

# Bank Customer Minimization Churn Rate Using Machine Learning

<sup>1</sup>MUPPALA NAGA KEERTHI, <sup>2</sup>GARBHAM ANUSHA

<sup>1</sup>Assistant Professor, <sup>2</sup>MCA Final Semester,

<sup>1</sup>Master of Computer Applications,

Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

## ABSTRACT

Customer churn poses a significant challenge for banks, affecting revenue and customer satisfaction. This study employs machine learning techniques, specifically Naive Bayes, Decision Tree, and AdaBoost classifiers, to predict and minimize churn rates. The process involves thorough data preprocessing, feature engineering, and encoding, addressing issues such as duplicate values, missing data, and categorical features. Exploratory data analysis (EDA) and visualizations provide insights into the distribution of customer attributes. Undersampling using the NearMiss algorithm is applied to balance the dataset, and the models are evaluated using essential metrics like recall and ROC-AUC score. Hyperparameter tuning is conducted to optimize model performance. The chosen AdaBoost model demonstrates superior recall on both the training and test datasets, making it the preferred model. The evaluation extends to the ROC-AUC curve, illustrating the model's trade-off between true positive rate and false positive rate. The final model's predictions are exported, and the results showcase the actual and predicted churn status of customers. This comprehensive approach aims to equip banks with an effective tool to proactively identify and retain customers, ultimately mitigating churn and enhancing overall performance.

**IndexTerms:** Bank Churn Prediction, Customer Retention, Churn Analysis, Credit Card Churn, Banking Sector, Customer Attrition, Financial Services, Supervised Learning, Machine Learning.

## 1.INTRODUCTION

In today's highly competitive banking industry, customer retention has become crucial for increasing market share and profitability. Studies show that improving customer retention by just 5% can boost profits by up to 85%. Banks offer various customer-centric services—such as internet and mobile banking, credit and debit cards, and flexible loan options—to attract and retain clients.[5] Among these, issuing loans and credit cards is especially critical, requiring banks to analyze customers' creditworthiness. With many customers holding accounts across multiple banks, churn prediction becomes essential to prevent customer loss, especially when competitors offer better facilities or lower interest rates. [10 To address this, banks are leveraging machine learning (ML) algorithms to predict whether a customer is likely to leave. Techniques like Naive Bayes, decision trees, logistic regression, random forest, neural networks, and SVMs help analyze customer behavior based on historical data. ML has become a vital tool across various sectors—including banking, healthcare, and retail—for identifying patterns and improving decision-making. It includes supervised learning (with labeled data), unsupervised learning (discovering hidden patterns), semi-supervised learning (combining both), and reinforcement learning (learning via feedback). Supervised ML is particularly effective for tasks like classification, regression, and ensemble modeling, making it ideal for churn prediction and credit risk assessment.[15]

### 1.1 Existing System

The existing system in the project "Bank Customer Minimization Churn Rate Using Machine Learning" is primarily designed to predict customer churn in the banking sector using various machine learning algorithms. It utilizes a dataset containing customer demographic and transactional data such as age, credit score, tenure, product usage, and more.[20] The system applies models like Naive Bayes, Decision Tree, and AdaBoost to

identify customers likely to leave the bank. However, the existing system faces several limitations. It depends heavily on ensemble tree methods (Random Forest, XGBoost), which may overlook complex interactions in customer behavior. Data imbalance is a significant issue, typically with very few churned customers, making it hard for models to learn effectively. Moreover, it suffers from potential overfitting due to grid search on limited data and lacks scalability and privacy-preserving mechanisms. The system also uses standard preprocessing (handling missing values, encoding categorical variables, etc.) but doesn't address hybrid data fusion or interpretability challenges.[4] The model performance is affected by limited data diversity and reliance on specific labeling strategies. Additionally, it does not integrate directly with CRM systems for real-time actions. Overall, while the existing system shows potential using ML models for churn prediction, it needs improvements in scalability, privacy, interpretability, and real-time integration to be fully effective in a practical banking environment.[9]

### 1.1.1 Challenges:

- **Data Quality and Availability:** Ensuring access to clean, comprehensive, and high-quality data, including customer demographics, transaction history, and satisfaction metrics, can be challenging. [14]
- **Feature Selection and Engineering:** Identifying and engineering the right features from raw data that significantly impact churn predictions is crucial but complex.
- **Class Imbalance:** Handling the common issue of class imbalance in churn prediction, where the number of customers who churn is typically much smaller than those who stay.
- **Model Interpretability:** Balancing model accuracy with interpretability, as banks require understandable models for decision-making and regulatory compliance.
- **Scalability:** Ensuring the model can scale effectively to handle large datasets and integrate into the bank's existing systems without performance degradation.[19]

