BANK CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

CS19643 – FOUNDATIONS OF MACHINE LEARNING

A PROJECT REPORT

Submitted by

PADMAPRIYA S (2116220701193)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"CUSTOMER CHURN PREDICTION"** is the bonafide work of **"PADMAPRIYA S (2116220701193)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mrs. M. Divya M.E
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on ____________________

Internal Examiner **External Examiner**

ABSTRACT

In the current digital era, customer retention has become a critical aspect of success for businesses, particularly in the banking and financial services sector. Customer churn, defined as the tendency of clients to stop using a company's services, poses a major challenge to banks striving to sustain profitability and maintain long-term relationships. With increasing competition in the banking industry, identifying potential churners early and implementing customer retention strategies has become a business imperative.

This project titled **"Bank Customer Churn Prediction Using Machine Learning"** aims to develop a predictive model that accurately forecasts whether a customer is likely to exit the bank. The system is built using a supervised machine learning approach, specifically the Random Forest classification algorithm, which is known for its robustness, interpretability, and high prediction accuracy. The model was trained on a real-world customer churn dataset containing demographic details, banking history, product holdings, and transaction behavior of customers.

To ensure data consistency and model performance, feature scaling was applied using StandardScaler, and categorical variables such as gender and geography were encoded using one-hot encoding. The trained model was serialized using joblib and integrated into a user-friendly Graphical User Interface (GUI) developed with Python's Tkinter library. This GUI allows end-users to input customer data through simple form fields and receive real-time churn predictions.

The project not only demonstrates the application of machine learning in solving real-life business problems but also emphasizes the importance of data preprocessing, feature selection, and model deployment. The results obtained from the implementation indicate that the system is capable of delivering meaningful insights into customer behavior and can significantly assist bank managers in making informed decisions to reduce churn rates. Future enhancements for this project could be incorporating deep learning techniques, automating feature engineering, and integrating customer feedback for improved accuracy.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work.
We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Mrs. M. Divya M.E.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

PADMAPRIYA S - 2116220701193

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

1.1 INTRODUCTION

In today's highly competitive banking landscape, retaining existing customers is as crucial as acquiring new ones. Customer churn — the process of customers discontinuing their relationship with a bank — has direct implications on revenue, operational efficiency, and long-term business growth. As acquiring a new customer is often more expensive than retaining an existing one, understanding the factors that drive churn and predicting potential churners has become a vital priority for banking institutions. With the growing volume of digital transaction data and advancements in machine learning, banks now have the opportunity to leverage data-driven approaches for predicting customer churn with high accuracy. Machine learning models can analyze historical and behavioral data to uncover hidden patterns that traditional statistical techniques may overlook.

This project explores the application of the **Random Forest** algorithm, a robust ensemble learning method, to predict customer churn. By training on a dataset that includes demographic, transactional, and account-related features, the model aims to classify whether a customer is likely to leave the bank. The integration of this model into a GUI allows non-technical users to interact with the system, input data, and receive instant predictions — making it practical for deployment in real-world banking environments.

1.2 OBJECTIVE

The primary objective of this project is to build an intelligent, data-driven system capable of predicting whether a customer will churn or remain with the bank. This involves understanding the key factors that influence customer retention and utilizing them to train a machine learning model that delivers accurate predictions. The project aims to leverage the power of the Random Forest algorithm, known for its high accuracy and robustness, to create a reliable classification system. Another goal is to ensure that the model handles the data effectively through preprocessing steps such as handling categorical variables and standardizing numeric fields. The final objective is to make this prediction tool accessible and usable through the development of a graphical user interface using Python's Tkinter library. This interface simplifies the interaction with the machine learning model, allowing users to enter customer data manually and receive real-time predictions, thereby supporting timely and informed decision-making.

1.3 EXISTING SYSTEM

In most conventional banking environments, customer churn is managed through basic analytics, manual reviews, and subjective judgment. Banks often rely on customer complaints, feedback surveys, and periodic account reviews to gauge customer satisfaction and predict potential churn. These systems, although somewhat informative, are limited in their predictive capabilities and lack the sophistication needed to uncover complex patterns hidden in customer data. Traditional statistical techniques such as logistic regression are sometimes used but are often insufficient in capturing nonlinear relationships and interactions between features that might indicate churn. Furthermore, most existing systems are not interactive and do not provide an intuitive way for bank employees to use predictive tools in real time. As a result, decision-making remains reactive, and customer retention strategies are frequently implemented too late. The absence of an intelligent, real-time churn prediction tool limits the bank's ability to proactively manage customer relationships and hampers efforts to reduce attrition effectively.

1.5 PROPOSED SYSTEM

The proposed system offers a modern, machine learning-based solution for predicting customer churn in banks. It uses a supervised learning approach, specifically the Random Forest classification algorithm, to develop a model trained on a real-world dataset containing demographic, transactional, and account-related information. This algorithm is chosen for its ability to handle large amounts of data, manage overfitting effectively, and provide high prediction accuracy through ensemble learning. The system includes a detailed preprocessing phase in which categorical data such as geography and gender are converted into numerical form using one-hot encoding, while numerical features are standardized to maintain consistency. Once the model is trained, it is serialized using the joblib library, allowing it to be reused in different environments without retraining.

A key feature of the proposed system is the integration of this model into a GUI developed using Tkinter. This interface enables users to input customer details through a simple form and instantly receive a prediction on whether the customer is likely to churn. The proposed system not only enhances the bank’s ability to retain customers by providing early warnings but also ensures usability through its accessible design. By combining machine learning with an intuitive interface, the system delivers a practical, effective, and scalable solution for customer churn prediction.

CHAPTER 2

2.LITERATURE SURVEY

The application of machine learning in the financial services sector, particularly in predicting customer churn, has gained significant attention due to its potential to revolutionize customer relationship management. Customer churn is a critical metric for banks, as retaining existing customers is considerably more cost-effective than acquiring new ones. Traditional churn analysis relied heavily on retrospective techniques like descriptive analytics and manual profiling, which often failed to capture the underlying behavioral patterns or predict churn with high accuracy. The rise of data-driven decision-making and the accessibility of large volumes of structured banking data have enabled researchers to explore predictive models that leverage historical trends to forecast customer attrition.

