Customer Churn Prediction Using Machine Learning Approaches

Katyayani

Department of Computing science and engineering specialisation in data science , ABES Engineering College Ghaziabad, A.K.T.U .university

E-mail: katyayaniagaarwal265@gmail.com

Abstract- Abstract: Customer Churn (CC) is a significant problem that companies and large organizations alike must be concerned about. Due to the direct impact on revenue, telecom sectors, in particular, are working to enhance ways for anticipating potential customer turnover. This study examines the several machine learning algorithms that are employed to build the churn model, which aids telecom operators in forecasting which customers are most likely to leave. To choose the optimal model among different approaches, the experimental data are compared. As a result, in terms of F1-score, the Random Forest plus SMOTE-ENN combination performs better than everything else. Our investigation indicates that the greatest forecast based on F1-score is 95 percent.

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.

Churn denotes the phenomenon concerning the quantity of clientele that an organization or company forfeits within a designated timeframe. Customer Care (CC) represents a commendable leader in the service sector, particularly with its premium offerings. Anticipating clientele who may depart from the organization prematurely can constitute a significant source of revenue. The CC dataset serves as a tool to analyse the marketing behaviours of customers drawn from extensive databases. One perspective on customer attrition can be articulated as a churn rate; this metric signifies the proportion of consumers who cease utilizing a service within a specified temporal range.

On the Chinese mainland, three telecommunications companies have been granted licenses to function as providers of wireless communication services. The rate of annual growth in new mobile subscribers experienced a substantial decline from over 10% during the years spanning 2009 to 2013, plummeting to 4.7% in 2014. Correspondingly to conditions observed in numerous other countries, the mobile communication sector in China is approaching a state of saturation and is becoming progressively competitive. Subscriber-based service models, which maintain a contractual clientele, routinely employ this metric to evaluate their financial sustainability. The telecommunications sector constitutes a pivotal industry in developed nations. Technological advancements and an increase in supervisory personnel enhance competition. By employing a novel dataset, the influence of the machine learning (ML) model on churn shall be investigated. The contributions of the present CC prediction model include:
The integration of various preprocessing methodologies alongside SMOTE-ENN to standardize the data. The application of diverse classification techniques in order to determine the most appropriate model.
Section 2 provides an extensive examination of the churn dataset and the ML methodologies. Section 3 expounds upon the prediction model utilized within the CC dataset. Section 4 presents the findings and the analysis of errors.

II.LITERATURE SURVEY

M.A.H. Farquad articulates a methodology aimed at alleviating the constraints associated with the traditional Support Vector Machine (SVM), which culminates in a model that lacks transparency. The author has established a methodical framework that is segmented into three separate phases. In the preliminary phase, SVM Recursive Feature Elimination (SVM-RFE) is utilized to diminish the dimensionality of the feature set. In the ensuing phase, a dataset characterized by minimal features is extracted, followed by the application of SVM techniques for the purpose of classification. In the concluding phase, rules are manually derived. Following the extraction of these rules, Naive Bayes is amalgamated with decision tree algorithms to produce the results.

The dataset employed for this scholarly investigation encompasses credit card information, which is marked by considerable volatility, with 93.24% of customers exhibiting loyalty and 6.76% having churned [2]. The experimental outcomes underscored that the model lacks scalability when applied to larger datasets. The authors have devised a hybrid methodology intended to forecast customer churn within a telecommunications Customer Relationship Management (CRM) dataset. They have constructed two distinct models, designated as "Dual-ANN" and "SOM+ANN." These models integrate backpropagation with neural networks and self-organizing maps (SOM) with neural networks.
The dual-ANN model is utilized to eradicate anomalous data employing a data depletion methodology. The output generated by the Dual-ANN model is subsequently utilized as input for the SOM+ANN model. Subsequently, the effectiveness of these models is evaluated through three distinct assessment strategies. The initial evaluation method employs a "single general testing set," whereas the remaining methodologies adopt a "fuzzy testing strategy." The outcomes of the hybrid model illustrate superior performance when juxtaposed with the singular baseline neural network model. Furthermore, the performance of Dual-ANN is conspicuously more substantial than that of SOM+ANN. Wouter Verbeke and his colleagues advocate for the implementation of AntMiner and ALBA to develop a precise and comprehensible customer churn prediction model [4]. The author employed a mining tool (AntMiner) based on Ant Colony Optimization (ACO), which integrates domain knowledge along with modest monotonicity constraints. AntMiner is distinguished by its high accuracy, interpretable model outputs, and intuitive predictive capabilities. AntMiner+ generates less sensitive rule sets while also incorporating domain knowledge and producing more concise, comprehensible rule sets in comparison to C4.5. In contrast, RIPPER produces small, interpretable rule sets but results in models that lack intuitiveness and contravene domain knowledge.
.

