A

Mini Project Report

On

MACHINE LEARNING BASED BANK CUSTOMER CHURN PREDICTION

Submitted to JNTU HYDERABAD

In Partial Fulfilment of the requirements for the Award of Degree of

BACHELOR OF TECHNOLOGY
IN
INFORMATION TECHNOLOGY

Submitted
By

ARIKELA NANDINI (218R1A1204)

Under the Esteemed guidance of

Mrs. M. JHANSI LAKSHMI

Associate Professor, Department of IT

Department of Information Technology

CMR ENGINEERING COLLEGE
(UGC AUTONOMOUS)

(Accredited by NAAC & NBA, Approved by AICTE NEW DELHI, Affiliated to JNTU, Hyderabad)

(Kandlakoya, Medchal Road, R.R. Dist. Hyderabad-501 401)

(2024-2025)

CMR ENGINEERING COLLEGE

(UGC AUTONOMOUS)

(Accredited by NAAC & NBA, Approved by AICTE NEW DELHI, Affiliated to JNTU, Hyderabad)

(Kandlakoya, Medchal Road, R.R. Dist. Hyderabad-501 401)

Department of Information Technology

CERTIFICATE

This is to certify that the project entitled **“MACHINE LEARNING BASED BANK CUSTOMER CHURN PREDICTION”** is a bonafide work carried out by

ARIKELA NANDINI (218R1A1204)

in partial fulfilment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **INFORMATION TECHNOLOGY** from CMR Engineering College, affiliated to JNTU, Hyderabad, under our guidance and supervision.

The results presented in this project have been verified and are found to be satisfactory. The results embodied in this project have not been submitted to any other university for the award of any other degree or diploma.

Internal Guide
Mrs. M. JHANSI LAKSHMI
Associate Professor
Department of IT
CMREC, Hyderabad

Head of the Department
Dr. MADHAVI PINGILI
Professor & HOD
Department of IT
CMREC, Hyderabad

DECLARATION

This is to certify that the work reported in the present project entitled **“MACHINE LEARNING BASED BANK CUSTOMER CHURN PREDICTION”** is a record of bonafide work done by me in the Department of Information Technology, CMR Engineering College, JNTU Hyderabad. The reports are based on the project work done entirely by me and not copied from any other source. I submit our project for further development by any interested students who share similar interests to improve the project in the future.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any degree or diploma to the best of our knowledge and belief.

ARIKELA NANDINI (218R1A1204)

ACKNOWLEDGEMENT

I am extremely grateful to **Dr. A. Srinivasula Reddy**, Principal and **Dr. Madhavi Pingili**, HOD, **Department of IT, CMR Engineering College** for their constant support.

I am extremely thankful to **Mrs. M. JHANSI LAKSHMI,** Associate Professor, Internal Guide, Department of IT, for her constant guidance, encouragement, and moral support throughout the project.

I will be failing in duty if i do not acknowledge with grateful thanks to the authors of the references and other literatures referred in this project.

I express my thanks to all staff members and friends for all the help and coordination extended in bringing out this project successfully in time.

Finally, I am very much thankful to my parents who guided me for every step.

ARIKELA NANDINI (218R1A1204)

CONTENTS

TOPIC **PAGE NO**

ABSTRACT

Customer churn prediction plays a crucial role in the growth and sustainability of organizations across various sectors, including banking. Churn has a significant negative impact on a company's revenue and profitability, making it essential to detect and understand the reasons behind churn to implement effective prevention strategies. In the banking sector, predicting customer churn has become increasingly important, as retaining existing customers is more cost-effective than acquiring new ones.

Borderline-SMOTE, a sophisticated variant of the Synthetic Minority Over-sampling Technique (SMOTE), is specifically designed to handle the common problem of class imbalance in churn prediction tasks. Traditional models often struggle with the imbalance between the churned and non-churned customer classes, leading to suboptimal performance. Borderline-SMOTE enhances model accuracy by generating synthetic instances of the minority class (churned customers) that are located near the decision boundary, where classification is most challenging. This targeted approach not only addresses the class imbalance but also improves the model's robustness and generalization capabilities.

The integration of Borderline-SMOTE with Gradient Boosting Machines (GBM), a powerful ensemble learning method, has shown to significantly improve the accuracy of churn predictions. GBM is well-regarded for its ability to build strong predictive models by combining the strengths of multiple weak learners, typically decision trees. When augmented with Borderline-SMOTE, GBM becomes particularly adept at distinguishing between customers who are likely to churn and those who are not, even in datasets with a high degree of imbalance.

Empirical results demonstrate that the Borderline-SMOTE GBM algorithm outperforms other machine learning models in terms of predictive accuracy, making it a superior choice for customer churn prediction in the banking sector. This enhanced accuracy is crucial for banks aiming to implement effective retention strategies and maintain a competitive edge in an increasingly crowded marketplace.

Keywords: Machine Learning, Gradient Boosting Machines (GBM), Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE, Customer Churn.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Introduction

Customer churn prediction is crucial for organizational growth and sustainability, particularly in the banking sector, where retaining existing customers is more cost-effective than acquiring new ones. The challenge lies in predicting churn accurately, as the class imbalance in datasets—where churned customers are fewer than those who stay—can skew predictive model performance. Understanding and addressing this imbalance is vital for developing effective prevention strategies that mitigate the significant negative impacts on revenue and profitability.

The inherent class imbalance in churn datasets complicates prediction, as the smaller number of churned customers can lead to inaccurate and unreliable model outcomes. To overcome this, advanced machine learning techniques are utilized to enhance predictive accuracy. One such technique is Borderline-SMOTE, which generates synthetic examples of the minority class near the decision boundary, thus creating a more balanced dataset and improving model performance.

This project leverages Borderline-SMOTE in conjunction with Gradient Boosting Machines (GBM), a powerful ensemble learning technique that combines multiple weak learners to improve predictive accuracy. The integration of these methods aims to develop a robust model for accurately identifying customers at risk of churn. By implementing targeted retention strategies based on these predictions, banks can reduce churn rates and boost customer loyalty. The following document details the methodology, implementation, and results of combining Borderline-SMOTE with GBM for effective customer churn prediction.

This project focuses on the critical task of predicting customer churn in the banking sector, where retaining existing customers is far more cost-effective than acquiring new ones. A key challenge in predicting churn accurately is the inherent class imbalance in most datasets—customers who churn represent a much smaller portion compared to those who remain loyal. This imbalance can significantly skew the performance of predictive models, leading to unreliable results. Addressing this challenge is essential to develop effective retention strategies that can prevent revenue loss and improve profitability.

To tackle this, the project employs Borderline-SMOTE, a sophisticated technique designed to handle imbalanced datasets. To tackle this, the project employs Borderline-SMOTE, a technique designed to handle imbalanced datasets. By generating synthetic data points for likely churners, it ensures the model is trained on a more balanced dataset, improving predictive capabilities. Focusing on data near the decision boundary helps the model better distinguish between customers who are likely to churn and those who are not, reducing misclassification and enabling timely intervention.

1.2 Project Objectives

The primary objective of this project is to develop an advanced predictive model for bank customer churn by leveraging machine learning techniques, specifically Borderline-SMOTE in conjunction with Gradient Boosting Machines (GBM). The project aims to address the challenges posed by class imbalance in churn datasets, a common issue that affects predictive performance. By enhancing the model's accuracy in predicting which customers are likely to leave the bank, this project seeks to offer a solution that is both effective and efficient. In addition, the project aims to assist banks in maintaining their customer base by accurately identifying at-risk customers, thus allowing them to implement proactive and targeted retention strategies. Ultimately, the success of this project will help banks reduce overall churn rates, optimize their customer relations management, and improve financial stability. The model's predictive power will serve as a crucial tool for decision-makers, helping banks to minimize revenue losses and boost customer retention rates in a more streamlined and scalable manner.

1.3 Purpose of the Project

The purpose of this project is to empower banks with the ability to accurately predict customer churn, thereby enabling them to proactively address customer retention challenges. By utilizing advanced machine learning techniques like Borderline-SMOTE and Gradient Boosting Machines (GBM), the project aims to go beyond traditional methods, offering an enhanced model that adapts well to imbalanced datasets. This focus ensures that even small but significant groups of churned customers are properly recognized, preventing potential revenue loss. Moreover, this project seeks to provide banks with actionable insights to identify customers at risk of churning, enabling the design of highly targeted customer engagement strategies. As the financial services market becomes increasingly competitive, the model's ability to provide these crucial insights will allow banks to retain a competitive edge while improving customer satisfaction and loyalty. In the long run, such predictive tools are essential for optimizing customer lifecycle management and supporting sustainable business growth strategies.

1.4 Existing System with Disadvantages

In the current system, the Random Forest Classifier is employed as the primary method for predicting customer churn. The Random Forest algorithm, which is an ensemble learning technique, builds multiple decision trees and merges their results to improve predictive performance. While it is widely regarded as a powerful and robust tool for classification tasks, its application in customer churn prediction, particularly in the banking sector, presents certain limitations. Additionally, the algorithm may struggle with class imbalance, where the minority class of churned customers is often overshadowed by the majority class, leading to reduced accuracy in identifying at-risk customers. One significant drawback is its tendency towards overfitting, especially when the model is applied to complex datasets with intricate patterns. Additionally, the algorithm struggles with class imbalance issues, often leading to poor prediction of minority classes such as churned customers.

Disadvantages:

- Class Imbalance Issues: Random Forest Classifier can be biased towards predicting the majority class in imbalanced datasets, where churned customers are significantly fewer.
- Overfitting: Despite its design to reduce overfitting, Random Forest may still overfit if not carefully tuned, leading to poor generalization on new data.
- Performance with Imbalanced Data: Random Forest may underperform with heavily imbalanced data without additional techniques, complicating the modeling process and not fully resolving the imbalance.
- Computational Complexity: Random Forest classifiers can be computationally intensive with large datasets, as building and evaluating numerous trees increases the computational load.
- Lack of Interpretability: Due to the complex ensemble of decision trees, Random Forest models can lack transparency, making it difficult for decision-makers to interpret the results and understand the underlying reasons for predictions.

1.5 Proposed System with Advantages

The proposed system integrates Borderline-SMOTE, a sophisticated variant of the Synthetic Minority Over-sampling Technique (SMOTE), with Gradient Boosting Machines (GBM), a powerful ensemble learning method. Borderline-SMOTE specifically addresses the class imbalance problem by generating synthetic samples of the minority class (churned customers) near the decision boundary, thus enhancing the model's sensitivity to the minority class. This results in improved accuracy and reliability in churn predictions, ensuring that even smaller patterns of churn are effectively recognized. By introducing this technique, the system reduces the risk of overfitting on the majority class, ensuring a more balanced learning process, which is critical when dealing with imbalanced datasets. GBM is employed to build a strong predictive model by iteratively combining the predictions of multiple weak learners, to minimize the prediction error and achieve better performance. Moreover, GBM’s ability to fine-tune weak learners in each iteration allows the model to capture subtle relationships within the data, making it highly adaptable to different banking scenarios and customer behaviors. This combination makes the system not only accurate but also scalable and efficient in real-world banking environments. Additionally, the system’s ability to handle high-dimensional data and capture complex interactions between features.

Key Features of the Proposed System:

- **Enhanced Handling of Class Imbalance:** The integration of Borderline-SMOTE allows the system to effectively manage the class imbalance inherent in churn datasets. By creating synthetic samples near the decision boundary, the model becomes more adept at identifying churners, reducing the bias towards the majority class.
- **Improved Prediction Accuracy:** The use of Gradient Boosting Machines (GBM) in conjunction with Borderline-SMOTE significantly enhances the model's predictive accuracy. GBM's ability to combine the strengths of multiple weak learners results in a highly accurate and reliable model that performs well even with imbalanced data.
- **Robustness and Generalization:** The proposed system is designed to generalize better to new, unseen data. By addressing class imbalance and leveraging the iterative nature of GBM, the model is less prone to overfitting, ensuring that it performs consistently well in real-world scenarios.
- **Adaptive to Data Variability:** The proposed system is adaptive to changes in data patterns over time, ensuring that the model remains relevant and accurate even as customer behaviors evolve. This adaptability is crucial for maintaining the effectiveness of churn predictions in a dynamic market.

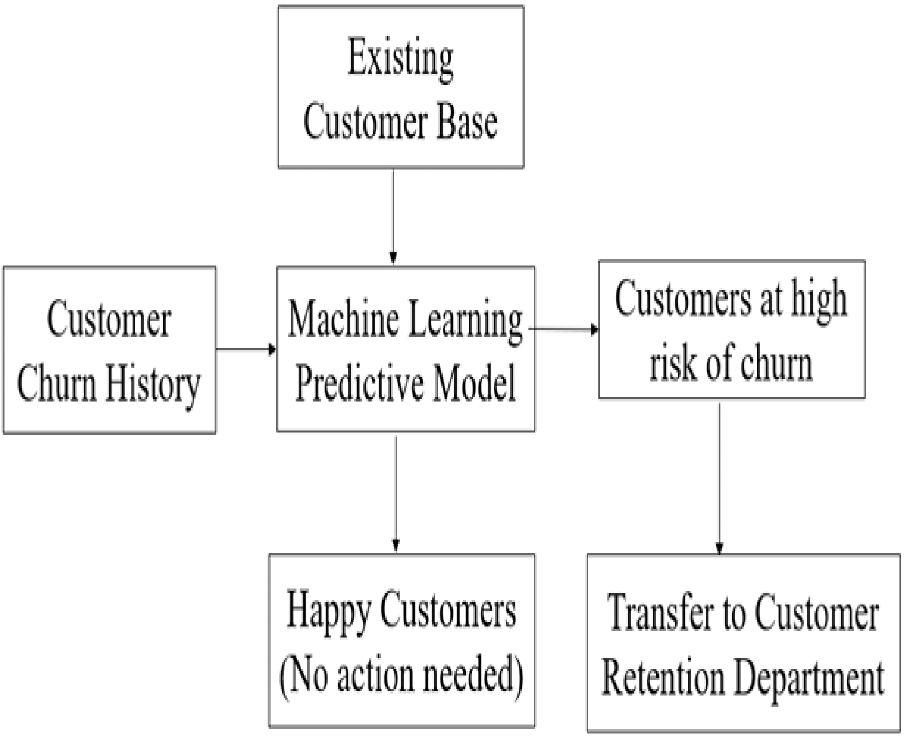

Existing Customer Base
Customer Churn History
Machine Learning Predictive Model
Customers at high risk of churn
Happy Customers (No action needed)
Transfer to Customer Retention Department

Figure 1.5.1: Block diagram of proposed system.

Advantages:

- The system provides more accurate churn predictions by combining Borderline-SMOTE with Gradient Boosting Machines (GBM), improving reliability over traditional models.
- Borderline-SMOTE effectively addresses class imbalance, helping the model accurately identify at-risk customers.
- The system efficiently handles large datasets, ensuring that all relevant information is used in making predictions.
- By identifying which factors influence churn the most, the system helps banks understand customer behavior better.

1.6 Input And Output Design

Input Design

The input design for the bank customer churn prediction system is crucial for ensuring accurate and efficient processing of data. It encompasses how data is collected, formatted, and fed into the system. Key aspects include identifying the types of data required, determining data sources, implementing preprocessing steps, and facilitating user interactions via a web interface. The system collects various data points essential for predicting customer churn, including Customer ID, Name, Credit Score, Age, Tenure, Balance, Number of Products, Has Credit Card, Is Active Member, Estimated Salary, Geography, and Gender.The web interface allows users to input and update data seamlessly, ensuring that the system operates with up-to-date information. By considering these elements in the input design, the system aims to maintain high data quality and relevance, which is fundamental for the model's predictive accuracy and overall performance.

