

International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Early Bank Customer Churn Risk Prediction and Decision Support System using Machine Learning

P. Neha Sri¹, M. Rushmika², S. Kiran Kumar³, K. Sri Vijaya⁴

UG Students, Department of Information Technology, Prasad V Potluri Siddhartha Institute of Technology,
Vijayawada, India¹²³

Assistant Professor, Department of Information Technology, Prasad V Potluri Siddhartha Institute of Technology,
Vijayawada, India⁴

ABSTRACT: Customer churn prediction has become an important task in the banking sector, as it helps reduce revenue loss and improve customer retention. In this work, a machine learning-based churn prediction and decision support system is developed using the publicly available BankChurners dataset, which contains over 10,000 customer records. Data preprocessing techniques were applied, and SMOTE was used to address class imbalance in the dataset. The data was divided into 80% for training and 20% for testing. Multiple machine learning models were evaluated, among which XGBoost showed the best performance. It achieved an accuracy of 97%, recall of 90%, and an AUC score of 0.99, making it effective in identifying customers who are likely to churn. The system also provides probability-based risk segmentation and includes a simplified Value at Risk (VaR) approach to estimate potential financial loss. In addition, a simulation-based decision support module is introduced to analyze different retention strategies, such as customer engagement programs and transaction-based incentives. A Streamlit dashboard was developed to enable real-time interaction and analysis. Overall, the proposed system supports proactive and data-driven decision-making for customer retention in banking applications.

KEYWORDS: Customer Churn; Machine Learning; XGBoost; SMOTE; Risk Prediction; Decision Support System.

I. INTRODUCTION

Customer churn refers to the situation where customers discontinue their relationship with a bank. In today's competitive financial environment, retaining existing customers is generally more cost-effective than acquiring new ones. Because of this, identifying customers who are likely to churn at an early stage has become increasingly important. Traditional churn prediction approaches mainly use binary classification and focus heavily on accuracy as the primary evaluation metric. However, these methods often do not handle class imbalance effectively and provide limited support for real business decisions. In many cases, they only indicate whether a customer may churn, without offering deeper insights into risk levels or financial impact.

To address these issues, this work presents a machine learning-based system that goes beyond simple classification. Instead of producing only binary outputs, the system generates probability-based risk scores, allowing customers to be grouped into different risk categories. In addition, a financial impact estimation is incorporated to support better decision-making. Unlike conventional approaches, the proposed system combines prediction, financial analysis, and simulation-based decision support within a single framework. The main contributions of this work include integrating churn prediction with financial risk estimation using Value at Risk (VaR), introducing probability-based segmentation of customers into multiple risk levels, and developing a simulation-based framework to evaluate different customer retention strategies.

II. LITERATURE REVIEW

Several studies have applied machine learning techniques to customer churn prediction. Gulati et al. demonstrated that using SMOTE helps handle class imbalance and improves prediction performance. Similarly, He Jin focused on classification-based churn prediction models, although the work did not include decision support components. Chawla et al. introduced SMOTE, which is now widely used to generate synthetic samples and improve the prediction of



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

minority classes. Chen and Guestrin proposed XGBoost, a scalable and efficient boosting algorithm that has shown strong predictive performance in many applications.

Although these studies have contributed significantly to improving prediction accuracy, most of them focus mainly on classification performance. They often do not include decision support mechanisms or financial impact analysis, which limits their usefulness in real-world banking scenarios.

To overcome these gaps, the proposed system combines churn prediction with financial risk estimation, probability-based risk segmentation, and a simulation-based decision support framework. This integrated approach provides more practical and business-oriented insights for customer retention.

III. PROPOSED SYSTEM

The proposed system is developed to predict customer churn and support decision-making through a structured machine learning pipeline. The workflow includes data preprocessing, handling class imbalance, model training, and prediction.

In the first step, the dataset is preprocessed by removing irrelevant features, handling missing values, and encoding categorical variables. To deal with class imbalance, SMOTE is applied, which helps improve the model’s ability to identify customers who are likely to churn.

