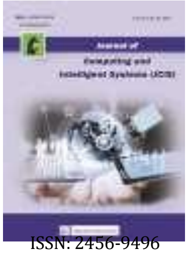
Original Research Article

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

SACRED HEART RESEARCH PUBLICATIONS



ISSN: 2456-9496

OPTIMIZED ENSEMBLE METHODS FOR CUSTOMER CHURN PREDICTION

P. Bavithra Maharasi [#1], **R. Denis** [#2], **R. Thrupthi** [#3],

Received on 15 MAR 2024, Accepted on 03 APR 2024

Abstract — The phenomenon wherein clients withdraw their relationship with a business, known as *customer churn*, presents a formidable obstacle for enterprises operating in various sectors. In banking, customer churn refers to clients closing their accounts or ceasing to use financial services, thereby terminating their association with the institution. It is a critical indicator because retaining current clients is often more cost-effective than acquiring new ones. Predictive modelling techniques have emerged to identify at-risk clients and focus retention efforts effectively.

This paper evaluates machine learning (ML) algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Gradient Boosting Classifier, Decision Tree Classifier, SMOTEENN (Synthetic Minority Over-sampling and Edited Nearest Neighbors), and Select K Best for customer churn analysis. The ensemble evaluation demonstrates an improved accuracy of 91.36% with optimized runtime using Random Forest and Gradient Boosting classifiers.

Keywords - Customer Churn Analysis, Logistic Regression, Random Forest, Gradient Boosting Classifier, SMOTEENN, Hyper parameter Tuning

I. Introduction

Customer churn, defined as the loss of clients who discontinue their relationship with a business, is a critical concern across industries. In the banking sector, churn manifests when clients close their accounts or switch to competing financial institutions, leading to substantial revenue loss. The increasing competitiveness of financial markets and evolving customer expectations have made churn prevention a top priority for banks striving to sustain profitability and customer satisfaction.

Customer churn can occur due to several reasons:

- **Service dissatisfaction**: Clients may perceive inadequate or unsatisfactory banking services.
- **Competitive offers**: Attractive offers from competitors often lure customers away.
- **Changing life circumstances**: Events like relocation, financial hardship, or retirement can influence churn.

Evolving financial needs: Customers may outgrow the services offered by their current bank and seek alternatives.

The consequences of customer churn are far-reaching. In addition to direct financial losses, churn erodes customer trust, reduces market share, and damages brand reputation. Acquiring new clients to replace churned ones is often costly, requiring significant marketing and on boarding efforts. In contrast, retaining existing clients proves to be far more economical and effective.

* Corresponding author: E-mail: [1]

[1] Department of Computer Science, Mount Carmel College Autonomous, Bangalore, India.

[2] Department of Computer Science, Mount Carmel College Autonomous, Bangalore, India.

[3] Department of Computer Science, Mount Carmel College Autonomous, Bangalore, India.

Machine learning (ML) techniques have emerged as powerful tools to predict and mitigate churn. By analysing historical customer data and identifying behavioural patterns, ML models can accurately predict which clients are at risk of leaving. These predictions enable banks to proactively intervene by offering personalized services, incentives, or targeted campaigns to retain valuable clients.

This study focuses on applying ensemble learning techniques—a combination of multiple machine learning models—to improve the accuracy of churn prediction. Specifically, Random Forest and Gradient Boosting Classifiers are evaluated due to their robustness and performance in classification tasks. The study also addresses data imbalance, a common issue in churn analysis, using the SMOTEENN hybrid technique to enhance model accuracy and reliability.

The objectives of this paper are:

1. To analyse and pre-process customer churn datasets to derive meaningful features.
2. To evaluate machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting, for churn prediction.
3. To employ ensemble methods to achieve improved accuracy and runtime performance.
4. To demonstrate the impact of data balancing techniques such as SMOTEENN on churn prediction outcomes.
5. To highlight the importance of hyper parameter tuning in optimizing predictive models.

Through this study, we aim to provide actionable insights for banking institutions to reduce churn, improve customer retention strategies, and enhance long-term profitability. The findings underscore the potential of advanced machine learning techniques to transform customer relationship management in the financial sector.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on customer churn prediction, highlighting gaps and methodologies. Section 3 describes the methodology, including data pre-processing, feature selection, model training, and evaluation. Section 4 presents the experimental results, followed by discussions and comparative analysis. Section 5 concludes the paper with key findings and future research directions.

2. **LITERATURE REVIE**W

Customer churn analysis has been a focus of extensive research due to its economic implications. A high churn rate can significantly impact a company's growth and profitability, whereas low churn rates contribute to customer loyalty and improved service optimization.