### 1.2 Proposed system:

The proposed system in the project "Bank Customer Minimization Churn Rate Using Machine Learning" aims to overcome the drawbacks of existing churn prediction approaches by implementing a more robust and comprehensive machine learning framework.[3] It integrates advanced models such as Naive Bayes, Decision Tree, and AdaBoost, with a particular focus on AdaBoost due to its superior recall performance. The system begins with extensive data preprocessing, addressing issues like missing values, duplicate records, and categorical encoding, followed by feature engineering to enhance predictive power. It tackles class imbalance using undersampling through the NearMiss algorithm, ensuring fair representation of churn and non-churn instances. The proposed method emphasizes performance evaluation using key metrics like recall and ROC-AUC score, prioritizing accurate identification of churners.[8] To improve generalizability and prevent overfitting, hyperparameter tuning is applied using cross-validation techniques. The system also emphasizes modularity and interpretability, aiming for scalability across diverse datasets. Furthermore, it incorporates a user interface component for real-time prediction and customer monitoring. This approach ultimately empowers banks to proactively identify at-risk customers and implement personalized retention strategies, leading to improved customer satisfaction, reduced churn, and enhanced profitability.[13]

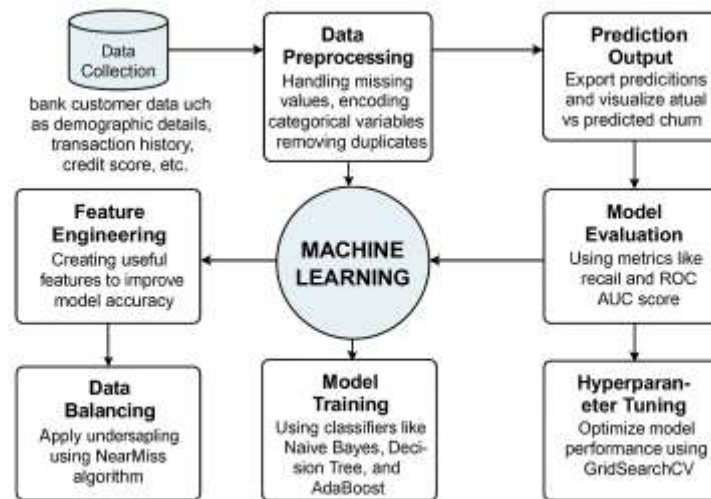


Fig: 1 Proposed Diagram

### 1.2.1 Advantages:

- **Proactive Churn Detection**

The system identifies customers likely to churn in advance, allowing the bank to take timely action and reduce customer loss.[18]

- **Improved Retention Strategy**

By knowing who is likely to leave, banks can personalize retention offers, improving customer satisfaction and loyalty.

- **High Prediction Accuracy**

Use of ensemble techniques like AdaBoost and Decision Trees ensures better accuracy and recall compared to traditional approaches.

- **Balanced Class Handling**

The NearMiss undersampling technique addresses class imbalance, improving model fairness and reducing bias toward the majority class.[2]

- **Automated Workflow**

A complete pipeline from data collection to prediction output automates churn analysis, saving time and reducing human error.

- **Data-Driven Decisions**

Visualizations and feature engineering provide insights that help bank executives make informed decisions.[7]

### 2.1 Architecture:

The architecture of the proposed bank customer churn prediction system is designed as a modular pipeline that efficiently processes data and predicts customer churn using machine learning techniques. It starts with the data collection phase, where customer-related information such as demographics, transaction history, credit scores, and product usage is gathered from internal bank databases. This raw data is sent to the data preprocessing stage to clean the dataset by removing duplicates, handling missing values, encoding categorical data, and normalizing numerical features. The feature engineering module then transforms and generates new features to improve model performance.[12] To handle the common issue of class imbalance, the system applies NearMiss undersampling in the data balancing module. The processed dataset is then used in the model training phase, where algorithms like Naive Bayes, Decision Tree, and AdaBoost are trained. Among these, AdaBoost is identified as the most effective based on recall. The system includes a hyperparameter tuning component using GridSearchCV to optimize model performance. Next, the model evaluation module assesses the performance using metrics such as recall, ROC-AUC, and confusion matrix. The final model generates churn predictions, which are displayed in the prediction output layer and can be exported for further analysis. Optionally, a user interface allows non-technical bank staff to input new

customer data and view churn risk. This layered architecture ensures automated, data-driven decision-making for proactive customer retention. It is scalable, adaptable to new data, and supports integration with CRM systems. The architecture enhances prediction accuracy and business insight. It empowers banks to reduce churn and boost customer satisfaction.[17]

