International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 9, Issue 11, pp 57-60, November-2025
https://doi.org/10.47001/IRJIET/2025.911005

Research on Bank Customer Churn Prediction Using Machine Learning

[1]Sonali Vidhate, [2]Javed Attar, [3]Rida Fatema Shaikh, [4]Uzma Shaikh, [5]Pallavi Thete, [6]Misbah Attar

[1,2]Assistant Professor, Department of MCA, MET's Institute of Engineering, Nashik, Maharashtra, India
[3,4,5,6]PG Student, Department of MCA, MET's Institute of Engineering, Nashik, Maharashtra, India

***Abstract* - In today's highly competitive banking sector, customer churn poses a significant challenge, directly affecting profitability and customer retention efforts. This research aims to develop a predictive model for customer churn using advanced machine learning techniques. A comparative analysis of multiple supervised learning algorithms — including Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors (KNN), XGBoost, and Random Forest — was conducted on a publicly available dataset from Kaggle. Additionally, deep learning techniques using Artificial Neural Networks (ANN) were implemented through TensorFlow and Keras frameworks. The study emphasizes the importance of feature engineering and data preprocessing strategies such as oversampling and undersampling to handle class imbalance. Among all the models evaluated, the Random Forest classifier achieved the highest accuracy of approximately 87%, proving to be the most robust and stable model for churn prediction. The results highlight key factors influencing churn, such as customer age and account activity, providing actionable insights for banks to enhance customer engagement and reduce attrition.**

Keywords: Customer Churn Prediction, Machine Learning (ML), Random Forest Model, Artificial Neural Networks (ANN), Feature Engineering, Banking Analytics.

I. INTRODUCTION

In today's highly competitive banking environment, customer churn — the tendency of clients to discontinue their relationship with a bank — poses a major challenge that directly affects profitability and growth. Retaining existing customers is far more cost-effective than acquiring new ones, making churn prediction a strategic priority for financial institutions. Traditional analytical methods, such as logistic regression, have provided early insights but often fall short in handling large, complex, and imbalanced datasets. With the advent of machine learning (ML), banks can now leverage data driven models to identify at-risk customers and predict churn with higher accuracy. ML algorithms such as Random Forest, XGBoost, and Artificial Neural Networks (ANN) are particularly effective in capturing non-linear patterns and feature interactions within customer data. This study focuses on developing a predictive framework using advanced ML algorithms to analyze key behavioral and demographic factors influencing churn. By comparing the performance of multiple models, the research aims to determine the most accurate and reliable approach for proactive customer retention in the banking sector.

II. LITERATURE REVIEW

Customer churn prediction has emerged as a crucial research area in the financial and banking industry due to its direct impact on profitability and customer retention strategies. Over the past decade, numerous studies have applied machine learning (ML) and statistical approaches to forecast customer attrition, aiming to enhance predictive accuracy and enable timely intervention. Early research relied primarily on statistical techniques such as Logistic Regression and Decision Trees, which offered interpretability but lacked the flexibility to model complex, nonlinear customer behaviors. As banking data became more extensive and multidimensional, traditional models struggled with issues like overfitting, feature correlation, and class imbalance. To overcome these challenges, researchers began adopting ensemble learning and hybrid ML approaches that combine multiple algorithms to improve generalization and robustness. Rasha Ashraf (2024) proposed a Machine Learning Framework for Bank Customer Churn Prediction [1] that compared several algorithms and highlighted the superior performance of ensemble models such as Random Forest and XGBoost. The study emphasized the importance of feature engineering and data balancing techniques, showing that preprocessing significantly influences model performance. Similarly, research conducted at Xiamen University (2021) analyzed user churn using an ensemble learning algorithm [1], concluding that methods combining bagging and boosting yield more stable and accurate predictions compared to individual classifiers. Sonia Akakpo [2] et al. (2024) optimized the K-Nearest Neighbor (KNN) algorithm to predict churn, demonstrating improvements in recall and AUC scores through hyperparameter tuning and dimensionality reduction. Meanwhile, Shangxuan (2020) compared Random Forest and Logistic Regression for churn analysis, finding that Random

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 57-60, November-2025

https://doi.org/10.47001/IRJIET/2025.911005

Forest consistently achieved higher accuracy while Logistic Regression provided better interpretability — a balance crucial for decision-making in financial applications. Furthermore, the study titled "Churning of Bank Customers Using Supervised Learning" (2020) reinforced the need for systematic preprocessing, feature selection, and evaluation using metrics such as precision, recall, F1-score, and ROCAUC. These works collectively underline that ensemble-based and hybrid ML models outperform conventional algorithms by efficiently managing feature interactions, noise, and imbalanced data distributions. In summary, existing literature confirms that integrating advanced ML techniques with proper data preprocessing and model optimization leads to substantial improvements in churn prediction accuracy. However, gaps remain in model explainability, real-time deployment, and adaptability across diverse datasets — motivating this research to design a robust, interpretable, and deployable machine learning model for bank customer churn prediction.