Multiple machine learning models were evaluated, and XGBoost achieved the best performance in terms of accuracy and recall. Instead of generating only binary outputs, the model produces churn probability scores. Based on these scores, customers are grouped into high-, medium-, and low-risk categories.

To further enhance decision-making, the system estimates financial impact using Value at Risk (VaR). This approach combines churn probability with customer credit value to approximate potential revenue loss.

In addition, a decision support component is included to suggest retention strategies such as targeted engagement campaigns and personalized incentives. A user-friendly dashboard is also developed to visualize predictions, feature importance, and risk insights, making it easier for decision-makers to interpret the results.

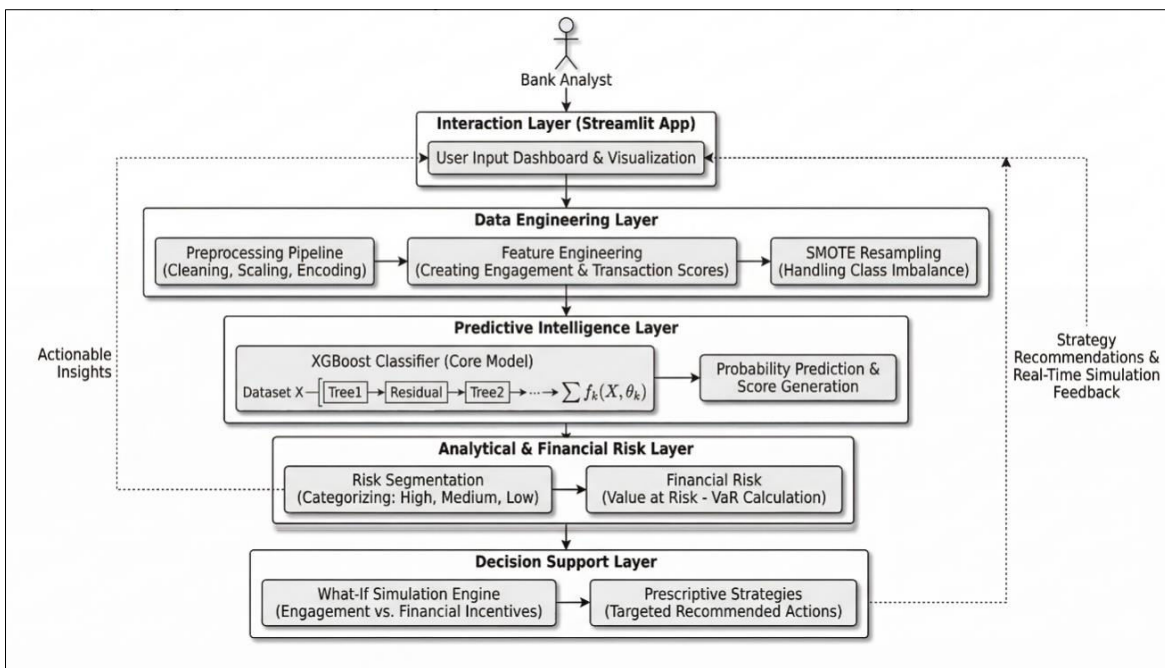


Fig.1. System Architecture



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. METHODOLOGY

A. Dataset Description

The dataset used in this study is the BankChurners dataset obtained from Kaggle, containing over 10,000 customer records with demographic, behavioral, and financial attributes related to banking services and customer activity.

B. Experimental Setup

The implementation was carried out using the Python programming language with libraries including Scikit-learn, XGBoost, Pandas, and Streamlit. Experiments were conducted on a standard laptop environment with sufficient computational capability for model training and evaluation. The dataset was divided into training and testing subsets to evaluate model performance and generalization ability.