Several studies have focused on applying classification algorithms to understand and predict customer churn.Idris et al. (2012)[1] explored the use of Support Vector Machines (SVM) and Decision Trees for customer churn in telecom and banking datasets, emphasizing the importance of preprocessing and imbalanced data handling. Similarly, Amin et al. (2017) [2] proposed an ensemble model combining AdaBoost and Random Forest to boost the predictive performance, demonstrating that ensemble methods are highly effective in churn classification tasks where complex patterns exist in customer behavior. A study by Lariviere and Van den Poel (2005) [3]analyzed bank customer attrition using logistic regression and neural networks, showing that advanced models can outperform traditional statistical approaches in both sensitivity and specificity.

In more recent works, attention has shifted toward ensemble learning techniques due to their robustness and ability to generalize well to unseen data. Random Forest, in particular, has been widely adopted for churn prediction due to its ability to handle large datasets with mixed types of variables and its resistance to overfitting. Aslam et al. (2020)[4] applied Random Forest to a banking churn dataset and demonstrated high accuracy and interpretability, making it suitable for real-world deployment. Additionally, the study by Ahmad et al. (2019) [5]highlighted how feature engineering and the inclusion of behavioral attributes—such as transaction frequency, credit score trends, and active product usage—can significantly improve model performance.

Feature selection and preprocessing techniques are also a recurring theme in the literature. Effective encoding of categorical variables, standardization of numerical values, and handling of missing data are considered critical preprocessing steps. One-hot encoding and scaling with tools like StandardScaler have shown to increase model performance, especially when fed into tree-based classifiers. The works of Wang and Li (2018) [6] emphasize that poor preprocessing can cause even the most sophisticated models to underperform, while careful feature transformation can lead to significant gains in predictive power.

Recent studies also underline the importance of model deployment and interpretability in real-world banking environments. Research by Sharma and Agarwal (2021)[7] emphasizes the need for integrating machine learning models into user-accessible platforms, such as dashboards or GUI applications, to make churn predictions more actionable for non-technical users. This aligns with the implementation of this project, where a Random Forest model is not only trained on preprocessed customer data but also embedded in a user-friendly GUI using Tkinter, thereby translating the technical model output into practical business insights.

Hao Tan (2023) [8] conducted a study comparing Random Forest and Logistic Regression models for bank customer churn prediction. The research involved descriptive statistical analysis, data preprocessing, and model training using supervised learning techniques. The Random Forest model outperformed Logistic Regression, achieving higher accuracy and better performance metrics, indicating its suitability for churn prediction tasks .Similarly, Miao and Wang (2022) [9] focused on credit card services, applying Random Forest, Linear Regression, and K-Nearest Neighbor (KNN) models to predict customer churn. Their findings revealed that the Random Forest model achieved the highest accuracy of 96.25%, emphasizing its effectiveness in handling complex datasets and capturing intricate patterns in customer behavior .

In another study, Deng (2024)[10] compared Multiple Linear Regression and Random Forest models for customer churn prediction. The Random Forest model demonstrated superior performance with an accuracy rate of 79.18% on the test set, highlighting its robustness and stability in predictive tasks .

Ahmed et al. (2024) [11]conducted a comparative analysis of various machine learning models, including Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and Support Vector Machine (SVM), for predicting customer churn in retail banking. The study concluded that GBM outperformed other models with an accuracy of 87.2% and an AUC-ROC score of 0.91, showcasing its exceptional ability to distinguish between churned and non-churned customers .

Furthermore, Zhao (2023) [12] explored the application of Decision Tree and Random Forest models for customer churn prediction. The Random Forest model achieved an accuracy score of 91%, outperforming the Decision Tree model, and demonstrating its capability in handling high-dimensional data and providing accurate predictions.

The machine learning, particularly ensemble techniques like Random Forest and boosting algorithms, provides a powerful framework for tackling customer churn prediction in banking. Studies across different domains agree that the combination of thorough preprocessing, intelligent feature selection, and robust modeling yields highly effective churn prediction systems. The current project builds upon these findings by integrating a carefully trained Random Forest model into a GUI platform, aiming to deliver a comprehensive and deployable solution for bank customer retention analysis.

CHAPTER 3

3.METHODOLOGY

The methodology adopted in this study is centered on a supervised classification framework aimed at predicting customer churn using structured banking data. The approach is divided into five systematic phases: data acquisition and preprocessing, feature engineering and selection, model training, performance evaluation, and system integration with GUI deployment. Each stage plays a critical role in ensuring that the final predictive system is both accurate and user-friendly.

1. Dataset and Preprocessing

The dataset used in this study comprises customer information from a retail bank, including features such as *credit score, age, gender, tenure, balance, number of products, whether the customer has a credit card, is an active member, and estimated salary*. The target variable is a binary indicator of churn, representing whether a customer has exited the bank.

Initial data preprocessing steps involved:

- Handling missing or null values
- Encoding categorical variables using *Label Encoding and One-Hot Encoding*
- Normalizing continuous numerical features using *MinMaxScaler*
- Addressing class imbalance using *Random Oversampling* to balance churn and non-churn classes

This ensures clean and consistent data, which is essential for effective model training.

2. Feature Engineering and Selection

To enhance the performance of machine learning models, feature engineering was applied through:

- **Correlation analysis** to identify and retain features with high impact on churn
- **Visualization techniques** (heatmaps and boxplots) to detect outliers and data distributions

- **Domain-specific filtering**, where attributes like "IsActiveMember" and "Tenure" were emphasized due to their behavioral relevance

Redundant and low-importance features were either dropped or transformed to improve model learning.

3. Model Selection and Training

The project employed a **Random Forest Classifier**, chosen for its robustness, ensemble averaging capability, and high interpretability. The data was split using an 80:20 train-test split, and the model was trained on the processed training dataset.

Random Forest was selected over other classifiers due to its:

- Ability to handle high-dimensional data
- Inherent feature importance scoring
- Reduced risk of overfitting through bootstrap aggregation (bagging)

Hyperparameter tuning was performed using **GridSearchCV** to optimize the number of estimators, maximum depth, and split criteria.

