iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** II **Month of publication:** February 2025

DOI: https://doi.org/10.22214/ijraset.2025.67002

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

Machine Learning in Bank Customer Churn Prediction: Improving Accuracy through Feature Engineering

Jadyn Dias[1], Pallavi Tawde[2]
[1]*Student, Department of MSc. IT, Nagindas Khandwala College, Mumbai, Maharashtra, India*
[2]*Assistant Professor, Department of BSc. IT and CS, Nagindas Khandwala College, Mumbai, Maharashtra, India*

Abstract: ***Customer churn is a critical challenge in the banking sector, where retaining existing customers is more cost-effective than acquiring new ones. This study explores the role of feature engineering and model optimization in enhancing machine learning-based churn prediction. Using a dataset of bank customers, various preprocessing techniques of categorical encoding, feature scaling, and feature selection were applied. Model optimization was performed using hyperparameter tuning technique of Randomized Search CV to refine key parameters. Further, the models were trained with only the most influential features of customer complaints, number of products, age, account activity, and credit card ownership, resulting in improved computational efficiency and interpretability while maintaining strong predictive performance. Multiple machine learning models were evaluated, with XGBoost emerging as the most effective due to its ability to handle complex patterns and structured data efficiently. The findings provide actionable insights for banks to implement targeted retention strategies, ultimately reducing churn rates and improving customer engagement.***

Keywords: Customer Churn, Machine Learning, Feature Engineering, Model Optimization, Banking Sector

I. INTRODUCTION

Customer churn, also known as customer attrition, refers to the phenomenon where customers stop using a company's products or services over a given period. In highly competitive markets, retaining customers is a crucial challenge for businesses, as acquiring new customers often incurs significantly higher costs than maintaining existing ones. Consequently, companies across various industries, including banking, telecommunications, e-commerce, and subscription-based services, invest heavily in customer retention strategies. Understanding the reasons behind customer churn and predicting it accurately can help businesses take proactive measures to reduce attrition and enhance customer satisfaction. In the highly competitive banking sector, customer retention is crucial for long-term business sustainability. Banks and financial institutions continuously strive to minimize customer churn the phenomenon where customers stop using a bank's services since acquiring a new customer is often more costly than retaining an existing one. Effective churn prediction allows banks to identify at-risk customers early and implement targeted retention strategies. Traditional churn prediction methods relied on manual statistical analysis, which often lacked precision due to limited data processing capabilities. However, modern ML techniques can handle vast datasets, extract hidden patterns, and provide more accurate predictions. This paper explores how feature engineering and model optimization contribute to the efficiency of churn prediction models in the banking domain.

A. Importance of Feature Engineering in Churn Prediction

Feature engineering is a critical step in ML-based churn prediction, as raw data often contains redundant or irrelevant attributes that may negatively impact model performance. In the banking sector, customer data includes demographic details, transaction history, account activity, credit scores, and product usage patterns. Identifying the most relevant features ensures that ML models focus on meaningful insights, reducing noise and improving predictive power.

B. Role of Model Optimization in Churn Prediction

Once relevant features are engineered, model optimization plays a crucial role in maximizing the performance of ML algorithms. Different models exhibit varying capabilities in handling structured banking data, requiring hyperparameter tuning to achieve optimal results.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

C. *Objectives*

1) To identify key features influencing customer churn.
2) To optimize machine learning models for accurate churn prediction.
3) To compare model performance using various evaluation metrics.

II. REVIEW OF LITERATURE

Taware et al. (2022) explored the application of machine learning algorithms for customer churn prediction in the banking sector. The study employed Logistic Regression and Naive Bayes models to analyze factors such as age, location, gender, and credit card usage to predict potential customer attrition. The findings emphasize the significance of machine learning techniques in improving customer retention strategies within financial institutions. [1]

Jiang (2024) explored customer churn prediction in the banking industry using supervised machine learning techniques. The study employed XGBoost, Random Forest, Support Vector Machine (SVM), and AdaBoost to identify at-risk customers based on factors such as age, location, account balance, and credit score. The research found that XGBoost provided the highest predictive accuracy (86.55%) and emphasized the importance of features such as the number of banking products, active membership status, and customer geography. [2]

