Customer Churn Prediction Using Machine Learning Approaches

Conference Paper · January 2023

DOI: 10.1109/ICECONF57129.2023.10083813

Srinivasan Rajendran

National University of Singapore

Rajeswari Devarajan

SRM Institute of Science and Technology

2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) | 979-8-3503-3436-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICECONF57129.2023.10083813

Customer Churn Prediction Using Machine Learning Approaches

R. Srinivasan[1,*], D. Rajeswari[2] and G. Elangovan[2]

[1]Department of Computing Technologies SRM Institute of Science and Technology Kattankulathur-603 203, India
[2]Department of Data Science and Business Systems SRM Institute of Science and Technology, Kattankulathur-603 203, India
*(*Corresponding Author)*
E-mail:srinirvs89@gmail.com drajiit@gmail.com Elangovg2@srmist.edu.in

Abstract-Customer Churn (CC) is a major issue and important concerns for large organizations and businesses alike. Telecom industries are attempting to improve methods to predict possible customer churn due to the immediate impact on revenue, particularly in the telecom sector. This paper discusses the various ML algorithms used to construct the churn model that helps telecom operators to predict customers who are likely to churn. The experimental results are compared to predict the best model among various techniques. As a result, the use of the Random Forest combined with SMOTE-ENN outperforms best result than other in terms of F1-score. According to our analysis, the maximum prediction is 95 percent based on F1-score.

Keywords—Class Imbalance; Machine Learning; Customer Churn; Entity Extraction

I.Introduction

Churn is a term that refers to the number of customers that an organization/company loses over a specific time period. CC is an appreciable heed in service section with high fierce services. Predicting customers who will leave the company early can be a large revenue source. CC dataset is used to examine the marketing tendency of customers from the large databases. One way to think about customer attrition is as a churn rate, it is the percentage of consumers that discontinue using a service within a specified time frame.

On the Chinese mainland, three mobile companies have been awarded licenses to operate as wireless communication providers. The pace of yearly growth of new mobile subscribers dropped dramatically from more than 10 % during the years 2009 to 2013 to 4.7 percent in 2014. Similar to the situation in a great number of other nations, the mobile communication industry in China is reaching its saturation point and becoming increasingly competitive. Subscriber-based service models, which have a contractual client base, frequently utilize this metric to assess their financial viability. Telecommunications is a major industry in developed nations. Technical improvements and more supervisors boost competition. Using a new dataset, the machine learning (ML) model impact on churn will be tested. The contributions present CC prediction model is

1. Incorporation of various preprocessing techniques along with SMOTE-ENN normalize the data
2. Applying different classification techniques to predict the suitable model

Section 2 gives the detailed review of churn dataset and ML techniques. Section 3 elaborates the prediction model used in CC dataset. Section 4 provides the results and the error analysis.

II.Literature Survey

M.A.H. Farquad [1] present an avenue to overwhelm the fault of general SVM that creates a black box model. The author constructed an approach that divides into three phases. In the 1st phase, SVM Recursive Feature Elimination (SVM-RFE) is hired to decrease the feature set. During 2nd phase, dataset with minimal features are extracted and then apply the SVM techniques to do the classification. In the last phase, rules are extracted manually. After extracting the rules, Naive Bayes is combined with Decision tree and produce the results.

The dataset used for this research work credit card details which is extremely unstable with 93.24% loyal and 6.76% churned customers [2]. The experimental showed that the model does not expandable to large datasets. The authors developed a hybrid approaches to predict CC in a telecom CRM dataset. The authors created a two different model named as “Dual-ANN” and “SOM+ANN”. These models are combining back-propagation with neural networks and self-arranging maps (SOM) with neural networks.

Dual-ANN are used to remove the odd data using data depletion method. The output of the Dual-ANN model is considered as an input to the SOM+ANN. Later, the performance of these models are evaluated using three testing strategies. The first testing method are “one general testing set” and rest of them are “fuzzy testing strategy”. The outcome of the hybrid model proves the better performance than the single baseline neural network model. In addition, Dual-ANN performance is more significant than SOM+ANN.

WouterVerbeke and his team suggest using Ant-Miner and ALBA to build accurate and comprehensible CC prediction model [4]. The author used a mining tool (Ant Miner) based on Ant Colony Optimization (ACO) that includes domain knowledge with modest monotonicity constraints. Ant-Miner has high accuracy, readable models, and intuitive predictive models. Ant-Miner+ produces less sensitive rule-sets, but it incorporates domain knowledge and produces smaller, more understandable rule-sets than C4.5. RIPPER produces small, understandable rule-sets but unintuitive models that breach domain knowledge.