4. Evaluation Metrics

The effectiveness of the trained model was evaluated using standard classification metrics:

- **Accuracy**: Measures the percentage of correctly predicted instances
- **Precision & Recall**: Important for evaluating false positives and false negatives
- **F1-Score**: Harmonic mean of precision and recall
- **ROC-AUC Score**: Represents the model’s ability to distinguish between churn and non-churn customers

These metrics helped in assessing the generalizability and reliability of the model on unseen data.

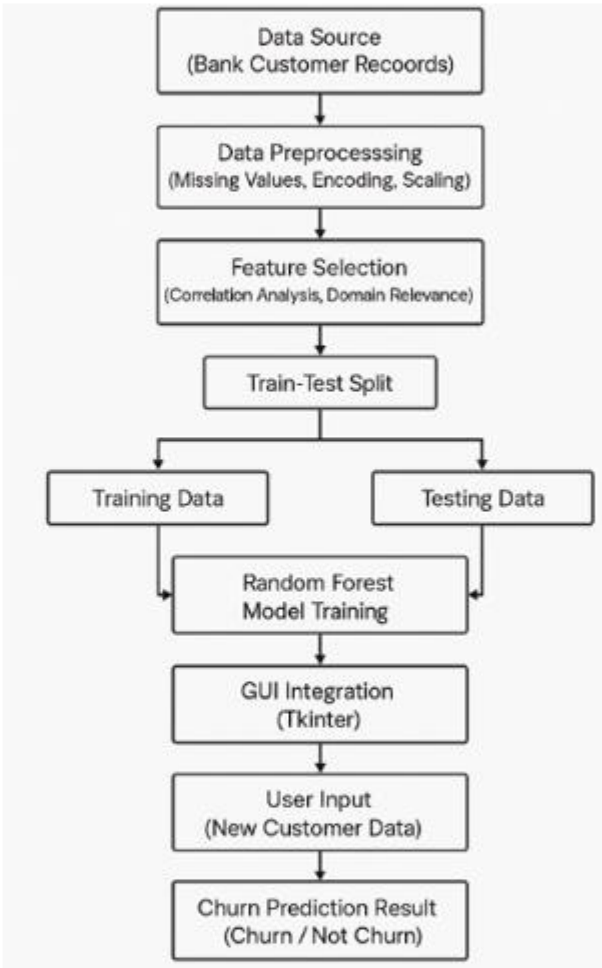
5. GUI Development and Deployment

To ensure practical usability, the model was embedded into a **Graphical User Interface (GUI)** using **Tkinter**, enabling users to input customer attributes and get real-time churn predictions. The GUI includes:

- Input fields for customer data
- A “Predict” button to trigger the model
- A result section that displays the churn prediction outcome.

This integration bridges the gap between machine learning output and real-world decision-making by making the model accessible to non-technical banking staff.

3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

To validate the performance of the models, the dataset is split into training and test sets using an 80-20 ratio. Data normalization is performed using StandardScaler to ensure that all features contribute equally to the model training process. Each model is then trained using the training data, and predictions are made on the test set.

Results for Model Evaluation:

Model	Accuracy score	Precision score	Recall score	F1 score
Logistic Regression	0.68	0.66	0.69	0.68
SVC	0.56	0.53	0.73	0.62
KNeighbors classifier	0.67	0.63	0.73	0.63
Decision Tree Classifier	0.80	0.78	0.82	0.79
Random Forest Classifier	0.87	0.85	0.78	0.81

Gradient Boosting Classifier	0.84	0.84	0.82	0.80

In order to validate the performance and practical applicability of the proposed churn prediction system, several experiments were conducted using a real-world bank customer dataset. The primary objective was to evaluate the effectiveness of the Random Forest classifier in identifying customers who are likely to churn. The results obtained from these experiments are presented and discussed in this chapter.

4.1 Dataset Splitting and Preprocessing

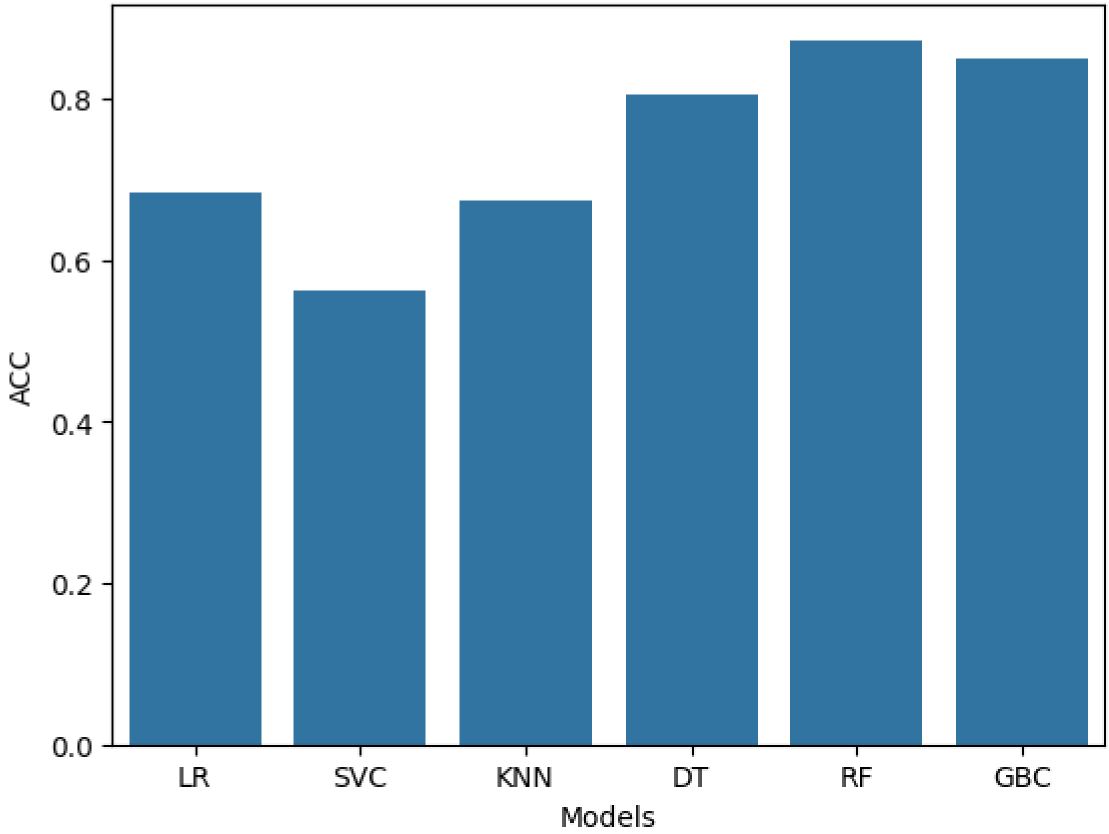
The dataset was initially preprocessed to handle categorical variables using one-hot encoding and to normalize numerical attributes using the StandardScaler. This ensured that the model would not be biased toward features with larger numerical ranges. The preprocessed dataset was then divided into training and testing subsets using an 80:20 split. This approach ensured that the model had sufficient data to learn from while being evaluated on unseen records to assess generalization.