Xiahou et al. (2022) proposed a customer churn prediction model for B2C e-commerce using a combination of K-means clustering and Support Vector Machine (SVM) classification. The study segmented customers into three categories based on shopping behaviours and applied SVM to predict churn. The research highlights the importance of time-based behavioural data, particularly night and evening purchases, in identifying at-risk customers, providing valuable insights for e-commerce businesses to enhance customer retention strategies. [3]

Agarwal et al. (2024) conducted an exploratory data analysis study in the banking and finance sector, focusing on customer churn and credit card usage patterns. The research employed descriptive statistics, data visualization, and correlation analysis to identify transaction behaviours, demographic influences, and credit limit distributions among customers. Findings highlight the value of EDA in uncovering customer retention insights and optimizing banking services, emphasizing data-driven decision-making for enhancing financial institution performance. [4]

Imani et al. (2023) investigated the impact of hyperparameter optimization and data sampling techniques on customer churn prediction in the telecommunications sector. The study employed various machine learning models, including Artificial Neural Networks, Decision Trees, Support Vector Machines, Random Forests, Logistic Regression, XGBoost, LightGBM, and CatBoost. The findings highlight the significance of combining hyperparameter optimization with data sampling strategies to enhance churn prediction accuracy. [5]

Vaduva et al. (2024) investigated customer churn detection in the banking sector using machine learning models combined with probability calibration techniques. The study utilized a synthetic dataset and applied SMOTETomek to balance class distribution. Two models, Random Forest and Light Gradient Boosting Machine were trained and evaluated using performance metrics such as precision, sensitivity, F1-score, and Brier score. [6]

III. METHODOLOGY

A. *Dataset Description*

The dataset used in this research is a secondary dataset from Kaggle datasets which consists of customer-related attributes. The dataset comprises 10,000 records, with each row representing a unique customer and multiple features describing their banking activity, personal information, and financial engagement.

B. *Data Preprocessing*

1) Data Cleaning: The data cleaning phase included removing unnecessary columns namely RowNumber, CustomerId, and Surname. Additionally, missing values were handled appropriately.
2) Encoding Categorical Variables: LabelEncoder was used to convert categorical variables of Geography, Gender, Card Type into numerical format.
3) Feature Scaling: StandardScaler was used to standardize numerical features so that better model performance can be achieved.

C. Model Training and Evaluation

1) Splitting Data

The dataset was split into 80% training and 20% testing using train_test_split with stratify=y to maintain class balance.

2) Machine Learning Models:

a) Logistic Regression: used for classification which models the probability of a binary outcome using a logistic function.
b) Decision Tree Classifier: splits data based on feature conditions, creating a tree-like structure to classify outcomes.
c) Random Forest Classifier: ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.
d) Support Vector Classifier (SVC): finds the optimal hyperplane to separate classes by maximizing the margin between them.
e) XGBoost Classifier: boosting algorithm known for its efficiency and performance in structured data classification tasks.
f) LightGBM Classifier: efficiently handles categorical features and reduces computation time by implementing a histogram-based learning approach.
g) CatBoost Classifier: improves accuracy by addressing overfitting and handling missing values effectively.

3) Model Evaluation Metrics

Model evaluation metrics assess performance. Accuracy shows overall correctness, precision minimizes false positives, recall measures sensitivity, F1-score balances both, and ROC-AUC evaluates class distinction using the ROC curve.

D. Feature Importance Analysis

To understand the impact of different features on the model's predictions, feature importance was analyzed. The model assigns an importance score to each feature based on how frequently it is used for splitting nodes in decision trees. Higher-ranked features indicate strong influence, guiding further feature selection and engineering efforts.

