Stochastic Modelling and Computational Sciences

PREDICTING CUSTOMER CHURN IN E-BANKING USING MACHINE LEARNING TECHNIQUES: AN ENSEMBLE APPROACH

Dr. Navneet Kaur
Professor, SGTBIMIT
drnavneet.sgtbimit@gmail.com

ABSTRACT

Customer churn prediction is critical for e-banking institutions aiming to retain clients and remain competitive. This study evaluates the effectiveness of various machine learning models—Logistic Regression, Random Forest, XGBoost, and CatBoost—in predicting customer churn on a real-world e-banking dataset. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Results indicate that CatBoost, combined with SMOTE, achieves the best balance of accuracy and recall for identifying churners. Feature importance analysis reveals that age, balance, and estimated salary are key predictors of churn. The findings provide actionable insights for e-banking customer retention strategies.

1. INTRODUCTION

In the digital era, e-banking has transformed the financial sector, offering convenience and accessibility to customers. However, the ease of switching providers has intensified competition, making customer retention a strategic priority. Predicting customer churn—the likelihood that a customer will leave the bank—is essential for proactive retention efforts. Traditional statistical methods often fall short in capturing the complex patterns underlying customer behavior. This paper explores the application of advanced machine learning (ML) techniques, with a focus on handling class imbalance, to improve churn prediction in e-banking.

2. WHY THIS STUDY

Customer churn poses a significant challenge for e-banking institutions, directly impacting profitability and long-term sustainability. In the highly competitive digital banking landscape, retaining existing customers is more cost-effective than acquiring new ones, making churn prediction a strategic priority. However, predicting churn is complex due to the multifaceted nature of customer behavior and the inherent imbalance in churn datasets, where churners form a minority.

Traditional statistical techniques such as logistic regression often fail to capture nonlinear relationships and complex interactions in the data, leading to suboptimal predictive performance. Moreover, class imbalance causes models to be biased toward the majority class, reducing the ability to identify churners effectively.

This study aims to:

- Evaluate and compare the predictive performance of multiple machine learning models, including Logistic Regression, Random Forest, XGBoost, and CatBoost, on a real-world e-banking churn dataset.
- Investigate the impact of Synthetic Minority Over-sampling Technique (SMOTE) in addressing class imbalance and improving model recall for churners.
- Provide actionable insights by identifying key features influencing churn, enabling banks to design targeted retention strategies.

By systematically benchmarking these models and methodologies, this research contributes to both academic knowledge and practical applications in customer retention within the e-banking sector.

3. LITERATURE REVIEW

Customer churn prediction has been a major research focus across industries such as telecommunications, retail, and banking. Early approaches relied on statistical models like logistic regression and decision trees (Huang et al.,

2012), but the rise of ensemble methods and deep learning has significantly improved predictive performance (Ebrah & Elnasir, 2019).

Recent studies emphasize the importance of addressing class imbalance, as churners typically represent a minority of the customer base. Techniques such as SMOTE and its variants are widely adopted to synthesize minority class samples and improve recall (Chawla et al., 2002; Musunuri, 2023). Ensemble models, such as Random Forest, Gradient Boosting, XGBoost, and CatBoost, have demonstrated superior performance in churn prediction tasks due to their ability to model complex, non-linear relationships and handle heterogeneous data (Suh, 2023; Pegah-Ardehkhani, 2023).

He et al. (2024) proposed a novel ensemble-fusion model, outlining the importance of end-to-end systems that integrate data collection, preprocessing, model construction, and intelligent deployment. Their work highlights the value of combining multiple models and advanced preprocessing to achieve robust, scalable churn prediction systems. Wagh et al. (2024) demonstrated the effectiveness of decision trees and ensemble models in the telecom sector, emphasizing the importance of visualization and interpretability for business adoption.

Moreover, business-focused literature stresses the need for actionable insights, not just predictions. Accurate churn prediction enables personalized retention campaigns and better resource allocation (RevPartners, 2025; Contentsquare, 2025). Explainable AI tools like SHAP and LIME are increasingly used to interpret model predictions and identify key churn drivers (Peng et al., 2023; Neptune.ai, n.d.).

