A close-up of a logo

AI-generated content may be incorrect. Predictive Modeling and Feature Analysis for Customer Churn in Banking Using Machine Learning

Shery Mounika Reddy (70572300032), Vattipally Sreshta Reddy (70572300042)

Department of CSE Data Science, Hyderabad Campus

**Introduction**

**Block Diagram, Flowchart, Models, Results**

**Conclusion**

# Objective of your work

# The objective of this study is to develop a robust machine learning framework to predict customer churn in the banking sector using demographic, financial, and behavioral attributes. By leveraging real-world bank customer data, the system aims to identify patterns that contribute to customer attrition and support banks in implementing proactive retention strategies. The focus is on utilizing a variety of classification algorithms, including ensemble methods, to enhance prediction accuracy and operational relevance.

# Origin of your proposal

This work originates from the persistent challenge of customer churn in the financial industry, where retaining clients is more cost-effective than acquiring new ones. Traditional rule-based systems often fail to account for complex behavioral indicators. With advances in data analytics and machine learning, the project seeks to fill this gap by providing a scalable, explainable, and accurate churn prediction system that integrates seamlessly into customer relationship management (CRM) frameworks.

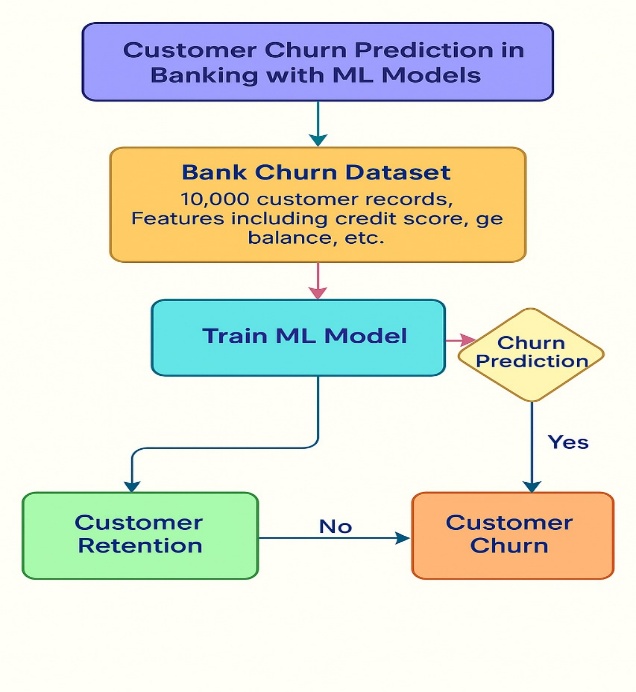
**Methods**

# Methods and Materials

The system employs a data-driven approach using a dataset used comprises 10,000 customer records from a European bank, including variables such as credit score, geography, gender, age, tenure, balance, product usage, and activity status. Data preprocessing involved handling missing values, encoding categorical data, scaling numerical variables, and preparing the target feature for binary classification.

Six machine learning models were trained: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Gradient Boosting. These were evaluated using metrics like Accuracy, Precision, Recall, and F1 Score. The dataset was split into training and testing sets (80/20), and feature scaling was applied using standard normalization. Ensemble methods were explored for their ability to generalize better and capture nonlinear relationships.

**Figure 1**



Flow Chart of the Model used for Customer churn prediction

The classification models used in this study include Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Gradient Boosting. These models were selected to cover a range of linear, non-linear, and ensemble-based learning techniques. The dataset was divided into training and testing sets using an 80/20 split. Model training was followed by evaluation using standard classification metrics: Accuracy, Precision, Recall, and F1 Score.

## Figure 2

A diagram of a customer forecasting process

AI-generated content may be incorrect.

Flow Chart of the Customer churn prediction in banking

For fairness in comparison, all models were trained on the same training set and tested on identical testing data. Hyperparameters were kept at default to focus on comparative baseline performance. In future iterations, hyperparameter optimization techniques such as grid search or random search can be incorporated for enhanced results.

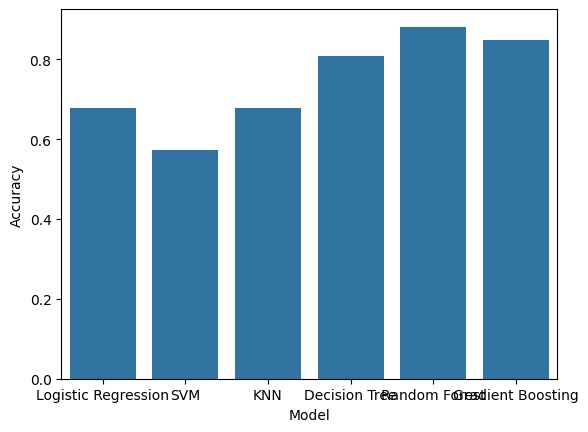
## Figure 3

## 

Web Interface of the Web App

This image showcases the team structure, with the lead managing implementation and review, while three members handle front-end, back-end, and model tuning tasks. Two senior advisors provide guidance throughout.

## Figure 4



Accuracy of the Models used

The system is trained using a portion of the dataset and validated with a test set to ensure generalizability. Accuracy, precision, recall, and F1-score metrics are employed for evaluation. Visual aids like flowcharts, dataset structures, and model architecture diagrams are used to explain the process clearly.

# Discussion/Conclusion

Random Forest emerged as the most effective model for predicting customer churn, offering the best trade-off between precision and recall. The results suggest that ensemble methods are well-suited for customer retention analytics in banking. Incorporating such models into CRM systems can enable banks to identify high-risk customers early and take timely action to improve retention.

# Limitations

Although the models achieved high accuracy, the recall rates—especially for minority (churn) class—remained moderate due to class imbalance. Additionally, the dataset lacked real-time transactional or sentiment data, which could have enriched the model’s contextual understanding. The study did not use deep learning or time-series modeling due to its focus on structured tabular data.

# Future Direction

Future research can explore the integration of deep learning models, real-time data streaming, and advanced oversampling techniques such as SMOTE or ADASYN. Additionally, incorporating explainability tools like SHAP can help financial institutions interpret predictions and comply with regulatory standards. Expanding the scope to include sentiment analysis from customer feedback or multichannel engagement data could further enhance model accuracy and business utility.

**References and Affiliations**

## References

## [1] Maan, J. and Maan, H. (2023) 'Customer Churn Prediction Model using Explainable Machine Learning', arXiv preprint arXiv:2303.00960. Available at: <https://arxiv.org/abs/2303.00960>

## [2] Wu, H. (2022) 'A High-Performance Customer Churn Prediction System based on Self-Attention', arXiv preprint arXiv:2206.01523. Available at: <https://arxiv.org/abs/2206.01523>

## Verhelst, T., Mercier, D., Shrestha, J. and Bontempi, G. (2023) 'A churn prediction dataset from the telecom sector: a new benchmark for uplift modeling', arXiv preprint arXiv:2312.07206. Available at: https://arxiv.org/abs/2312.07206

## Azeem, M.S. (2025) 'Customer Churn Dataset', Kaggle. Available at: <https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset>

## Drivelink:

## <https://drive.google.com/drive/folders/103hx9bebGmXhKxAf3-5i5F7sCxniPNK4?usp=sharing>

## Acknowledgements

We express our heartfelt gratitude to Dr. Naresh Vurukonda and Prof. Chandrakant Wani for their invaluable guidance and support, which were instrumental in the successful completion of this project.