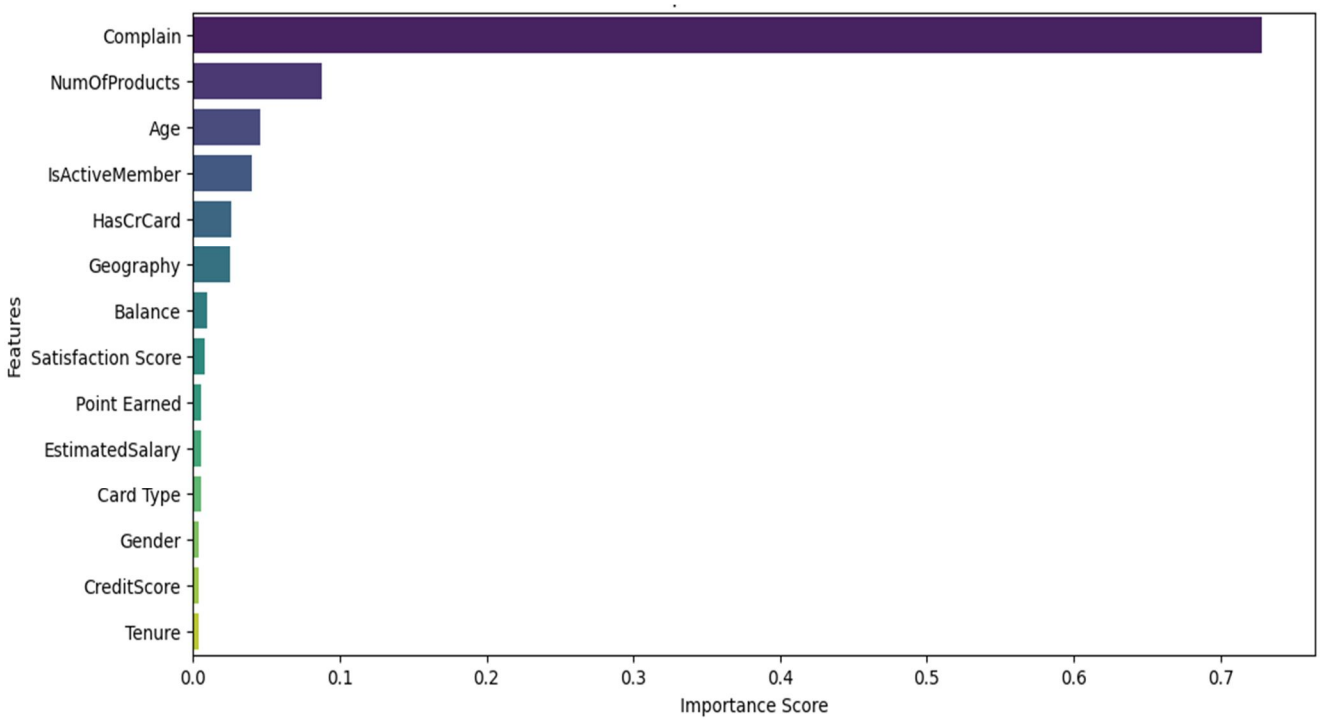


Figure 1: Important Features

The bar plot above illustrates the relative importance of different features in predicting customer churn. The x-axis represents the importance score, while the y-axis lists the features ranked by their contribution to the model's predictions. Features with higher scores have a greater impact on the model's decision-making process.

E. Key Influential Features and Their Impact

From the visualization, the most significant features influencing churn prediction include:

- Complain: This is the most dominant factor, indicating that customers who have lodged complaints are significantly more likely to churn. This suggests dissatisfaction as a major driver of customer attrition.
- NumOfProducts: The number of products a customer holds plays a crucial role, where customers with fewer products are more prone to leaving, possibly due to lower engagement with the company's services.
- Age: Older customers may be more likely to leave due to changes in financial priorities, product suitability, or switching to competitors offering better benefits.
- IsActiveMember: Less active members exhibit a higher probability of churn, emphasizing the importance of customer engagement in retention strategies.
- HasCrCard: Credit card ownership appears to influence churn likelihood, possibly linked to financial activity and customer spending behaviour.