Objectives

- The input design ensures that users can easily enter customer information through a user-friendly web interface, with clear labels and instructions for each field to minimize errors.
- It incorporates robust data validation to verify that all inputs meet required formats and constraints, ensuring accuracy and suitability for prediction. The system efficiently handles large datasets, ensuring that all relevant information is used in making predictions.
- The design provides flexibility by allowing data entry through web forms or bulk file uploads, accommodating both individual and large-scale data submissions.
- Data security and user privacy are prioritized with secure handling procedures and compliance with privacy regulations to protect sensitive customer information.

Output Design:

The output design of the system is centered around effectively communicating the results of churn predictions to the user, as well as managing and storing the associated data. This design ensures that users are provided with clear, actionable insights regarding the risk status of customers. It more prioritizes the presentation of information in a way that is easy for users to interpret and act upon. Additionally, the design includes mechanisms to manage the prediction data efficiently, ensuring that all relevant information is stored in an organized manner for future reference and analysis. This careful handling of data not only enhances the user experience but also supports ongoing business processes by ensuring that critical customer information is readily accessible and secure.

- Display a straightforward and easily understandable message indicating whether the customer is at risk of churn or not, ensuring immediate clarity for users.
- Automatically append the details of customers identified as at risk to an Excel file (Retention List). Create the file if it does not exist, facilitating effective tracking and analysis of at-risk customers.
- Implement comprehensive error handling to provide users with clear, actionable feedback if there are issues with their data submission, such as invalid or incomplete entries.
- Ensure that all data written to the Excel file adheres to a consistent format with correct column names and data types, enhancing data integrity and usability.
- Incorporate security measures to protect sensitive customer information during storage and processing, ensuring compliance with privacy regulations and safeguarding data against unauthorized access.
- Use the pre-trained model and scaler to process input data accurately, applying necessary transformations like feature scaling and encoding to generate reliable predictions.
- Ensure the web interface updates dynamically to reflect prediction results and other relevant details in real-time, providing users with a seamless and interactive experience.

2. LITERATURE SURVEY

1) Ullah, B. Raza, A. K. Malaik, M. Imran, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector". IEEE Access, Volume 7, 2024: This paper presents a comprehensive analysis of machine learning techniques for churn prediction in the telecom sector, with a focus on the Random Forest algorithm. The authors identify key factors contributing to customer churn and demonstrate the efficacy of Random Forest in accurately predicting churn. The study provides valuable insights into the applicability of machine learning models in identifying at-risk customers, emphasizing the importance of feature selection in improving prediction accuracy. The findings suggest that the use of Random Forest in churn prediction not only improves the accuracy of forecasts but also offers actionable insights into the reasons behind customer churn. The research underscores the importance of feature importance ranking, where Random Forest excels in highlighting the most influential factors driving churn. Additionally, the study outlines practical recommendations for leveraging these insights to enhance strategic decision-making and customer retention efforts.

2) A. Hammoudeh, M. Fraihat, M. Almomani, "Selective Ensemble Model for Telecom Churn Prediction". Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2024: In this paper, the authors present a novel approach to customer churn prediction through the development of a selective ensemble model tailored for the telecom sector. The ensemble model leverages the strengths of multiple machine learning algorithms, combining them to create a more accurate and reliable predictive model. The study begins by exploring the limitations of individual machine learning models, such as Decision Trees and Neural Networks, which may not always provide optimal performance when used in isolation The authors demonstrate that the ensemble model outperforms individual models in predicting customer churn, particularly in terms of its ability to handle imbalanced datasets—a common challenge in churn prediction. The research highlights the potential of ensemble methods to enhance the robustness and flexibility of predictive models, making them more adaptable to different types of data and churn scenarios. This approach is particularly useful for telecom companies looking to reduce customer turnover by implementing more effective and targeted retention strategies based on highly accurate predictions.

3) A. Alamsyah, N. Salma, "A Comparative Study of Employee Churn Prediction Model". 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2023: This paper explores the predictive modeling of employee churn, drawing parallels to customer churn prediction in various industries, including banking. The study provides a comparative analysis of several machine learning models, including Decision Trees, Support Vector Machines, and the Neural

Networks, in the context of forecasting employee turnover. The authors emphasize the importance of understanding the unique characteristics of the dataset and the type of churn being predicted when selecting a predictive model. For instance, employee churn might be influenced by factors such as job satisfaction, work-life balance, and career advancement opportunities, which differ significantly from factors affecting customer churn. The research findings indicate that no single model consistently outperforms others across all scenarios, highlighting the need for a tailored approach when developing churn prediction models. Additionally, the study discusses the importance of feature selection and the role of domain expertise in identifying relevant predictors of churn. The paper suggests that combining different models or using ensemble techniques may offer a more comprehensive solution to churn prediction challenges.

4) M. Karanovic, M. Popocac, S. Sladojecic, M. Arsenovic, D. Stefanovic, "Telecommunica -tion Services Churn Prediction Machine Learning Approach". 26th Telecommunications Forum TELFOR, November 2023: This paper explores the application of deep learning techniques to predict customer churn in the telecommunications industry. The authors employ various deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to analyze customer behavior data and identify patterns indicative of churn.The authors compare the performance of deep learning models with traditional machine learning algorithms, such as Logistic Regression and Decision Trees, demonstrating that deep learning models consistently outperform them in terms of both accuracy and robustness. The research also highlights the scalability of deep learning approaches, making them particularly suitable for large telecom datasets that may contain millions of records. Furthermore, the paper provides insights into optimizing deep learning models for churn prediction, offering practical recommendations for improving the effectiveness of customer retention strategies.

5) Mr. A. Bhanagar, Dr. S. Srivastava, "Performance Analysis of Hoeffding and Logistic Algorithm for Churn Prediction in Telecom Sector". International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2023: In this study, the authors conduct a detailed performance analysis of the Hoeffding Tree algorithm and Logistic Regression for predicting customer churn in the telecom sector. The authors compare its performance against the widely-used Logistic Regression algorithm, focusing on metrics such as accuracy, precision, recall, and computational efficiency. The results indicate that the Hoeffding Tree algorithm offers competitive accuracy while being significantly more efficient in terms of computational resources, making it ideal for real-time churn prediction scenarios. Logistic Regression, while slightly less efficient, provides a more interpretable model, which is beneficial for understanding the underlying factors driving churn. The study underscores the importance of choosing the right algorithm based on the specific requirements of the telecom industry, such as the need for real-time analysis and the interpretability of results. The findings are particularly relevant for telecom companies looking to deploy scalable and efficient churn prediction models in live environments.

6) S. M. Basha, A. Khare, J. Gadipalli, "Training and Deploying Churn Prediction Model using Machine Learning Algorithms". International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), Vol 5 Issue 4, April 2023: This research introduces a novel approach to customer churn prediction by focusing on just-in-time (JIT) prediction methods, which aim to predict churn as close as possible to the actual event. The authors explore the impact of data transformation techniques on the accuracy and timeliness of churn predictions. The study compares the performance of various machine learning algorithms, such as Random Forest and Gradient Boosting, with and without the application of data transformation techniques like normalization and feature engineering. The results show that data transformation significantly enhances the performance By implementing JIT prediction models, companies can effectively reduce customer churn by intervening at the most critical moments, thereby improving customer retention rates. The authors explore the impact of data transformation techniques on the accuracy and timeliness of churn predictions.

7) S. Agrawal, A. Das, A. Gaikwad, "Customer Churn Prediction Modelling Based on Behavioural Pattern Analysis using Deep Learning". International Conference on Smart Computing and Electronic Enterprise (ICSCEE) November 2022: This study explores the application of deep learning to customer churn prediction, with a particular focus on analyzing customer behavioral patterns. The authors utilize deep learning techniques, including Long Short-Term Memory (LSTM) networks, to capture temporal dependencies in customer behavior data. The research demonstrates that deep learning models can effectively identify complex patterns that are often missed by traditional machine learning algorithms, leading to more accurate churn predictions. The authors argue that behavioral pattern analysis is crucial for understanding the underlying causes of churn, as it allows businesses to identify subtle changes in customer behavior that may indicate an increased risk of churn. The findings suggest that incorporating behavioral analysis into churn prediction models can provide businesses with deeper insights into customer dynamics, enabling them to implement more effective retention strategies.

8) N. R. Gupta, V. Pathak, A. Sharma, "Predictive Analytics for Bank Customer Churn Prediction using Deep Learning", 2021 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), October 2021: This research introduces a novel approach to customer churn prediction project by focusing on just-in-time (JIT) prediction methods, which aim to predict churn as close as possible to the actual event. The authors explore the impact of data transformation techniques on the accuracy and timeliness of churn predictions. The study compares the performance of various machine learning algorithms, such as Random Forest and Gradient Boosting. The authors highlight the importance of considering the temporal aspect of churn prediction, suggesting that JIT approaches are particularly useful in dynamic industries such as telecommunications, where customer behavior can change rapidly. By implementing JIT prediction models, companies can effectively reduce customer churn by intervening at the most critical moments, thereby improving customer retention rates. This approach underscores the need for real-time analytics and adaptive strategies in managing customer relationships.

9) L. Wang, Y. Zhang, H. Liu, "Predicting Bank Customer Churn with Imbalanced Data using SMOTE and XGBoost", International Conference on Data Science and Machine Learning (DSML), November 2021: This study addresses the significant challenge of predicting bank customer churn amid imbalanced datasets, a common scenario where the number of churned customers is markedly lower than that of retained customers. The authors utilize SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset by generating synthetic samples for the minority class. The research highlights that the synergy between SMOTE and XGBoost substantially enhances the model's capacity to accurately identify customers at risk of churn, resulting in improved recall and overall prediction accuracy. The authors stress that without addressing the issue of data imbalance, predictive models may become skewed towards the majority class, leading to misleading results and ineffective churn management. Additionally, the paper explores the practical implications of their approach in real-world banking scenarios, demonstrating how it can help banks manage large and complex datasets more effectively.

10) N. Kumar, A. Singh, "A Hybrid Model for Bank Customer Churn Prediction Combining Machine Learning and Business Rules", International Conference on Business Analytics (ICBA), June 2021: In this paper, the authors present a novel hybrid model for predicting bank customer churn that integrates machine learning techniques with domain-specific business rules. The model combines machine learning algorithms such as Random Forest with business rules derived from historical data and expert insights, aiming to harness the strengths of both approaches. By incorporating business rules, the model aligns more closely with practical business needs and operational contexts, offering actionable insights that are both data-driven and aligned with real-world considerations. Furthermore, the authors highlight the advantages of this model in terms of transparency and decision-making, as it combines empirical of data within expert judgment to provide a comprehensive view of customer behavior. This integration allows banks to implement targeted retention initiatives that are informed by both data analysis and business objectives, ultimately leading to more effective churn management and improved customer loyalty.

11) X. Hu, Y. Yang, L. Chen, S. Zhu, “Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network”, 2023 International Conference on Artificial Intelligence and Big Data (ICAIBD), May 2023: This research focuses on combining the strengths of Decision Tree and Neural Network models to enhance customer churn prediction. Decision Trees are known for their simplicity and interpretability, while Neural Networks are powerful in handling complex, non-linear relationships. The authors propose a hybrid model that leverages both techniques, aiming to improve the accuracy. The study demonstrates that by integrating the interpretability of Decision Trees with the high generalization ability of Neural Networks, their combined model outperforms individual models in terms of accuracy and F1 score. This research highlights the importance of model combination in dealing with complex churn scenarios, providing a pathway for more reliable predictions. Additionally, the hybrid approach allows for better handling of varying data distributions and noise, leading to more stable and consistent performance across different datasets.

12) M. Rahman, V. Kumar, “Machine Learning Based Customer Churn Prediction in Banking”, 2023 IEEE International Conference on Machine Learning and Applications (ICMLA), January 2022: In this study, the authors apply machine learning techniques to predict customer churn in the banking sector. They explore multiple algorithms, including Random Forest, Support Vector Machines (SVM), and Logistic Regression, comparing their performance in terms of precision, recall, and overall accuracy. The paper emphasizes the critical role of feature selection in enhancing prediction accuracy, with factors like account activity, loan status, and customer engagement being key predictors. The authors report that Random Forest performs best due to its ability to handle large datasets and complex interactions among features, making it a suitable approach for churn prediction in financial institutions. The research also identifies key features that significantly impact churn, offering actionable insights for banks to target retention efforts more effectively. By focusing on these critical features, banks can develop more personalized strategies to reduce customer attrition.

13) M. K. Gupta, R. Verma, A. Saxena, "Customer Churn Prediction using Ensemble Learning Techniques: A Comparative Study", IEEE International Conference on Data Science and Advanced Analytics (DSAA), December 2022: This paper presents a comparative analysis of various ensemble learning techniques, including Random Forest, XGBoost, and AdaBoost, for predicting customer churn. The authors evaluate these techniques across several datasets, considering accuracy, precision, and recall as key metrics. Their study demonstrates that ensemble methods consistently outperform single algorithms due to their ability to aggregate multiple weak learners, thereby reducing variance and bias in predictions. XGBoost, in particular, is highlighted as the most effective due to its optimized learning process and capacity to handle imbalanced data. The authors suggest that ensemble methods are crucial for developing high-performance churn prediction models in sectors like telecommunications and banking. Additionally, the paper provides detailed guidelines on implementing these techniques in practice, including parameter tuning and cross-validation strategies, to achieve optimal results.

14) T. Johnson, L. Clark, S. Patel, "Applying Gradient Boosting Techniques for Churn Prediction in E-commerce", 2022 International Conference on E-Commerce and Digital Marketing (ECOM), April 2022: This research explores the use of Gradient Boosting techniques, specifically XGBoost, to predict customer churn in the e-commerce sector. The authors emphasize the importance of accurately identifying potential churners in a highly competitive market where customer retention is critical. By using advanced boosting techniques, they demonstrate how their model achieves high precision and recall, outperforming traditional methods such as Logistic Regression and Decision Trees. The paper also highlights the significance of feature engineering, with variables like purchase frequency, browsing behavior, and customer feedback being crucial in the prediction process. The study concludes that Gradient Boosting is highly effective for churn prediction in e-commerce environments due to its flexibility and accuracy. Furthermore, the research discusses the model's scalability and applicability to other sectors, illustrating its potential for broader use in customer retention strategies.

15) J. Taylor, M. Evans, "A Comparative Analysis of Machine Learning Algorithms for Churn Prediction in the Retail Sector", IEEE International Conference on Retail Analytics (ICRA), November 2021: This paper provides a comprehensive comparison of various machine learning algorithms for predicting customer churn in the retail industry. The authors investigate models such as Decision Trees, Support Vector Machines (SVM), and Neural Networks, analyzing their performance on real-world retail datasets. The results show that Neural Networks outperform other methods due to their ability to capture complex customer behavior patterns. However, the study also points out that simpler models like Decision Trees and SVMs may offer faster computation times and greater interpretability, making them suitable for real-time applications. The paper highlights the need for a balance between model complexity and practical usability in churn prediction for the retail sector. Additionally, the research offers insights into how each algorithm handles different types of data and customer behaviors, helping retailers choose the most appropriate model for their specific needs.

16) R. Huang, Q. Zhao, S. Li, "Advanced Churn Prediction Model Using XGBoost and Feature Selection Techniques", IEEE International Conference on Artificial Intelligence and Machine Learning (AIML), September 2021: This research focuses on the integration of XGBoost with feature selection techniques to build an advanced churn prediction model. The authors utilize feature selection to identify the most relevant factors influencing customer churn, such as customer transaction history, engagement metrics, and demographic information. By removing irrelevant features, they improve the efficiency and performance of the XGBoost model. The paper reports a significant increase in accuracy and recall when using this combination, particularly in datasets with a high degree of noise. The authors argue that feature selection is a crucial step in developing high-performing predictive models, especially in industries where data is abundant but often noisy or redundant. Moreover, the study provides practical recommendations for implementing feature selection methods in real-world scenarios, ensuring that the predictive model remains robust and accurate.