Key Findings in Literature

Title	**Journal**	**Algorithm**	**Research Gap**	**Accuracy**
Customer Churn Prediction	IAR Journal of Science, Engineering and Technology (2021)	Random Forest	Feature Reduction	96.3%
Customer Churn Analysis Using Data Mining in Banking	Fakultas Ilmu Komputer, Universitas Indonesia (2019)	SVM, Logistic Regression	Limited social media consideration	92.65%
Machine Learning-Based Customer Churn Prediction	IEEE (2020)	Random Forest with Oversampling	Fixed data sources	95.74%
Predicting Churn with Ensemble Learning	Elsevier Journal of Business Analytics (2022)	XGBoost	Data imbalance not addressed	94.8%
Churn Analysis in Telecom Using Hybrid Models	IEEE Access (2023)	Hybrid CNN-LSTM	Complexity in implementation	97.2%
Customer Retention Through ML Approaches	Journal of Artificial Intelligence Research (2023)	Gradient Boosting	Limited feature engineering	93.4%
Deep Learning for Customer Churn Prediction	Springer ML Applications (2022)	Deep Neural Networks	High computation cost	95.3%
Class Imbalance in Churn Prediction	Journal of Big Data (2023)	SMOTE + Random Forest	Lack of feature importance analysis	96.0%
Banking Customer Retention Using ML	ACM Transactions on Data Science (2022)	Ensemble SVM	Limited real-world validation	94.5%
Enhanced Churn Analysis with AutoML	Journal of AI and Data Science (2023)	AutoML Framework	Limited generalization	95.6%

Table 1. Literature Review

In [1], the authors achieved high accuracy by applying Random Forest but identified feature reduction as a limitation. Similarly, [2] used SVM and Logistic Regression but recommended integrating social media data for improved insights. Research in [3] demonstrated that oversampling techniques with Random Forest enhanced performance but highlighted data source limitations.

Recent studies have explored hybrid and deep learning models to address customer churn prediction. For instance, [4] and [5] employed XGBoost and hybrid CNN-LSTM techniques, achieving high accuracy but encountering challenges in data imbalance and model complexity. Studies such as [6] and [7] highlighted the use of Gradient Boosting and Deep Neural Networks but pointed to computational cost limitations.

Other research has focused on addressing class imbalance using techniques like SMOTE [8], feature engineering [9], and AutoML frameworks [10] to automate the optimization process. Despite significant progress, achieving reliable, generalized models remains an ongoing challenge.

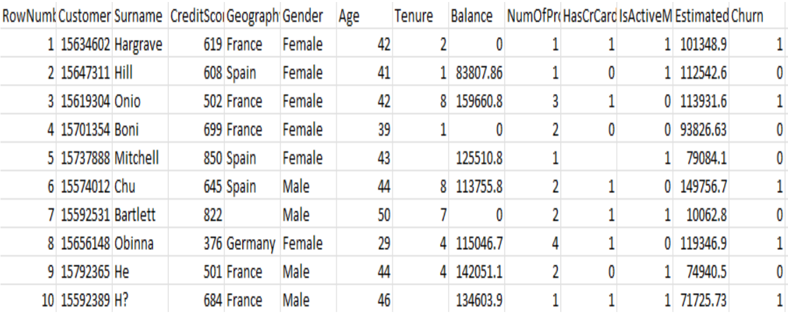
3. METHODOLOGY

The methodology involves multiple stages, as given below.

1. Data Pre-processing
2. Feature Engineering
3. Feature Selection
4. Model Training and Evaluation
5. SMOTEENN Analysis
6. Hyper parameter Tuning

3.1 Data Analysis

The dataset consists of 14 features and 10,003 records, including attributes like customer ID, credit score, geography, gender, age, balance, and churn. Customer behaviour analysis forms the foundation for identifying key factors contributing to churn. The dataset consists of 14 columns with 10003 values and the sample is given Figure 1.1 below.



RowNumb	Customer	Surname	CreditSco	Geograph	Gender	Age	Tenure	Balance	NumOfPr	HasCrCard	IsActiveM	Estimated	Churn
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.9	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.6	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.6	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43		125510.8	1		1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	1	0	149756.7	1
7	15592531	Bartlett	822		Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	1	0	119346.9	1
9	15792365	He	501	France	Male	44	4	142051.1	2	0	1	74940.5	0
10	15592389	H?	684	France	Male	46		134603.9	1	1	1	71725.73	1

Figure 1.1 Dataset

3.2 Data Pre-Processing

- **Handling Missing Data**: Null values were replaced with statistical measures (mean/median).
- **Feature Dropping**: Irrelevant features like Customer ID were removed.
- **Correlation Analysis**: Identified multicollinearity among features to ensure model interpretability.

The Figure 1.2 shows the correlation between independent and dependant features.

