problems. Supervised learning approaches the data by mapping the input data into desired outputs. The learning approach is to learn from the given set of data i.e initially the model is trained with training set. Later, the model is tested with test data in the testing phase. Supervised learning primary goal is to get the computer learn our classification system and find out better insights from the test data given. Few approaches in supervised learning is as follows, Linear classifiers, KNN, support vector classifier, K- means, decision trees etc. The algorithms used are KNN, SVM, Decision Tree, Random forest, Logistic Regression, LightGBM machine learning algorithms.

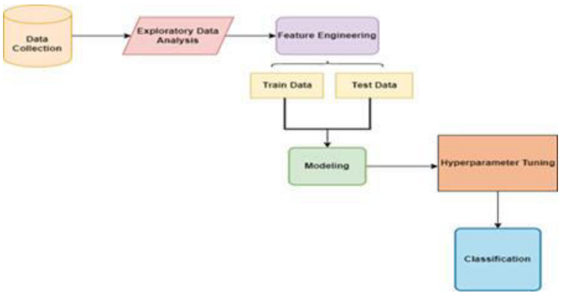


Figure 1.Proposed Methodology

K-Nearest Neighbor (KNN): The KNN method is one of the easiest and most efficient non-parametric ways of classification, based on supervised learning [13]. KNN works by identifying the k nearest samples from an existing dataset and when a new unknown sample appears, classify the new sample in the most similar class. That is, the classification algorithm determines the test sample group by the k training samples that are the nearest neighbors to the test sample and assign it to the class with the highest likelihood.

Support Vector Machine (SVM): It is possibly the most notable Supervised Learning computation, which is utilized for Classification comparably as Regression issues. Regardless, generally, it is utilized for Classification issues in Machine Learning. The objective of the SVM calculation is to make the best line or choice breaking point that canisolate n-layered space into classes so we can undoubtedly put the new part in the right course of action later on . This most ideal choice breaking point is known as a hyperplane.

Decision Tree (DT): A decision tree is a procedure that slices a collection of data into various branch-like segments [20]. A tree of decisions is easy to read. This advantage makes explanations for the model simple. While another algorithm (like a neural network) can generate a much more accurate model in a given scenario, a decision tree could be trained to predict the neural network's predictions, thus opening up the neural network's"black box". Another benefit is that, in the correlation between the target variables and the predictor variables it can model a high degree of nonlinearity. A decision tree is composed of two major strategies [21]; Tree creation and Classification.

Random Forest (RF): Breiman [22] presented RF as an ensemble classifier for tree learners. The method employs several decision trees so that each tree relies on the values of an individually selected random vector with the same distribution for all trees. Right choice for the tendency of decision trees to overfit their training collection. In short, Random forests are actually a way to combine many deep decision trees which are learned on various sections of the same dataset with the target of decreasing the variance. The real advantage of using RF is it comes with quite high dimensional data, with no need to perform dimensionality reduction and feature selection. The training rate is also higher and eases to use in parallel models.

Logistic Regression (LR): Logistic regression which is quite similar to linear regression helps in finding discrete outcomes. It is tend to be less likely in over fitting which is important while working with huge data sets in churn prediction. It helps in relating coefficients and predict outcomes of dependable variables. It is a class estimation model which uses a single estimator to build a logistic regression model. It usually defines a boundary between classes so as to state the probabilities of the classes depending on the distance of the boundary. There exists two extremes (0 & 1) which help in defining the probabilities and grew larger when the data set is big. It works well in predicting categorical value and continuous values.

Light Gradient Boosting Machine (LGBM): LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage. It uses two novel techniques: **Gradient-based One Side Sampling** and **Exclusive Feature Bundling (EFB)** which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking classification and many other machine learning tasks.

Experimental Results

To evaluate the performance of algorithms, we used different metrics. Most of them are based on the confusion matrix. Confusion matrix is a tabular representation of a classification model performance on the test set, which consists of four parameters: true positive, false positive, true negative, and false negative. Accuracy is the best measurement in determining the best machine learning algorithms. It can be defined as the total number of correct predictions of data out of total data in the test data set. Usually accuracy is the best performing metrics while comparing classification models.