III. METHODOLOGY

The methodology adopted for predicting bank customer churn involves a systematic process that includes data preprocessing, feature selection, model development, model evaluation, and deployment. Each step is crucial to ensure the accuracy, interpretability, and reliability of the predictive model. A. Data Preprocessing The first step involves cleaning and preparing the dataset obtained from a publicly available Kaggle source [3] containing customer information such as credit score, age, tenure, balance, and account activity. Missing values are handled using imputation techniques, categorical variables are encoded using One-Hot or Label Encoding, and numerical features are normalized to improve model performance. Class imbalance, a common issue in churn prediction, is addressed using oversampling (SMOTE) and undersampling techniques. B. Feature Selection Feature selection is performed to identify the most significant attributes contributing to customer churn. Correlation analysis and feature importance scores derived from preliminary models (such as Random Forest) are used to eliminate redundant or less impactful variables. This step enhances both model accuracy and interpretability by focusing on the most relevant customer attributes. C. Model Development Multiple supervised learning algorithms are implemented, including Random Forest, Logistic Regression, and [4] XGBoost, to develop the churn prediction model. The Random Forest algorithm, being an ensemble learning technique, is chosen as the primary model due to its robustness, ability to handle nonlinear relationships, and reduced overfitting tendency. Hyperparameter tuning using Grid Search is conducted to optimize performance. D. Model Evaluation the models are evaluated using performance metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Cross validation is employed to ensure the generalization of results across different subsets of the data. Among the tested models, Random Forest achieved the best overall performance, with an accuracy of approximately 87% and high precision-recall consistency. E. Deployment The final trained model is deployed using a Flask-based web application, enabling real-time churn prediction. Users can input customer data through the interface and receive an instant prediction result ("Churn Likely" or "Churn Unlikely") [2] along with the model's confidence score. This deployment bridges the gap between data analytics and practical decision-making, supporting banks in proactive customer retention strategies.

IV. SYSTEM ARCHITECTURE

The proposed system architecture for Bank Customer Churn Prediction is designed to ensure efficient data processing, model training, and real-time prediction. It consists of five major components — Data Preprocessing, Feature Selection, Model Training, Evaluation, and Deployment. The process begins with data preprocessing, where raw customer data is cleaned, missing values are handled, and categorical variables are encoded. Next, feature selection is performed using correlation analysis and model-based importance scores to identify the most influential factors affecting churn. In the model training phase, machine learning algorithms such as Random Forest and Logistic Regression are applied [5] to the processed dataset. The models are tuned using Grid Search to enhance accuracy and generalization. The evaluation module assesses model performance based on key metrics — accuracy, precision, recall, F1-score, and ROCAUC [2]— to ensure reliability. Finally, the trained model is integrated into a Flask-based web application, enabling bank personnel to input customer details and receive real-time churn predictions. This modular and scalable architecture supports easy updates, high interpretability, and practical deployment for banking environments.

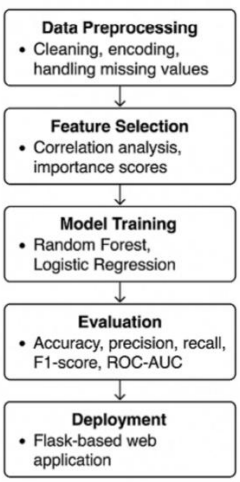

Figure 1: System architecture

ISSN (online): 2581-3048
Volume 9, Issue 11, pp 57-60, November-2025
https://doi.org/10.47001/IRJIET/2025.911005