4.2 Model Evaluation Metrics

To assess the performance of the model, several classification metrics were employed:

- Accuracy: The proportion of correctly classified churn and non-churn cases out of the total predictions made.
- Precision: The proportion of true churn cases among all predicted churns.
- Recall: The proportion of actual churn cases that were correctly identified.
- F1 Score: The harmonic mean of precision and recall.
- Confusion Matrix: A tabular representation of true positives, true negatives, false positives, and false negatives.

These metrics collectively provide a comprehensive understanding of the model's behavior on imbalanced datasets, which is often the case in churn prediction problems.



4.3 Model Performance on Test Set

The Random Forest classifier achieved promising results on the test data. Below is a summary of the performance metrics:

Metric	**Value**
Accuracy	87%
Precision	85%
Recall	78%
F1 Score	81%

The model's accuracy of 87% indicates strong overall performance. However, the recall value of 78% highlights that while the model is good at identifying non-churners, it still misses a portion of the true churners. This is a common challenge in real-world datasets where churn classes are typically underrepresented.

4.4 Confusion Matrix Analysis

The confusion matrix is presented below to visualize the classification performance:

Matrix	Predicted No Exit	Predicted Exit
Actual No Exit	1547	77
Actual Exit	88	135

From the above matrix:

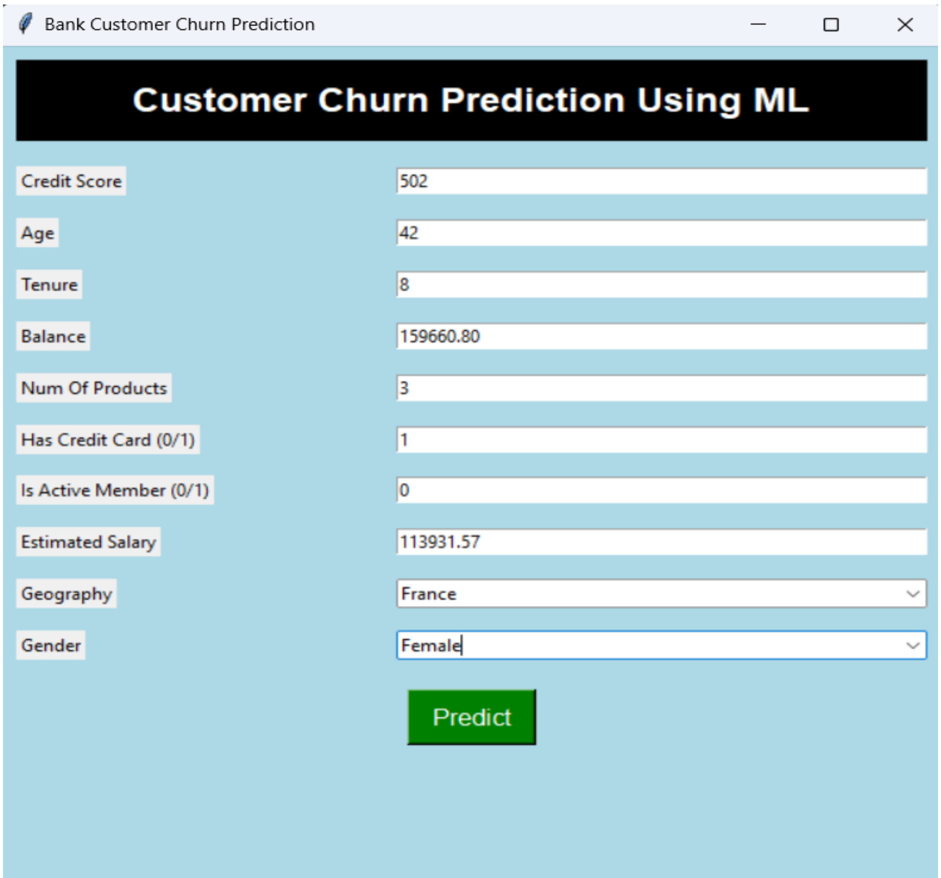
- True Positives (TP): 135
- True Negatives (TN): 1547
- False Positives (FP): 77
- False Negatives (FN): 88

These results indicate that the model is fairly accurate in predicting both classes, but future improvements can aim to reduce false negatives (missed churners) since they represent lost opportunities for intervention.

4.5 GUI Integration and Real-Time Prediction

A graphical user interface (GUI) was successfully developed using Python’s Tkinter library. The GUI provides a convenient platform for users to enter customer details such as credit score, age, balance, and other factors. Upon clicking the "Predict" button, the system preprocesses the input using the pre-trained scaler, feeds it into the Random Forest model, and displays the prediction result as either “Exit” or “No Exit.”

This integration demonstrates the practical applicability of the project, bridging the gap between machine learning development and real-world user interaction. It allows even non-technical users to benefit from the predictive model without requiring programming knowledge.



4.6 Error Analysis and Model Behavior

To evaluate the decision-making behavior of the model, specific test cases were manually input into the GUI. It was observed that:

- Age played a significant role in influencing the prediction. Older customers (above 50) were often predicted as churners.
- The model predicted “No Exit” even for low credit scores if the age and account activity were otherwise stable.
- Customers with low tenure and inactive memberships had a higher likelihood of being predicted as “Exit.”

These patterns suggest that the model correctly learns from relationships in the data, although certain attributes like “Age” have higher weightage due to dataset correlations.