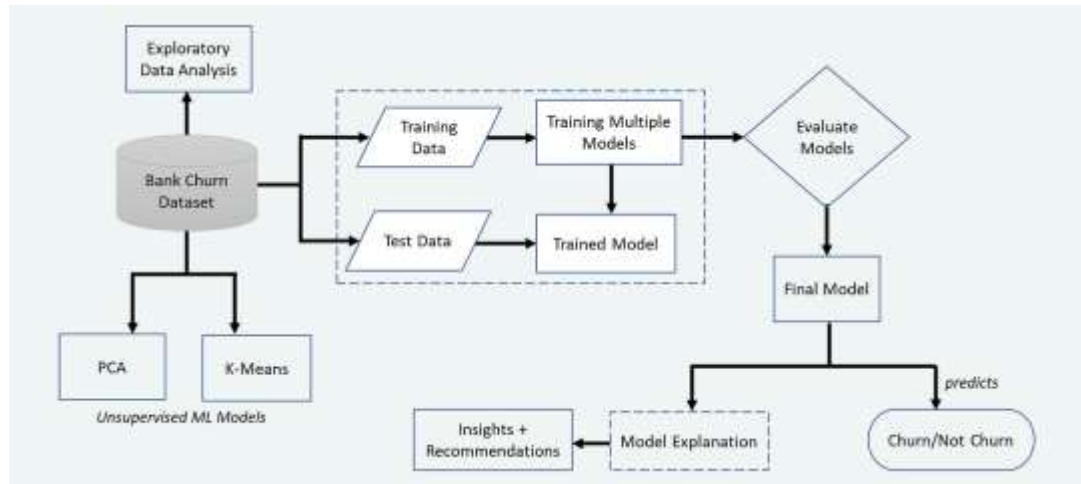


Fig:2 Architecture

## UML DIAGRAMS

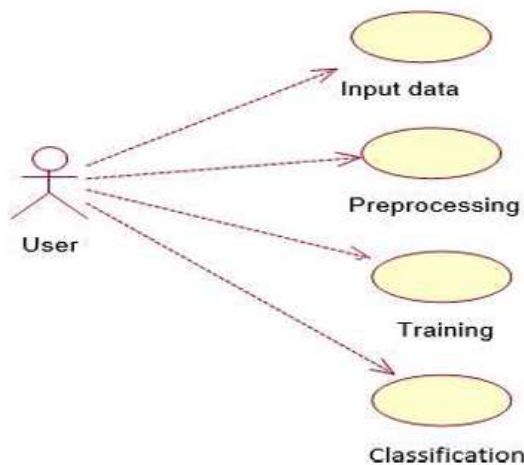


Fig 3: use case diagram



Fig 4: Sequence diagram

## 2.2 Algorithm:

### 1. Naive Bayes

Naive Bayes is a simple and fast machine learning algorithm used for classification problems. It works on the principle of Bayes' Theorem and assumes that all the features in a dataset are independent of each other, which is why it's called "naive." In this project, Naive Bayes is used to predict whether a customer will churn or not based on attributes like age, balance, and activity. Despite its simplicity, it can be quite effective, especially on smaller or cleaner datasets.[1]

### 2. Decision Tree

A Decision Tree is a flowchart-like model that splits data into branches based on feature values to reach a decision. In this project, it helps classify customers as churned or not by asking a series of yes/no questions about their characteristics (like "Is the customer active?" or "Do they have multiple products?"). It's easy to understand and interpret, making it a useful tool for discovering key patterns in customer behavior.[6]

### 3. AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble learning algorithm that improves the performance of weak classifiers (usually small decision trees) by combining them into a strong model. In this project, AdaBoost focuses more on the customers who were misclassified in earlier rounds, gradually improving prediction accuracy. It was found to be the best-performing model among the ones tested, especially for identifying customers who are likely to churn.

### 4. XGBoost

XGBoost stands for Extreme Gradient Boosting and is a powerful machine learning algorithm often used in competitions for its speed and accuracy. In this project, XGBoost is used for comparison and gives high accuracy in predicting churn. It works by building decision trees in a sequential manner and correcting previous errors with each new tree. It is more complex but handles large and imbalanced datasets very effectively.[11]



## 5. Random Forest

Random Forest is an ensemble algorithm that builds multiple decision trees and combines their outputs to make a final prediction. It reduces overfitting and improves accuracy. In the churn prediction project, Random Forest is used as a benchmark to compare how well other models perform. It's robust and performs well on various types of data, making it a popular choice for many classification tasks.