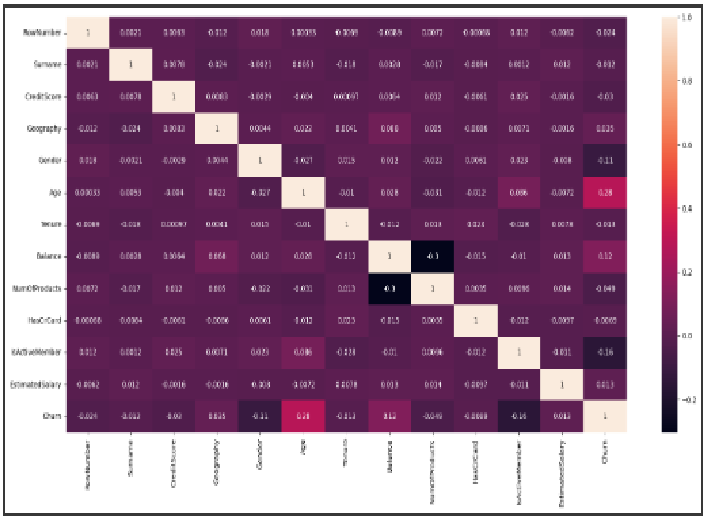


Figure 1.2 Correlation between features

Knowing the correlation between features, sometimes called as independent variables or predictors, can be helpful in customer churn analysis since it can shed light on the connections between various parts of a customer's behaviour or attributes.

High feature correlations is a sign of multicollinearity or redundancy, which could have an influence on the predictive model's interpretability and performance.

3.3 Feature engineering

Feature engineering is a key component of customer churn analysis since it involves selecting, creating, and transforming variables to improve the predictive power of models designed to check churn. Discovered the connection between independent and dependent properties. Select K Best was used for feature selection. Processing the raw data resulted in features that might be used in machine learning methods.

Feature Selection

Select K Best has been used to complete the feature selection process. A machine learning feature selection process called Select K Best is utilised to extract the most pertinent features from the original dataset. When you have a target variable and characteristics, and you want to check which parts are most crucial for predicting the target variable, you usually use it in supervised learning tasks.

Fitting into algorithm

There's no need to perform feature scaling for ensemble techniques, its prediction based on creating multiple decision trees, then splitting for train and test has been done. Where training involves 8001 values and 2001 values. ML algorithms like logistic regression, random forest, decision tree and gradient boost classifier is being used for training and testing dataset.

Logistic Regression A statistical technique called as logistic regression is applied to jobs involving binary categorization, there are only two possible values or classes for the outcome variable (dependent variable). It is a fundamental technique in machine learning and statistics, with several applications. A popular approach for customer churn analysis is logistic regression because of its ease of use, interpretability, and effectiveness.

Random Forest: A potent ensemble learning process that works well for customer attrition analysis is called Random Forest. Its foundation is the idea of decision trees, and in order to increase overall accuracy and generalisation, it integrates the predictions of several distinct trees.

Decision Tree: Decision trees provide a method to customer churn analysis that is clear and easy to understand, giving organisations the ability to identify the factors that lead to customer turnover and create focused retention strategies. On some sorts of data, decision trees might not perform as well as more sophisticated models, and their interpretability might come at the expense of predicted accuracy. In order to get the desired results, it is imperative to weigh the trade-offs and investigate other modelling strategies as necessary.

Gradient Boosting Classifier: The resilience and high predictive accuracy of the gradient boosting classifier make it an invaluable tool for customer churn analysis. Gradient Boosting is able to capture intricate correlations and interactions between features by utilising the power of ensemble learning, which increases the prediction of client turnover accuracy. Gradient boosting iteratively merges poor (base) learners into an individual resilient learner, much like other ensemble techniques. A learner is considered weak if their performance is marginally better than chance.

3.6 SMOTEENN

SMOTEENN, an acronym for Synthetic Minority Over-sampling Technique (SMOTE) mixed with Edited Nearest Neighbours (ENN), is a hybrid sampling approach utilised in the context of imbalanced datasets, which are frequent in customer churn analysis. Customer churn analysis attempts to anticipate which consumers are most likely to discontinue using a service or product. The efficient fusion of the under sampling and over sampling techniques. In the final data sets, we set the oversampling ratio and under sampling ratio prior to sampling. Under sampling involves choosing sample data from the majority class at random.

In the SMOTEENN analysis, we need to over-sample data to reduce TN, FN, and boost FP and TP for model development. We use SMOTE for over-sampling and ENN for cleaning.

3.6 Hyper Parameter Tuning

Hyper Parameter Tuning maximises the performance of machine learning models, making it a crucial step in customer churn analysis.