C. Methodology

The proposed system follows a structured machine learning pipeline consisting of the following steps:

- 1. Data Preprocessing:** The dataset was cleaned by removing irrelevant features and handling missing values. Categorical variables were converted into numerical format using appropriate encoding techniques to prepare the data for model training.
- 2. Handling Class Imbalance:** SMOTE (Synthetic Minority Over-sampling Technique) was applied to address class imbalance by generating synthetic samples for the minority churn class, improving model learning and prediction capability.
- 3. Model Training:** Multiple machine learning models were trained and evaluated, among which XGBoost demonstrated superior performance. Hyperparameters including learning rate, maximum depth, number of estimators, subsample ratio, and column sampling ratio were tuned empirically to optimize recall while preventing overfitting and improving generalization.
- 4. Prediction and Risk Segmentation:** The trained model generated churn probability scores for each customer. Based on probability values, customers were categorized into high-, medium-, and low-risk groups. A classification threshold of 0.40 was selected to prioritize recall and ensure effective identification of churn-prone customers.
- 5. Financial Risk Estimation:** Financial impact was estimated using a simplified Value at Risk (VaR) approximation defined as:

$$\text{VaR} = \text{Churn Probability} \times \text{Credit Limit} \quad (1)$$

This formulation provides an intuitive estimate of potential revenue loss associated with customer churn and assists in prioritizing high-value customers for retention strategies.

- 6. What-if Simulation for Decision Support:** A simulation module was developed to evaluate retention strategies by modifying selected customer attributes such as transaction count, credit limit, and activity level. The updated inputs were passed to the trained model to recalculate churn probability. Comparison between original and simulated outcomes enables data-driven evaluation of customer retention strategies.

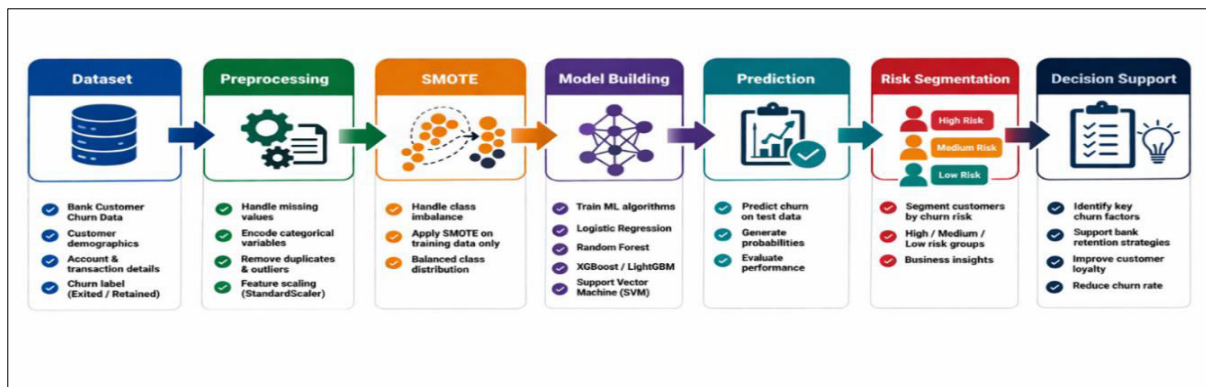


Fig.2. Machine Learning Workflow



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. RESULTS AND DISCUSSION

A. Model Performance Output

A comparative performance analysis of multiple machine learning models was conducted using accuracy, precision, recall, and AUC score. The evaluation results for all models are summarized in Table I.

Table I: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	AUC
Logistic Regression	0.832	0.487	0.837	0.922
Decision Tree	0.918	0.697	0.862	0.954
Random Forest	0.95	0.818	0.886	0.984
SVM	0.665	0.3	0.815	0.795
Naive Bayes	0.707	0.327	0.785	0.825
KNN	0.811	0.453	0.852	0.89
Gradient Boosting	0.962	0.871	0.892	0.989
XGBoost	0.969	0.905	0.905	0.992

A comparative analysis was performed before and after applying SMOTE to address class imbalance. As shown in Fig. 3, SMOTE improved recall across all models, enhancing identification of churn-prone customers. Among the evaluated models, XGBoost achieved the best performance with 97% accuracy and 90% recall. The confusion matrix in Fig. 4 shows a high true positive rate, confirming the effectiveness of the proposed approach.