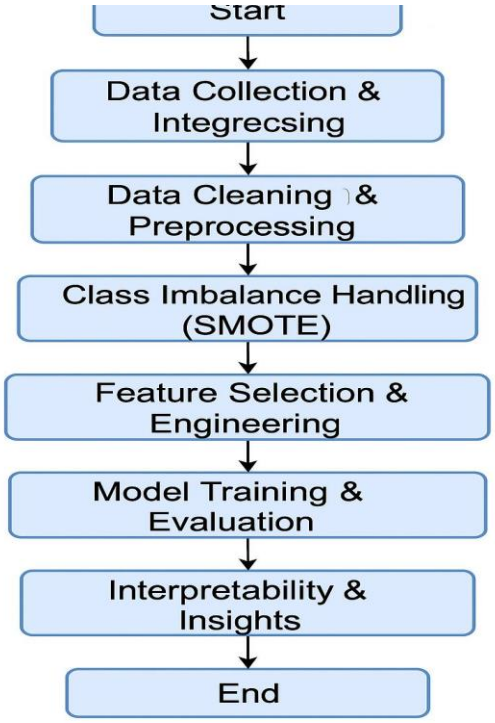
4. PROPOSED METHODOLOGY

4.1 Overview

The proposed methodology is an end-to-end pipeline for customer churn prediction in e-banking, leveraging ensemble machine learning models and advanced data preprocessing. The process consists of the following stages:

1. Data Collection & Integration: Aggregation of customer demographic, behavioral, and transactional data from e-banking systems.
2. Data Cleaning & Preprocessing: Handling missing values, encoding categorical variables, and feature scaling.
3. Exploratory Data Analysis (EDA): Statistical analysis and visualization (e.g., pie charts, bar graphs) to understand churn distribution and feature relationships.
4. Class Imbalance Handling: Application of SMOTE to balance the dataset.
5. Feature Selection & Engineering: Identification and creation of relevant features using correlation analysis and domain knowledge.
6. Model Training & Evaluation: Training multiple models (Logistic Regression, Random Forest, XGBoost, CatBoost), hyperparameter tuning, and evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
7. Interpretability & Insights: Feature importance analysis and visualization for business interpretability.
8. Deployment & Monitoring: Integration of the best-performing model into business processes, with ongoing monitoring and refinement.

4.2 Flowchart



4.3 Mathematical Formulation

The key evaluation metrics are defined as:

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

$$[\text{Recall} = \frac{TP}{TP + FN} \quad \text{Precision} = \frac{TP}{TP + FP}]$$