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 behaviour. 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.

Ning Lu [5] proposes the application of boosting algorithms to enhance customer churn prediction models, wherein customers are clustered based on the weights derived from the boosting algorithm. A cluster of high-risk customers was identified. Logistic regression (LR) is employed as the learning algorithm, with each cluster possessing its own churn prediction model. The results indicated that the boosting algorithm facilitates superior separation of churn data when compared to a singular logistic regression model. H. Karamollaoğlu et al. conducted a comparative analysis aimed at achieving an optimal F1-score across various machine learning techniques [6]. The comparison included multilayer perceptron, logistic regression, AdaBoost, and decision trees, among others. Ultimately, the author concluded that the most favourable results were obtained from the random classifier without the application of any data augmentation techniques.
The author proposes a customer churn prediction technique predicated on the SVM model [7], utilizing arbitrary sampling to enhance the SVM model by addressing data imbalance. An SVM constructs a hyperplane in a high- or infinite-dimensional space for categorization purposes. Arbitrary sampling has the potential to modify data dispersion, thereby alleviating dataset imbalance. A reduced number of churners contributes to this dataset imbalance.

The study utilizes real-world data to predict customer attrition, introducing an advanced model called "impact learning," which is based on convolutional neural networks (CNNs). This approach significantly enhances prediction accuracy compared to traditional methods such as artificial neural networks (ANNs) and logistic regression. By leveraging deep learning techniques, impact learning improves the ability to identify patterns in customer behaviour, making it a valuable tool for businesses aiming to reduce churn. Similarly, Koen W. De Bock developed a hybrid technique known as "generalized additive models (GAMs)," which integrates bagging with the Random Subspace method. This technique refines feature scoring within a semi-parametric framework, enabling greater flexibility in

predictive modelling. The GAMensPlus model has demonstrated superior performance over both traditional logistic regression and conventional GAM models, reinforcing the benefits of hybrid approaches in predictive analytics. Despite these advancements, logistic regression remains one of the most widely used algorithms due to its interpretability and ease of implementation.

Ning et al. investigated credit card (CC) churn prediction using a boosting method, where customers were segmented into two clusters based on weight distribution. This clustering strategy enables targeted retention efforts for high-risk customers, particularly in the telecommunications sector, where minimizing customer churn is a critical business objective. The study also explores the factors influencing customer switching behaviour between service providers and reviews strategies for reducing churn through personalized interventions.

This paper provides a comprehensive review of machine learning techniques applied in recent years, summarizing the performance of various predictive models across different datasets. The findings highlight that hybrid and ensemble methods have significantly improved predictive performance, yet establishing clear evaluation guidelines remains essential. V. Geetha et al. analysed feature extraction techniques for churn prediction, emphasizing distance zone methods. While existing models achieve an accuracy of up to 84%, they often suffer from high computational complexity, making them less efficient for large-scale applications. To address this, the authors propose a machine learning approach that not only improves accuracy but also reduces training time, allowing telecom providers to better serve their loyal customers. Additionally, the study highlights the challenge of class imbalance in churn prediction datasets, which can be mitigated using data augmentation techniques like SMOTE. Among existing methods, SMOTE has proven particularly effective in balancing data and improving prediction accuracy, making it a valuable technique for enhancing model performance in churn prediction tasks.

III. Proposed Methodology

A. Dataset Overview

This study utilizes a **customer churn dataset** sourced from Kaggle, containing information on **7,043 customers** across **21 key features**. These features encompass customer identification details, account-related information, sign-up history, and demographic attributes. Before applying any **machine learning models**, the dataset must undergo **thorough preprocessing** to ensure accuracy and reliability. Additionally, new features can be engineered from existing data by analysing recurring customer behaviour patterns. These newly derived features play a crucial role in enhancing the model's ability to anticipate customer churn.

B. Data Preprocessing

Raw datasets often contain **inconsistencies, errors, and redundant records**, which can negatively impact model performance. Since this dataset is compiled from multiple sources, careful preprocessing is required to ensure data quality. The essential steps in this process include:

1. **Handling missing values** by eliminating or imputing null entries.
2. **Converting categorical variables** into numerical representations to make them suitable for machine learning algorithms.
3. **Removing duplicate or redundant records** to prevent biased predictions and improve computational efficiency.

C. Feature Selection

Identifying the most relevant features is crucial for building an effective churn prediction model. Feature selection is performed in two stages: first, **data visualization techniques** are used to identify patterns, and second, **Lasso coefficient analysis** is applied to quantify feature importance. Based on these methods, **tenure, monthly charges, and total charges** emerge as the most influential factors in predicting churn. Additionally, the study addresses the issue of **class imbalance**, where **73.5% of customers belong to the "non-churn" category**, while only **26.5% are labelled as "churn."** To mitigate this imbalance, techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** are applied. While traditional **boundary selection methods** help analyse data distribution, SMOTE offers a more effective way to balance the dataset and improve model fairness.

