Understanding Customer Retention Patterns in Retail Banking Using Data-Driven Approaches

V. Manikyala Rao, P. Akhila , Bodanapu Jyothiswaroop
G Dileep Kumar, Jaajala Snehasri

Department of Computer Science and Engineering
(Artificial Intelligence and Machine Learning)
B V Raju Institute of Technology, Narsapur, India
bvrit@bvrit.ac.in,
WWW home page: https://bvrit.ac.in//
Contributing authors: manikyalarao.v@bvrit.ac.in, akhila.p@bvrit.ac.in,
bodanapu.j.swaroop@gmail.com, dileep89787@gmail.com,
jaajalasnehasri14@gmail.com

Abstract. Customer churn has been one of the biggest concerns for retail banks, since it is much costlier to acquire a fresh customer compared to holding on to an existing one. This paper offers a data-intensive explanation for understanding customer retention trends and machine learning models for churn predictions. A systematic approach has been adopted on a retail banking dataset holding close to 10,000 customer data points, summarized with demographics, behavior, and financial parameters. An exploratory data analysis has been carried out on the dataset to identify important features for determining customer churn, showing robust correlations with age, account engagement, balance amounts, and product utilization factors. Various machine learning models, such as Logistic Regression, Decision Trees, Random Forest, and XGBoost models, have been compared for their efficiency in churn predictions using Recall, F1 Score, and ROC-AUC Score, showing better results for churn predictions using the Random Forest model. Additionally, it has been able to yield predictive outputs indicating customer churn in terms of low-risk, grey, and high-risk customers. Eventually, interactive visualization dashboards in Power BI tools have been prepared for business-level decision-making based on analytical interpretations, improving strategic planning, and customer handling in retail banking operations ...

Keywords: Customer Churn Analysis, Retail Banking, Exploratory Data Analysis, Machine Learning, Churn Prediction, Risk Segmentation, Business Intelligence, Power BI

1 Introduction

Customer retention has become an important factor for profitability in the highly competitive retail banking industry because associated clients have easy access to digital banking services, leading to customer churn, which has become common

for financial institutions. Customer retention is more cost-effective compared to customer acquisition, according to research, emphasizing its importance in detecting customer churn at its initial stages for intervention and interaction with clients.

The traditional methods for managing customer churn involve the use of retroactive reporting and/or business rules, which are not capable of recognizing the inherent sophisticated behavior patterns hidden in the customer dataset. The emerging trend of obtaining large banking datasets has provided the possibility of the application of data intelligence techniques, namely exploratory data analysis and machine learning methods, for uncovering the underlying behavior patterns of customer churn decisions.

Machine learning provides the capability for evaluating multiple features of the customer, including demographic information and usage trends, to determine the predicted probabilities of churn. Most approaches reported in current literature focus on increasing the model accuracy. However, accuracy is less about meeting the particular demand for churn prevention than recall. In particular, because missing predictions for actual churners can result in expensive lost revenue, recall requirements may be more important than accuracy. Probabilities can be valued less than hard predictions.

In this study, a holistic framework for churn analysis in retail banking has been proposed, incorporating EDA, various machine learning algorithms, and risk segmentation based on probabilities. In contrast to simply predicting churn occurrence, in this work, customers can be grouped into various risk areas, helping in early detection for retention purposes. Moreover, Power BI tools have been utilized to provide business-ready insights based on analytical outputs, helping in decision-making and managing business relationships. In banking today, how various data-driven techniques can be used to better retain customers has been well illustrated in the proposed framework.

2 Related Work

Customer churn prediction has been widely researched in the banking and financial services sector due to the critical implications of the results on the profit margins and customer lifetime value of the organizations. For the initial research on customer churn prediction, the techniques used included statistical and rule-based methods such as decision rules and logistic regression. Even though the results were interpretable, the technique was limited because they do not handle the nonlinear relationships in the complex customer data.