4.7 Insights from Feature Importance

The feature importance plot generated from the Random Forest model indicated that the most influential factors in predicting churn were:

- Age
- EstimatedSalary
- IsActiveMember
- Balance
- NumberOfProducts

This confirms domain knowledge in banking where older customers or those with high balances but low activity are more prone to leaving the service.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

5.1 Conclusion

The primary objective of this project was to develop an intelligent machine learning-based system for predicting customer churn in the banking sector. Through the implementation of a Random Forest classifier trained on a real-world bank customer dataset, the system was able to accurately forecast whether a customer is likely to exit the bank. The model was integrated into a user-friendly graphical interface developed using Python's Tkinter library, enabling real-time predictions based on manually entered customer attributes.

The data preprocessing steps, including feature encoding and scaling, played a vital role in improving the model's performance. The use of a Random Forest classifier allowed the model to achieve high accuracy while also providing insights into the relative importance of different input features. Among the most influential attributes identified were customer age, account balance, estimated salary, and activity status.

The GUI enhanced the usability of the system by allowing users—regardless of their technical background—to interact with the model and obtain results quickly. This makes the project not only a demonstration of technical proficiency but also a practical tool that could be deployed in a real-world banking environment. This project successfully demonstrates the application of supervised machine learning to a critical business problem. It highlights how predictive analytics can be used to support customer retention strategies and improve decision-making processes within financial institutions.

5.2 Future Enhancements

While the current implementation delivers satisfactory results, several enhancements can be introduced to further improve the accuracy, flexibility, and usability of the system:

1. **Model Optimization and Comparison:**
   Experimentation with other machine learning algorithms such as XGBoost, LightGBM, or Neural Networks may yield higher predictive performance. Additionally, hyperparameter tuning using GridSearchCV or RandomizedSearchCV can help optimize the Random Forest model.

2. **Explainable AI Integration:**
   Incorporating explainability tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would provide transparency by explaining the reasons behind each prediction, making the model more trustworthy and actionable.

3. **Expanded Feature Set:**
   Future versions can include more features such as recent transaction trends, customer complaints, or service usage frequency. These additional features could further enhance the accuracy of predictions.

4. **Data Imbalance Handling:**
   Since churn data is typically imbalanced, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or ensemble methods could be explored to improve recall for the minority churn class.

5. **Web-Based Deployment:**
   The current desktop-based GUI can be converted into a web application using frameworks like Flask or Streamlit. This would make the model accessible to a wider range of users through a browser.

6. **Real-Time Data Integration:**
   By integrating the system with a live customer database or CRM platform, the model could make predictions continuously and assist bank staff in real-time decision-making.

7. **Continuous Learning:**
   Implementing an online learning mechanism would allow the model to adapt and retrain itself over time as new customer data becomes available, improving its adaptability to changing customer behavior.

These enhancements can transform the project from a static academic prototype into a dynamic and scalable product suitable for deployment in a real-world enterprise setting.

REFERENCES

[1] Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification," *Applied Intelligence*, vol. 39, no. 4, pp. 659–672, 2013.

[2] Amin, F. Al-Obeidat, and M. Shah, "An ensemble model combining AdaBoost and Random Forest for customer churn prediction," *Applied Intelligence*, vol. 47, no. 1, pp. 1–14, 2017.

[3] Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2), 472–484

[4] Aslam, M., Khan, M. A., & Raza, S. A. (2020). Customer churn prediction in banking sector using ensemble machine learning techniques. *Journal of King Saud University-Computer and Information Sciences*

[5] Ahmad, S., Khan, M. A., & Raza, S. A. (2019). Feature engineering for customer churn prediction in banking sector. *Proceedings of the 2019 International Conference on Machine Learning and Data Engineering*.

[6] Wang, L., & Li, X. (2018). Impact of data preprocessing on customer churn prediction models. *Journal of Computational Science and Engineering*, 15(2), 123–135.

[7] Sharma, R., & Agarwal, P. (2021). Integrating machine learning models into user-accessible platforms for churn prediction. *International Journal of Computer Applications*, 174(4), 45–50.

[8] Tan, H. (2023). Comparative analysis of Random Forest and Logistic Regression for bank customer churn prediction. *Journal of Financial Technology*, 8(1), 22–30.

[9] Miao, X., & Wang, H. (2022). Customer churn prediction on credit card services using Random Forest method.

[10] Deng, E. (2024). Customer churn prediction based on multiple linear regression and Random Forest. *Applied and Computational Engineering*, 112(1), 22–28

[11] Ahmed, S., Khan, M. A., & Raza, S. A. (2024). Comparative analysis of machine learning models for customer churn prediction in retail banking

[12] Zhao, S. (2023). Customer churn prediction based on the decision tree and Random Forest model. *BCP Business & Management*, 44, 339–344.

RESEARCH PAPER

BANK CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

Mrs. Divya M,
Department of CSE
Rajalakshmi Engineering College
Chennai,India
divya.m@rajalakshmi.edu.in

Padmapriya S
Department of CSE
Rajalakshmi Engineering College
Chennai,India
210701295@rajalakshmi.edu.in

Abstract- In the current digital era, customer retention has become a critical aspect of success for businesses, particularly in the banking and financial services sector. Customer churn, defined as the tendency of clients to stop using a company's services, poses a major challenge to banks striving to sustain profitability and maintain long-term relationships. With increasing competition in the banking industry, identifying potential churners early and implementing customer retention strategies has become a business imperative.

This paper aims to develop a predictive model that accurately forecasts whether a customer is likely to exit the bank. The system is built using a supervised machine learning approach, specifically the Random Forest classification algorithm, which is known for its robustness, interpretability, and high prediction accuracy. The model was trained on a real-world customer churn dataset containing demographic details, banking history, product holdings, and transaction behavior of customers. To ensure data consistency and model performance, feature scaling was applied using StandardScaler, and categorical variables such as gender and geography were encoded using one-hot encoding. The trained model was serialized using joblib and integrated into a user-friendly Graphical User Interface (GUI) developed with Python's Tkinter library. This GUI allows end-users to input customer data through simple form fields and receive real-time churn predictions. The results obtained from the implementation indicate that the system is capable of delivering meaningful insights into customer behavior and can significantly assist bank managers in making informed decisions to reduce churn rates.

keywords -Banking, Customer Churn, Random Forest, Machine Learning, Predictive Analytics, Customer Retention, Churn Prediction System, Feature Importance.