1) Correlation Analysis of Complain and Churn

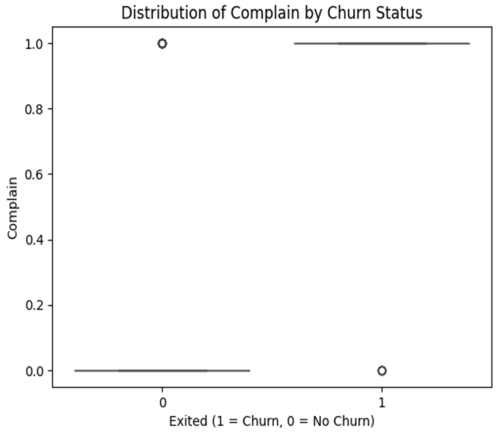


Figure 2: Complain and Churn

A significant portion of churned customers (Exited = 1) have filed complaints, highlighting a strong association. In contrast, most non-churned customers (Exited = 0) have not complained, suggesting higher satisfaction. Complaints strongly predict churn, likely due to dissatisfaction or unresolved issues. Addressing complaints promptly can help retain at-risk customers.

2) Correlation Analysis of Number of Products and Churn:

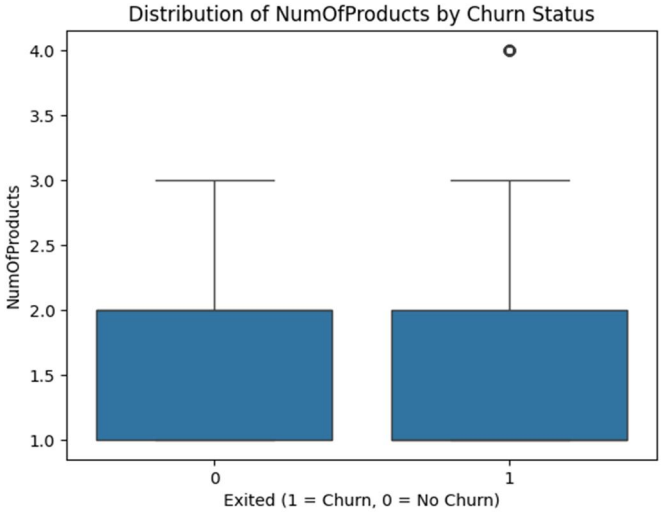


Figure 3: Number of Products and Churn

The median NumOfProducts is similar for both groups, with most customers holding 1 or 2 products. A small number of churned customers with 4 products appear as outliers. Most customers have 1 or 2 products, showing stability, while those with 4 products in the churned group suggest exceptions. This indicates that having multiple products does not guarantee retention. Customers with only one product are more likely to churn due to lower engagement. While 2-3 products provide stability, some high-product customers still leave, possibly due to dissatisfaction or better alternatives.

3) Correlation Analysis of Age and Churn:

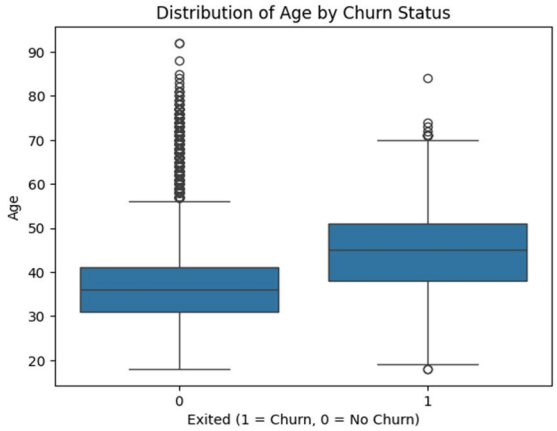


Figure 4: Age and Churn

Churned customers tend to be older, with a higher median age than non-churned customers. Younger customers, mostly aged 30–40, are more likely to stay. The age range for churned customers is wider, with both very young and very old individuals churning. While most elderly customers leave, some remain loyal, as seen in the non-churned outliers (60+ years). Older customers (40–50+) are more likely to churn, possibly due to retirement or reduced financial activity. Younger customers (30–40) are more likely to stay, though some exceptions exist.