However, with the advancement in the machine learning algorithm, different researchers have attempted to apply supervised machine learning algorithms to the churn modeling task in the retail banking sector. Logistic regression has been cited as the base algorithm on account of its interpretability and simplicity. However, different researchers have pointed towards the superior performance capabilities of tree-based machinelearning algorithms like decision trees and random forest as compared to linear algorithms in the task of modeling relationships

among the behavior of different customers. Ensemble algorithms like the Random Forest and Gradient Boosting Algorithm have shown superiority based on ROC AUC values and/or the recall measure.

Recent research has pointed out that accuracy is not enough when it comes to the evaluation of churn prediction models, as the churn dataset is usually imbalanced, with fewer customers churning than being retained. The meaningful metrics in these scenarios would then be recall, F1-score, and ROC-AUC, since these measures put their focus on the correct identification of customers who are most likely to leave. Research indicates that focusing on recall allows a reduction in false negatives, which refer to those customers who churn without being detected; this could mean potential loss in revenue for banks.

Aside from predicting churn, there has been an increased interest in customer segmentation, which offers better targeting and more actionable decisions. Some recent studies adopted clustering methods to group customers by behavior, while others used the output-the churn probability scores-of machine learning models to divide customers into categories of risk. However, most of these works focused primarily on enhancing predictive performance and did not provide a structured approach that linked risk segmentation to practical business use.

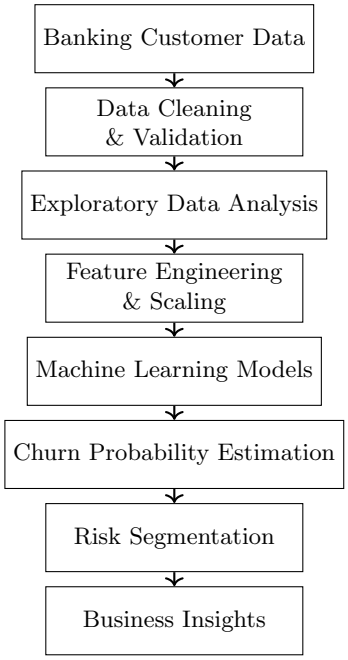
Despite mention of business intelligence systems and dashboards in some existing studies, very less evidence exists on complete integration of outcomes of machine learning techniques with interactive visualization systems in order to make a concrete decision support system. As a consequence, there is a gap in decision-making in business, as analytical insights are not related to concrete business decisions. For this reason, this study will propose a complete system that integrates Exploratory Data Analysis, different machine learning models, probability-driven risk segmentation, and Power BI systems in order to identify a possible risk of consumers churning in retail banking.

Table 1: Summary of related work in banking churn prediction

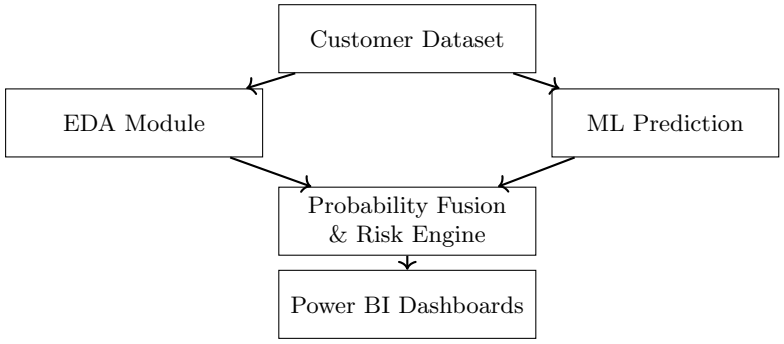
Study Focus	**Techniques Used**	**Key Findings**	**Limitations**
Traditional churn analysis	Logistic regression, rule-based methods	Identified basic churn drivers	Limited handling of non-linear patterns
ML-based churn prediction	Decision tree, Random forest, XGBoost	Improved recall and ROC-AUC	Focused mainly on prediction accuracy
Metric-oriented studies	Recall, F1-score, ROC-AUC	Highlighted importance of recall	Limited business integration
Customer segmentation	Clustering, probability thresholds	Enabled risk-based grouping	Often disconnected from prediction models
BI and visualization	Statistical reports, dashboards	Improved interpretability	Rare integration with ML outputs
Proposed approach	EDA + ML + probability-based risk zones + Power BI	Early churn detection and actionable insights	Designed for retail banking context

3 Methodology

This section presents the proposed end-to-end framework for analyzing customer churn in retail banking. The methodology integrates data preprocessing, exploratory data analysis, supervised machine learning models, probability-based risk segmentation, and business intelligence visualization to support proactive retention strategies.