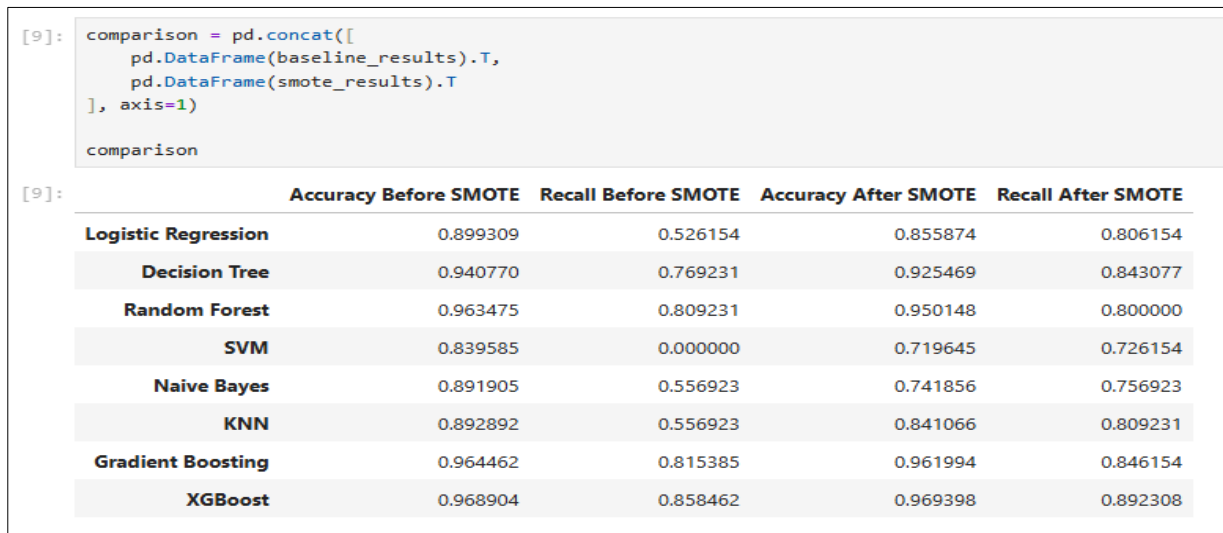


Fig.3. Performance Comparison Before and After SMOTE



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

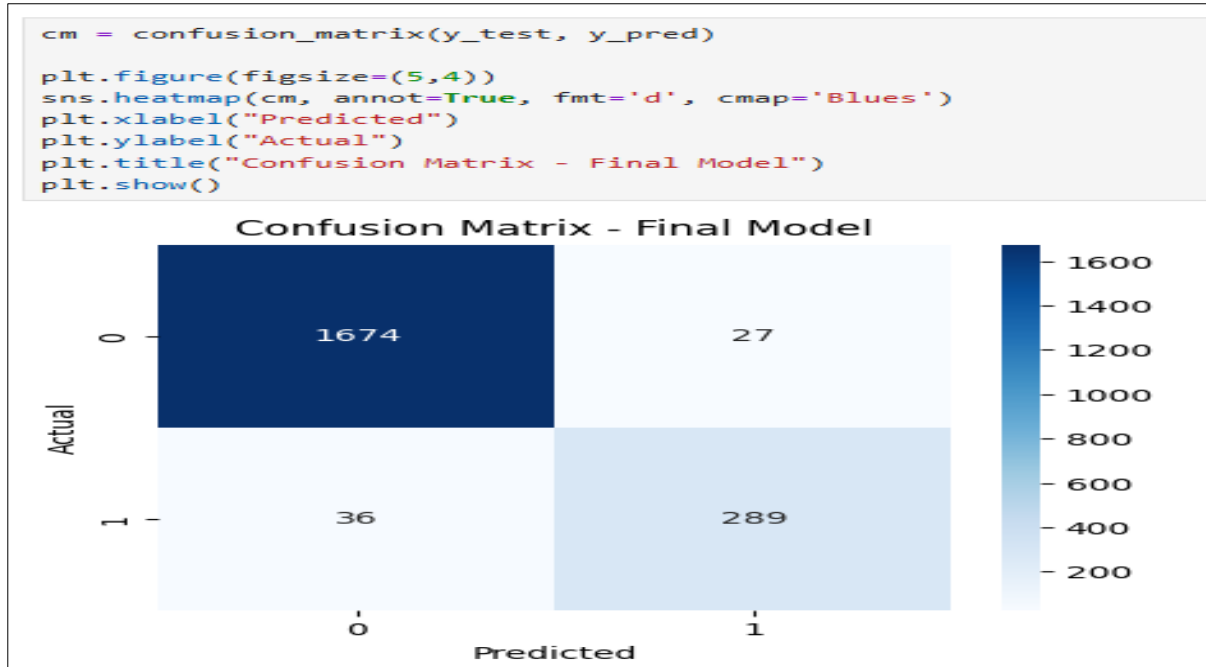


Fig.4. Confusion Matrix

B. Feature Importance Visualization

Feature importance analysis, illustrated in Fig. 5, indicates that transaction count, inactivity period, and relationship count are the most influential factors affecting customer churn. Customers exhibiting lower engagement and reduced transaction activity show a higher likelihood of churn.