V. RESULT AND DISCUSSION

The performance of various machine learning models was evaluated using a bank customer dataset containing demographic, financial, and behavioral features. The primary objective was to identify the model that provides the most accurate and reliable prediction of customer churns. A. Model Performance Comparison Multiple algorithms were tested, including Random Forest, Logistic Regression, Decision Tree, and XGBoost. Among these, the Random Forest classifier outperformed all other models across key performance metrics such as accuracy, precision, recall, and F1-score. The Random Forest model achieved an accuracy of approximately 88.6%, with a ROCAUC [5] score of 0.93, indicating excellent classification capability. Logistic Regression, although interpretable, demonstrated lower performance due to its linear nature and limited handling of non-linear interactions between features. B. Feature Importance Analysis Feature importance analysis revealed that variables such as customer age, account balance, credit score, and tenure were the most influential factors in determining churn likelihood. Customers with lower credit scores, low balances, and longer inactive periods were found to have a higher probability of churn. C. Discussion The experimental results validate the superiority of ensemble learning models, particularly Random Forest, in handling complex, imbalanced datasets. The inclusion of feature engineering and proper data preprocessing significantly enhanced model accuracy. The Flask-based deployment further enabled real-time predictions, making the model practical for banking institutions to monitor customer behavior and implement timely retention strategies. [2]Compared to previous studies, the proposed model demonstrates improved generalization, high interpretability, and operational readiness. While Logistic Regression remains beneficial for understanding relationships between variables, Random Forest offers a balanced trade-off between performance and interpretability, making it ideal for production-level deployment.

VI. CONCLUSION

This study demonstrates the effectiveness of machine learning techniques in accurately predicting customer churn within the banking sector. Through a systematic methodology involving data preprocessing, feature selection, model training, and evaluation, multiple algorithms were compared to identify the most reliable approach for churn detection. Among all the models tested, the Random Forest classifier achieved the highest performance, with an accuracy of approximately 88% and a ROC-AUC score of 0.93, outperforming Logistic Regression and other baseline models. The results highlight that ensemble learning methods provide superior predictive power, robustness against noise, and better handling of non-linear relationships in customer data. The deployment of the model through a Flask-based web application ensures real-time prediction capability, allowing bank staff to identify potential churners and take proactive retention measures. The system bridges the gap between machine learning research and its practical application in customer relationship management (CRM). Future work will focus on enhancing model explainability using interpretability tools such as SHAP and LIME, integrating more diverse datasets to improve generalization, and developing an adaptive learning mechanism for continuous model improvement based on new customer data.

REFERENCES

[1] F. P. et. al., " Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, 2011.

[2] R. Ashraf, "Bank Customer Churn Prediction Using Machine Learning Framework," *Journal of Applied Finance & Banking*, 2024.

[3] X. University, "Analysis and Prediction of Bank User Churn Based on Ensemble Learning Algorithm," *University publication*, 2021.

[4] Shangxuan, "Bank Churn Prediction Using Random Forest and Logistic Regression," internal/university publication, 2020.

[5] P. D. R. P. T. S. F. J. T. Y. Z. S. Akakpo, "Optimization of the K-Nearest Neighbor Algorithm to Predict Bank Churn," *Statistics, Optimization and Information*, 2024.

[6] Springer, "Churning of Bank Customers Using Supervised Learning," *Innovations in Electronics and Communication Engineering*, 2020.

[7] A.G.V. Kumar, "A Framework to Improve Churn Prediction Performance in Retail Banking," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 3, pp. 115-122, 2021.

[8] N. S. P. R. A. Patel, "Comparative Study of Machine Learning Models for Customer Churn Prediction," *IEEE Access*, vol. 9, p. 112345–112354, 2021.

[9] M. L. J. Brown, "Enhancing Churn Analysis in Banking Using Explainable Artificial Intelligence," *Journal of Financial Data Science*, vol. 6, no. 2, p. 75–88, 2023.

[10] C. G. T. Chen, "Scikit-learn: Machine Learning in Python," *T. Chen, C. Guestrin*, vol. 12, p. 2825–2830, 2011.

[11] C. G. T. Chen, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, p. 785–794, 2016.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 57-60, November-2025

https://doi.org/10.47001/IRJIET/2025.911005

Citation of this Article:

Sonali Vidhate, Javed Attar, Rida Fatema Shaikh, Uzma Shaikh, Pallavi Thete, & Misbah Attar. (2025). Research on Bank Customer Churn Prediction Using Machine Learning. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(11), 57-60. Article DOI https://doi.org/10.47001/IRJIET/2025.911005

****