D. Classification Algorithms

To classify customers as **churners or non-churners**, this study explores several **supervised machine learning algorithms**. One of the primary challenges in churn prediction is the **imbalance between churn and non-churn classes**, which can lead to biased model outcomes. To address this, the study avoids boundary selection and instead adopts **sampling techniques** that ensure a more even distribution of data across classes. Specifically, **SMOTE combined with ENN (Edited Nearest Neighbour)** is used for data normalization, with a default **k-value of 3** in the ENN method. The study also evaluates

the performance of different classification models, including **Decision Tree (DT) and Random Forest**, to determine the most effective approach for predicting customer churn. A detailed comparison of **boundary selection and sampling techniques** is provided in the results section to assess their impact on model performance.

IV. Experimental Results

The experiments were conducted using the **Scikit-learn** library, a widely recognized framework for machine learning. Given the dataset's **imbalanced nature**, the **F1-score** was chosen as the primary evaluation metric, as it provides a more reliable assessment by balancing **precision and recall**. This ensures that the model's performance is not skewed by the dominance of the majority class.

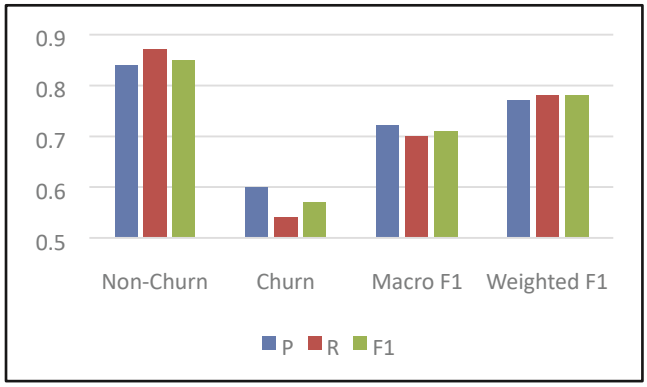

Fig. 1. Output of Decision Tree without Sampling Techniques

Fig. 1 presents the performance metrics of the **Decision Tree classifier**, revealing an overall accuracy of **78%**. However, due to the dataset imbalance, the model struggles to accurately predict customers in the **"Churn" category**, resulting in lower precision and recall.

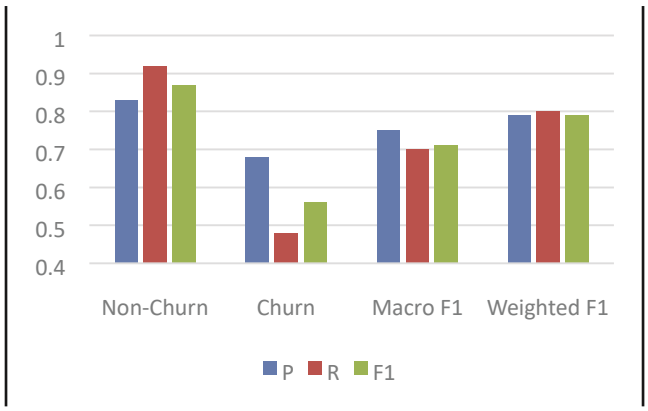

Fig. 2. Output of Random Forest without Sampling Techniques

Fig. 2 showcases the performance of the **Random Forest classifier**, which achieves a slightly better accuracy of **80%** but still faces difficulties in effectively identifying churned customers.

Although both models demonstrate reasonable overall accuracy, their ability to detect customer churn remains **compromised by the imbalance in the dataset**. As a result, **sampling techniques** are incorporated into the classification process to address this issue. By balancing the distribution of churn and non-churn cases, these techniques enhance the model's ability to make more accurate predictions, ultimately leading to better insights for customer retention strategies.