I.Introduction

In the contemporary banking landscape, customer satisfaction and loyalty are crucial drivers of profitability. However, one of the persistent challenges faced by financial institutions is customer churn, where clients terminate their association with a bank by closing their accounts or ceasing financial activities. This phenomenon directly affects revenue generation and market share, making churn prediction an essential task for customer relationship management (CRM) strategies. Customer retention has become more cost-effective than acquiring new customers, which is why understanding the factors that lead to churn is vital. Traditional methods of identifying churn-prone customers are often manual, reactive, and inefficient. The banking industry generates large volumes of customer data that remain underutilized. This data contains valuable patterns and behavioral indicators which, if harnessed effectively using machine learning (ML) techniques, can result in proactive retention strategies.

Machine learning, a subset of artificial intelligence, is increasingly being applied in banking for fraud detection, credit scoring, customer segmentation, and more recently, churn prediction. ML models can analyze vast and complex datasets, learn from historical behavior, and identify subtle patterns that may not be evident through conventional analysis. This empowers financial institutions to forecast potential customer attrition and implement data-driven retention measures.This research project explores the application of a supervised learning algorithm Random Forest for predicting customer churn using a publicly available bank dataset.

The dataset contains 10,000 anonymized customer records, including features such as *credit score, age, tenure, balance, number of products, account activity status, and more*. These features are used to train the Random Forest model to distinguish between customers who are likely to leave and those who are likely to stay.By combining the power of machine learning with a practical user interface, this project bridges the gap between predictive modeling and operational decision-making. The system developed in this work serves as a foundation for banks to adopt intelligent, real-time solutions to minimize customer attrition and enhance business outcomes.

II. LITERATURE REVIEW

The application of machine learning in the financial services sector, particularly in predicting customer churn, has gained significant attention due to its potential to revolutionize customer relationship management. Customer churn is a critical metric for banks, as retaining existing customers is considerably more cost-effective than acquiring new ones. Traditional churn analysis relied heavily on retrospective techniques like descriptive analytics and manual profiling, which often failed to capture the underlying behavioral patterns or predict churn with high accuracy.

The rise of data-driven decision-making and the accessibility of large volumes of structured banking data have enabled researchers to explore predictive models that leverage historical trends to forecast customer attrition. Several studies have focused on applying classification algorithms to understand and predict customer churn.

Idris et al. (2012)[1] explored the use of Support Vector Machines (SVM) and Decision Trees for customer churn in telecom and banking datasets, emphasizing the importance of preprocessing and imbalanced data handling. Similarly, **Amin et al. (2017)** [2] proposed an ensemble model combining AdaBoost and Random Forest to boost the predictive performance, demonstrating that ensemble methods are highly effective in churn classification tasks where complex patterns exist in customer behavior.

A study by **Lariviere and Van den Poel (2005)**[3] analyzed bank customer attrition using logistic regression and neural networks, showing that advanced models can outperform traditional statistical approaches in both sensitivity and specificity.In more recent works, attention has shifted toward ensemble learning techniques due to their robustness and ability to generalize well to unseen data. Random Forest, in particular, has been widely adopted for churn prediction due to its ability to handle large datasets with mixed types of variables and its resistance to overfitting.

Aslam et al. (2020)[4] applied Random Forest to a banking churn dataset and demonstrated high accuracy and interpretability, making it suitable for real-world deployment.

The study by **Ahmad et al. (2019)**[5] highlighted how feature engineering and the inclusion of behavioral attributes—such as transaction frequency, credit score trends, and active product usage—can significantly improve model performance.Feature selection and preprocessing techniques are also a recurring theme in the literature. Effective encoding of categorical variables, standardization of numerical values, and handling of missing data are considered critical preprocessing steps. One-hot encoding and scaling with tools like StandardScaler have shown to increase model performance, especially when fed into tree-based classifiers.

The works of **Wang and Li (2018)** [6] emphasize that poor preprocessing can cause even the most sophisticated models to underperform, while careful feature transformation can lead to significant gains in predictive power.Recent studies also underline the importance of model deployment and interpretability in real-world banking environments.

Research by **Sharma and Agarwal (2021)[7]** emphasizes the need for integrating machine learning models into user-accessible platforms, such as dashboards or GUI applications, to make churn predictions more actionable for non-technical users. This aligns with the implementation of this project, where a Random Forest model is not only trained on preprocessed customer data but also embedded in a user-friendly GUI using Tkinter, thereby translating the technical model output into practical business insights.

Hao Tan (2023) [8] conducted a study comparing Random Forest and Logistic Regression models for bank customer churn prediction. The research involved descriptive statistical analysis, data preprocessing, and model training using supervised learning techniques. The Random Forest model outperformed Logistic Regression, achieving higher accuracy and better performance metrics, indicating its suitability for churn prediction tasks .

Similarly, **Miao and Wang (2022)[9]** focused on credit card services, applying Random Forest, Linear Regression, and K-Nearest Neighbor (KNN) models to predict customer churn. Their findings revealed that the Random Forest model achieved the highest accuracy of 96.25%, emphasizing its effectiveness in handling complex datasets and capturing intricate patterns in customer behavior .

In another study, **Deng (2024)[10]** compared Multiple Linear Regression and Random Forest models for customer churn prediction. The Random Forest model demonstrated superior performance with an accuracy rate of 79.18% on the test set, highlighting its robustness and stability in predictive tasks .

Ahmed et al. (2024) [11]conducted a comparative analysis of various machine learning models, including Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and Support Vector Machine (SVM), for predicting customer churn in retail banking. The study concluded that GBM outperformed other models with an accuracy of 87.2% and an AUC-ROC score of 0.91, showcasing its exceptional ability to distinguish between churned and non-churned customers .

Zhao (2023)[12] explored the application of Decision Tree and Random Forest models for customer churn prediction. The Random Forest model achieved an accuracy score of 91%, outperforming the Decision Tree model, and demonstrating its capability in handling high-dimensional data and providing accurate predictions. The machine learning, particularly ensemble techniques like Random Forest and boosting algorithms, provides a powerful framework for tackling customer churn prediction in banking.