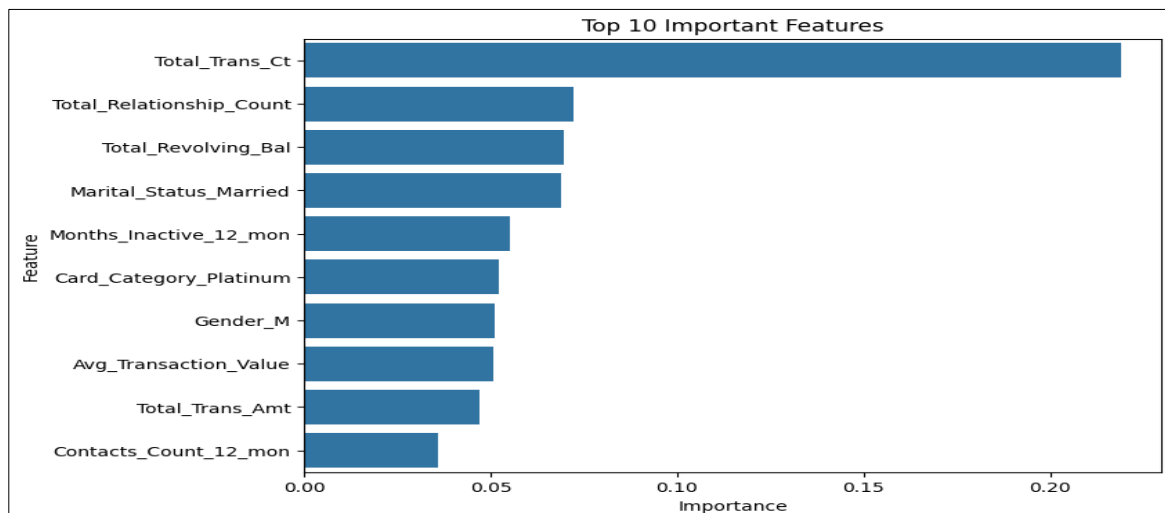


Fig.5. Feature Importance

C. Dashboard Output

An interactive Streamlit dashboard was developed to visualize churn probability, risk segmentation, financial impact estimation, and recommended retention actions. As shown in Figs. 6–8, the dashboard enables real-time analysis and supports decision-making through intuitive visual representations of model outputs and risk insights.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