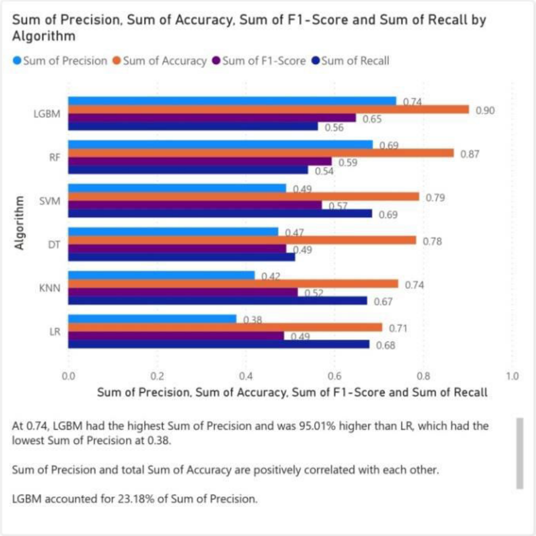
$$\text{Accuracy} = \frac{\text{Total number of correct predictions}}{\text{Total number of predictions}}$$

In classification algorithms the predictions quality can be measured by using classification report. Classification report consists of four main metrics. All the four metrics are calculated based on true positive, false positive, true negative and false negatives. The four metrics are Precision, Recall, F1-score and Accuracy. Recall answers the question How many variables belonging to the reference Class is correctly identified? F1-Score is a metric which is a harmonic mean of precision and recall. The metric is used to measure an overall model performance. The performance measures are represented in Table 2

Table 2.Performance measures of different classifiers

Algorithm	Precision	Recall	F1-Score	Accuracy
KNN	0.420691	0.674081	0.517820	0.744133
SVM	0.491419	0.685224	0.572237	0.791333
DT	0.473790	0.511771	0.491784	0.784400
RF	0.686758	0.540921	0.594289	0.869333
LR	0.379344	0.679299	0.486794	0.708133
LGBM	0.739747	0.563191	0.648862	0.903733

When the preprocessing of the data has been completed, the data will be in operational form. And the features which are obtained after preprocessing are taken for the remaining study. Among that, 80% of the data will be used for training and the remaining 20% will use for testing as random. The classifiers will be used alone and along with the specified feature selection methods. Different classifiers are used such as KNN, SVM, Logistic Regression, Decision Tree, Random Forest, and LGBM. All these algorithms are tested and the accuracies are calculated. W found the highest accuracy is achieved by the light gradient-boosting machine algorithm. The results graphs are denoted in Figure 2.



At 0.74, LGBM had the highest Sum of Precision and was 95.01% higher than LR, which had the lowest Sum of Precision at 0.38.

Sum of Precision and total Sum of Accuracy are positively correlated with each other.

LGBM accounted for 23.18% of Sum of Precision.

Figure 2. Graphical representation of performance measures

Conclusion

While the banking sector is considered, like any other organization, customer engagement has become one of the primary concerns. To resolve this crisis, banks need to identify customer churn possibilities as quickly as possible. There are various studies ongoing in banking churn prediction. Different entities measure the churn rate of customers in various ways using different bits of data or information. The need for a system that can forecast the client churning in banking in a generalized way in the early stages is really important. The system needs to work with fixed and potential data sources that are independent of any service provider. And also the model must be in a form that; can use minimal information and can give maximum throughput for the prediction. This study focuses to fulfill these needs. The purpose of this study is to build the most appropriate model to predict client churn in a Bank in the early stages. The model examined KNN, SVM, DT, LGBM, RF, and

LR classifiers under different conditions for this study. A better result is achieved when using the LGBM classifier.

References

[1] S. De, P. Prabu and J. Paulose, "Effective ml techniques to predict customer churn", 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 895-902, 2021.

[2] LSTM Model to Predict Customer Churn in Banking Sector with SMOTE Data Preprocessing Jesmi Latheef; Vineetha S; 2021.