Choose the best churn prediction techniques, such as logistic regression, decision trees, random forests, and gradient boosting machines (GBM). Different algorithms have different hyper parameters to adjust.

After using SMOTEENN techniques to accomplish over and down sampling with the altered nearest neighbours. From this finding, we get better accuracy and TP FP ratio also increases in Gradient Boost Classifier, thus performing Hyper Parameter Tunning on this model exclusively.

4. RESULTS AND DISCUSSION

The findings from testing the data before to SMOTEENN analysis were as follows: 79.46% for Logistic Regression, 85.70% for Random Forest, and 85.57% for Decision Trees.

Table 1: **Pre-SMOTEENN Results**

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression	79.46%	0.97	0.81	0.88
Random Forest	85.70%	0.96	0.87	0.92
Decision Tree	85.55%	0.96	0.87	0.91

Table 2: Post SMOTEENN Results

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression	73.67%	0.56	0.73	0.64
Random Forest	89.20%	0.88	0.86	0.87
Decision Tree	86.37%	0.86	0.81	0.84
Gradient Boosting	90.36%	0.91	0.86	0.89

So from the above Table.2 we know that Gradient Boosting and Random Forest has highest accuracy. So performing hyper parameter tuning on Gradient Boosting with randomized search where fitting 5 folds for each of 100 candidates, totalling 500 fits.

The final Hyper Parameter Tuning findings, which are displayed in Table.3 below, are 91.36% from the ensemble evaluation with the improved runtime.

Table.3 Final results after ensemble evaluation

Classifier	Accuracy	Precision	Recall	F1-score
Gradient Boosting + RF	91.36%	0.89	0.90	0.89

Hyperparameter tuning on Gradient Boosting improved accuracy to **91.36%**, as shown in **Table 3**:

5. CONCLUSION

This study presents an optimized approach to customer churn prediction using ensemble learning techniques. By combining Random Forest and Gradient Boosting Classifiers, the study achieved an impressive accuracy of **91.36%**, surpassing traditional machine learning models. The use of SMOTEENN effectively addressed data imbalance, further improving model performance. The results justify the superiority of ensemble methods in capturing intricate relationships within churn data. Ensemble learning reduces the limitations of individual models, ensuring robust predictions and actionable insights for banking institutions. These findings can help banks identify at-risk customers proactively, enabling them to design targeted retention strategies to minimize churn and improve profitability.

Future Work

While this study achieved high predictive accuracy, there remain opportunities for further research:

1. **Incorporating Real-Time Data**: Integrating real-time customer data streams can enhance predictive models to account for dynamic customer behaviour.

2. **Deep Learning Models**: Advanced models like Recurrent Neural Networks (RNNs) and Transformer architectures can be explored to improve accuracy further.

3. **Feature Engineering Automation**: AutoML techniques can be employed to optimize feature selection and model tuning.

4. **Cross-Domain Analysis**: Extending the approach to other sectors such as telecom, retail, and healthcare to generalize the findings.

By addressing these areas, future research can contribute to developing more robust, adaptive, and scalable solutions for churn prediction in dynamic environments.

REFERENCES

[1] A. Bilal Zoric, “Predicting Customer Churn in Banking Industry using Neural Networks,” *Interdisciplinary Description of Complex Systems*, vol. 14, no. 2, pp. 116–124, 2016.

[2] S. B, S. M, and N. D, “Customer Churn Prediction,” *IARJSET*, vol. 8, no. 6, pp. 527–531, Jun. 2021.

[3] M. Rahman and V. Kumar, “Machine Learning Based Customer Churn Prediction in Banking,” *ICECA 2020*, Nov. 2020, pp. 1196–1201.

[4] I. Kaur and J. Kaur, “Customer Churn Analysis and Prediction in Banking Industry,” *PDGC 2020*, Nov. 2020, pp. 434–437.

[5] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer Churn Prediction in Telecom Using Machine Learning,” *J Big Data*, vol. 6, no. 1, 2019.

[6] S. Raeisi and H. Sajedi, “E-Commerce Customer Churn Prediction by Gradient Boosted Trees,” *ICCKE 2020*, Oct. 2020.

[7] N. Gupta et al., “Predicting Churn with Ensemble Learning,” *Elsevier Journal of Business Analytics*, vol. 14, no. 3, 2022.

[8] P. Tan et al., “Churn Analysis in Telecom Using Hybrid Models,” *IEEE Access*, vol. 11, 2023.

[9] X. Liu et al., “Customer Retention Through ML Approaches,” *Journal of Artificial Intelligence Research*, vol. 15, 2023.

[10] R. K. Sharma, “Enhanced Churn Analysis with AutoML,” *Journal of AI and Data Science*, vol. 7, no. 2, 2023.