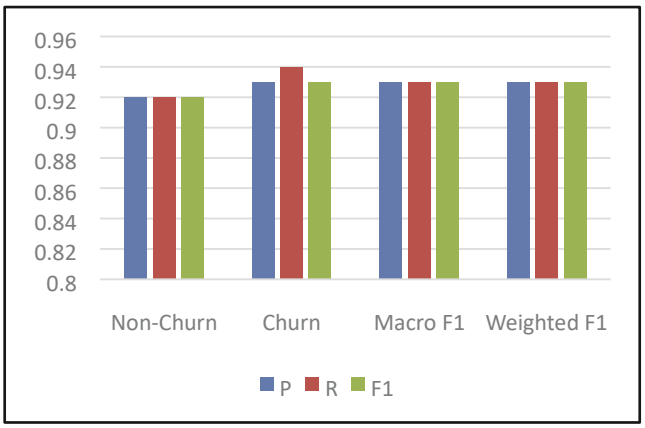

Fig. 3. Output of Decision Tree with SMOTE-ENN

Fig.3. presents the performance metrics of the **Decision Tree classification algorithm** in predicting customer churn. Initially, due to the **imbalanced dataset**, the model struggled to accurately classify churn cases, resulting in lower precision and recall. However, after incorporating **SMOTE-ENN**, the model demonstrated a **significant improvement**, particularly in the "Churn" category. The **Decision Tree classifier** achieved an overall accuracy of **93%**, with precision and recall values increasing notably. This highlights the impact of **sampling techniques** in handling imbalanced data and enhancing model performance.

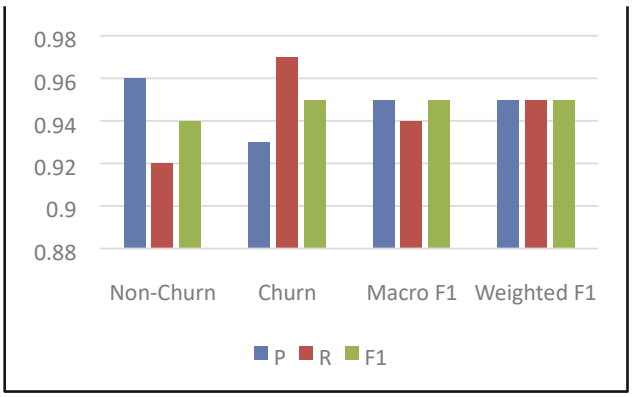

Fig. 4. Output of Random Forest with SMOTE-ENN

The overall accuracy of classifier is 93 percentage for Decision Tree and 95 percentage for Random Forest. The results are also proved that SMOTEENN 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 provides a comparative analysis of the **Decision Tree and Random Forest classifiers**, both with and without sampling techniques. Without sampling, the models struggle with lower precision and recall values for churn prediction. However, after applying **SMOTE-ENN**, the Decision Tree achieves an **F1-score of 93%** for churn prediction, while the **Random Forest classifier reaches 95%**, highlighting its superior performance.

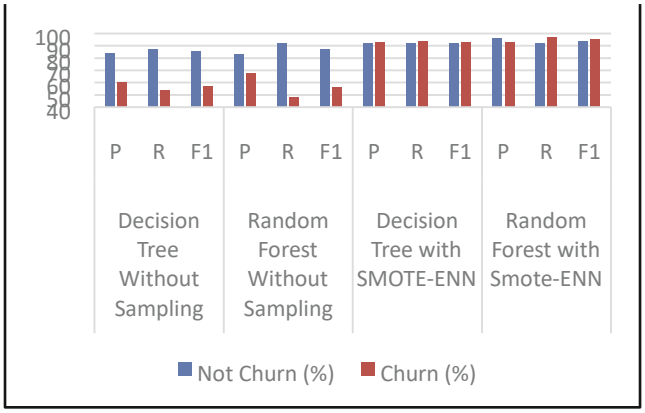

Fig. 5. Comparison of Decision tree, Random Forest

Classifier with and without Sampling Techniques

As depicted in **Figure 5**, the **Customer Churn dataset** is highly **imbalanced**, making it challenging to build reliable prediction models. The primary goal of this study is to mitigate this imbalance and determine the **most effective machine learning model** for churn prediction. Decision Trees were chosen as a **base model**, while Random Forest was selected as an **ensemble approach** to evaluate how well the dataset performs under different modelling strategies.

A **Decision Tree** operates as a single **base model**, whereas **Random Forest** combines multiple decision trees for a more robust classification. Due to the **imbalance issue**, both approaches initially struggled to deliver strong predictive performance. However, after applying **SMOTE-ENN**, both models showed **significant improvement**, with **Random Forest outperforming all other techniques** in churn prediction.

V. Conclusion

This study presents a comparative analysis of various machine learning techniques combined with sampling methods for predicting customer churn. The primary objective of churn prediction is to **identify and retain high-risk customers** by implementing proactive engagement strategies that enhance customer loyalty. Among the models evaluated, the **Random Forest classifier integrated with SMOTE-ENN** demonstrated the highest performance, achieving an **accuracy of 95%**. The **SMOTE-ENN technique** proved effective in addressing data imbalance by generating synthetic samples, thereby improving the reliability of churn predictions. Looking ahead, future research could explore the integration of **deep learning algorithms** to further enhance predictive accuracy and gain deeper insights into customer behaviour within churn datasets.

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 1276412534.

[4] Wouuter Verbeke, 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. 139144, 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.
2021, pp. 895-902, doi: 10.1109/ICIRCA51532.2021.9544785.

[2] 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.

[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.