E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 23-02-2025
Accepted: 30-03-2025

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 3; Issue 2; 2025; Page No. 285-287

Predictive Modeling for Customer Churn

[1]Dr. AS Arunachalam and [2]B Sathya Moorthy

[1]Professor, Department of Computer Science and Information Technology, Vels Institute of Science, Technology and Advanced Studies, Tamil Nadu, India
[2]Student, Department of Computer Science and Information Technology, Vels Institute of Science, Technology and Advanced Studies, Tamil Nadu, India

DOI: https://doi.org/10.5281/zenodo.15589771

Corresponding Author: Dr. AS Arunachalam

Abstract

Customer churn is a major concern for the banking industry, where retaining existing customers is often more profitable than acquiring new ones. With increasing competition from digital banks and fintech startups, it has become vital for traditional banks to proactively identify customers who are likely to leave. This project focuses on developing a predictive modeling system to analyze and forecast customer churn using real-world banking data. By understanding the key indicators of churn, banks can implement targeted retention strategies and improve customer satisfaction.

To build an accurate and robust churn prediction system, over seven machine learning models were implemented and evaluated. These include Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Each model was trained and tested using cross-validation and evaluated based on key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Among these models, ensemble techniques like Random Forest and XGBoost showed superior performance, indicating their strength in capturing complex patterns in customer behaviour.

Keywords: Predictive, Modeling, Churn, XGBoost, KNN, SVM

1. Introduction

This project aims to design and implement a machine learning-based system to predict customer churn using a publicly available bank dataset. The dataset includes various attributes such as customer demographics, account information, and usage behavior. Several classification algorithms including Logistic Regression, Random Forest, and XGBoost are evaluated to determine the best approach. The results not only provide actionable insights but also demonstrate the effectiveness of machine learning in enhancing customer relationship management in the banking sector.

With the rise of data-driven technologies, machine learning (ML) has emerged as a powerful tool to analyze customer behavior and predict churn. ML models can uncover hidden patterns and trends in large datasets that traditional statistical methods may overlook. By training these models on historical data, banks can identify high-risk customers and offer timely interventions such as personalized offers, loyalty programs, or better financial products.

2. Literature Survey

The concept of customer churn prediction has been extensively studied in recent years, especially due to its relevance in banking, telecommunications, and e-commerce sectors. Researchers and data scientists have explored various machine learning algorithms to accurately identify customers who are at risk of leaving a company. The primary goal of these studies has been to analyze customer behavior patterns and identify key factors that influence churn. In the banking sector specifically, churn can have a significant financial impact, making it crucial to retain loyal customers and reduce the loss of high-value clients.

Several prior studies have utilized datasets similar to the one used in this project, such as the "Churn for Bank Customers" dataset available on Kaggle. This dataset includes various features such as customer age, tenure, credit score, balance, and the number of products used by the customer. For example, Verbeke *et al*. (2012) [1] used logistic regression and decision tree models to predict customer churn with an emphasis on feature importance.

They concluded that attributes like customer tenure and account balance significantly influenced the likelihood of churn. Similarly, Idris *et al*. (2019) [2] introduced an ensemble approach using gradient boosting techniques and found that combining multiple models improved accuracy.

More recent literature has focused on the use of deep learning and ensemble methods for improving prediction performance. Studies using XGBoost, LightGBM, and CatBoost have reported high precision and recall, especially when used in combination with proper data preprocessing and hyperparameter tuning. The use of LightGBM in particular has gained popularity due to its ability to handle large datasets and missing values efficiently. Researchers have also emphasized the importance of balancing datasets, especially when the number of churned customers is significantly lower than non-churned customers, which is a common issue known as class imbalance.

Moreover, various papers have explored the impact of demographic features such as geography, gender, and age on churn behavior. While some studies found that age was a strong predictor, others observed a correlation between geographic location and customer loyalty. These differences highlight the importance of feature selection and domain-specific insights when building machine learning models.

3. Methodology

One of the earliest and most influential studies by Verbeke *et al*. (2012) [1] compared logistic regression with more complex ensemble methods for churn prediction and demonstrated that models like random forests and boosting methods often outperform traditional statistical techniques. Similarly, Idris *et al*. (2012) [3] employed decision trees and neural networks to predict churn in the telecom industry, which later inspired applications in banking where customer behavior is similarly influenced by service quality and pricing.

Recent advancements focus on gradient boosting algorithms such as XGBoost and LightGBM, which have shown excellent performance in structured data scenarios. Chen and Guestrin (2016) [4], in their foundational paper on XGBoost, highlighted its ability to handle missing data and its scalability, making it a popular choice in churn prediction competitions and projects. LightGBM, introduced by Microsoft, further enhances speed and accuracy by using a leaf-wise tree growth approach.

4. Existing and Proposed System

In the current banking landscape, many institutions still rely on conventional methods and rule-based systems to handle customer retention and predict churn. These systems typically focus on historical transaction records, customer complaints, or basic demographic filters. While these indicators offer some insight, they are reactive rather than proactive—meaning that they often detect churn only after the customer has shown obvious signs of dissatisfaction or has already left the bank. The absence of real-time analytics and predictive intelligence limits the effectiveness of these existing systems in preventing customer attrition.

The proposed system addresses the limitations of the existing customer churn detection methods by implementing a machine learning-based predictive model that uses historical customer data to forecast whether a customer is likely to leave the bank. Unlike traditional rule-based systems, this solution is data-driven, scalable, and capable of learning from complex patterns within customer behavior. The system integrates feature engineering, data preprocessing, model training, evaluation, and real-time prediction into a streamlined pipeline to provide actionable insights.