4) Correlation Analysis of Is Member Active and Churn:

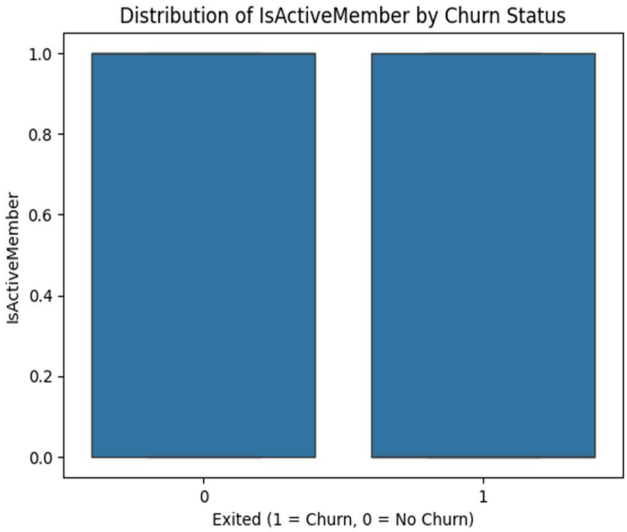


Figure 5: Is Member Active and Churn

The distribution of IsActiveMember is balanced across both churned and non-churned customers, suggesting it is not a strong churn predictor. Some churned customers were active members, indicating that activity alone does not prevent churn. The IsActiveMember variable shows weak correlation with churn, meaning that while customer engagement is important, it is not a decisive factor in predicting churn.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

5) Correlation Analysis of whether Customer has Credit Card and Churn:

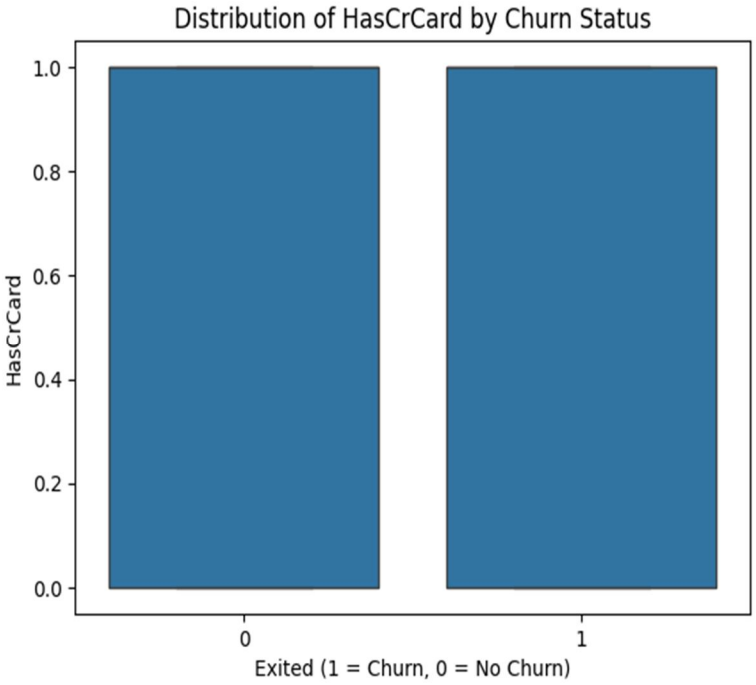


Figure 6: Has Credit Card and Churn

The correlation between HasCrCard and churn is weak, indicating that having a credit card does not significantly impact customer retention. The correlation value remains close to zero. The distribution of HasCrCard is nearly identical for both churned and non-churned customers. This suggests that credit card ownership is not a distinguishing factor in churn behaviour. Customers with and without credit cards churn at similar rates, unlike factors such as age or account balance. Having a credit card alone does not strongly influence churn decisions.

Training with Selected Features

The models were trained using only the top important features:

- Complain
- NumOfProducts
- Age
- IsActiveMember
- HasCrCard

F. Hyperparameter Optimization

Hyperparameter tuning is crucial in improving a model's performance by optimizing its parameters to achieve the best possible results. RandomizedSearchCV was employed to fine-tune the hyperparameters of the XGBoost classifier, allowing for efficient exploration of a wide search space while reducing computational cost.

Best Hyperparameters Identified

After running RandomizedSearchCV, the following optimal hyperparameters were selected:

Table 1: Best hyperparameters identified

Hyperparameter	n_estimators	learning_rate	max_depth	min_child_weight	subsample	colsample_bytree
Optimized Value	1000	0.3	7	1	0.5	0.8

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

IV. RESULTS

A. *Model Performance Evaluation:*

To assess the effectiveness of different machine learning models, we evaluated them using multiple performance metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC Score. The table below presents a comparative analysis of the results obtained for each model.