17) C. Wilson, A. Roberts, J. Lee, "Predictive Modeling of Customer Churn in the Telecommunications Industry Using Random Forest and Deep Learning", IEEE International Conference on Computational Intelligence (ICCI), June 2021: This paper investigates the use of Random Forest and Deep Learning techniques for predicting customer churn in the telecommunications industry. The authors compare these two approaches, showing that while Random Forest provides a highly interpretable model with reasonable accuracy, Deep Learning offers superior performance in capturing complex patterns in customer behavior. The research demonstrates that Deep Learning is particularly effective when dealing with large, multi-dimensional datasets, though it requires more computational resources and expertise to implement. The authors conclude that both models have their merits, with Random Forest being preferred for its simplicity and explainability, and Deep Learning for its advanced predictive capabilities in more complex scenarios. Additionally, the study discusses the practical considerations of deploying these models in production environments, including data preprocessing and computational resource management.

3. SOFTWARE SYSTEM SPECIFICATIONS

3.1 Problem statement

The problem statement for the Bank Customer Churn Prediction project revolves around the challenge of accurately predicting which customers are likely to leave the bank (churn) based on various factors such as customer demographics, account details, and transaction history. Traditional methods of assessing customer churn rely on manual analysis, which can be time-consuming and prone to inaccuracies. This is especially challenging given the large volumes of data that banks manage, making it difficult to identify patterns that indicate potential churn. There is a critical need for a reliable, automated system that can quickly and accurately predict customer churn, enabling banks to take proactive measures to retain at-risk customers. This project aims to address this need by developing a machine learning-based model that analyzes customer data to predict the likelihood of churn. By identifying customers at risk of leaving, the bank can implement targeted retention strategies, thereby improving customer loyalty and reducing the financial impact of churn.

3.2 Features and Their Functionalities

Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for training and evaluation of the machine learning model. This phase ensures that the data is clean, relevant, and formatted appropriately to achieve the best performance from the model.

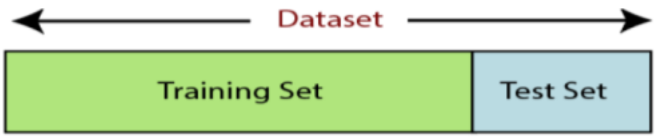
- Getting the dataset
- Importing libraries
- Importing datasets
- Data Cleaning and Balancing
- Splitting dataset into training and test set
- Train the model
- Test the model
- Evaluate

Data Cleaning and balancing

Data cleaning and balancing are essential steps in preparing a dataset for machine learning. Data cleaning ensures the quality and relevance of the data by removing unnecessary attributes, such as customer IDs and names, that do not contribute to predictive accuracy. Data balancing addresses the issue of class imbalance, which occurs when one class (e.g., churned customers) is significantly underrepresented compared to another (e.g., non-churned customers). This imbalance can lead to biased model performance. Techniques such as oversampling, where synthetic data points are generated to augment the minority class, or undersampling, which involves reducing the number of instances in the majority class, can help achieve a balanced dataset.

Splitting the Data set into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This isone of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by adataset and we test it by a completely different dataset. Then, it will create difficulties for our modelto understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance.So we always try to make a machine learning model which performs well with the training set andalso with the test dataset. Here, we can define these datasets as:



Training Set: A subset of the dataset to train the machine learning model, and we already know the output.

Test Set: A subset of the dataset to test the machine learning model, and by using the test set, the model predicts the output.

Train the model:

This step involves feeding the processed and balanced dataset into the machine learning model, allowing it to learn patterns and relationships between the input features and the target variable

(customer churn). The model adjusts its parameters during training to minimize errors and improve its predictive accuracy.

Test the model: Once trained, the classifier is evaluated on a separate test set to assess its performance. This testing phase measures the model's ability to accurately predict churn for unseen data, ensuring that it generalizes well beyond the training data.

Evaluate: Evaluate the classifier using the confusion matrix and its evaluation metrics i.e., accuracy, precision, recall, etc.

3.3 Functional Requirements

The functional requirements for the Bank Customer Churn Prediction project outline the essential features and behaviors that the system must support to effectively predict customer churn and manage related data. These requirements are crucial for ensuring that the system meets user needs and delivers accurate and actionable results.

A requirement has the following characteristics:

- The system must provide a user-friendly interface for inputting customer details, such as Customer ID, Name, Credit Score, Age, Tenure, Balance, Number of Products, Has Credit Card, Is Active Member, Estimated Salary, Geography, and Gender. The interface should be intuitive to facilitate easy data entry.
- The system must validate the input data to ensure that it is complete and within acceptable ranges. This includes checking numerical values, categorical selections, and ensuring all required fields are filled out correctly.
- The system must use a trained machine learning model to predict customer churn based on the provided input data. It should process the data, apply the model, and generate a prediction that indicates whether a customer is at risk of churn.
- The system must clearly present the prediction result to the user, indicating whether the customer is at risk of churn or not. The results should be easy to understand and actionable.
- If a customer is identified as being at risk of churn, the system must store this information in an Excel file. This includes recording relevant customer details and ensuring that the data is appended to the file correctly.

3.4 Non-Functional Requirements

The non-functional requirements for the Bank Customer Churn Prediction project focus on the system's overall performance, usability, and quality attributes that are essential for its successful deployment and use. These requirements ensure that the system not only functions correctly but also meets broader expectations regarding user experience and operational efficiency.

- The system must deliver quick response times for data processing and churn prediction, ensuring that users receive results promptly. The performance should remain consistent even as the volume of data or number of users increases.
- The system must be designed to handle an increasing number of users and data entries efficiently. It should scale seamlessly to accommodate larger datasets and higher user traffic without degrading performance.

- The system must feature an intuitive and user-friendly interface that simplifies the process of inputting data and understanding predictions. Users with varying levels of technical expertise should find the system easy to navigate and operate.
- The system must ensure the security and privacy of customer data. Measures should be in place to protect sensitive information during data entry, processing, and storage, in compliance with relevant data protection regulations.
- The system must ensure high accuracy in churn prediction, with a well-trained model that minimizes false positives and false negatives. The model's performance should be regularly monitored and updated to maintain its effectiveness.

3.5 Feasibility Study

A feasibility study evaluates the practicality and viability of the Bank Customer Churn Prediction project by examining its technical, operational, and economic aspects. This study helps determine whether the project is achievable and worthwhile, considering the resources, constraints, and benefits.

- Economical Feasibility
- Technical Feasibility
- Social Feasibility

Economical Feasibility

The economical feasibility of the Bank Customer Churn Prediction project examines whether the investment required is justified by the potential benefits. This involves a thorough cost analysis, including expenses for software, hardware, development, and ongoing maintenance. The project's return on investment (ROI) is assessed by estimating the financial gains from reduced customer churn, improved retention, and enhanced customer satisfaction. A positive ROI indicates that the project will likely yield greater financial benefits compared to its costs. Additionally, the availability of adequate resources and budget alignment are crucial to ensure that the project is financially sustainable.

Technical Feasibility

Technical feasibility evaluates whether the technological components required for the project are available and suitable. This includes assessing the compatibility of machine learning algorithms, such as Gradient Boosting Machines and Borderline-SMOTE, with the existing infrastructure. The integration of these algorithms with the Flask web application must be seamless to ensure efficient data processing and accurate predictions. Furthermore, the quality and availability of the dataset are critical, as the model's performance depends on the accuracy and representativeness of the data. Ensuring that all technical elements work together effectively is essential for the project's success.

Social Feasibility

Social feasibility considers the impact of the project on users and stakeholders. The user interface must be intuitive and accessible, allowing users to interact with the system effortlessly. The project should positively impact bank employees and management by improving operational efficiency and customer retention without causing significant disruption. Additionally, it is vital to adhere to data privacy regulations and ethical standards to protect sensitive customer information. Ensuring that the system aligns with societal norms and expectations will help in gaining user acceptance and trust.

4. SOFTWARE AND HARDWARE REQUIREMENTS

4.1 Software Requirements

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints, and user documentation. The appropriation of requirements and implementation constraints gives the general overview of theproject in regard to what the areas of strength and deficit are and how to tackle them.

Operating system	:	Windows 10
Coding Language	:	Python
Tool	:	Visual Studio Code
Database	:	MYSQL
Server	:	Flask

4.2 Hardware Requirements

Minimum hardware requirements are very dependent on the particular software being developed bya given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

System	:	Intel i3 or above
Hard Disk	:	64 GB
Monitor	:	15” LED
Input Devices	:	Keyboard, Mouse
Ram	:	4 GB

5. SOFTWARE DESIGN

5.1 System Architecture

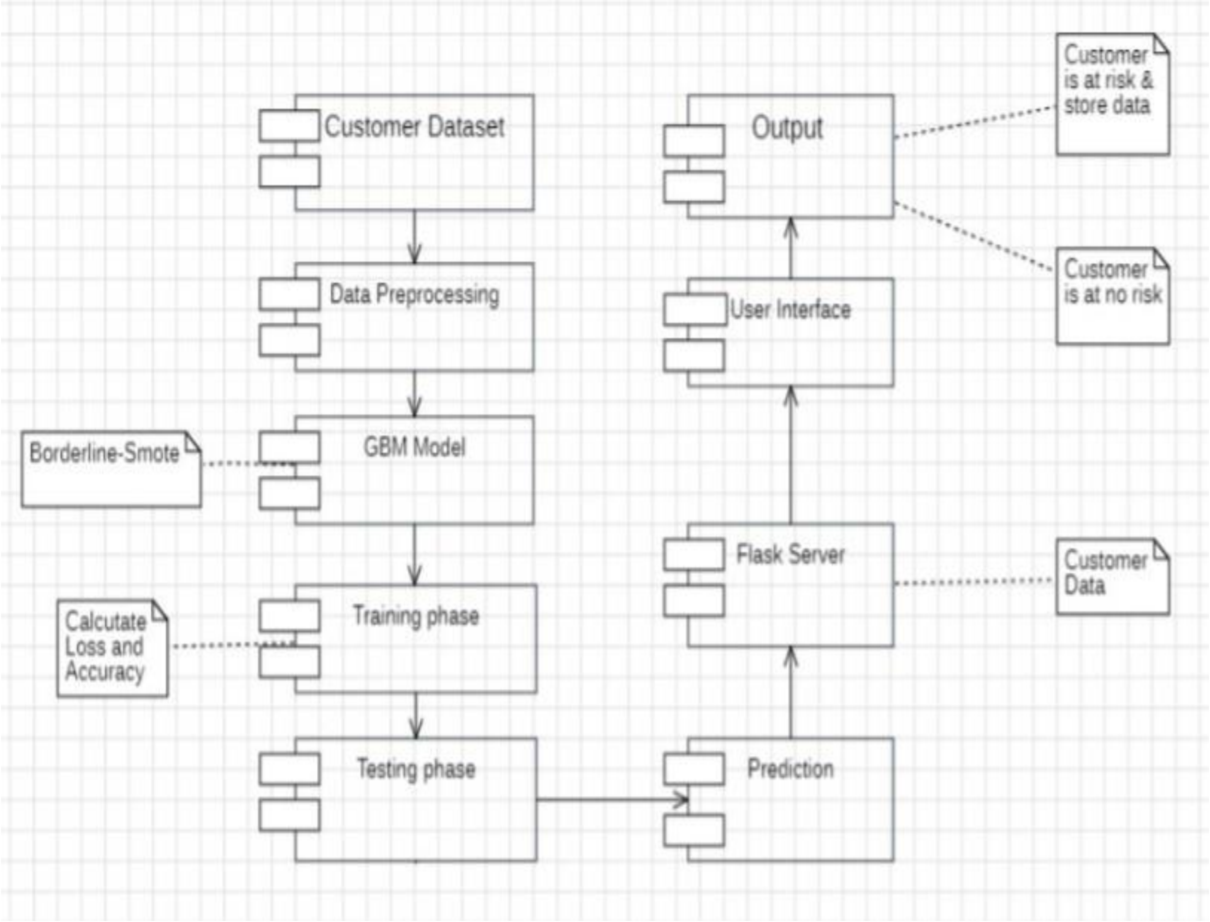


Figure:5.1 System Architecture

The system architecture for the Bank Customer Churn Prediction project defines the framework for how various components work together to deliver accurate churn predictions. It begins with the user interface (UI), where users input customer data through a web form. This data is sent to the web server (Flask Application), which processes the inputs and coordinates interactions between the user and the machine learning model. The data preprocessing module prepares the raw data by performing tasks like cleaning and normalization to ensure it is suitable for analysis. The machine learning model, specifically a Borderline-SMOTE Gradient Boosting Machine (GBM), then uses this prepared data to generate churn predictions. Finally, the data storage component manages the storage of prediction results and customer data, saving at-risk customer information into an Excel file for future reference and analysis. This architecture ensures a seamless flow of data and accurate predictions while maintaining effective data management and user interaction.

5.2 Dataflow Diagram

1. The DFD is also called a bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformationsthat are applied as data moves from input to output.
4. DFD is also known as a bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.
5. DFDs help identify inefficiencies in the system by visualizing how data moves and is processed, which can expose redundancies, bottlenecks, and other performance issues, ultimately providing a basis for targeted improvements to streamline operations and enhance overall system efficiency.
6. The data flow diagram facilitate communication among stakeholders by offering a clear, visual representation of system processes and data flows, thereby helping to ensure that all parties have a shared understanding of how the system functions and how information is exchanged.
7. DFDs are hierarchical in nature, allowing them to be broken down into progressively more detailed diagrams, which supports a top-down approach to understanding and managing system complexity by progressively revealing finer levels of detail.
8. DFDs are invaluable in both system design and analysis, as they provide a structured framework for examining how a system operates, how data flows between components, and how modifications might impact these flows, thus aiding in the development of efficient and effective system solutions.

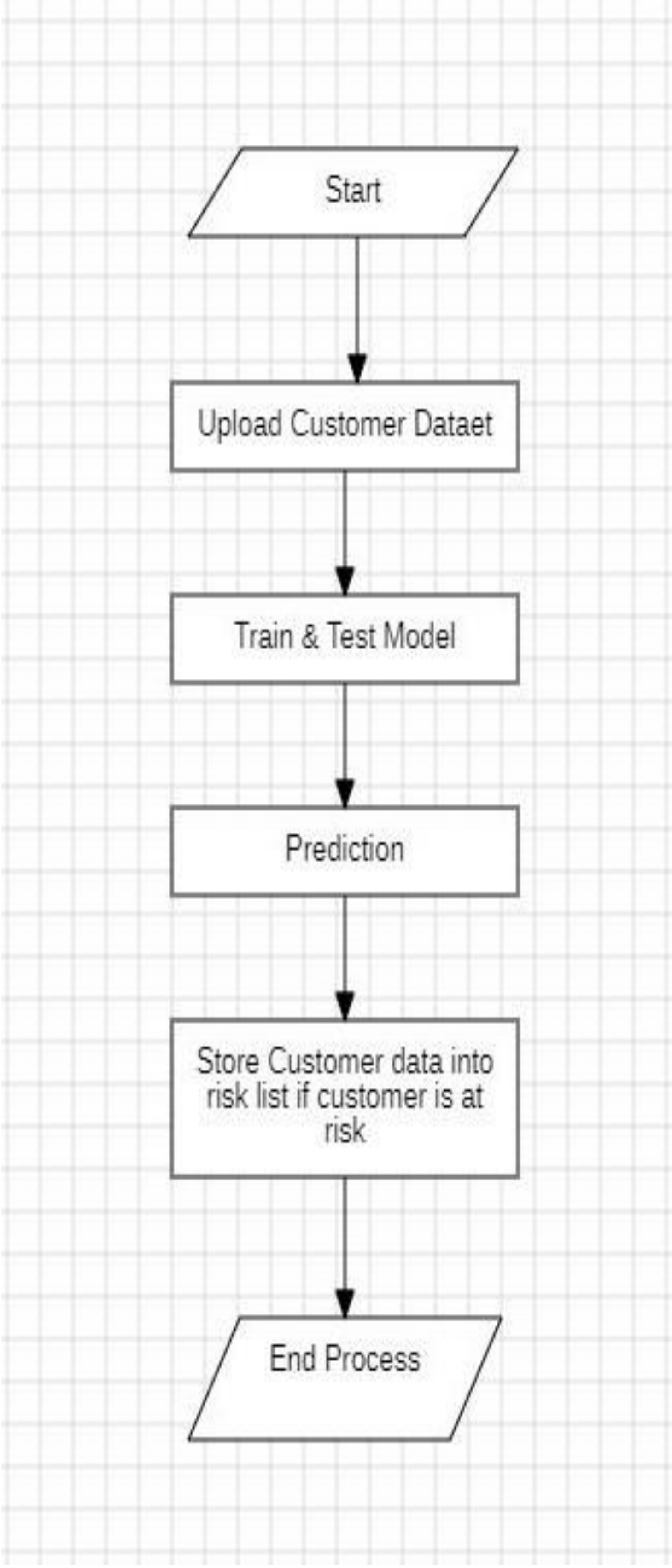


Figure:5.2 Dataflow Diagram

5.3 UML Diagrams

UML is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems. UML was created by the Object Management Group (OMG) and UML 1.0 specification draft was proposed to the OMG in January 1997. Behavioral UML diagrams and Structural UML diagrams. Behavioral diagrams focus on representing the dynamic aspects of a system, while Structural diagrams emphasize the static aspects, such as the architecture and relationships between components. Together, these categories provide a comprehensive view of both how the system operates and how it is constructed.

There are several types of UML diagrams and each one of them serves a different purpose regardless of whether it is being designed before the implementation or after (as part of documentation). UML has a direct relation with object-oriented analysis and design. After some standardization, UML has become an OMG standard. The two broadest categories that encompass all other types are:

- Behavioral UML diagram
- Structural UML diagram.

As the name suggests, some UML diagrams try to analyses and depict the structure of a system or process, whereas other describe the behavior of the system, its actors, and its building components.

Goals: The Primary goals in the design of the UML are as follows:

- Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
- Provide extendibility and specialization mechanisms to extend the core concepts.
- Be independent of particular programming languages and development process.
- Provide a formal basis for understanding the modeling language.
- Encourage the growth of tools market.
- Support higher level development concepts such as collaborations, frameworks, patterns and components.
- Integrate best practices.

The different types are as follows:

- Sequence diagram
- Use case Diagram
- Activity diagram
- Class diagram
- Collaboration diagram

Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e., the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

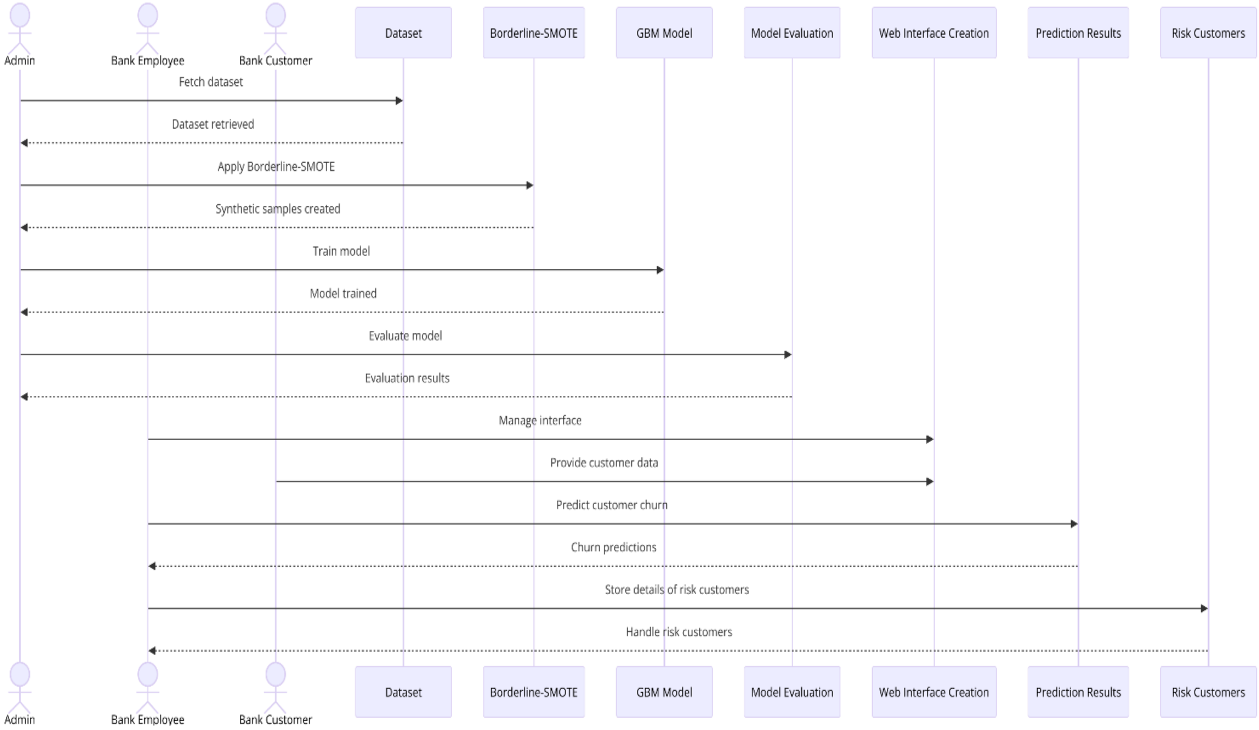


Figure 5.3.1 Sequence Diagram

List of actions

Admin:

- Fetches the dataset.
- Develops the prediction model.
- Creates web interface.

Bank Employee:

- Manages the web interface.
- Takes customer data.
- Handles risk customers details.

Bank Customer:

- Provides customer data.

System:
System will give the output as Bank employee enters details of customers according to the given data.

Result:
As per user enters the data it will give customer is at risk or not.

Use Case Diagram

A use case diagram at its simplest is the representation of a user's interaction with the system that shows the relationship between the user and the different use case in which the user is involved. A use case diagram is used to structure of the behavior thing in a model. The use cases are represented by either circles or ellipses.

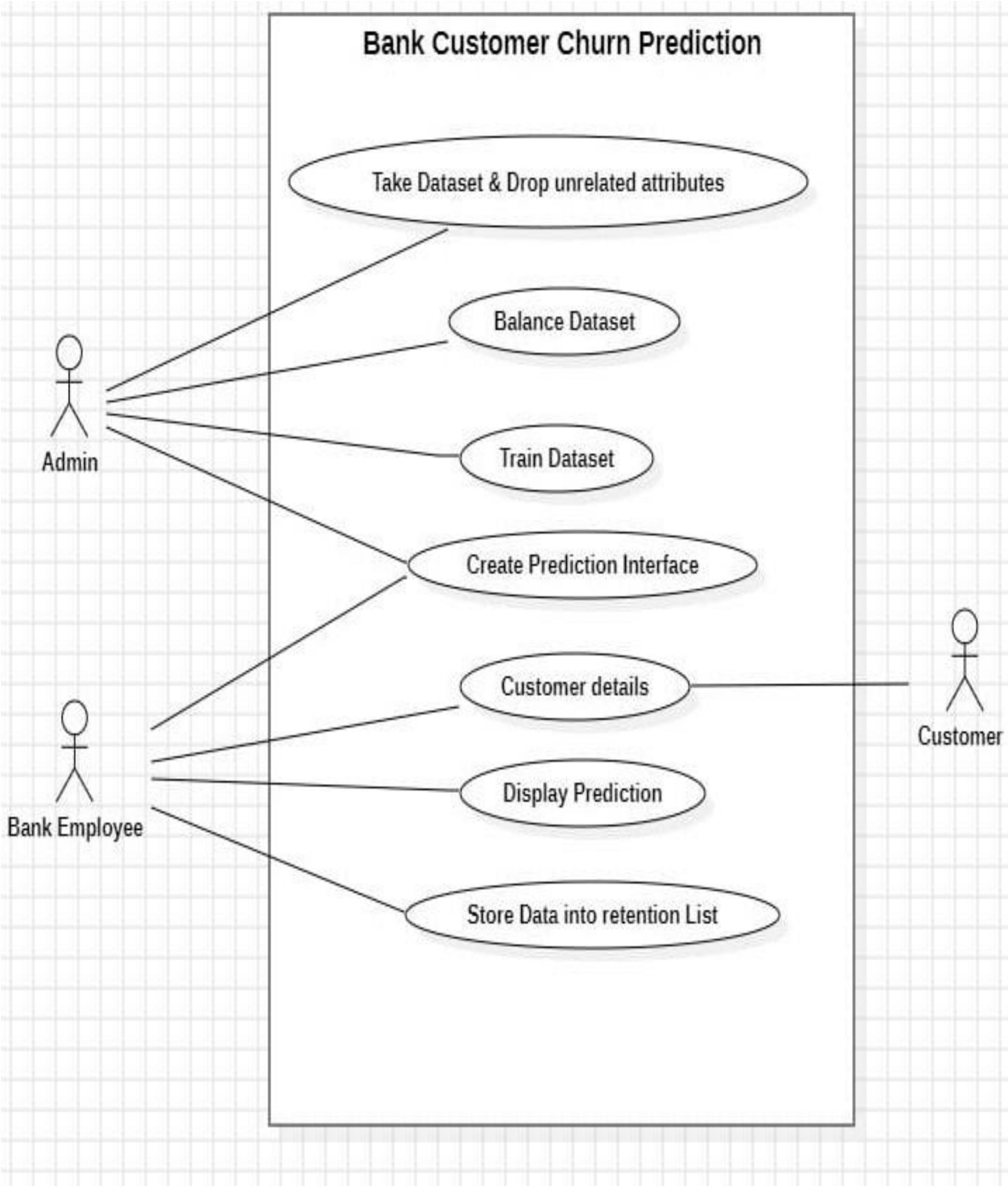

Bank Customer Churn Prediction
Take Dataset & Drop unrelated attributes
Balance Dataset
Admin
Train Dataset
Create Prediction Interface
Customer details
Customer
Bank Employee
Display Prediction
Store Data into retention List

Figure 5.3.2 Use Case Diagram

Activity Diagram

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system.Activity diagram is basically a flowchart to represent the flow from one activity to another activity.This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc.

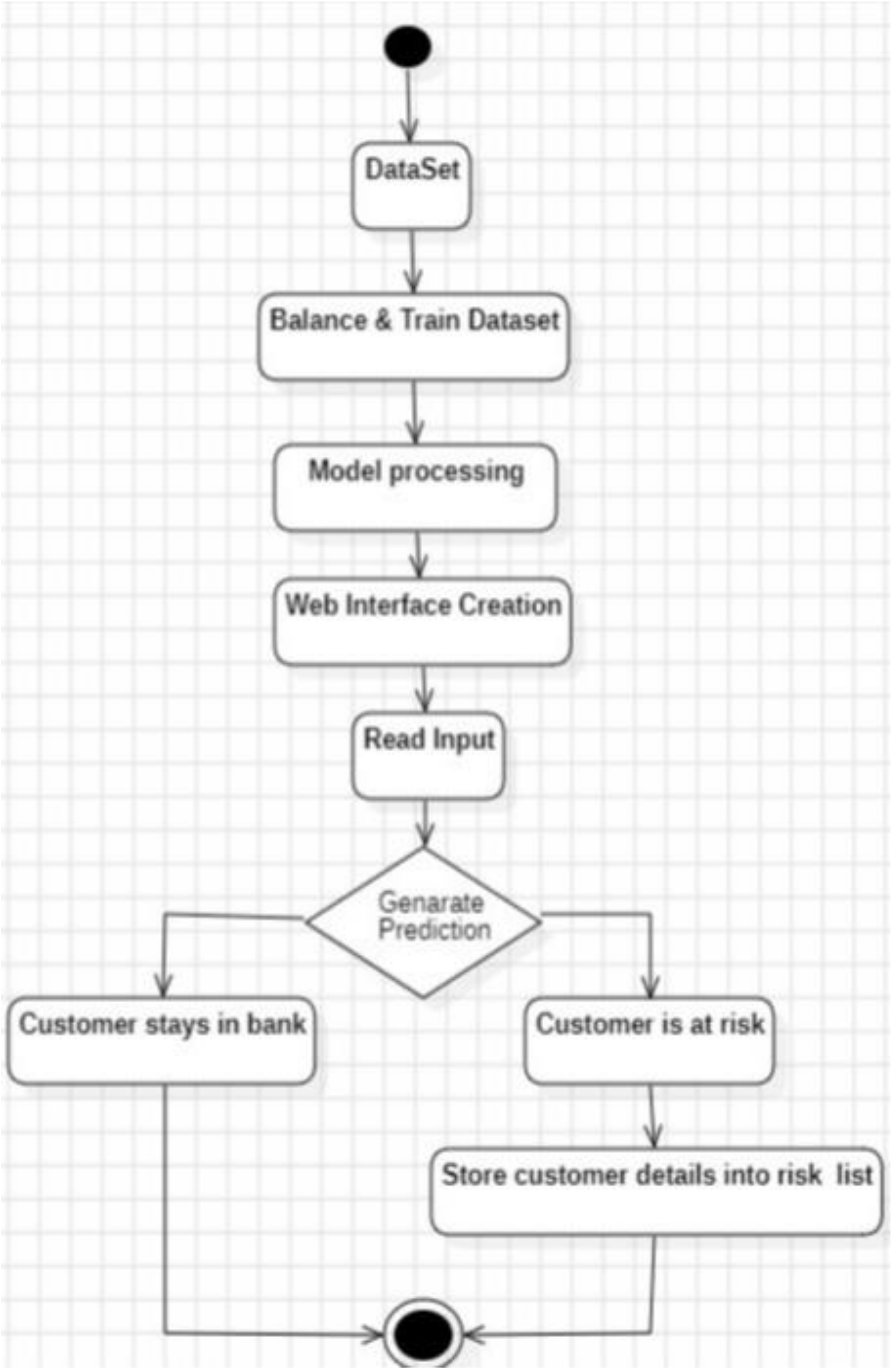


Figure 5.3.3 Activity Diagram

Class Diagram

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

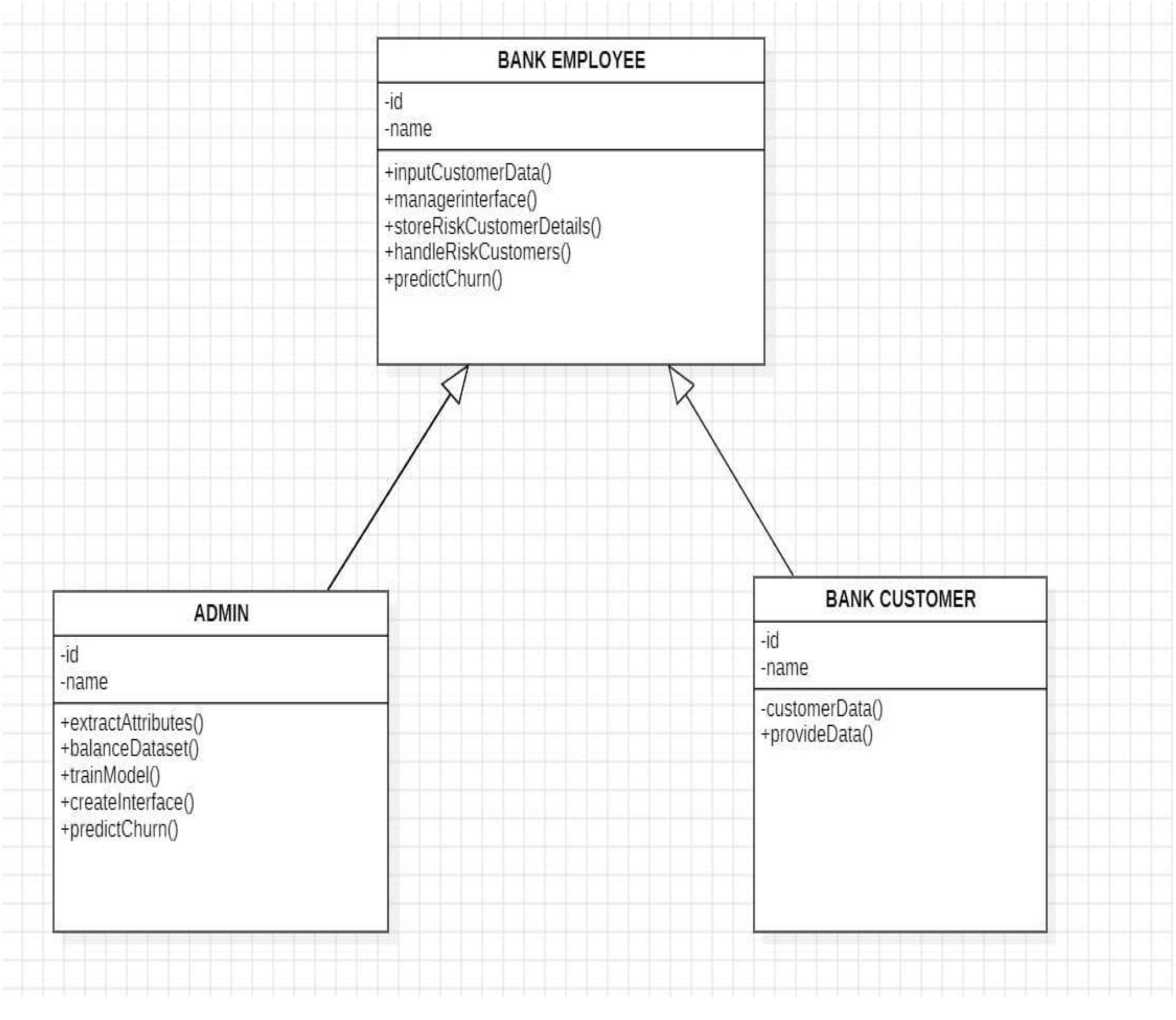


Figure 5.3.4 Class Diagram

6. CODING AND ITS IMPLEMENTATION

6.1 Source code

```
from flask import Flask, request, render_template
import joblib
import pandas as pd
from sklearn.preprocessing import StandardScaler
import os
app = Flask(__name__)
# Load the saved model and scaler
model = joblib.load('/mnt/data/borderline_smote_gbm_model.pkl')
# Load and preprocess the data to fit the scaler
data = pd.read_csv('/mnt/data/Churn_Modelling.csv')
data = data.drop(['RowNumber', 'CustomerId', 'Surname'], axis=1)
data = pd.get_dummies(data, drop_first=True)
X = data.drop('Exited', axis=1)
sc = StandardScaler()
sc.fit(X)
# Feature names for scaling
feature_names = X.columns
# Path for the Excel file
excel_file_path = 'risk_customers.xlsx'
def append_to_excel(df, file_path):
 # Append data to the Excel file, creating the file if it doesn't exist
    if os.path.exists(file_path):
      existing_df = pd.read_excel(file_path)
            new_df = pd.concat([existing_df, df], ignore_index=True)
           new_df.to_excel(file_path, index=False)
       else:
           df.to_excel(file_path, index=False)
```

```
@app.route('/', methods=['GET', 'POST'])
def home():
    prediction = ""
    if request.method == 'POST':
        # Get form data
        customer_id = int(request.form['CustomerID'])
        name = request.form['Name']
        p1 = int(request.form['CreditScore'])
        p2 = int(request.form['Age'])
        p3 = int(request.form['Tenure'])
        p4 = float(request.form['Balance'])
        p5 = int(request.form['NumOfProducts'])
        p6 = int(request.form['HasCrCard'])
        p7 = int(request.form['IsActiveMember'])
        p8 = float(request.form['EstimatedSalary'])
        p9 = int(request.form['Geography'])
        p10 = int(request.form['Gender'])
        # Determine geography values
        Geography_Germany = 1 if p9 == 1 else 0
        Geography_Spain = 1 if p9 == 2 else 0
        Geography_France = 1 if p9 == 3 else 0
        # Create a DataFrame with the correct feature names
         input_data = pd.DataFrame([[p1, p2, p3, p4, p5, p6, p7, p8,
         Geography_Germany, Geography_Spain, p10]],
        columns=feature_names)
        input_scaled = sc.transform(input_data)
        result = model.predict(input_scaled)
         prediction = "The Customer is at Risk" if result == 1 else "The Customer is at No Risk"
        # If customer is at risk (result == 1), store in Excel
        if result == 1:
            risk_data = pd.DataFrame([{
                'Customer ID': customer_id,
                'Name': name,
                'Credit Score': p1,
                'Age': p2,
                'Tenure': p3,
                'Balance': p4,
                'Number of Products': p5,
                'Has Credit Card': p6,
                'Is Active Member': p7,
                'Estimated Salary': p8,
                'Geography': p9,
                'Gender': p10
            }])
            append_to_excel(risk_data, excel_file_path)
    return render_template('index.html', prediction=prediction)
if __name__ == '__main__':
    app.run(debug=True)
```

6.2 Implementation

6.2.1 Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy thatemphasizes code readability, notably using significant whitespace.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this; while it may seem like an all but useless metric, it reflects the amount of code one has to scan, read, and understand to troubleshoot problems or tweak behaviors. The rapid development capabilities, ease of learning for programmers from other languages, and extensive standard library are key factors where Python excels. Additionally, Python's rich ecosystem of third-party libraries and frameworks further accelerates development and simplifies complex tasks, making it a preferred choice for many developers.

All its tools have been quick to implement, saving a lot of time, and several of them have later been patched and updated by the community. This collaborative nature, coupled with Python's design philosophy, contributes to its robustness and versatility.

6.2.2 Libraries & Modules Used in Project

Scikit-learn

Scikit-learn is a comprehensive machine learning library in Python that provides a wide array of tools for data mining, analysis, and model building. It is designed for simplicity and efficiency, making it ideal for both beginners and experienced developers. Built on top of essential libraries like NumPy, SciPy, and Matplotlib, Scikit-learn supports various machine learning algorithms such as classification, regression, clustering, and dimensionality reduction. In your project, Scikit-learn is essential for tasks like pre-processing the customer data, training machine learning classifiers (such as Gradient Boosting), and validating model performance. Its robust functions for splitting data, cross-validation, and model evaluation ensure that the predictive model generalizes well to unseen data, thereby accurately predicting customer churn.

NumPy

NumPy is the core library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a wide array of mathematical functions to operate on these arrays. In your project, NumPy is essential for performing efficient numerical operations, such as handling matrix transformations and performing mathematical calculations needed during the data pre-processing phase. NumPy arrays form the backbone of data that is used in machine learning models, and they integrate seamlessly with other libraries like Scikit-learn and Pandas. Whether calculating correlations, standardizing data, or performing statistical operations, NumPy ensures that the underlying computations are efficient.

Pandas

Pandas is an open-source data manipulation and analysis library that offers powerful data structures like DataFrame, which is perfect for handling large, structured datasets. In your project, Pandas is invaluable for loading, cleaning, and transforming customer data, such as handling missing values, filtering records, and creating new features. It also allows you to easily merge, reshape, and group your data, simplifying the process of analyzing customer demographics and behaviors that contribute to churn. Additionally, Pandas is useful for exploring the dataset, running statistical summaries, and preparing the data before it is fed into machine learning models for prediction. Its intuitive syntax and extensive functionality make it a crucial tool for handling complex data operations in large-scale machine learning projects like customer churn prediction.

Matplotlib

Matplotlib is a versatile 2D plotting library that enables the creation of high-quality visualizations, which are critical for understanding data trends and communicating findings. In your project, Matplotlib is used to visualize customer demographics, churn rates, and model performance metrics. You can generate a variety of charts, including line plots, bar graphs, scatter plots, and histograms, to gain insights into customer behavior and the effectiveness of your churn prediction model. Visualization is key during exploratory data analysis, where understanding patterns in customer attributes like age, balance, and product usage helps in crafting a robust prediction model. Matplotlib's ability to create detailed, customizable visualizations helps in effectively communicating insights and refining your churn prediction model.

XGBoost

XGBoost is a highly efficient and flexible machine learning library that implements the gradient boosting algorithm. Known for its speed and accuracy, XGBoost is widely used in classification tasks, especially in competitions and large-scale industrial applications. In your project, XGBoost is a core component for training the churn prediction model, leveraging its ability to handle large datasets and complex interactions between variables. XGBoost's regularization capabilities make it less prone to overfitting, ensuring that the model performs well on unseen data. Its gradient boosting approach optimizes the predictive accuracy of the model, allowing you to make highly reliable predictions about customer churn.

Imbalanced-learn

Imbalanced-learn is a specialized Python library designed to address the challenges of working with imbalanced datasets. In customer churn prediction, the dataset is often skewed, with many more customers who stay than those who churn. Imbalanced-learn provides various techniques, including Borderline-SMOTE, to rebalance the dataset by oversampling the minority class (churned customers). This ensures that the predictive model does not become biased toward the majority class and can accurately identify customers at risk of leaving. By using Imbalanced-learn, your project improves the reliability of churn predictions, making it possible to take appropriate actions for at-risk customers.

Bank Employee Module

The Bank Employee module is designed to streamline interactions between bank employees and the data systems that predict customer churn. Employees use this module to input essential customer information such as personal details, financial data, and interaction history into the system. The module also provides an interface for managing customer interactions, which includes tracking and updating customer information over time. An important feature of the Bank Employee module is its ability to flag customers who are at risk of churning, storing their details securely in a dedicated list for future action.

This helps the employee focus on customers who are more likely to leave the bank, allowing them to intervene and offer retention strategies. Additionally, the module integrates with a machine learning model that processes the customer data and predicts churn. Based on this prediction, the bank employees can take appropriate measures to retain customers. The module's user-friendly interface simplifies data management, ensures efficient handling of customer churn risks, and helps optimize the overall customer retention process.

Admin Module

The admin module provides essential tools for the backend management of the churn prediction system. Admins have the ability to extract relevant attributes from large customer datasets, ensuring that the machine learning model is trained on the most relevant data points. The module also allows administrators to balance the dataset, which is crucial for ensuring accurate predictions. If the dataset is unbalanced, it could lead to biased predictions, where the model might incorrectly predict churn for certain groups of customers. The module includes features for monitoring and evaluating model performance to enable continuous improvement based on real-world data and feedback.

The admin module also handles the training of the machine learning model by applying data preprocessing techniques and fine-tuning model parameters to achieve the best possible prediction accuracy. Once the model is trained, it interfaces with the employee module, allowing for real-time churn predictions to be made based on incoming customer data. Additionally, admins can create and manage the user interface that employees will interact with, ensuring that the system is intuitive and efficient for daily operations. The admin module plays a pivotal role in maintaining the integrity and accuracy of the churn prediction system.

Bank Customer Module

The Bank Customer module is the foundation of customer data input for the entire system. It allows for the collection of important customer details such as demographic information, transaction history, account details, and behavior patterns. This data is critical for the churn prediction model, as it provides the information required to assess whether a customer is likely to leave the bank. Customers can provide data either through direct interactions with bank employees or via automated systems. The Bank Customer module works in tandem with the Bank Employee module, where employees input customer data and update it as necessary. The system ensures that the data collected is accurate, up-to-date, and comprehensive, covering all aspects of customer behavior. Furthermore, the Bank Customer module is designed to securely store sensitive customer information, ensuring compliance with data privacy regulations. This data serves as the backbone of the churn prediction model, which processes it to identify patterns that could indicate a risk of churn. The ability of this module to provide accurate and detailed customer data is key to the success of the prediction model and the bank’s retention strategies.

7. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each testtype addresses a specific testing requirement.

7.1. Types Of Tests

Unit testing

Unit testing involves validating individual components of the project such as data preprocessing, feature engineering, model training, and prediction. Each module, including the functions for balancing data using Borderline-SMOTE and training the Gradient Boosting Machine (GBM), should be tested independently to ensure they behave correctly. This test ensures that each small unit of code works as intended and produces the expected output when given valid inputs.

Integration testing

Integration testing ensures that individual modules work together seamlessly. In the context of this project, it’s essential to confirm that data preprocessing integrates with the model training and prediction pipeline. The workflow from data input, processing, prediction, and saving at-risk customers should be tested as an integrated process. This phase ensures that the machine learning model correctly interacts with other components, such as user interfaces and data storage

Functional test

Functional testing verifies that the system behaves according to the defined requirements. Key functionalities to be tested include the system's ability to handle both valid and invalid inputs, such as credit scores and balances, and ensure that valid predictions for customer churn are generated. Furthermore, functional testing will ensure that the list of at-risk customers is correctly generated and stored, while no action is taken for customers predicted to retain.

Functional testing is centered on the following items:

Valid Input: Ensuring valid customer data is accepted and processed correctly.

Invalid Input: Ensuring invalid inputs, such as missing or incorrect data, are rejected to prevent errors and maintain the accuracy and reliability of the system's predictions and processes.

Functions: Verifying that all functions, including data processing and prediction.

Output: Ensuring churn predictions are accurate and risk customer data is properly handled.

Systems/Procedures: Checking that system interfaces and procedures function seamlessly.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows data fields,predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Testing

System testing validates that the entire churn prediction system, including the frontend interface, model, and backend data storage, works cohesively. This test ensures that the whole system meets user expectations, accurately predicts churn, and handles real-world inputs without failure. This involves end-to-end testing of the entire workflow, from data input to churn prediction and risk customer storage. System testing also verifies the system's performance under various conditions, ensuring it can handle high volumes of data and user interactions while maintaining accuracy and efficiency in predictions.

White Box Testing

This involves testing the internal workings of the model, ensuring that each step (data preprocessing, feature scaling, model training, etc.) functions according to the intended logic. It helps uncover any internal issues within the code or model workflow. Additionally, White Box Testing verifies the correctness of algorithms used for feature selection and the implementation of various machine learning techniques. By examining the internal code structure, this testing method ensures that all components interact as expected, and any logic errors or inefficiencies can be identified and corrected.

Black Box Testing

Black box testing examines the functionality of the system without considering its internal logic. The focus is on the input and output—verifying that when the user inputs customer data, the system accurately predicts churn and stores high-risk customers. The tester is unaware of the internal processes like data transformations or model training. This type of testing ensures that the system meets the requirements from an end-user perspective, including how well it handles different types of input data and how effectively it manages and reports the predicted outcomes. It also evaluates the system's overall usability and robustness under various scenarios.

Test strategy and approach:

Field testing will be performed manually and functional tests will be written in detail.

Test objectives:

- All field entries must work properly.
- Pages must be activated from the identified link.

- The entry screen, messages and responses must not be delayed.

Integration Testing

Integration testing ensures that individual modules work together seamlessly. In the context of this project, it’s essential to confirm that data preprocessing integrates with the model training and prediction pipeline. The workflow from data input, processing, prediction, and saving at-risk customers should be tested as an integrated process. This phase ensures that the machine learning model correctly interacts with other components, such as user interfaces and data storage mechanisms.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

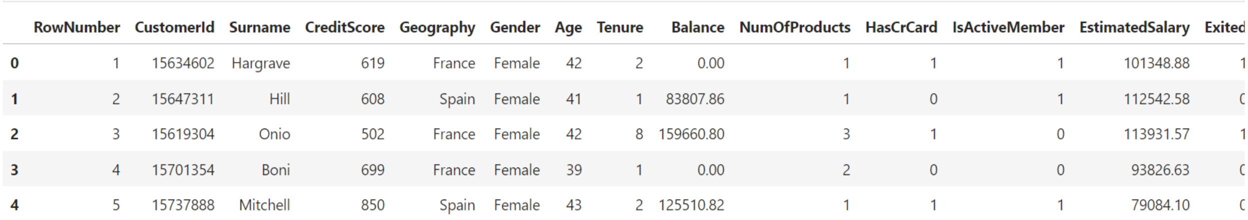
Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2 Test Cases:

S.no	Test Case	Excepted Result	Result	Remarks(IF Fails)
1.	Data uploading	Initially data need to be uploaded.	Pass	Executed successfully
2.	Data Splitting and balancing	After Uploading the Data, it should divideinto train and test data. Data should balanced.	Pass	Dataset balanced Successfully
3.	Applying Model	Model should generate a .pkl file	Pass	Algorithm implemented
4.	Giving Input	Input should be Uploaded.	Pass	Input uploaded successfully
5.	Prediction	After submit of customer data, it should generate the prediction	Pass	Successfully done the prediction .
6.	If customer is at risk store data into a list	Customer data should be stored in list	Pass	Stored successfully
7.	Prediction	Taking customer data from trained dataset and expecting correct output	Fail	Prediction is not correct.

Table no 7.2 Test Cases

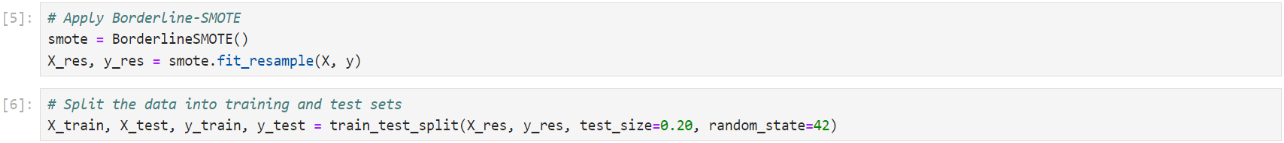
Test Case 1:



	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Figure 7.2.1: Test Case 1

Test Case 2:

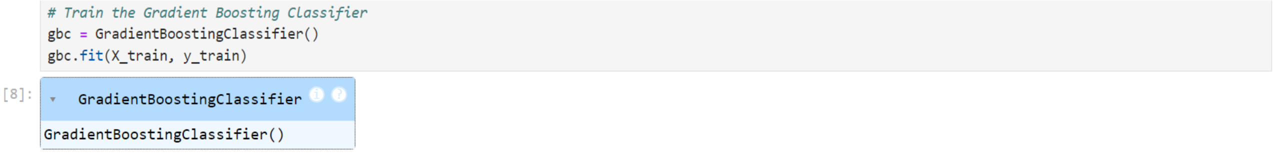


```
[5]: # Apply Borderline-SMOTE
     smote = BorderlineSMOTE()
     X_res, y_res = smote.fit_resample(X, y)

[6]: # Split the data into training and test sets
     X_train, X_test, y_train, y_test = train_test_split(X_res, y_res, test_size=0.20, random_state=42)
```

Figure 7.2.2: Test Case 2

Test Case 3:

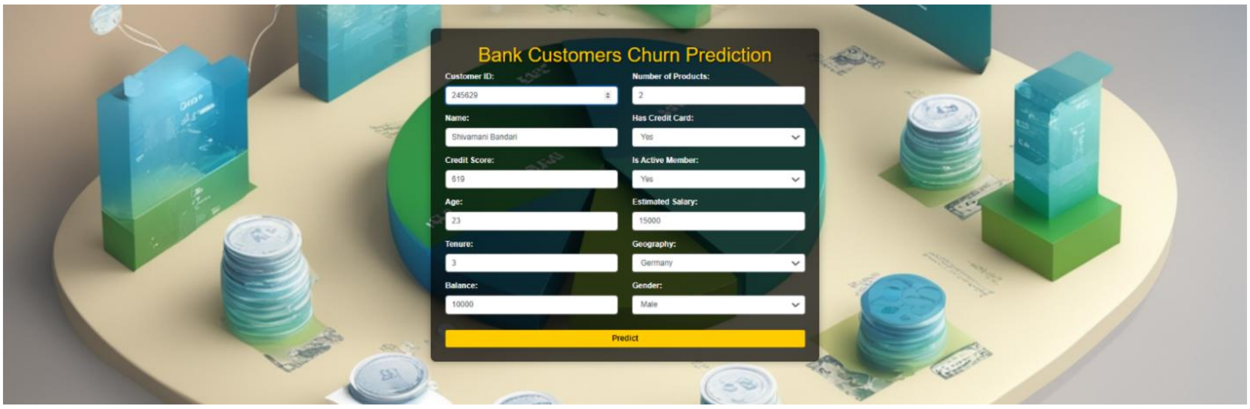


```
     # Train the Gradient Boosting Classifier
     gbc = GradientBoostingClassifier()
     gbc.fit(X_train, y_train)

[8]: ▾ GradientBoostingClassifier
     GradientBoostingClassifier()
```

Figure 7.2.3: Test Case 3

Test Case 4:



Bank Customers Churn Prediction

Customer ID: 245629

Number of Products: 2

Name: Shivamani Bandari

Has Credit Card: Yes

Credit Score: 619

Is Active Member: Yes

Age: 23

Estimated Salary: 15000

Tenure: 3

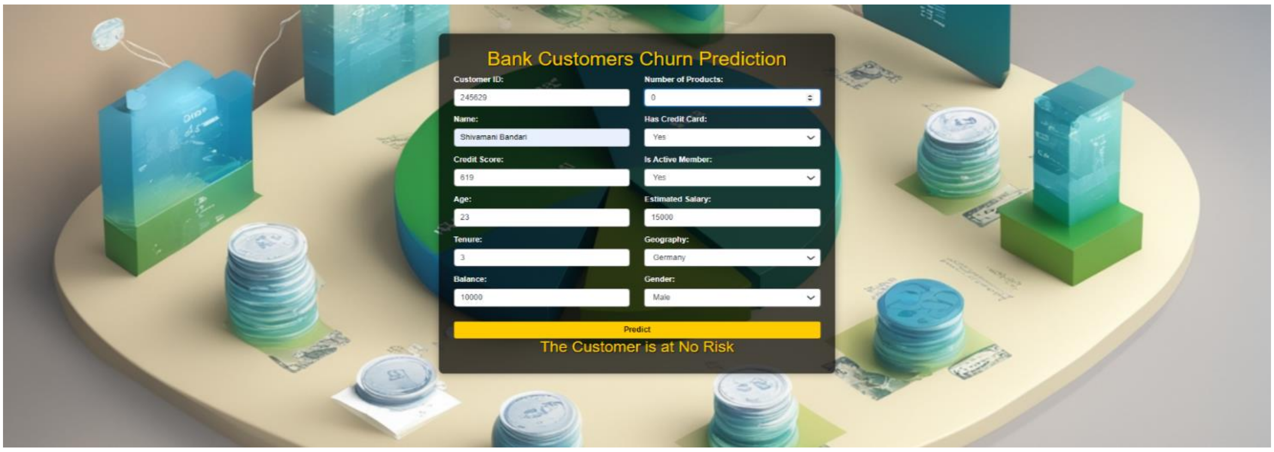
Geography: Germany

Balance: 10000

Gender: Male

Figure 7.2.4: Test Case 4

Test Case 5:



Bank Customers Churn Prediction

Customer ID: 245629

Name: Shivamani Bandari

Credit Score: 619

Age: 23

Tenure: 3

Balance: 10000

Number of Products: 0

Has Credit Card: Yes

Is Active Member: Yes

Estimated Salary: 15000

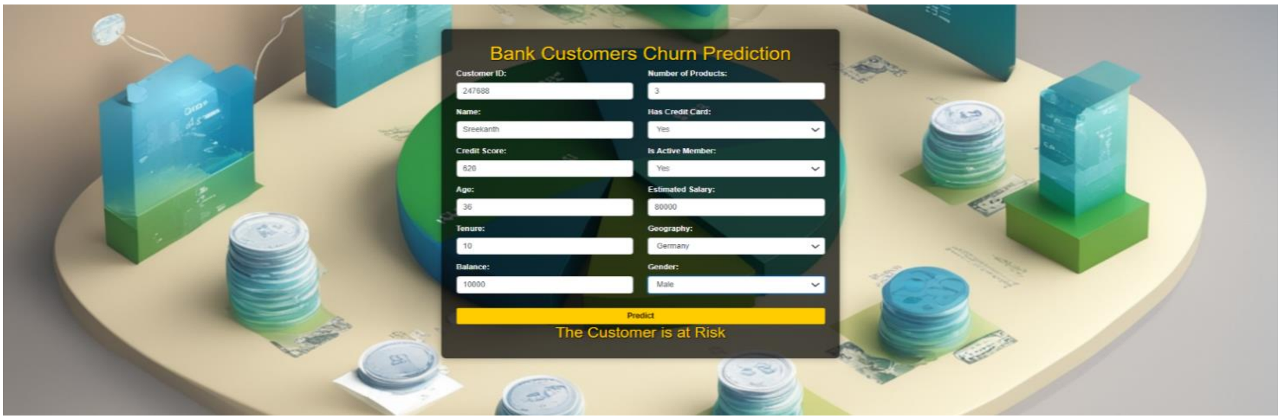
Geography: Germany

Gender: Male

The Customer is at No Risk

Figure 7.2.5: Test Case 5

Test Case 6:



Bank Customers Churn Prediction

Customer ID: 247688

Name: Sreekanth

Credit Score: 620

Age: 36

Tenure: 10

Balance: 10000

Number of Products: 3

Has Credit Card: Yes

Is Active Member: Yes

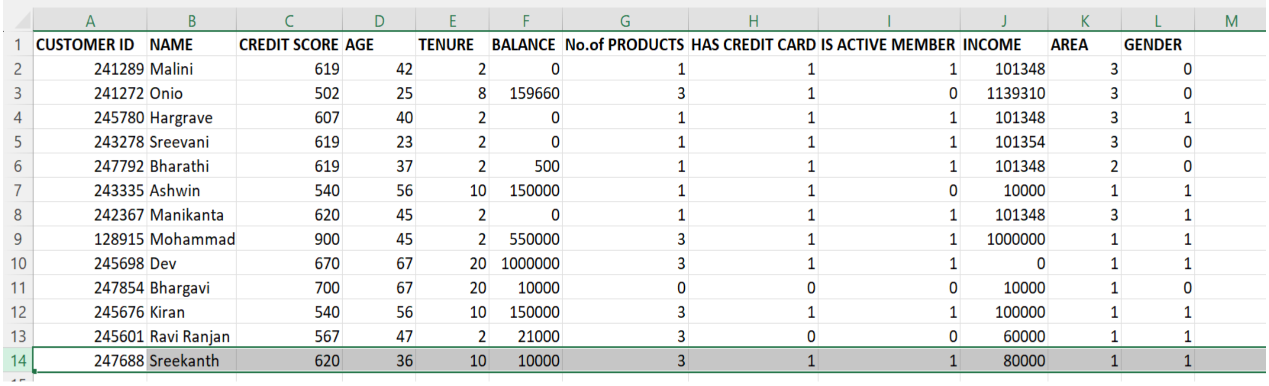
Estimated Salary: 80000

Geography: Germany

Gender: Male

The Customer is at Risk

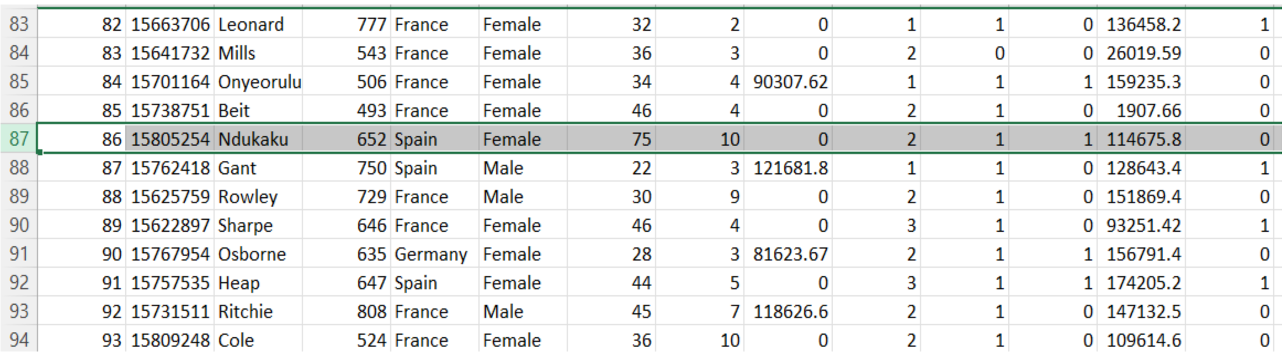
Figure 7.2.6: Test Case 6



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CUSTOMER ID	NAME	CREDIT SCORE	AGE	TENURE	BALANCE	No.of PRODUCTS	HAS CREDIT CARD	IS ACTIVE MEMBER	INCOME	AREA	GENDER	
2	241289	Malini	619	42	2	0	1	1	1	101348	3	0	
3	241272	Onio	502	25	8	159660	3	1	0	1139310	3	0	
4	245780	Hargrave	607	40	2	0	1	1	1	101348	3	1	
5	243278	Sreevani	619	23	2	0	1	1	1	101354	3	0	
6	247792	Bharathi	619	37	2	500	1	1	1	101348	2	0	
7	243335	Ashwin	540	56	10	150000	1	1	0	10000	1	1	
8	242367	Manikanta	620	45	2	0	1	1	1	101348	3	1	
9	128915	Mohammad	900	45	2	550000	3	1	1	1000000	1	1	
10	245698	Dev	670	67	20	1000000	3	1	1	0	1	1	
11	247854	Bhargavi	700	67	20	10000	0	0	0	10000	1	0	
12	245676	Kiran	540	56	10	150000	3	1	1	100000	1	1	
13	245601	Ravi Ranjan	567	47	2	21000	3	0	0	60000	1	1	
14	247688	Sreekanth	620	36	10	10000	3	1	1	80000	1	1	

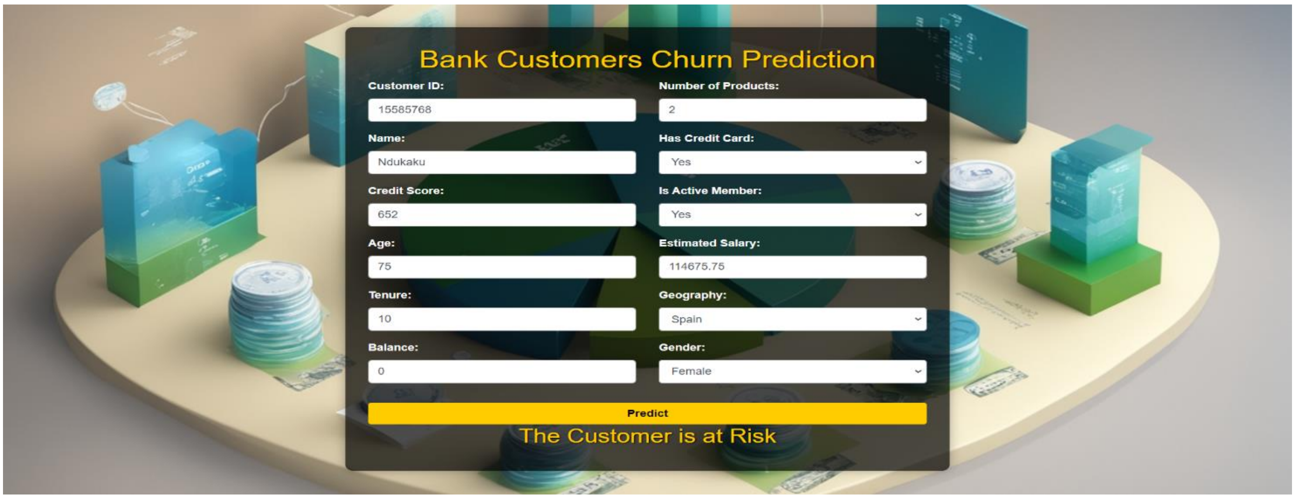
Figure 7.2.7: Excel sheet status after successfull prediction

Test Case 7:



83	82	15663706	Leonard	777	France	Female	32	2	0	1	1	0	136458.2	1
84	83	15641732	Mills	543	France	Female	36	3	0	2	0	0	26019.59	0
85	84	15701164	Onyeorulu	506	France	Female	34	4	90307.62	1	1	1	159235.3	0
86	85	15738751	Beit	493	France	Female	46	4	0	2	1	0	1907.66	0
87	86	15805254	Ndukaku	652	Spain	Female	75	10	0	2	1	1	114675.8	0
88	87	15762418	Gant	750	Spain	Male	22	3	121681.8	1	1	0	128643.4	1
89	88	15625759	Rowley	729	France	Male	30	9	0	2	1	0	151869.4	0
90	89	15622897	Sharpe	646	France	Female	46	4	0	3	1	0	93251.42	1
91	90	15767954	Osborne	635	Germany	Female	28	3	81623.67	2	1	1	156791.4	0
92	91	15757535	Heap	647	Spain	Female	44	5	0	3	1	1	174205.2	1
93	92	15731511	Ritchie	808	France	Male	45	7	118626.6	2	1	0	147132.5	0
94	93	15809248	Cole	524	France	Female	36	10	0	2	1	0	109614.6	0

Figure 7.2.8: Excel sheet status after unsuccessfull prediction



Bank Customers Churn Prediction

Customer ID: 15585768

Number of Products: 2

Name: Ndukaku

Has Credit Card: Yes

Credit Score: 652

Is Active Member: Yes

Age: 75

Estimated Salary: 114675.75

Tenure: 10

Geography: Spain

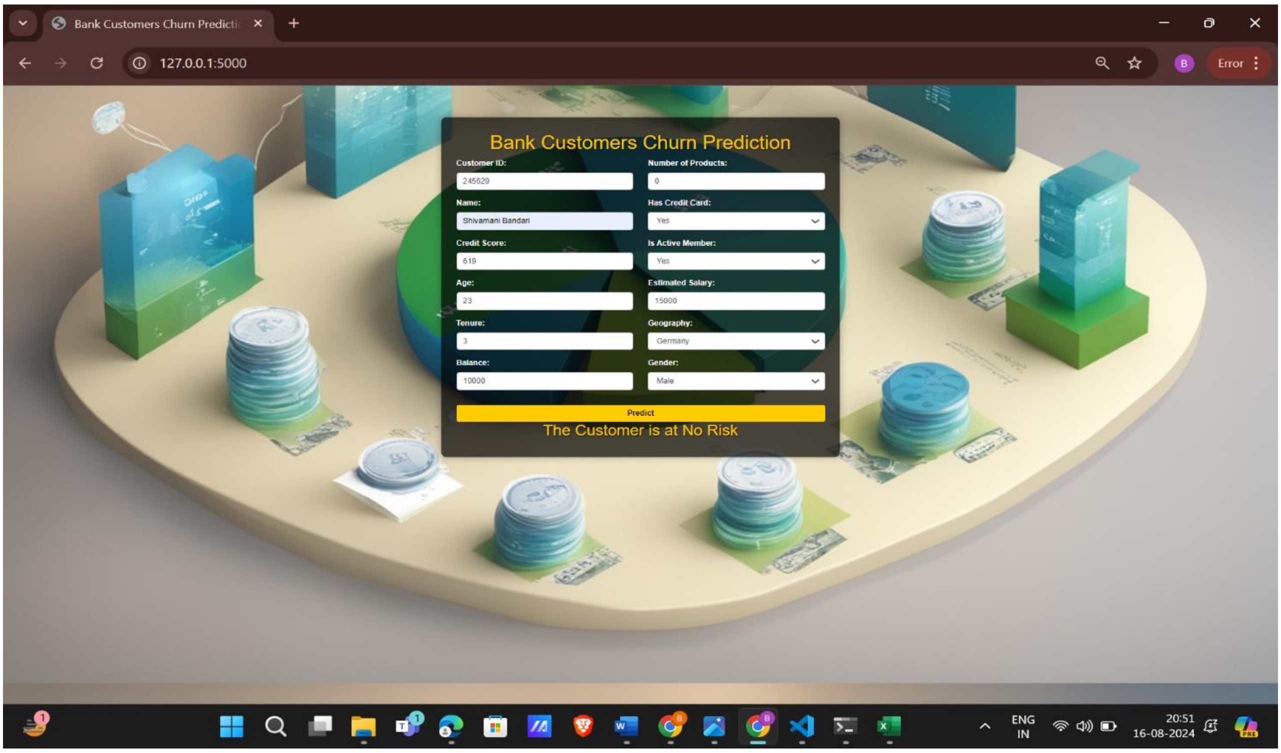
Balance: 0

Gender: Female

The Customer is at Risk

Figure 7.2.9: Test Case 7

8. OUTPUT SCREENS



Bank Customers Churn Prediction

Customer ID: 245629

Number of Products: 0

Name: Shivamani Bandari

Has Credit Card: Yes

Credit Score: 619

Is Active Member: Yes

Age: 23

Estimated Salary: 15000

Tenure: 3

Geography: Germany

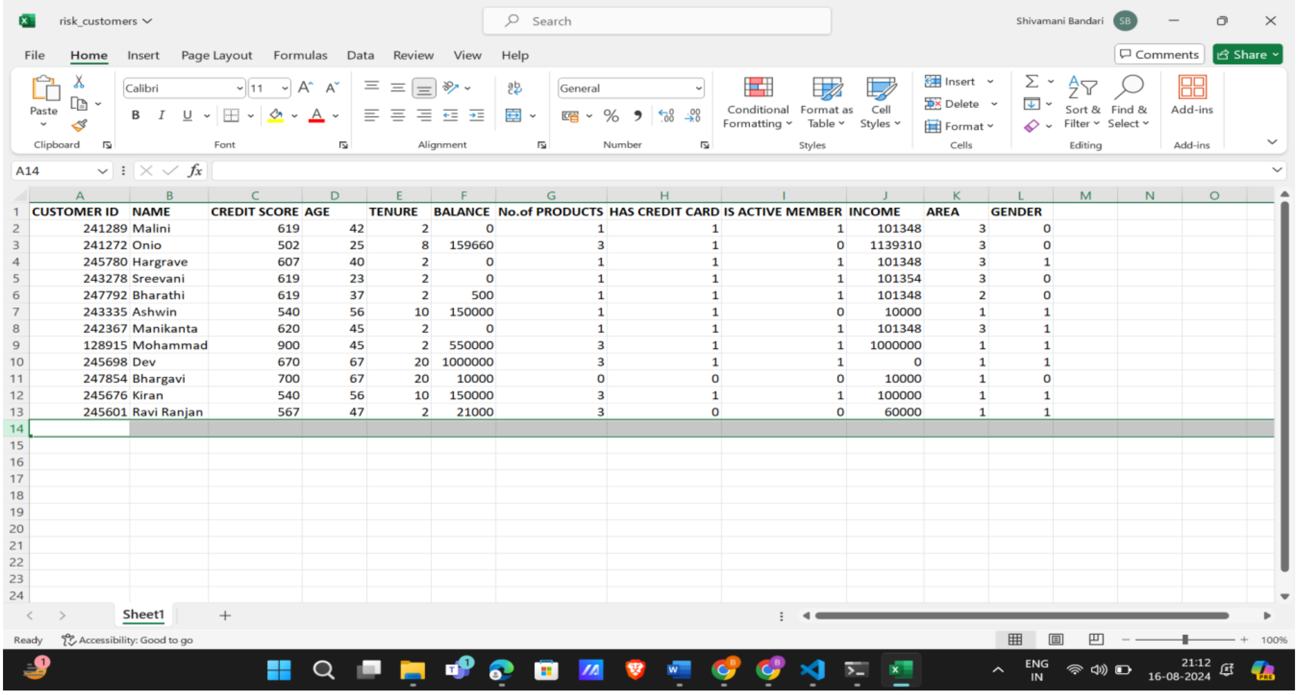
Balance: 10000

Gender: Male

Predict

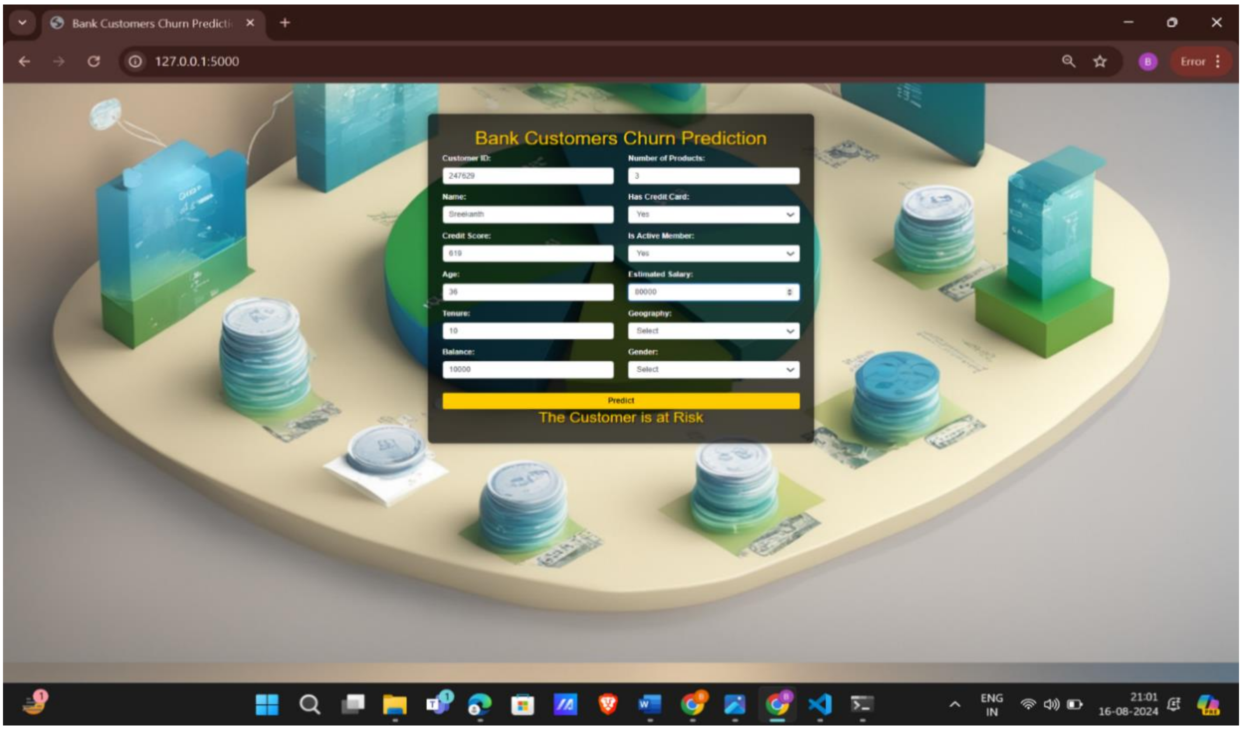
The Customer is at No Risk

Figure 8.1: Prediction of Non-Risk customer



CUSTOMER ID	NAME	CREDIT SCORE	AGE	TENURE	BALANCE	No.of PRODUCTS	HAS CREDIT CARD	IS ACTIVE MEMBER	INCOME	AREA	GENDER
241289	Malini	619	42	2	0	1	1	1	101348	3	0
241272	Onio	502	25	8	159660	3	1	0	1139310	3	0
245780	Hargrave	607	40	2	0	1	1	1	101348	3	1
243278	Sreevani	619	23	2	0	1	1	1	101354	3	0
247792	Bharathi	619	37	2	500	1	1	1	101348	2	0
243335	Ashwin	540	56	10	150000	1	1	0	10000	1	1
242367	Manikanta	620	45	2	0	1	1	1	101348	3	1
128915	Mohammad	900	45	2	550000	3	1	1	1000000	1	1
245698	Dev	670	67	20	1000000	3	1	1	0	1	1
247854	Bhargavi	700	67	20	10000	0	0	0	10000	1	0
245676	Kiran	540	56	10	150000	3	1	1	100000	1	1
245601	Ravi Ranjan	567	47	2	21000	3	0	0	60000	1	1

Figure 8.2: No action will be taken for Non-Risk customer



Bank Customers Churn Prediction

Customer ID: 247629

Number of Products: 3

Name: Sreekanth

Has Credit Card: Yes

Credit Score: 619

Is Active Member: Yes

Age: 36

Estimated Salary: 80000

Tenure: 10

Geography: Select

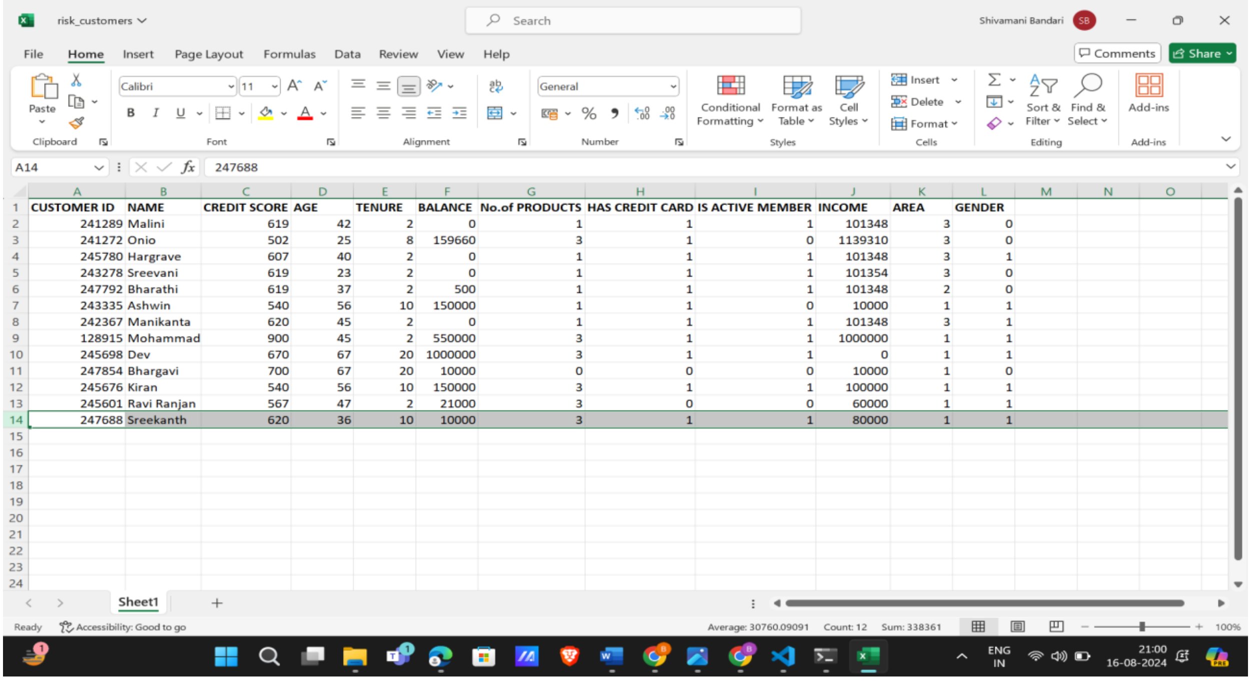
Balance: 10000

Gender: Select

Predict

The Customer is at Risk

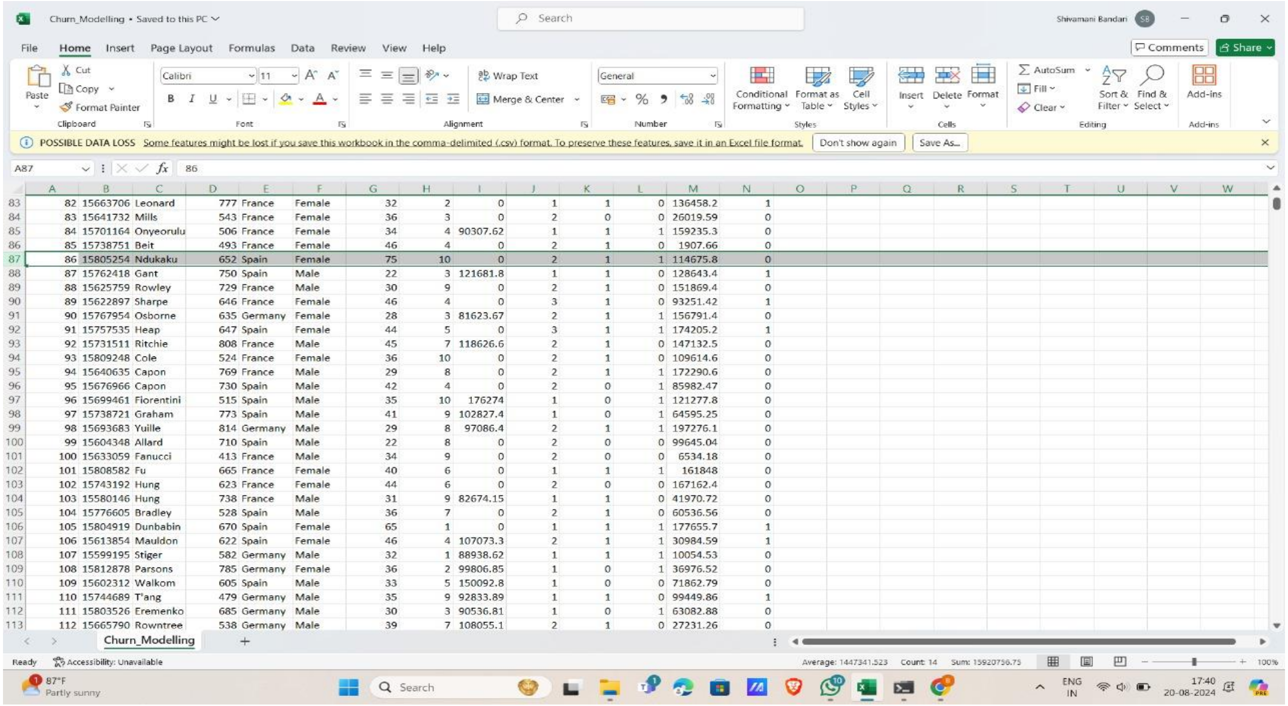
Figure 8.3: Prediction of Risk customer



A14 247688

CUSTOMER ID	NAME	CREDIT SCORE	AGE	TENURE	BALANCE	No.of PRODUCTS	HAS CREDIT CARD	IS ACTIVE MEMBER	INCOME	AREA	GENDER
241289	Malini	619	42	2	0	1	1	1	101348	3	0
241272	Onio	502	25	8	159660	3	1	0	1139310	3	0
245780	Hargrave	607	40	2	0	1	1	1	101348	3	1
243278	Sreevani	619	23	2	0	1	1	1	101354	3	0
247792	Bharathi	619	37	2	500	1	1	1	101348	2	0
243335	Ashwin	540	56	10	150000	1	1	0	10000	1	1
242367	Manikanta	620	45	2	0	1	1	1	101348	3	1
128915	Mohammad	900	45	2	550000	3	1	1	1000000	1	1
245698	Dev	670	67	20	1000000	3	1	1	0	1	1
247854	Bhargavi	700	67	20	10000	0	0	0	10000	1	0
245676	Kiran	540	56	10	150000	3	1	1	100000	1	1
245601	Ravi Ranjan	567	47	2	21000	3	0	0	60000	1	1
247688	Sreekanth	620	36	10	10000	3	1	1	80000	1	1

Figure 8.4: Action will be taken for Risk customer



Churn_Modelling • Saved to this PC

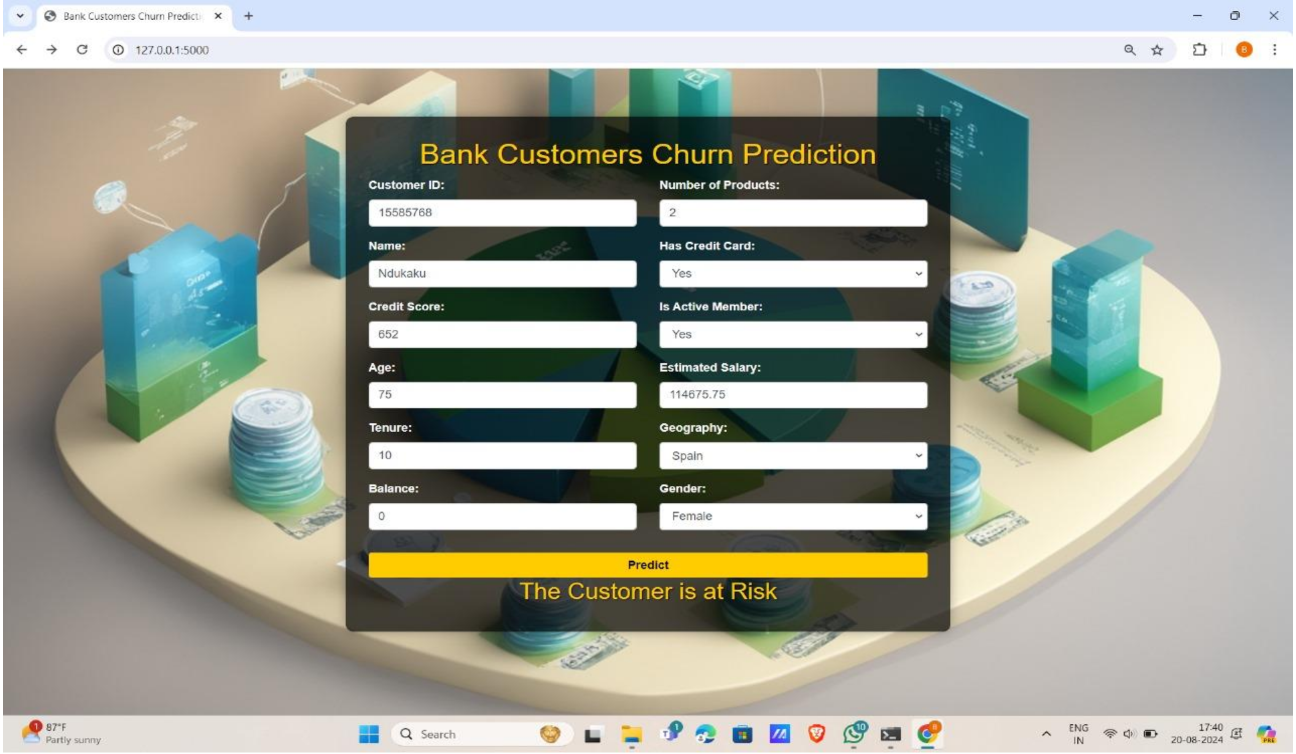
POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A87 86

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
83	82	15663706	Leonard	777	France	Female	32	2	0	1	1	0	136458.2	1
84	83	15641732	Mills	543	France	Female	36	3	0	2	0	0	26019.59	0
85	84	15701164	Onyeorulu	506	France	Female	34	4	90307.62	1	1	1	159235.3	0
86	85	15738751	Beit	493	France	Female	46	4	0	2	1	0	1907.66	0
87	86	15805254	Ndukaku	652	Spain	Female	75	10	0	2	1	1	114675.8	0
88	87	15762418	Gant	750	Spain	Male	22	3	121681.8	1	1	0	128643.4	1
89	88	15625759	Rowley	729	France	Male	30	9	0	2	1	0	151869.4	0
90	89	15622897	Sharpe	646	France	Female	46	4	0	3	1	0	93251.42	1
91	90	15767954	Osborne	635	Germany	Female	28	3	81623.67	2	1	1	156791.4	0
92	91	15757535	Heap	647	Spain	Female	44	5	0	3	1	1	174205.2	1
93	92	15731511	Ritchie	808	France	Male	45	7	118626.6	2	1	0	147132.5	0
94	93	15809248	Cole	524	France	Female	36	10	0	2	1	0	109614.6	0
95	94	15640635	Capon	769	France	Male	29	8	0	2	1	1	172290.6	0
96	95	15676966	Capon	730	Spain	Male	42	4	0	2	0	1	85982.47	0
97	96	15699461	Fiorentini	515	Spain	Male	35	10	176274	1	0	1	121277.8	0
98	97	15738721	Graham	773	Spain	Male	41	9	102827.4	1	0	1	64595.25	0
99	98	15693683	Yuille	814	Germany	Male	29	8	97086.4	2	1	1	197276.1	0
100	99	15604348	Allard	710	Spain	Male	22	8	0	2	0	0	99645.04	0
101	100	15633059	Fanucci	413	France	Male	34	9	0	2	0	0	6534.18	0
102	101	15808582	Fu	665	France	Female	40	6	0	1	1	1	161848	0
103	102	15743192	Hung	623	France	Female	44	6	0	2	0	0	167162.4	0
104	103	15580146	Hung	738	France	Male	31	9	82674.15	1	1	0	41970.72	0
105	104	15776605	Bradley	528	Spain	Male	36	7	0	2	1	0	60536.56	0
106	105	15804919	Dunbabin	670	Spain	Female	65	1	0	1	1	1	177655.7	1
107	106	15613854	Mauldon	622	Spain	Female	46	4	107073.3	2	1	1	30984.59	1
108	107	15599195	Stiger	582	Germany	Male	32	1	88938.62	1	1	1	10054.53	0
109	108	15812878	Parsons	785	Germany	Female	36	2	99806.85	1	0	1	36976.52	0
110	109	15602312	Walkom	605	Spain	Male	33	5	150092.8	1	0	0	71862.79	0
111	110	15744689	T'ang	479	Germany	Male	35	9	92833.89	1	1	0	99449.86	1
112	111	15803526	Eremenko	685	Germany	Male	30	3	90536.81	1	0	1	63082.88	0
113	112	15665790	Rowntree	538	Germany	Male	39	7	108055.1	2	1	0	27231.26	0

Churn_Modelling

Figure 8.5: Taking a customer details from trained dataset and expecting correct output.



Bank Customers Churn Prediction

Customer ID: 15585768

Name: Ndukaku

Credit Score: 652

Age: 75

Tenure: 10

Balance: 0

Number of Products: 2

Has Credit Card: Yes

Is Active Member: Yes

Estimated Salary: 114675.75

Geography: Spain

Gender: Female

Predict

The Customer is at Risk

Figure 8.6: The output is not correct for this input taken from trained dataset.

9. CONCLUSION

In the Bank Customer Churn Prediction project, the primary goal is to accurately predict which customers are likely to churn, enabling the bank to take proactive steps to retain them. To accomplish this, the project utilizes advanced techniques such as Borderline-SMOTE and Gradient Boosting Machine (GBM). Borderline-SMOTE, or Synthetic Minority Over-sampling Technique, is employed to address the class imbalance that often occurs in churn prediction datasets, where the number of churned customers is significantly smaller compared to those who stay. By generating synthetic samples near the decision boundary, where misclassification is most likely, Borderline-SMOTE enhances the model's ability to distinguish between churned and non-churned customers, leading to improved model performance.

GBM is a powerful ensemble learning method that builds a series of decision trees sequentially, with each tree correcting errors made by the previous ones. This iterative process helps to increase the model's accuracy and is particularly effective in managing complex datasets with intricate feature relationships. In this project, the data is first preprocessed and balanced using Borderline-SMOTE to ensure that the GBM model can learn from a well-represented dataset. The model is then trained on this balanced data, capturing subtle patterns and relationships that are critical for accurate churn prediction.

After training, the model is tested on a separate dataset to evaluate its performance and ensure it can accurately predict churn for new, unseen data. Additionally, the project features a mechanism to store details of customers identified as at risk of churn, enabling the bank to focus its retention efforts where they are most needed. Non-risk customers receive no further action, optimizing resource allocation. By combining Borderline-SMOTE and GBM, the project delivers a robust solution for predicting customer churn, helping the bank retain valuable customers and achieve long-term profitability.

10. FUTURE ENHANCEMENTS

Future enhancements to the Bank Customer Churn Prediction project could significantly boost its accuracy and relevance. Exploring and integrating advanced machine learning models, such as deep learning techniques, may better capture complex data patterns. Additionally, refining the feature engineering process could reveal new insights and improve model performance. Another key improvement is the integration of real-time data processing, allowing the model to dynamically predict churn as new data is received, making the system more adaptable to evolving customer information.

Expanding the model to include external data sources—such as economic indicators, social media sentiment, and customer interaction history—could offer a more comprehensive view of churn factors, leading to more accurate predictions and targeted retention strategies. Furthermore, incorporating explainability methods like SHAP (Shapley Additive Explanations) would enhance the transparency and interpretability of the model’s predictions, thereby building trust with stakeholders and supporting better decision-making. These enhancements would collectively improve the project's effectiveness and utility in the banking sector.

.

11. REFERENCES

1. I. Ullah, B. Raza, A. K. Malaik, M. Imran, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector”, *IEEE Access*, Volume 7, 2024

2. A. Hammoudeh, M. Fraihat, M. Almomani, “Selective Ensemble Model for Telecom Churn Prediction”, *Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2024.

3. A. Alamsyah, N. Salma, “A Comparative Study of Employee Churn Prediction Model”, *4th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, 2024.

4. M. Karanovic, M. Popocac, S. Sladojecic, M. Arsenovic, D. Stefanovic, “Telecommuni cation Services Churn Prediction Deep Learning Approach”, *26th Telecommunications Forum TELFOR 2018*, November 2023.

5. A. Bhanagar, S. Srivastava, “Performance Analysis of Hoeffding and Logistic Algorithm for Churn Prediction in Telecom Sector”, *2023 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*

6. S. M. Basha, A. Khare, J. Gadipalli, “Training and Deploying Churn Prediction Model using Machine Learning Algorithms”, *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, Vol 5 Issue 4, April 2023.

7. S. Agrawal, A. Das, A. Gaikwad, “Customer Churn Prediction Modelling Based on Behavioural Pattern Analysis using Deep Learning”, *International Conference on Smart Computing and Electronic Enterprise (ICSCEE 2023).*

8. N. R. Gupta, V. Pathak, A. Sharma, “Predictive Analytics for Bank Customer Churn Prediction using Deep Learning”, *International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, October 2023.

9. L. Wang, Y. Zhang, H. Liu, “Predicting Bank Customer Churn with Imbalanced Data using SMOTE and XGBoost”, *International Conference on Data Science and Machine Learning (DSML)*, November 2023.

10. N. Kumar, A. Singh, “A Hybrid Model for Bank Customer Churn Prediction Combining Machine Learning and Business Rules”, *2021 International Conference on Business Analytics (ICBA)*, June 2023.

11. X. Hu, Y. Yang, L. Chen, S. Zhu, “Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network”, *2023 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, May 2023.

12. M. Rahman, V. Kumar, “Machine Learning Based Customer Churn Prediction in Banking”, *2023 IEEE International Conference on Machine Learning and Applications (ICMLA)*, January 2022.

13. M. K. Gupta, R. Verma, A. Saxena, "Customer Churn Prediction using Ensemble Learning Techniques: A Comparative Study", *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, December 2022.

14. T. Johnson, L. Clark, S. Patel, “Applying Gradient Boosting Techniques for Churn Prediction in E-commerce”, *2022 International Conference on E-Commerce and Digital Marketing (ECOM)*, April 2022.

15. **J. Taylor, M. Evans**, “A Comparative Analysis of Machine Learning Algorithms for Churn Prediction in the Retail Sector”, *IEEE International Conference on Retail Analytics (ICRA)*, November 2021.

16. R. Huang, Q. Zhao, S. Li, “Advanced Churn Prediction Model Using XGBoost and Feature Selection Techniques”, *IEEE International Conference on Artificial Intelligence and Machine Learning (AIML)*, September 2021.

17. C. Wilson, A. Roberts, J. Lee, “Predictive Modeling of Customer Churn in the Telecommunications Industry Using Random Forest and Deep Learning”, *IEEE International Conference on Computational Intelligence (ICCI)*, June 2021.