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, vol. 6, no. 1,pp. 1–24, 2019.

[4] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, "A survey on churn analysing various business domains," IEEE Access, vol. 8, pp. 220 816–220 839, 2020.

[5] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods", Procedia Computer Science, vol. 167, pp. 706-716, 2020.

[6] L. Almuqren, F. S. Alrayes and A. I. Cristea, "An empirical study on customer churn behaviours prediction using arabic twitter mining approach", Future Internet, vol. 13, no. 7, pp. 175, 2021. Show in Context Google Scholar.

[7] I. Brânduşoiu, G. Toderean, and H. Beleiu, "Methods for churn prediction in thepre-paid mobile telecommunications industry," in 2016 International Conferenceon Communications (COMM), 2016, pp. 97–100.

[8] P. Routh, A. Roy and J. Meyer, "Estimating customer churn under competing risks", Journal of the Operational Research Society, vol. 72, pp. 1138-1155, 2021.

[9] Jesmi, Vineetha S, "Predicting Customer Loyalty in Banking Sector with Mixed Ensemble Model and Hybrid Model", Springer (2020) (Communicated).

[10] Bao, F., Wu, Y., Li, Z., Li, Y., Liu, L., & Chen, G. (2020). Effect Improved for High-Dimensional and Unbalanced Data Anomaly Detection Model Based on KNN-SMOTE-LSTM.Complexity, 2020.

[11] Sabbeh, Sahar F, "Machine-learning techniques for customer retention: A14 comparative study", International Journal of Advanced Computer Science and Applications 9.2 (2018).

[12] Afaq Alam Khan, Sanjay Jamwal, and M.M.Sepehri, " Applying data mining to customer churn prediction in an Internet Service Provider", International Journal of Computer Applications Volume 9No.7, 2010.

[13] S. Bharadwaj, B. S. Anil, A. Pahargarh, A. Pahargarh, P. S. Gowra and S. Kumar, "Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron(MLP)& quot;, Proc. 2nd Int. Conf. Green Comput. Internet Things ICGCIoT 2018, pp. 436-438, 2018.

[14] S. H. Dolatabadi and F. Keynia, " Designing of customer and employee churn prediction model based on data mining method and neural predictor", 2nd Int. Conf. Comput. Commun. Syst. ICCCS 2017, pp. 74-77, 2017.

[15] G. Khanvilkar and D. Vora, Activation Functions and Training Algorithms for Deep Neural Network, no. 4, pp. 98-104, 2018.

[16] A. P. Jagadeesan, " Bank customer retention prediction and customer ranking based on deep neural networks & quot;, Int. J. Sci. Dev. Res., 2020.

[17] Analysis and prediction of bank user chum based on ensemble learning algorithm Yihui Deng; Dingzhao *Li;* Lvqing Yang*;* Jintao *Tang;* Jiangsheng Zhao; 2021

[18] Machine Learning Based Telecom-Customer Churn Prediction Pushkar Bhuse; Aayushi Gandhi; Parth Meswani; Riya Muni; Neha Katre; *2020*

[19] Manas Rahman; V Kumar, Machine Learning Based Customer Churn Prediction In Banking; 2020.

[20] Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example Zhao Jing; Dang Xing-hua; 2008.

[21] S.F.Bischof, T.M.Boettger, and T. Rudolph, "Curated subscription commerce: A theoretical conceptualization," Journal of Retailing and ConsumerServices, vol. 54, p. 101822, 2020.

[22] A. Deligiannis and C. Argyriou, "Designing a real-time data-driven customer churn risk indicator for subscription commerce." International

Journal of Information Engineering & Electronic Business, vol. 12, no. 4, 2020.

[23] S. Khodabandehlou and M. Z. Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," Journal of Systems and Information Technology, 2017.

[24] A. Saran Kumar and D. Chandrakala, "A survey on customer churn prediction using machine learning techniques," International Journal of Computer Appli- cations, vol. 975, p. 8887, 2016.

[25]https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn ataset?resource=download.