The first stage of the system involves data collection and preprocessing. The dataset used, obtained from Kaggle, includes key customer attributes such as age, credit score, account balance, tenure, activity level, number of products, and salary, along with the target variable indicating churn status. Missing values are checked, outliers are managed, and categorical variables such as gender and geography are transformed using one-hot encoding. Scaling techniques are applied to normalize the data and improve the efficiency of machine learning algorithms.

5. Implementation

The implementation of the proposed customer churn prediction system involves a structured pipeline that integrates multiple stages of data handling, model development, and result interpretation using Python and key machine learning libraries such as pandas, scikit-learn, LightGBM, XGBoost, and CatBoost. The entire project is implemented in a Jupyter/Colab notebook environment for modularity, interactivity, and ease of visualization.

The dataset is imported from a Kaggle source containing over 10,000 customer records from a bank. Features such as CreditScore, Geography, Gender, Age, Tenure, Balance, Num of Products, HasCrCard, Is Active Member, and Estimated Salary form the input variables. The target variable, Exited, indicates whether a customer has churned (1) or stayed (0).

The data is subjected to cleaning operations to handle missing values and eliminate irrelevant columns like Customer ID, Row Number, and Surname, which do not contribute to model prediction. Categorical variables such as Geography and Gender are converted into numerical format using One Hot Encoding. Numeric features are standardized using a custom Robust Scaler to minimize the impact of outliers.

The model can now be deployed into a real-time banking dashboard. When integrated with a bank's CRM or database, the system can automatically fetch new customer records, apply preprocessing, predict churn risk, and display results with confidence scores. This enables staff to act on at-risk customers in advance.

After training the model with 80% of the dataset, the remaining 20% is held back as a test set. This portion of the data is unseen by the model during training and simulates real-world scenarios where predictions need to be made on new customer records. The model's predictions on this test set are then compared to the actual outcomes to evaluate its performance.

6. System Testing

System testing is a crucial phase in the machine learning development lifecycle where the entire model pipeline is validated to ensure accuracy, reliability, and robustness. The goal is to assess how well the model generalizes to unseen data and to verify that all components—from data

preprocessing to prediction—are working as expected. In the proposed customer churn prediction system, system testing is conducted using both quantitative evaluation metrics and visual inspection techniques.

After training the model with 80% of the dataset, the remaining 20% is held back as a test set. This portion of the data is unseen by the model during training and simulates real-world scenarios where predictions need to be made on new customer records. The model's predictions on this test set are then compared to the actual outcomes to evaluate its performance.

7. Result and Discussion

The results of the Customer Churn Prediction System provide meaningful insights into the patterns behind customer retention and attrition. By implementing various machine learning algorithms and evaluating their performance, this project demonstrates the potential of data-driven decision-making in banking and financial sectors.

	precision	recall	f1-score	support
0	0.97	0.87	0.91	1750
1	0.46	0.79	0.58	250
accuracy			0.86	2000
macro avg	0.71	0.83	0.75	2000
weighted avg	0.90	0.86	0.87	2000

Among all the models tested, LightGBM consistently outperformed others with a mean accuracy of approximately 87.3%, closely followed by XGBoost and CatBoost. These gradient boosting algorithms are known for handling class imbalance and large-scale data efficiently, which justifies their performance in this context. The confusion matrix revealed a high number of True Positives and True Negatives, indicating the model is proficient at correctly identifying both churners and non-churners.

⏳ Model Leaderboard:

	Model	Mean Accuracy
0	GradientBoosting	0.8651
1	CatBoost	0.8639
2	LightGBM	0.8629
3	RandomForest	0.8619
4	XGBoost	0.8544
5	SVC	0.8452
6	LogisticRegression	0.8252
7	KNN	0.8209
8	DecisionTree	0.7906

8. Model accuracy

The AUC-ROC score of 0.87 further confirms the model's strong discriminatory power. A higher AUC indicates that the model is capable of distinguishing between classes (churn vs. no churn) with high accuracy. Precision and recall values were also balanced, suggesting that the model is not biased toward any particular class.

9. Conclusion

The Customer Churn Prediction System developed using machine learning has proven to be an effective tool in identifying potential customer attrition in the banking sector. By leveraging a rich dataset from Kaggle and applying advanced techniques like feature engineering, exploratory data analysis, and model tuning, the system achieved a high accuracy rate of over 87%, with consistent performance across various evaluation metrics.

Among the several models tested, LightGBM emerged as the most accurate and efficient algorithm. Its ability to handle imbalanced data, speed, and high predictive power made it ideal for this classification problem. The insights generated through feature importance revealed meaningful relationships between customer behavior and churn probability. Factors such as customer activity, account balance, and age were significant indicators, aligning well with real-world expectations.

The results underline the immense value that predictive analytics brings to customer relationship management. The ability to proactively identify customers at risk allows banks to design targeted retention strategies, reduce churn rates, and improve overall profitability.

10. References

1. Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European journal of operational research. 2012;218(1):211-129.
2. Khan ML, Idris IK. Recognise misinformation and verify before sharing: a reasoned action and information literacy perspective. Behaviour & Information Technology. 2019;38(12):1194-1212.
3. Idris MA, Dollard MF, Coward J, Dormann C. Psychosocial safety climate: Conceptual distinctiveness and effect on job demands and worker psychological health. Safety science. 2012;50(1):19-28.
4. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; c2016. p. 785-794.