Table 2: Performance of models implemented

	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.9985	0.9975	0.9950	0.9963	0.9989
Decision Tree	0.9970	0.9950	0.9901	0.9926	0.9944
Random Forest	0.9985	0.9975	0.9950	0.9963	0.9981
SVM	0.9985	0.9975	0.9950	0.9963	0.9972
XGBoost	0.9985	0.9975	0.9950	0.9963	0.9974
LightGBM	0.9980	0.9950	0.9950	0.9950	0.9976
CatBoost	0.9985	0.9975	0.9950	0.9963	0.9976

B. *Overall Comparison of Models*

1) Consistency Across Models: Most models demonstrated strong performance, with accuracy values above 99.7%, indicating that the dataset and preprocessing steps were well-structured for classification tasks.

2) Logistic Regression, Random Forest, XGBoost, and CatBoost consistently ranked among the top-performing models, achieving the highest accuracy (0.9985) and balanced precision, recall, and F1-score.

3) Decision Tree Performance: The Decision Tree classifier exhibited the lowest accuracy (0.9970) and the lowest ROC-AUC score (0.9944), suggesting a relatively higher likelihood of overfitting or misclassification compared to ensemble methods.

4) LightGBM Performance: LightGBM achieved slightly lower accuracy (0.9980) than other ensemble models but maintained a balanced precision and recall, making it a competitive option for structured tabular data.

5) SVM and XGBoost: Both models performed exceptionally well, with high accuracy, precision, and recall, indicating their robustness in handling classification tasks.

C. *Best Model among the models implemented*

Among the models tested, XGBoost is the most effective due to the following reasons:

1) High Accuracy & Generalization: XGBoost achieved an accuracy of 0.9985, making it one of the best-performing models.

2) Balanced Performance Across Metrics: It maintained high precision (0.9975), recall (0.9950), and F1-score (0.9963), demonstrating a strong balance between correctly identifying positive cases and minimizing false negatives.

3) Superior Handling of Complex Data: XGBoost is an ensemble learning algorithm that excels in handling complex relationships within data, making it well-suited for the structured dataset used in this study.

4) Efficiency & Scalability: Compared to other models, XGBoost efficiently handles large datasets, minimizes overfitting through built-in regularization, and supports parallel processing, making it a practical choice for real-world applications.