$$[\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}]$$

where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

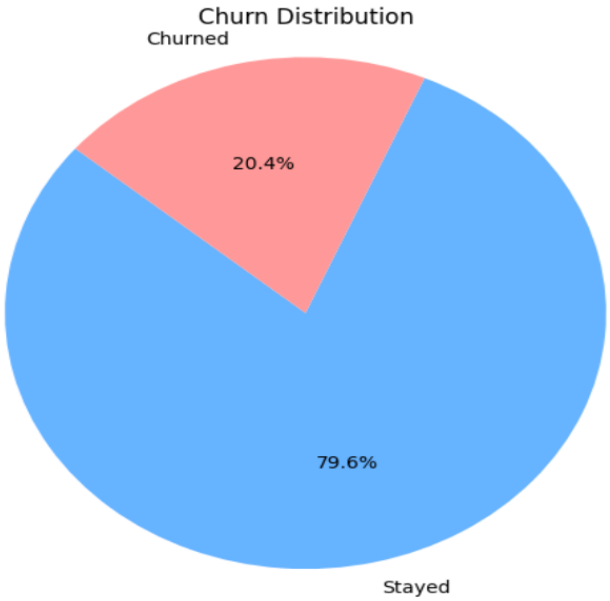
5. RESULTS

5.1 Model Performance

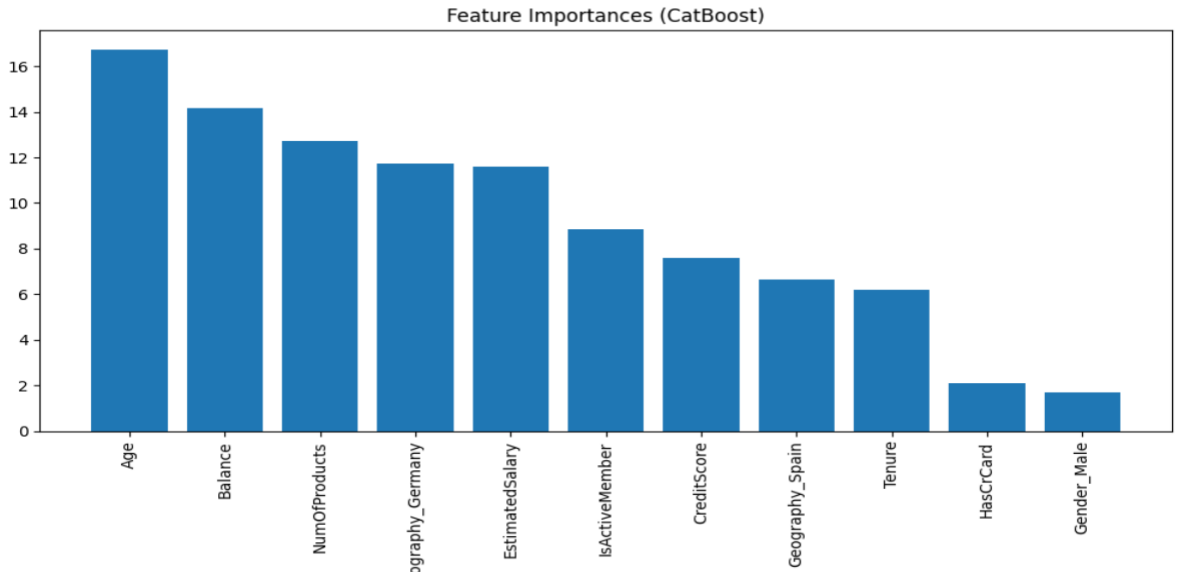
Model	Accuracy	Recall (Churned)	F1-score (Churned)
LogisticRegression	0.645	0.68	0.44
RandomForest	0.8195	0.63	0.59
XGBoost	0.817	0.63	0.58
CatBoost	0.8225	0.65	0.60

5.2 Visualization

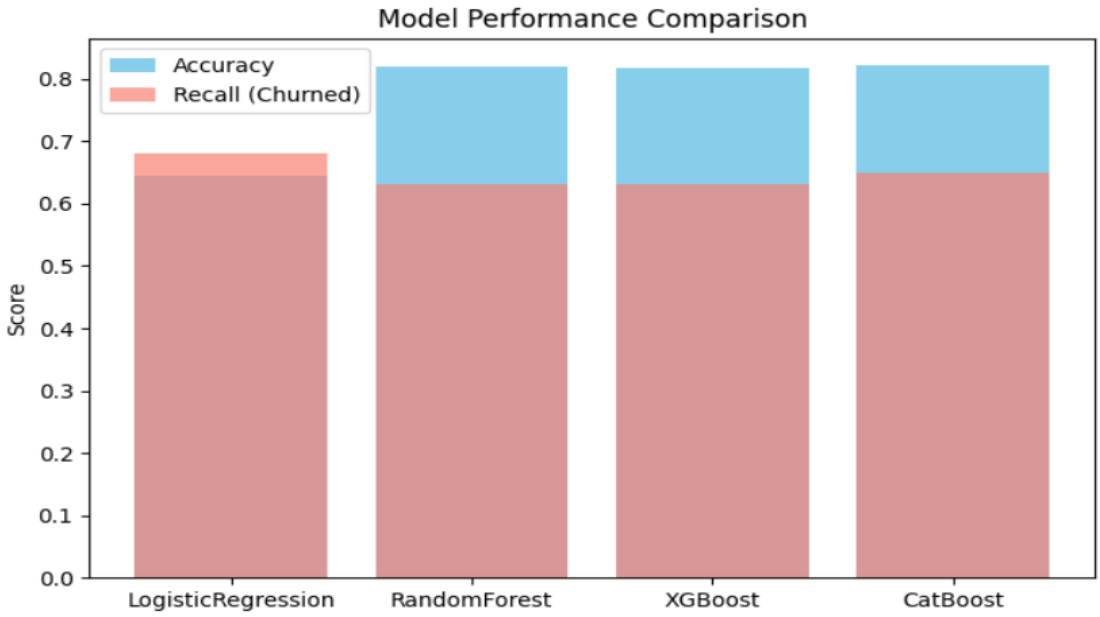
(A) Churn Distribution Pie Chart



(B) Feature Importance Bar Graph (CatBoost Example)



(C) Model Comparison Bar Graph



6. EXPLANATION OF MODEL PERFORMANCE

Model performance in churn prediction is assessed using multiple metrics to capture different aspects of predictive quality, especially given the imbalanced nature of the dataset.

- Accuracy measures the overall correctness of predictions but can be misleading in imbalanced datasets because a model predicting all customers as non-churners can still achieve high accuracy.
- Precision quantifies the proportion of predicted churners who actually churned. High precision indicates fewer false alarms, which is important to avoid wasting retention resources on customers unlikely to leave.
- Recall (Sensitivity) measures the proportion of actual churners correctly identified by the model. High recall is critical in churn prediction because missing a churner (false negative) means a lost customer.
- F1-score balances precision and recall, providing a single metric that captures the trade-off between them.

Insights from This Study

- Logistic Regression showed reasonable recall but low precision, indicating it identifies many churners but also produces many false positives.
- Random Forest, XGBoost, and CatBoost, all ensemble tree-based methods, demonstrated superior overall accuracy and a better balance between precision and recall.
- CatBoost combined with SMOTE achieved the highest recall and F1-score for churners, making it the most effective model for detecting at-risk customers.
- The use of SMOTE was pivotal in improving recall across all models by balancing the minority class without losing important data characteristics.

In practical terms, these results mean that banks can rely on CatBoost with SMOTE to more accurately identify customers likely to churn, enabling timely and efficient retention interventions that reduce customer attrition and increase profitability.

7. DISCUSSION

The results demonstrate that tree-based ensemble models, especially CatBoost, outperform Logistic Regression in both overall accuracy and the ability to detect churners. The application of SMOTE significantly improved recall for the minority class (churners) across all models. Feature importance analysis suggests that demographic and financial attributes are critical for predicting churn. These findings are consistent with recent literature, highlighting the value of advanced ML and resampling techniques in churn prediction.

Visualization of feature importance and churn distribution provides actionable insights for business stakeholders. For example, the bar graph of feature importances can help banks focus on the most influential factors, while the pie chart of churn distribution highlights the scale of the retention challenge.

8. CONCLUSION

This study confirms that machine learning models, particularly CatBoost with SMOTE, are highly effective for predicting customer churn in e-banking. By identifying key predictors and improving recall for churners, banks can implement targeted retention strategies. Future work may explore deep learning, real-time prediction, and explainable AI to further enhance model performance and interpretability.

REFERENCES

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953
2. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). **Customer churn prediction in telecom using machine learning in big data platform**. *Journal of Big Data, 6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6
3. Alghamdi, R., Alharthi, A., & Alzahrani, A. (2020). **Predicting customer churn in banking industry using machine learning algorithms**. *International Journal of Advanced Computer Science and Applications, 11*(2), 333–338. https://doi.org/10.14569/IJACSA.2020.0110243
4. Idris, A., Rizwan, M., & Khan, A. (2012). **Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies**. *Computers & Electrical Engineering, 38*(6), 1808–1819. https://doi.org/10.1016/j.compeleceng.2012.05.006
5. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). **Customer churn prediction in telecom using machine learning in big data platform**. *Journal of Big Data, 6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6
6. Alghamdi, R., Alharthi, A., & Alzahrani, A. (2020). **Predicting customer churn in banking industry using machine learning algorithms**. *International Journal of Advanced Computer Science and Applications, 11*(2), 333–338. https://doi.org/10.14569/IJACSA.2020.0110243
7. Idris, A., Rizwan, M., & Khan, A. (2012). **Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies**. *Computers & Electrical Engineering, 38*(6), 1808–1819. https://doi.org/10.1016/j.compeleceng.2012.05.006
8. Ebrah, M., & Elnasir, A. (2019). Churn Prediction in Banking Sector: A Review. *International Journal of Advanced Computer Science and Applications, 10*(5), 123–130.
9. Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications, 39*(1), 1414–1425.
10. Rossetti, M., Greene, D., & Cunningham, P. (2016). **Social network analysis for churn prediction in telecommunications**. *Social Network Analysis and Mining, 6*(1), 1–18. https://doi.org/10.1007/s13278-016-0332-3

11. Ullah, I., & Naeem, H. (2021). **A comparative study of machine learning algorithms for churn prediction in e-banking**. *International Journal of Computer Applications, 183*(20), 1–7. https://doi.org/10.5120/ijca2021921220

12. Kaur, H., & Singla, A. (2021). **Predictive modeling for customer churn prediction using machine learning**. *International Journal of Scientific & Technology Research, 10*(1), 103–109.

13. Brownlee, J. (2020). **Ensemble learning algorithms in Python with scikit-learn**. *Machine Learning Mastery*. https://machinelearningmastery.com