(a) Overall System Flow



(b) System Architecture

Fig. 1: Flow and architecture diagrams of the proposed customer churn analysis framework

3.1 System Overview

The proposed framework follows a structured format that begins with raw banking customer data and ends with informative business insights. The workflow includes data preprocessing, exploratory analysis, feature engineering, churn prediction using multiple machine learning models, churn probability estimation, customer risk segmentation, and visualization using Power BI dashboards.

3.2 Data Preprocessing

Let the dataset be represented as:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in R^m$ denotes the feature vector of the i^{th} customer and $y_i \in \{0, 1\}$ represents the churn label, with 1 indicating customer churn.

Identifier attributes such as *RowNumber*, *CustomerId*, and *Surname* are removed to prevent information leakage. The dataset is examined for missing values, duplicate records, and data type inconsistencies. No missing values are observed.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed to identify patterns and relationships between customer attributes and churn behavior. Both univariate and bivariate analyses are conducted on key features such as age, balance, credit score, activity status, and number of products.

For a numerical feature X_j, the mean and standard deviation are computed as:

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, \quad \sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \mu_j)^2}$$

Feature relationships are further analyzed using correlation analysis to identify churn-influencing variables.

3.4 Feature Engineering and Scaling

Categorical features such as gender and geography are transformed using label encoding. Numerical features are standardized using z-score normalization to ensure uniform feature contribution during model training:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

3.5 Churn Prediction Models

Multiple supervised machine learning models are trained and evaluated to predict customer churn.

Logistic Regression Logistic Regression estimates churn probability using the sigmoid function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

where β represents the model coefficients.

Decision Tree Classifier Decision Trees recursively split the feature space based on impurity reduction. The Gini Index is used to measure node impurity:

$$Gini = 1 - \sum_{k=1}^{C} p_k^2$$

where p_k denotes the probability of class k.

Random Forest Random Forest is an ensemble learning technique that constructs multiple decision trees using bootstrap sampling and random feature selection. The final prediction is obtained through majority voting:

$$\hat{y} = mode\{h_t(x)\}_{t=1}^{T}$$

where h_t represents the t^{th} decision tree and T is the total number of trees.

Random Forest is selected as the final model due to its superior recall and ROC-AUC performance, making it suitable for churn prediction tasks.

XGBoost XGBoost optimizes an additive objective function defined as:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k)$$

where $l(\cdot)$ is the loss function and $\Omega(\cdot)$ represents the regularization term.

3.6 Model Evaluation Metrics

Model performance is evaluated using standard classification metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are used to assess discriminative performance.

3.7 Churn Probability Estimation

Instead of relying solely on binary predictions, the final Random Forest model outputs churn probabilities:

$$P_i = P(y_i = 1 \mid x_i)$$

which provide a continuous measure of churn risk for each customer.

3.8 Risk-Based Customer Segmentation

Customers are segmented into risk categories based on churn probability thresholds:

$$Risk(i) = LowRisk, \quad P_i < \theta_1 GrayZone, \quad \theta_1 \leq P_i < \theta_2 HighRisk, \quad P_i \geq \theta_2$$

This segmentation enables early identification of churn-prone customers and supports targeted retention strategies.

3.9 Business Intelligence and Visualization

The churn probabilities, risk segments, and feature importance results are integrated into interactive Power BI dashboards. These dashboards provide business-ready insights and support strategic decision-making by enabling stakeholders to monitor churn trends and design proactive customer retention actions.