Ning Lu [5] suggests using boosting algorithms to improve customer churn prediction models in which customers are clustered based on the boosting algorithm's weight. A high-risk customer cluster was found. Logistic regression (LR) is used as a learner, and each cluster has a churn prediction model. The results showed that boosting algorithm separates churn data better than a single logistic regression model. H. Karamollaoğlu et al., had conducted the comparative analysis to predict the good f1-score in various ML techniques [6]. The compassion had on Multilayer sensors, Logistic Regression, AdaBoost, Decision Tree, etc. Finally, the author concluded the best result obtained from the random classifier without applying any data augmentation methods

The author anticipated a CC prediction technique based on SVM model [7] and used arbitrarily sampling to improve SVM model by considering data imbalance. A SVM builds a hyper-plane in a high- or infinite-dimensional space for categorizing. Arbitrary sampling can change data spread to reduce dataset imbalance. Fewer churners cause dataset imbalance.

Ssu-Han Chen [8] established an interesting mechanism on the gamma Cumulative SUM (CUSUM) chart to monitor the Inter Arrival Times (IAT) of customer using a limited mixture model for the reference value and decision interval of the chart with ranked Bayesian model to capture different customers. Recently, a parallel time interval variable to IAT, tracks recent login behavior. The graphical interface is an added benefit of this research work. The results showed that gamma CUSUM has a higher accuracy rate (ACC) than exponential CUSUM and a longer Average Time to Signal (ATS). Rotation Forest and Rot boost were recommended by Koen W. De Bock [9] for churn prediction. An ensemble classifier fuses the outputs of several member classifiers. Rotation Forests require feature extraction to train base classifiers. Combining Rotation Forest and AdaBoost [10]. Rotation Forests won. Rot Boost improves AUC and top decile lift precision over Rotation Forests. On Rot Boost and Rotation Forest classification presentations, they compare PCA, ICA, and SRP.

The author used an actual data to forecast client attrition. The author incorporated the new model named as “impact learning” that’s derived from CNN. This helps to improve the performance in client attrition. The experimental result helps to improve the outcome compared with ANN and Logistic regression [11]. Koen W. De Bock [12] developed a hybrid technique named as “generalized additive models (GAMs)” for predicting the CC dataset. This hybrid approach combines the bagging concept with Random Subspace method. This method extended the semi-parametric approach to generalize the feature score. GAMensPlus gives a good performance compared to logistic regression and GAM. Logistic regression is most preferable algorithm in the prediction process [13].Ning et al. [14] investigated the CC prediction and suggested boosting method to improve the performance. The author suggested splitting customers into two clusters based on the boosting algorithm's weight. The suggested model fits an Implementation Zone where high-churn customers can be harrassed for retention actions. The author clearly identifies the reason for the customer switching between multiple service providers in the telecommunication industry. Researchers offered many actions to prevent porting in the telecommunication industries. This research work presents a survey about different steps to avoid CC [15].

The goal of this paper is to explore ML techniques used by experts in recent years [16]. For a better understanding of the field, the paper summarizes the prediction methods performance applied to the different datasets. The analysis helps to find the hybrid and ensemble methods have helped improve model performance in many ways, but it is important to have clear guidelines on how to measure how well a model works. V.Geetha et al., analyzed the feature extracted in the CC prediction process in the state-of-art methods [17]. The characteristics of the existing techniques are explored with the help of distance zone methods. Though the existing system produces good accuracy of 84%, it takes more time to train the model. The time complexity of the process is very high in the distance zone methods. The author proposed a ML approach that's helps to increase the accuracy and takes less time to evaluate the model compared to the existing approaches. The efficiency of the proposed method guarantees that telecom sector provides the right services to the non-churning customers. After analyzing the existing approaches, CC dataset suffers with the data imbalance problem. Existing approaches uses different augmentation methods are used to normalize the prediction process. As far my knowledge, SMOTE method helps to normalize the imbalanced data and produces good result compared to other techniques.