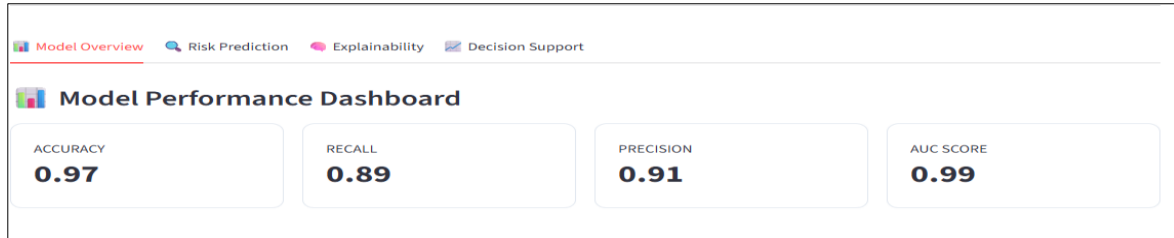


Fig.6. Dashboard Output

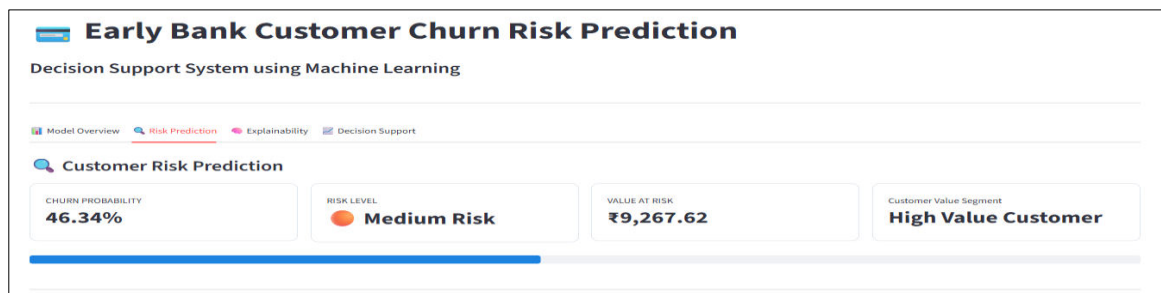


Fig.7. Risk Prediction Output

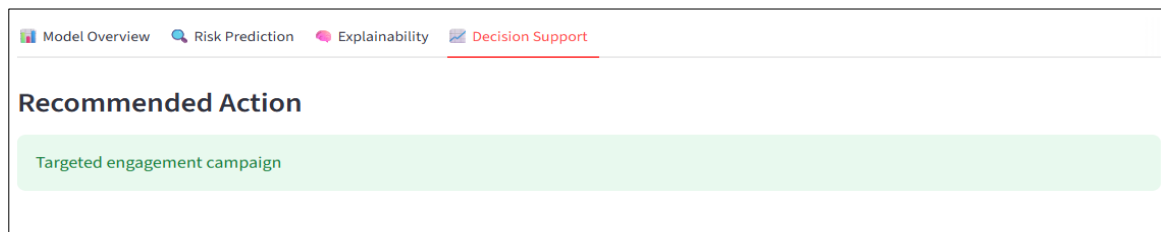
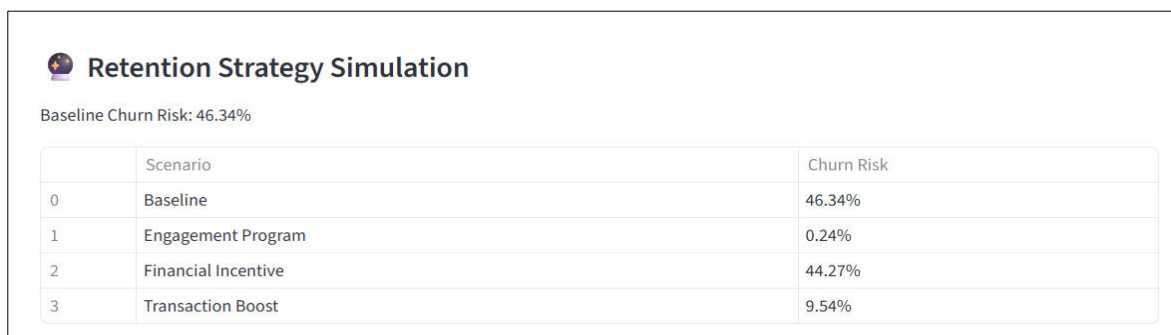


Fig.8. Decision Support Output

D. What-if Simulation Results

The simulation module evaluates different customer retention strategies by modifying selected customer attributes and recalculating churn probability. As illustrated in Fig. 9, engagement-based interventions demonstrate a greater reduction in churn probability compared to financial incentives alone, highlighting the importance of customer activity and interaction in retention planning.