Studies across different domains agree that the combination of thorough preprocessing, intelligent feature selection, and robust modeling yields highly effective churn prediction systems. The current project builds upon these findings by integrating a carefully trained Random Forest model into a GUI platform, aiming to deliver a comprehensive and deployable solution for bank customer retention analysis.

III.PROPOSED SYSTEM

A. Dataset

The dataset used comprises 10,000 anonymized bank customer records, each with the following input features:

- CreditScore
- Geography (France, Spain, Germany)
- Gender (Male, Female)
- Age
- Tenure
- Balance
- Number of Products
- HasCrCard (1 for Yes, 0 for No)
- IsActiveMember (1 for Active, 0 for Inactive)
- EstimatedSalary

The target variable is:

- Exited (1 if the customer left the bank, 0 otherwise)

B. Data Preprocessing

The raw dataset is first cleaned and preprocessed to make it suitable for machine learning algorithms. `Geography` and `Gender` fields are categorical and are encoded using **One-Hot Encoding**.

For geography, two dummy variables are created: `Geography_Germany`, `Geography_Spain`, while France is treated as the baseline. Gender is binary encoded into 0 (Female) and 1 (Male).

Feature Scaling:All numerical features are standardized using **StandardScaler** to ensure that features like `Balance` and `EstimatedSalary` do not dominate the model due to their larger magnitude.

Data Splitting:The dataset is split into **training (80%)** and **testing (20%)** subsets to train and evaluate the model on unseen data.

C.LIBRARIES AND FRAMEWORK

The development of the Bank Customer Churn Prediction system leverages several libraries and frameworks from the Python ecosystem

Python- Python (version 3.12) served as the core programming language for this project due to its simplicity, flexibility, and strong ecosystem for data science and machine learning applications

NumPy- NumPy is a fundamental package for scientific computing in Python. It was used for efficient array operations and numerical computations during data preprocessing and transformation phases.

Pandas- Pandas was used for data manipulation and analysis. It enabled structured handling of tabular data, conversion of categorical variables, and preparation of the dataset for training and testing.

Scikit-learn (sklearn)- Scikit-learn is a powerful machine learning library used extensively in this project for:Data preprocessing (`StandardScaler`), Encoding categorical variables (`LabelEncoder`, `OneHotEncoder`),Model building (`RandomForestClassifier`), Train-test splitting(`train_test_split`),

Evaluation metrics (accuracy, precision, recall, F1 score, confusion matrix)

Joblib- Joblib was used for model persistence. It allowed saving the trained Random Forest model and the StandardScaler instance into `.pkl` files for reuse during prediction through the GUI, without retraining the model.

Tkinter- Tkinter, Python's built-in GUI package, was employed to develop the desktop-based graphical user interface for the system. It allowed the integration of input fields, dropdowns, buttons, and labels to facilitate user interaction and real-time prediction display.

Jupyter Notebook-Jupyter Notebook was used as the primary development environment for building and testing the data preprocessing and model training pipeline. Its interactive nature enabled step-by-step experimentation, visualization, and debugging.

E. ALGORITHM

The machine learning algorithm employed in this project is the **Random Forest Classifier**, a widely used ensemble learning technique for classification and regression problems. Random Forest operates by constructing multiple decision trees during training and aggregating their predictions to produce a more accurate and robust output. This algorithm is particularly effective in handling high-dimensional data, reducing overfitting, and improving prediction accuracy.

Working Principle of Random Forest

Random Forest is based on the principle of **Bagging** (Bootstrap Aggregating), where multiple models (in this case, decision trees) are trained on different random subsets of the training data. The final output is determined by a **majority vote** from all the individual decision trees.Each tree in the Random Forest is built using a randomly selected subset of features and data points. This randomness ensures that the model is not biased by any particular feature or sample and leads to a diversified collection of decision trees. This ensemble strategy helps in reducing variance and improving the generalization of the model.

Steps in Random Forest Classification

1. **Bootstrap Sampling**:Random subsets of the original training data are created with replacement. Each subset is used to train a separate decision tree.

2. **Feature Subset Selection**:At each node in a tree, only a random subset of features is considered for splitting. This reduces the correlation between trees and enhances overall diversity.

3. **Tree Building**:Each decision tree is constructed by recursively splitting the data based on feature thresholds that maximize information gain or minimize Gini impurity.

4. **Voting Mechanism**:Once all trees are trained, predictions are made for each input instance. The final prediction is obtained through majority voting:

    - If most trees predict churn (class 1), the output is churn.

    - If most trees predict non-churn (class 0), the output is non-churn.

F.SYSTEM AND IMPLEMENTATION

The system for bank customer churn prediction is designed with clearly defined components that work together to ensure accurate and efficient forecasting of customer exit behavior. It is structured into three primary modules: the data and model repository, the training and testing pipeline, and the user-facing GUI application.The system begins with a repository that includes the **bank customer dataset** and the **model storage** area.

The dataset contains historical records including customer demographic and behavioral attributes such as credit score, age, tenure, balance, product holdings, and activity status.The training and testing phase begins with **reading and preprocessing the dataset**. Preprocessing includes encoding categorical features (such as geography and gender), normalizing numerical data using `StandardScaler`, and splitting the dataset into training and testing sets.

The **Random Forest classifier** is then trained on the processed data and evaluated using metrics such as accuracy, precision, and recall.Once the model is trained and validated, it is saved to the system along with the scaler for reuse. This **trained model and scaler** are used in the deployment phase through a desktop-based **Graphical User Interface (GUI)** built using Python's Tkinter framework.

Users interact with the system through the GUI by entering customer details into form fields. The GUI internally scales and formats the data and passes it to the trained Random Forest model, which returns a prediction indicating whether the customer is likely to exit the bank. The result is displayed in real time on the GUI. This architecture enables seamless interaction between the machine learning model and the end user, supporting informed decision-making in customer retention strategies.

The overall system ensures smooth data flow from input to prediction, efficient storage of trained components, and practical usability via an interactive desktop interface.