III. Proposed Methodology

A. Dataset Details

The dataset is obtained directly from the Kaggle. The dataset consist of 7043 customers and each column consist of 21 features. The dataset explains the customer id, sign up details, customer account information, and demographic information about the customers. The dataset has to be preprocessed properly before applying the supervised classification techniques. The new features can be created from existing nature from the recurrent usage of peoples. These features are necessary to determine the usage of customer in advance and it is should be much needed information for the model.

B. Data Preprocessing

The data consists of ambiguities, errors, redundancy which needs to be cleaned before applying the prediction model. The data are aggregated from the multiple sources and then it should be cleaned properly. Because uncleansed data may also affect the accuracy.

1. Elimination of null values from the dataset
2. Transforming categorical values into numerical values
3. Eliminating redundant data

C. Feature Selection

Feature selection is an important task to find the customer is churn or not.Feature selection can be done on two aspects before applying the classification algorithms. First, data's are plotted using visualization techniques and next is used to find the value using "Lasso coefficient". Three features are (tenure, monthly charges, total charges) selected after applying the two different types of feature selection methods. The boundary selection is also much needed task in the lasso coefficient process. This paper also addresses the sampling techniques to solve the imbalance problem in the classification process. The boundary selection helps to find the data is distributed properly or not. Instead of using boundary selection, sampling techniques are also useful to solve imbalancing problem. Out of 7000 data, 73.5 % data are "Non churn" and rest of them are "Churn" data.

D. Classification algorithm

They are various supervised classification algorithms are attempted to classify the churn or not. Another factor affect the CC dataset is class imbalance problem. This research works avoids to select the boundary selection process and select the sampling techniques to distribute the data properly in all classes. According to churn dataset, the data falls under the category of "Churn" is minority class. Sampling techniques helps to equalize the data in the minority class [18]. This research work uses the SMOTE technique to normalize the data in all classes. Specifically, this research using SMOTE with ENN (Edited Nearest Neighbor) method to normalize the data. The default value of k=3 is used in ENN method. This paper compares the boundary selection with sampling techniques in the result section. This research work address the DT and Random Forest techniques to analyses the performance CC prediction.

IV. Experimental Results

The experiment are carried out in the Scikit-learn library. Since it is an imbalanced dataset, F1- score will be the proper metric for evaluation various classification algorithms. F1-score will be the harmonic mean of precision and recall value.

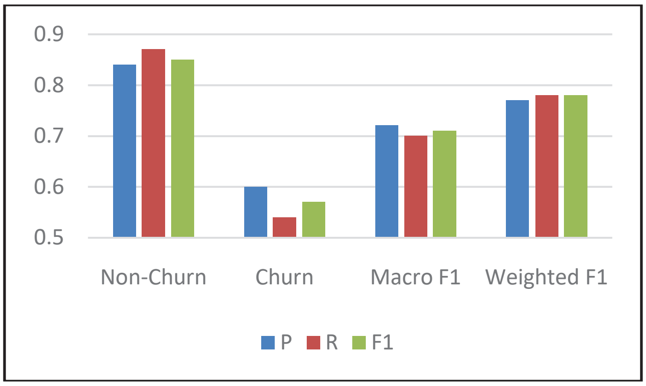

Fig. 1. Output of Decision Tree without Sampling Techniques

Fig. 1 describes the performance metrics of the decision tree classification algorithm. The output of the algorithm shows poor precision and recall value in the "Churn" category due to imbalancing problem.

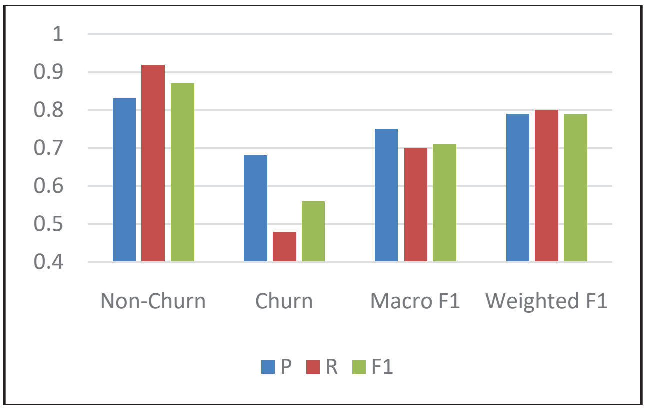

Fig. 2. Output of Random Forest without Sampling Techniques

Fig. 2 describes the performance metrics of the random forest classification algorithm. The output of the algorithm shows poor precision and recall value in the "Churn" category due to imbalancing problem. The overall accuracy of classifier is 78 percentage for Decision Tree and 80 percentage for Random Forest. Since the accuracy is good, the Churn category data is not predicted properly due to imbalancing. To avoid imbalancing, sampling techniques is incorporated in the classifier.