V. CONCLUSION

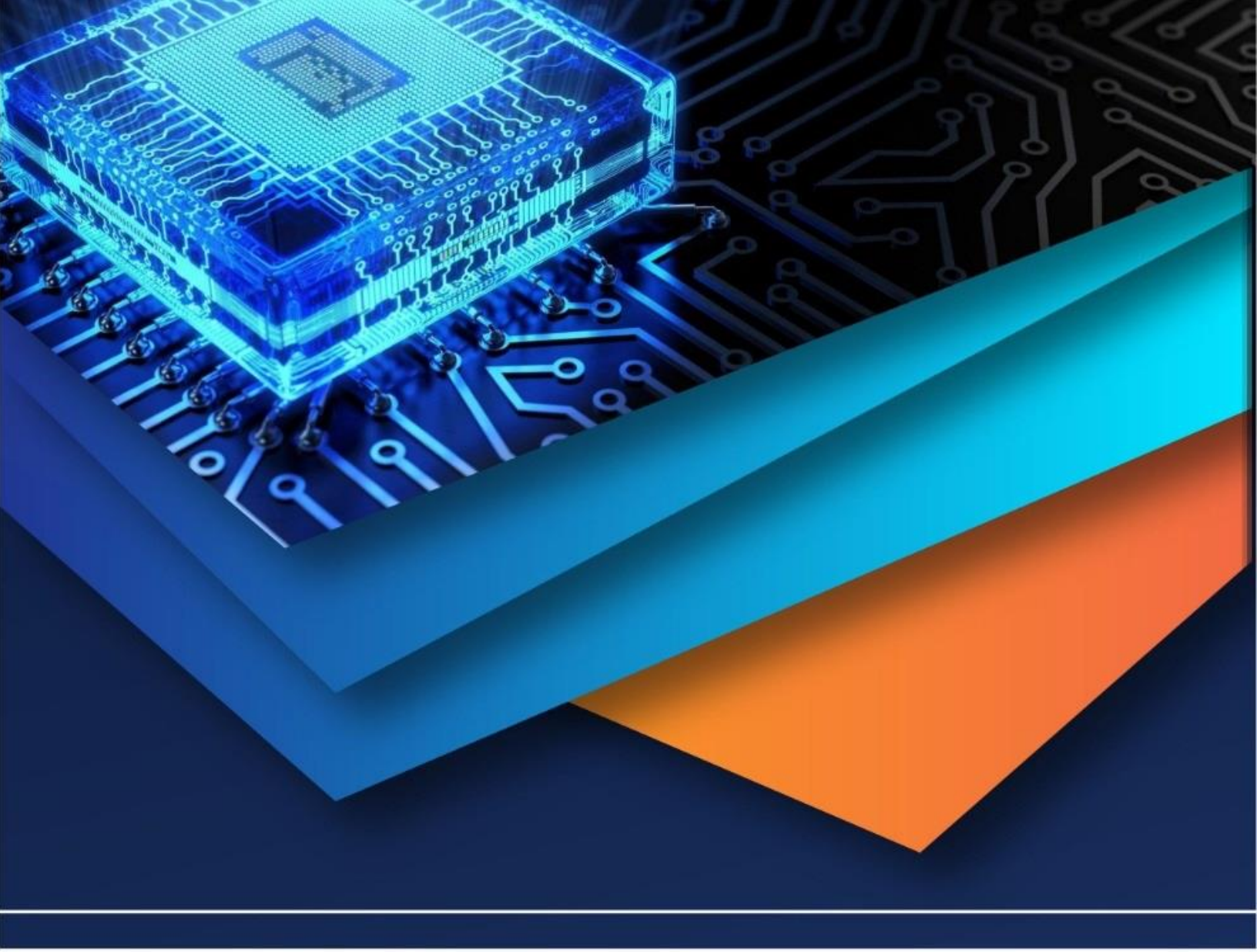
This study successfully achieved its objectives in predicting customer churn in the banking sector. Key features influencing churn of customer complaints, number of products, age, account activity, and credit card ownership were identified, providing actionable insights for retention strategies.

Machine learning models were optimized using RandomizedSearchCV for hyperparameter tuning and feature selection, improving computational efficiency while maintaining predictive performance. Model comparisons showed that XGBoost outperformed other classifiers, demonstrating strong precision, recall, and F1-score, making it the most effective for churn prediction.

These findings emphasize the role of feature engineering and optimization in building scalable, interpretable models for proactive customer retention. Banks can leverage these insights to minimize churn through targeted interventions. Future research can explore deep learning and real-time prediction techniques to further enhance churn modeling capabilities.

REFERENCES

[1] Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A. L. N., & Srivastav, V. K. (2022, October). Customer-Churn Prediction Using Machine Learning. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 893-899). IEEE.
[2] Jiang, S. (2024). Customer Churn Prediction In Banking Industries: Supervised Machine Learning Approach (Doctoral dissertation, UCLA).
[3] Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458-475.
[4] Agarwal, A., Prabha, S., & Yadav, R. (2024). Exploratory Data Analysis for Banking and Finance: Unveiling Insights and Patterns. arXiv preprint arXiv:2407.11976.
[5] Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis. Technologies, 11(6), 167.
[6] Văduva, A. G., Oprea, S. V., Niculae, A. M., Bâra, A., & Andreescu, A. I. (2024). Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques. Electronics, 13(22), 4527.
[7] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management, 7(1), 7-16.
[8] Bogaert, M., & Delaere, L. (2023). Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art. Mathematics, 11(5), 1137.
[9] de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. Neural Computing and Applications, 34(14), 11751-11768.
[10] Harbaugh, S. P. (2022). Predicting Retail Bank Customer Churn Using Push, Mooring, and Temporal Inputs with Machine Learning (Doctoral dissertation, Northcentral University).
[11] Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. International Journal of Intelligent Networks, 4, 145-154.
[12] https://www.kaggle.com/datasets/marusagar/bank-customer-attrition-insights



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY