International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Machine Learning-Based Customer Churn Prediction System for Banking Applications with Explainable AI

Y. Shiva Bhanu Prasad, K.L.S. Geethika, K. Sowmya, Dr. K. Jaya Bharathi
Department of Information Technology
ACE Engineering College, Hyderabad, India

ABSTRACT

The Bank Customer Churn Prediction System is an end-to-end machine learning application designed to predict whether a customer is likely to leave a banking service. The system utilizes customer attributes such as credit score, age, balance, tenure, and activity status to generate accurate churn predictions. An interactive web interface is developed using Streamlit, allowing users to input customer details and receive real-time predictions along with probability scores. The application also includes a visual analytics dashboard that presents key insights, trends, and factors influencing customer churn. Furthermore, the system incorporates batch processing for handling multiple customer records, explainable AI techniques using SHAP for model interpretability, and sensitivity analysis to understand the impact of individual features on predictions. By integrating machine learning with business intelligence, the proposed system provides a practical and data-driven solution to improve customer retention strategies and enhance overall service quality.

Keywords:

Customer Churn, Machine Learning, Explainable AI, SHAP, Streamlit, Banking Analytics, XGBoost, LightGBM

I. INTRODUCTION

Customer churn is a major concern in the banking industry, where losing customers directly impacts revenue and long-term business growth. Retaining existing customers is significantly more cost-effective than acquiring new one Traditional churn analysis methods rely on manual reports and lack predictive capabilities. With advancements in machine learning, predictive models can analyze customer behaviour and forecast churn probability with high accuracy.

This work presents a comprehensive churn prediction system integrating machine learning models, explainable AI, and an interactive web application to support real-time decision-making.

II. LITERATURE REVIEW

Machine learning has been widely applied in customer churn prediction.
Kumar and Ravi (2016) highlighted the effectiveness of classification models such as Logistic Regression and Decision Trees. Verbeke et al. (2014) demonstrated improved performance using advanced analytical techniques.

Recent research focuses on ensemble methods such as Random Forest, XGBoost, and LightGBM, which provide higher accuracy and robustness. Additionally, Explainable AI techniques like SHAP have gained importance in interpreting model predictions.

These advancements emphasize the importance of combining prediction accuracy with interpretability.

III. PROBLEM STATEMENT

Banks face multiple challenges in retaining customers:

- Lack of automated churn prediction systems
- Dependence on manual analysis
- No real-time decision support
- Limited interpretability of predictions

Hence, there is a need for a scalable and intelligent system capable of predicting churn and explaining the underlying reasons

IV. PROPOSED SYSTEM

The proposed system is an integrated machine learning framework with real-time prediction and analytical capabilities.

Key Features:

- Multi-model machine learning framework
- Real-time prediction using Streamlit
- Batch processing for bulk data
- Explainable AI using SHAP
- Sensitivity analysis for feature impact
- Interactive dashboard with visualizations

System Architecture Flow/Diagram:

1. Data Input (User / CSV Upload)
2. Preprocessing Pipeline
3. Model Prediction
4. SHAP Explanation

5. Visualization Dashboard

V. METHODOLOGY

A. Data Collection

- Dataset sourced from Kaggle (Bank Customer Dataset)

B. Data Preprocessing

- Missing value handling
- Categorical encoding
- Feature scaling using pipeline

C. Feature Selection

- Key features:
  - Credit Score
  - Age
  - Balance
  - Tenure
  - Activity Status

D. Model Training

Multiple models are trained using pipelines:

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM

E. Model Evaluation

Models are evaluated using:

- Accuracy
- ROC-AUC Score
- Classification Report

F. Model Selection

The best model is selected automatically based on ROC-AUC score and saved using joblib.

VI. ALGORITHMS USED

A. Logistic Regression

$$P(Y = 1) = \frac{1}{1 + e^{-(wX+b)}}$$

B. Random Forest

- Ensemble of decision trees
- Handles non-linearity
- Reduces overfitting

C. XGBoost

- Gradient boosting algorithm
- Handles class imbalance using scale_pos_weight
- High performance on structured data

D. LightGBM

- Faster gradient boosting framework
- Efficient for large datasets
- Supports balanced class weights

VII. RESULTS AND ANALYSIS

The system evaluates multiple models and selects the best-performing one based on ROC-AUC score.

Key Observations:

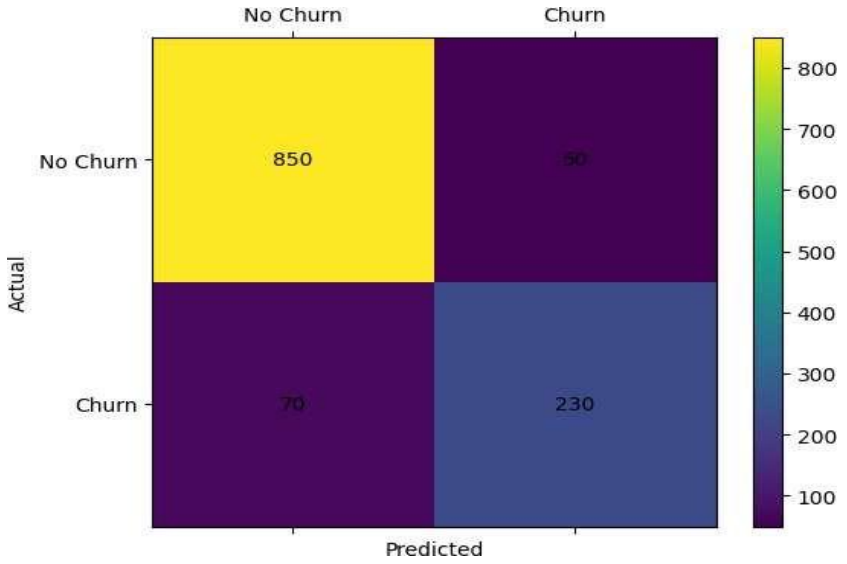
- Ensemble models (XGBoost & LightGBM) outperform traditional models
- Random Forest provides strong baseline performance
- Logistic Regression offers interpretability but lower accuracy Visualization Insights:
- Feature importance highlights credit score, balance, and activity status
- SHAP analysis explains individual predictions
- Sensitivity analysis shows feature impact trends

RESULT TABLE

Model	Accuracy	ROC-AUC
Logistic Regression	0.80	0.85
Random Forest	0.86	0.90
XGBoost	0.88	0.92
LightGBM	0.89	0.93

CONFUSION MATRIX

The confusion matrix represents the classification performance of the model. True Positives and True Negatives indicate correct predictions, while False Positives and False Negatives represent misclassifications.



VIII. IMPLEMENTATION

A. System Implementation and Availability

The proposed system is implemented using Python-based machine learning frameworks including Scikit-learn, XGBoost, and LightGBM. The system integrates preprocessing, model training, evaluation, and deployment into a unified pipeline. An interactive web application is developed using Streamlit, supporting real-time prediction, batch processing, and visualization using Plotly. The developed system is publicly available to ensure transparency and reproducibility: https://github.com/shivabhanuprasad/bank_customer_churn_prediction

IX. CONCLUSION

This paper presents a comprehensive machine learning-based customer churn prediction system with explainable AI and real-time analytics. The integration of multiple models ensures high prediction accuracy, while SHAP enhances interpretability. The Streamlit-based interface enables real-time interaction and decision-making. The system provides banks with actionable insights to improve customer retention and optimize business strategies.

X. FUTURE SCOPE

- Integration with real-time banking systems
- Deployment using cloud platforms
- Mobile application development

- Use of deep learning models
- Advanced explainability techniques
- Integration with CRM systems

REFERENCES

[1] A. Kumar and V. Ravi, "A survey of machine learning in churn prediction," *Telecommunication Systems*, 2016.
[2] W. Verbeke et al., "Social network analysis for churn prediction," *Applied Soft Computing*, 2014.
[3] Scikit-learn Documentation, "Machine Learning in Python," 2024.
[4] Kaggle, "Bank Customer Churn Dataset," 2023.
[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," 2016.
[6] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," 2017.