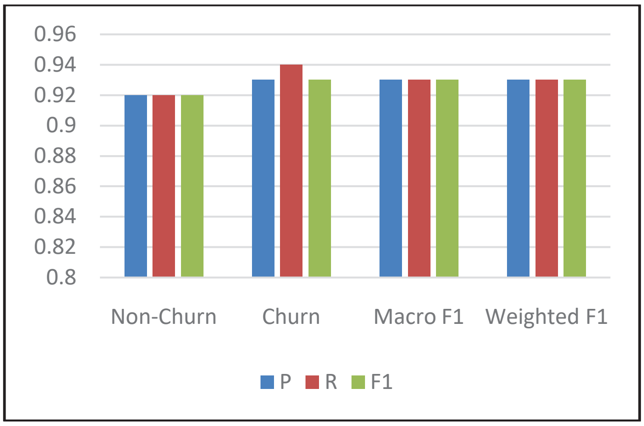

Fig. 3. Output of Decision Tree with SMOTE-ENN

Fig. 3 describes the performance metrics of the decision tree classification algorithm. The output of the algorithm shows better precision and recall value in the "Churn" category due to imbalancing problem. Fig. 4 describes the performance metrics of the Random Forest classification algorithm. The output of the algorithm shows better precision and recall value in the "Churn" category.

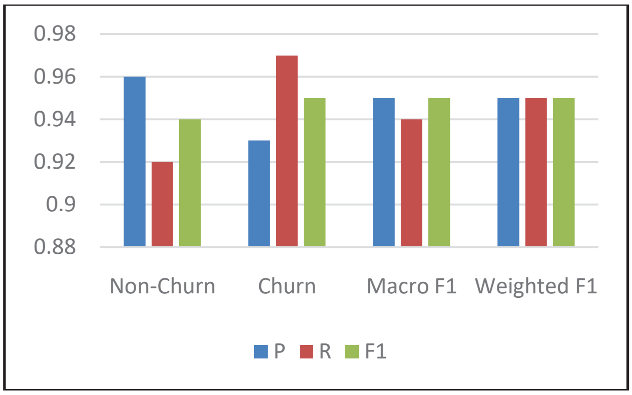

Fig. 4. Output of Random Forest with SMOTE-ENN

The overall accuracy of classifier is 93 percentage for Decision Tree and 95 percentage for Random Forest. The results are also proved that SMOTE-ENN better recall and precision value.

Table 1 Comparison of Decision tree, Random Forest Classifier with and without Sampling Techniques

Techniques	Metrics	Not Churn (%)	Churn (%)
Decision Tree Without Sampling	P	84	60
	R	87	54
	F1	85	57
Random	P	83	68

Forest Without Sampling	R	92	48
	F1	87	56
Decision Tree with SMOTE-ENN	P	92	93
	R	92	94
	F1	92	93
Random Forest with Smote-ENN	P	96	93
	R	92	97
	F1	94	95

Table 1 and Figure 5 clearly depicts the comparison of Decision Tree and Random Sampling Techniques. Customer Churn dataset is very popular imbalanced dataset. The objective of the work is to solve the imbalancing problem in the customer Churn Dataset and identify the best Machine learning model. The reason for choosing the decision tree and random forest is to check the dataset is performing in base model and ensembling process.

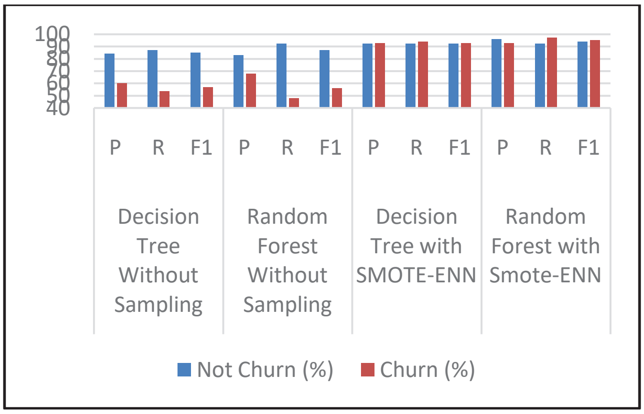

Fig. 5. Comparison of Decision tree, Random Forest Classifier with and without Sampling Techniques

Decision tree is a single base model and random forest is the combination of several decision trees. Due to imbalancing problem, single base model and ensemble approaches performs poor results. After applying the sampling techniques (SMOTE-ENN), both decision tree and Random forest produce good results. Though, random forest outperforms the best result compared to all other approaches.