4 Results

Overall, the results demonstrate that customer churn in retail banking is strongly influenced by demographic, behavioral, and product-related factors. Exploratory analysis reveals higher churn rates among older and inactive customers, while machine learning evaluation confirms that ensemble models provide improved recall.

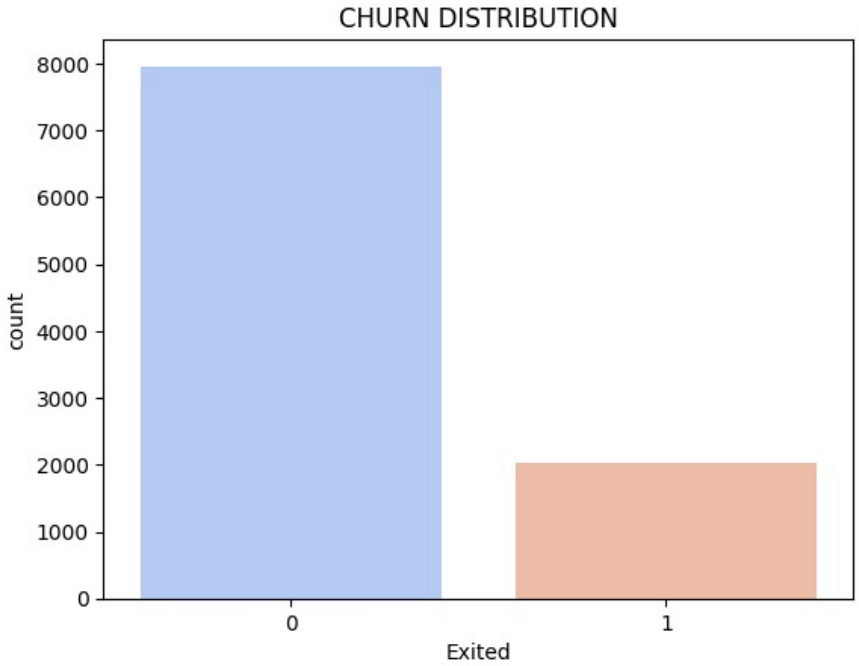


Fig. 2: Distribution of churned and retained customers

Figure 2 shows the distribution of customers based on churn status. It is observed that retained customers significantly outnumber churned customers, indicating a clear class imbalance in the dataset. This motivates the use of recall-focused evaluation metrics in churn prediction models.

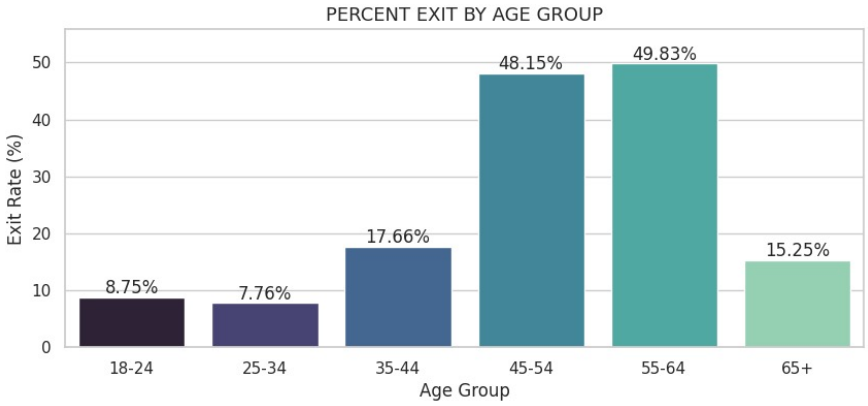


Fig. 3: Exit rate across different customer age groups

Figure 3 illustrates that churn rates increase significantly with age, particularly for customers between 45 and 64 years, highlighting age as a strong demographic factor influencing customer churn.