International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

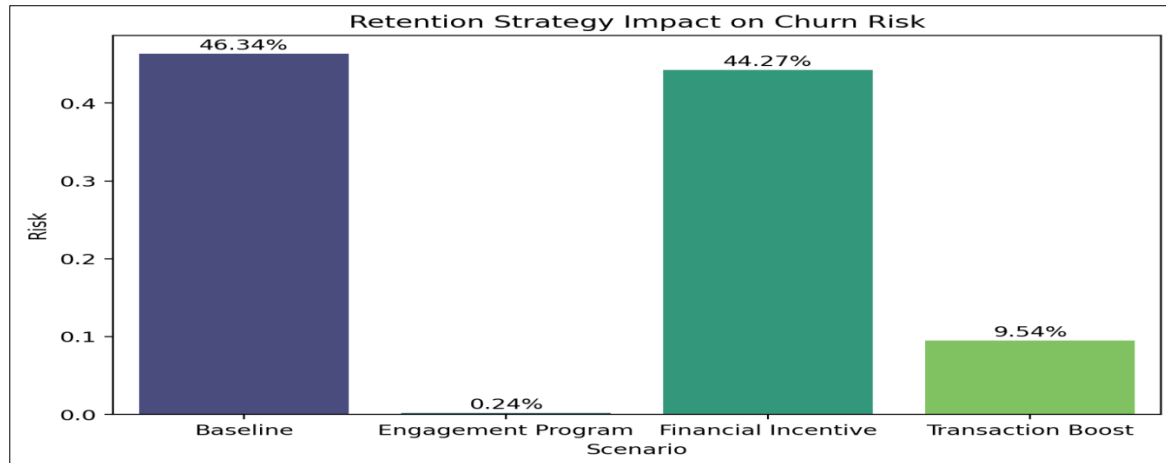


Fig.9. What-if Simulation Output

VI. CONCLUSION

This paper presents a machine learning-based churn prediction and decision support system for banking applications. The use of SMOTE along with XGBoost helps achieve strong predictive performance, especially in identifying customers who are likely to churn.

The proposed approach goes beyond traditional prediction by combining probability-based risk segmentation, simplified financial risk estimation, and simulation-based strategy evaluation. This makes it possible to support proactive and data-driven customer retention.

Overall, the system offers a practical and scalable solution for reducing churn and improving decision-making in real-world banking environments. The experimental results show that integrating predictive analytics with decision support components improves the practical usefulness of churn prediction systems.

VII. FUTURE SCOPE

Future work can focus on improving model interpretability by using techniques such as LIME or SHAP, which can provide clearer explanations at the individual customer level. This would help in understanding why a customer is predicted to churn and support better decision-making. The system can also be extended to real-time deployment using cloud platforms, allowing churn prediction based on live customer data. This would make the solution more practical for real-world applications.

In addition, deep learning models such as RNN and LSTM can be explored to capture temporal patterns in customer behavior over time. Financial analysis can be further enhanced by incorporating Customer Lifetime Value (CLV) to support long-term strategic decisions.

Finally, sentiment analysis using natural language processing (NLP) on customer feedback can be included to improve prediction performance by considering customer opinions and experiences.

REFERENCES

- [1] A. Gulati, et al., "Bank Customer Churn Prediction Using Machine Learning: A Comparative Study with SMOTE-Based Class Balancing," Proc. IEEE ICCES, 2025.
- [2] H. Jin, "Bank Customer Churn Prediction Based on Machine Learning Models," Proc. 9th Int. Conf. Economic Management and Green Development, 2025. doi: 10.54254/2754-1169/2025.LH23973.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- [3] “Customer Churn Prediction Model using Machine Learning,” International Journal of Computer Science Trends and Technology, 2023.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
- [5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
- [6] F. Pedregosa, et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [7] C. Molnar, Interpretable Machine Learning, 2nd ed., 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [8] “BankChurners Dataset,” Kaggle, 2024. [Online]. Available: <https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers>



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details