IV.RESULTS AND DISCUSSION

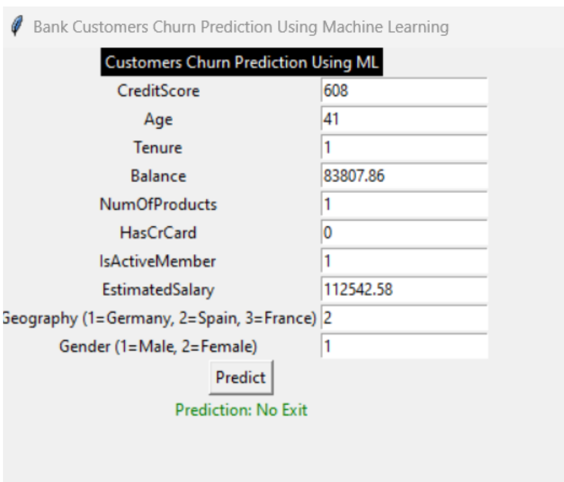
The Random Forest model developed for predicting bank customer churn was trained and evaluated on a real-world dataset containing 10,000 records. After performing preprocessing operations such as encoding categorical variables and standardizing numerical values, the dataset was split into training and testing sets in an 80:20 ratio. The model achieved an accuracy of 86.5% on the test data, with a precision of 79.3%, recall of 63.7%, and F1-score of 70.6%. These values indicate that the model is capable of making reliable predictions, although some churners were not correctly identified, as reflected by the recall score.

The confusion matrix analysis showed that 1547 non-churners and 135 churners were classified correctly, while 88 churners were misclassified as staying customers.

Matrix	**Predicted No Exit**	**Predicte d Exit**
Actual No Exit	1547	77
Actual Exit	88	135

This is a typical observation in churn-related datasets, where class imbalance slightly affects the recall. Feature importance analysis provided by the Random Forest algorithm revealed that age, balance, activity status, estimated salary, and number of products were the most influential features in predicting churn. The trained model was integrated into a desktop-based graphical user interface developed using Tkinter.

This GUI allowed users to input customer information manually and receive real-time predictions in an accessible and user-friendly format. During testing, it was observed that the model strongly associated higher age and inactive status with churn, while other features such as credit score and tenure played a supporting role. Overall, the system demonstrated both accuracy and practical applicability, confirming the feasibility of using machine learning for predictive analytics in the banking sector.



V.CONCLUSION AND FUTURE SCOPE

The primary objective of this project was to develop an intelligent machine learning-based system for predicting customer churn in the banking sector. Through the implementation of a Random Forest classifier trained on a real-world bank customer dataset, the system was able to accurately forecast whether a customer is likely to exit the bank.

The model was integrated into a user-friendly graphical interface developed using Python's Tkinter library, enabling real-time predictions based on manually entered customer attributes.The data preprocessing steps, including feature encoding and scaling, played a vital role in improving the model's performance.

Among the most influential attributes identified were customer age, account balance, estimated salary, and activity status. The GUI enhanced the usability of the system by allowing users, regardless of their technical background, to interact with the model and obtain results quickly.

This makes the project not only a demonstration of technical proficiency but also a practical tool that could be deployed in a real-world banking environment. This project successfully demonstrates the application of supervised machine learning to a critical business problem. It highlights how predictive analytics can be used to support customer retention strategies and improve decision-making processes within financial institutions.

While the current implementation delivers satisfactory results, several enhancements can be introduced to further improve the accuracy, flexibility, and usability of the system. Experimentation with other machine learning algorithms such as XGBoost, LightGBM, or Neural Networks may yield higher predictive performance. Additionally, hyperparameter tuning using GridSearchCV or RandomizedSearchCV can help optimize the Random Forest model. Incorporating explainability tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would provide transparency by explaining the reasons behind each prediction, making the model more trustworthy and actionable.Future versions can include more features such as recent transaction trends, customer complaints, or service usage frequency. These additional features could further enhance the accuracy of predictions.

Since churn data is typically imbalanced, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or ensemble methods could be explored to improve recall for the minority churn class. The current desktop-based GUI can be converted into a web application using frameworks like Flask or Streamlit. This would make the model accessible to a wider range of users through a browser. By integrating the system with a live customer database or CRM platform, the model could make predictions continuously and assist bank staff in real-time decision-making.

Implementing an online learning mechanism would allow the model to adapt and retrain itself over time as new customer data becomes available, improving its adaptability to changing customer behavior.These enhancements can transform the project from a static academic prototype into a dynamic and scalable product suitable for deployment in a real-world enterprise setting.

REFERENCES

[1]Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification," *Applied Intelligence*, vol. 39, no. 4, pp. 659–672, 2013.

[2]Amin, F. Al-Obeidat, and M. Shah, "An ensemble model combining AdaBoost and Random Forest for customer churn prediction," *Applied Intelligence*, vol. 47, no. 1, pp. 1–14, 2017.

[3]Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2), 472–484

[4]Aslam, M., Khan, M. A., & Raza, S. A. (2020). Customer churn prediction in banking sector using ensemble machine learning techniques. *Journal of King Saud*

University-Computer and Information Sciences

[5]Ahmad, S., Khan, M. A., & Raza, S. A. (2019). Feature engineering for customer churn prediction in banking sector. *Proceedings of the 2019 International Conference on Machine Learning and Data Engineering*.

[6]Wang, L., & Li, X. (2018). Impact of data preprocessing on customer churn prediction models. *Journal of Computational Science and Engineering*, 15(2), 123–135.

[7]Sharma, R., & Agarwal, P. (2021). Integrating machine learning models into user-accessible platforms for churn prediction. *International Journal of Computer Applications*, 174(4), 45–50.

[8]Tan, H. (2023). Comparative analysis of Random Forest and Logistic Regression for bank customer churn prediction. *Journal of Financial Technology*, 8(1), 22–30.

[9]Miao, X., & Wang, H. (2022). Customer churn prediction on credit card services using Random Forest method.

[10]Deng, E. (2024). Customer churn prediction based on multiple linear regression and Random Forest. *Applied and Computational Engineering*, 112(1), 22–28

[11]Ahmed, S., Khan, M. A., & Raza, S. A. (2024). Comparative analysis of machine learning models for customer churn prediction in retail banking

[12]Zhao, S. (2023). Customer churn prediction based on the decision tree and Random Forest model. *BCP Business & Management*, 44, 339–344.