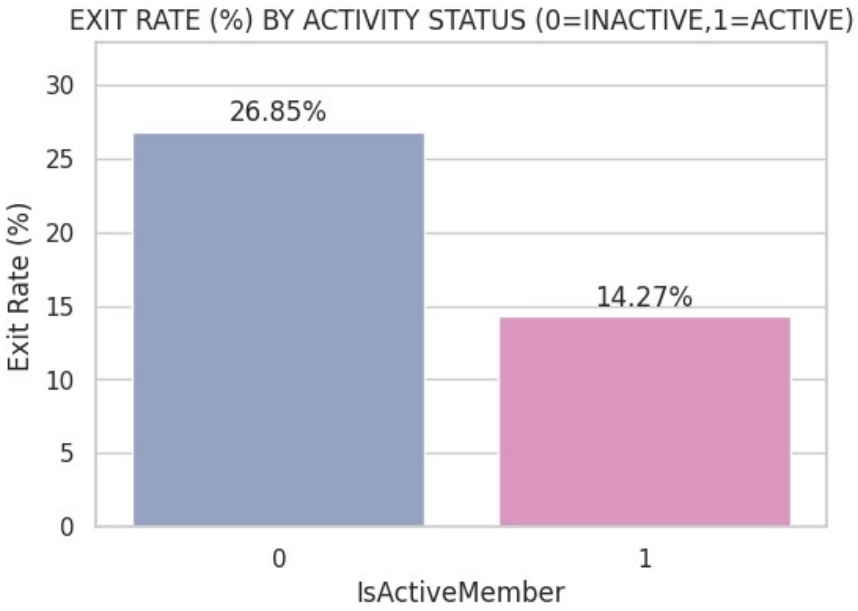


Fig. 4: Exit rate by customer activity status

Figure 4 compares churn behavior between active and inactive customers. Inactive customers exhibit a substantially higher exit rate, indicating that customer engagement plays a critical role in retention.

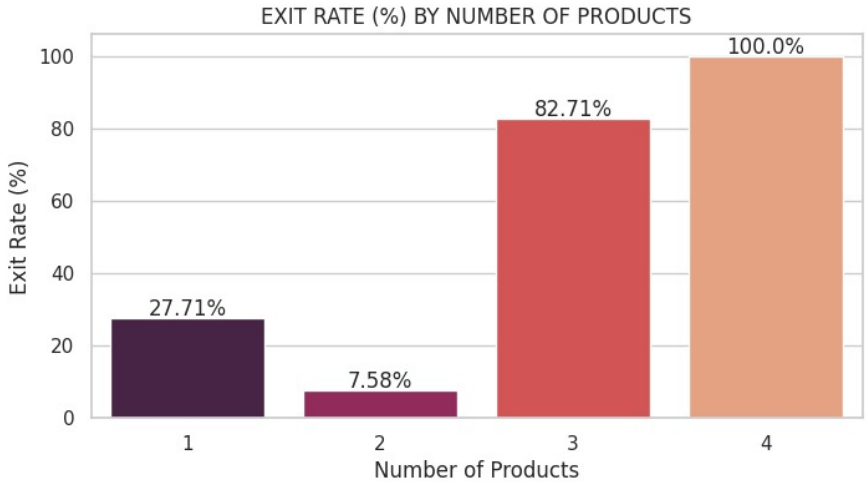


Fig. 5: Exit rate based on number of banking products held

Figure 5 shows a non-linear relationship between product ownership and churn behavior. Customers with either very low or very high numbers of products demonstrate higher exit rates, while moderate product usage is associated with better retention.

	model	accuracy	precision	recall	f1	roc_auc
0	Random Forest	0.8625	0.782051	0.449631	0.570983	0.854673
1	XGBoost	0.8490	0.682927	0.481572	0.564841	0.832834
2	Decision Tree	0.7580	0.445078	0.766585	0.563177	0.830742
3	Logistic Regression	0.7135	0.387228	0.700246	0.498688	0.777138

Fig. 6: Performance comparison of machine learning models

Figure 6 compares the performance of multiple machine learning models using accuracy, precision, recall, F1-score, and ROC-AUC. Random Forest achieves a favorable balance between recall and ROC-AUC, making it suitable for early identification of churn-prone customers.