V. Conclusion

The paper provides a comparison study on ML methods with sampling techniques in the process of CC prediction. The target of CC prediction technique is to retain customers at the highest risk of churn by proactively engaging with them by different methods like customers can be stick to the particular company. Random Forest combined with SMOTE-ENN achieves the output with the accuracy of 95%. SMOTE-ENN helps to replicate the data for the customer prediction model. The future enhancement of this research work is to incorporate the deep learning algorithms and analyses the performance of customer churn dataset.

References

[1] Farquad, H. &Vadlamani, Ravi &Surampudi, Bapi. (2014). Churn Prediction using Comprehensible Support Vector Machine: an Analytical CRM Application. Applied Soft Computing. 19. 10.1016/j.asoc.2014.01.031.

[2] Kumar, Dudyala& Ravi, Vadlamani. (2008). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies. 1. 4-28. 10.1504/IJDATS.2008.020020.

[3] Chih Fong Tsai, "Customer churn prediction through the hybrid neural networks", Expert Systems with Applications 12764-12534.

[4] WouuterVerbeke, Bart- Baesens "Constructing intelligible customer churn prediction models with advanced rule induction techniques", Expert Systems with Applications 2378–2394.

[5] Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang "A Customer Churn Prediction Model in Telecom Industry Using Boosting", IEEE Transactions on Industrial Informatics, vol. 10, no. 2, may 2014.

[6] H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru, "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 139-144, doi: 10.1109/UBMK52708.2021.9558876.

[7] R.V.S. Rohit, D. Chandrawat and D. Rajeswari, "Smart Farming Techniques for New Farmers Using Machine Learning", Proceedings of 6th International Conference on Recent Trends in Computing, vol. 177, 2021.

[8] Ssu-Han Chen, "The gamma CUSUM chart method for online customer churn prediction", Electronic Commerce Research and Applications, 17 (2016) 99–111.

[9] Koen W. De Bock, Dirk Van den Poel, "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction", Expert Systems with Applications 38 (2011) 12293–12301.

[10] D. Sikka, Shivansh, R. D and P. M, "Prediction of Delamination Size in Composite Material Using Machine Learning," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1228-1232, doi: 10.1109/ICEARS53579.2022.9752123.

[11] M. D. S. Rahman, M. D. S. Alam and M. D. I. Hosen, "To Predict Customer Churn By Using Different Algorithms," 2022 International Conference on Decision Aid Sciences and Applications (DASA), 2022, pp. 601-604, doi: 10.1109/DASA54658.2022.9765155.

[12] Koen W. De Bock, Dirk Van den Poel, "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models", Expert Systems with Applications 39 (2012) 6816– 6826.

[13] Sangamnerkar, S., Srinivasan, R., Christhuraj, M.R., Sukumaran, R., " An ensemble technique to detect fabricated news article using machine learning and natural language processing techniques", 2020 International Conference for Emerging Technology, INCET 2020, 2020, 9154053

[14] L. Ning, L. Hua, L. Jie, Z. Guangquan, "A customer churn prediction model in telecom industry using boosting", IEEE Trans. Ind. Inform. 10 (2014) 1659– 1665.

[15] K. Goyal, K. Kanishka, K. Vasisth, S. Kansal and R. Srivastava, "Telecom Customer Churn Prediction: A Survey," 2021 3rd International Conference on Advances in Computing,

Communication Control and Networking (ICAC3N), 2021, pp. 276-280, doi: 10.1109/ICAC3N53548.2021.9725621.

[16] S. De, P. P and J. Paulose, "Effective ML Techniques to Predict Customer Churn," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 895-902, doi: 10.1109/ICIRCA51532.2021.9544785.

[17] V. Geetha, A. Punitha, A. Nandhini, T. Nandhini, S. Shakila and R. Sushmitha, "Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262288.

[18] Srinivasan, R., Subalalitha, C.N. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distrib Parallel Databases (2021). https://doi.org/10.1007/s10619-021-07331-4.