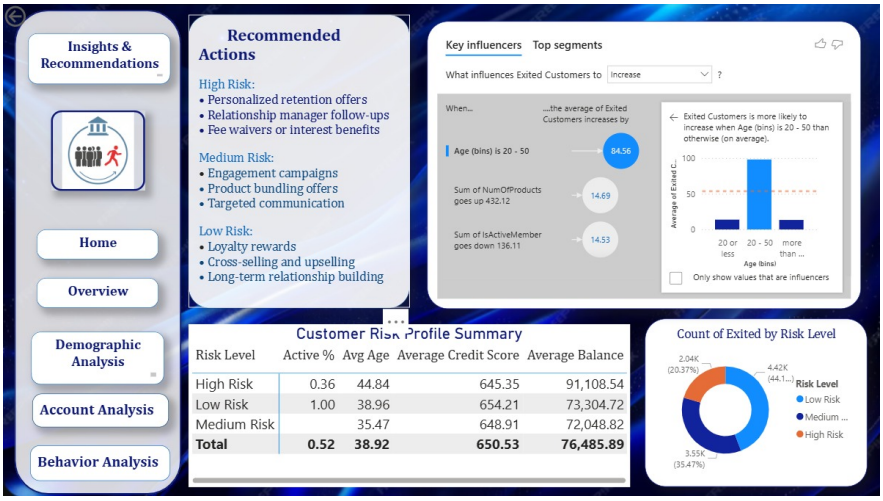


Fig. 7: Power BI dashboard for churn risk segmentation and actionable insights

Figure 7 presents an integrated business intelligence dashboard that combines churn probability, risk segmentation, and key influencing factors. The dashboard supports data-driven decision-making by enabling targeted retention strategies for high- and medium-risk customer segments.

5 Limitations

Although the proposed framework for customer churn analysis is proved to be effective, there are some limitations that need to be mentioned as well. The problem tackled in this study only involves one dataset from retail banks, which might affect the generalization capability of the analysis to other banking organizations. The data used in this study are historical data according to customer behavior, in which the dynamic data related to customer interactions are not reflected in real time, which might affect churn behavior too. Although the result obtained by the Random Forest classifier has high values in terms of accuracy, the black-box nature of this method might make it less explainable in data analysis compared to other models that require fewer computations. The Power BI tools used in this study are based on data that are in static form, meaning the data are not related to real churn analysis in real time either.

6 Conclusion

In the proposed paper, a comprehensive data-intensive methodology is presented for CH visualization and forecasting to the retail banking industry by using exploratory data analysis techniques, machine learning models, and business intelligence tools and techniques. It was proved that CH visualization using a proposed methodology highlighted that customer age, status, activity, and balance are key factors for CH visualization or CH occurrence. Among various models used for CH visualization and accuracy regarding recall and ROC-AUC value, the random forest model is the best model and hence can also be effectively used for predictive modeling for CH visualization/occurrence. By using CH visualization and risk-related CH segmentation, proactive CH prevention and retention are quite possible and feasible. Also, by using business intelligence tools and techniques such as Power BI tools and techniques, data display for improved business intelligence can be effectively done.

References

1. A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financial Innovation*, vol. 2, no. 10, pp. 1–18, 2016.
2. C. Luong, N. Luong, T. Tran, *et al.*, "Application of machine learning techniques for customer churn prediction in the banking sector," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 20, pp. 165–186, 2025.
3. A. Dal Pozzolo, G. Bontempi, and M. Snoeck, "Rebalancing strategies for customer churn prediction in banking environments," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 68–75, 2015.
4. J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.
5. R. Prakash, *Customer Churn Prediction in Retail Banking Using Machine Learning Techniques*, Master's thesis, National College of Ireland, Dublin, Ireland, 2023.
6. Y. Lu, *et al.*, "Propension to customer churn in a financial institution: a machine learning approach," *PLOS ONE*, vol. 17, no. 3, e0264823, 2022.
7. Microsoft Corporation, "Power BI documentation: Data visualization and business intelligence for decision support," Available: https://learn.microsoft.com/power-bi. Accessed: 2024.