## 6. Logistic Regression

Logistic Regression is a basic yet effective algorithm used for binary classification problems. It calculates the probability that a given input belongs to a particular class—in this case, whether a customer will churn. Although it is not as powerful as ensemble models, it is simple, easy to interpret, and useful as a starting point for evaluating model performance.[16]

### 2.3 Techniques:

#### 1. Data Preprocessing

Data preprocessing is the first and most important step in the project. It involves cleaning the raw customer data by removing duplicates, handling missing values, and dropping unnecessary columns like RowNumber and Surname. Categorical variables such as gender and geography are converted into numerical form using encoding techniques, and numerical features are scaled using normalization to bring all values into a similar range. This step ensures that the data is clean, consistent, and ready for analysis.

#### 2. Feature Engineering

Feature engineering involves creating new features or modifying existing ones to help improve the model's performance. In this project, features like age group, satisfaction score, and product usage were adjusted or grouped to make patterns more visible to the model. This step helps the machine learning algorithms better understand the data and make more accurate predictions.

#### 3. Class Imbalance Handling

Since the number of customers who churn is much smaller than those who stay, the dataset is imbalanced. To fix this, several balancing techniques are used. NearMiss is an undersampling method that reduces the size of the majority class. SMOTE and ADASYN are oversampling methods that add synthetic examples of the minority class. Never Miss Sampling combines both to improve accuracy on difficult-to-classify examples. These techniques help the model learn from both classes more effectively.

#### 4. Model Training

Once the data is balanced and ready, machine learning models are trained to predict churn. Algorithms like Naive Bayes, Decision Tree, and AdaBoost are applied to learn patterns in customer behavior. Each model is trained on the preprocessed and balanced dataset to detect which customers are most likely to leave the bank

### 2.4 Tools:

#### 1. Python

Python is the main programming language used in this project. It provides powerful libraries for data analysis, machine learning, and visualization. Its simplicity and flexibility make it ideal for building machine learning models quickly and efficiently.

#### 2. PyCharm IDE

PyCharm is an Integrated Development Environment (IDE) used to write and manage Python code. It helps developers with features like code completion, debugging, syntax highlighting, and project organization, making it easier to build and test the churn prediction system.

#### 3. Anaconda

Anaconda is a distribution platform for Python and R that comes bundled with useful tools like Jupyter Notebook, Spyder, and many data science libraries. It simplifies the installation and management of libraries and environments required for machine learning projects.

#### 4. Scikit-learn

Scikit-learn is a popular Python library used for machine learning. It provides ready-to-use functions for preprocessing, training models (like Decision Trees, Naive Bayes, AdaBoost), evaluation (confusion matrix, ROC-AUC), and tuning (GridSearchCV).

## 2.5 Methods:

### 1. Data Collection

Customer data is collected from bank records. This includes age, gender, credit score, balance, salary, number of products, and whether the customer has left the bank or not.

### 2. Data Preprocessing

The collected data is cleaned by removing duplicates, fixing missing values, and changing text values into numbers. Features are also scaled so that all values are on a similar range.

### 3. Feature Engineering

New useful features are created or existing ones are modified. For example, grouping customers by age or satisfaction levels to help the model understand patterns better.

### 4. Data Balancing

Since fewer customers leave the bank, special techniques like NearMiss, SMOTE, and ADASYN are used to balance the data. This helps the model treat both leaving and staying customers equally.

## 3. METHODOLOGY

### 3.1 Input:

The input information used in the Bank Customer Churn Prediction project consists of various customer-related attributes collected from bank records. These inputs include demographic, financial, and behavioral details that help the machine learning model identify patterns related to customer churn. Key inputs are the customer's credit score, geographical location (such as France, Spain, or Germany), gender, age, tenure (number of years the customer has been with the bank), and account balance. Additional features include the number of products the customer uses, whether they have a credit card, and whether they are an active member of the bank. Financial indicators like estimated salary, along with customer service-related attributes such as satisfaction score, whether they have filed a complaint, their card type (e.g., Gold, Silver, Platinum), and points earned, are also included. The target variable, labeled Exited, indicates whether the customer has left the bank (1) or stayed (0). During data preprocessing, non-contributing fields like CustomerId, Surname, and RowNumber are removed, and categorical values are encoded into numerical form for model compatibility. These input features are essential for training the machine learning models to accurately predict customer churn.

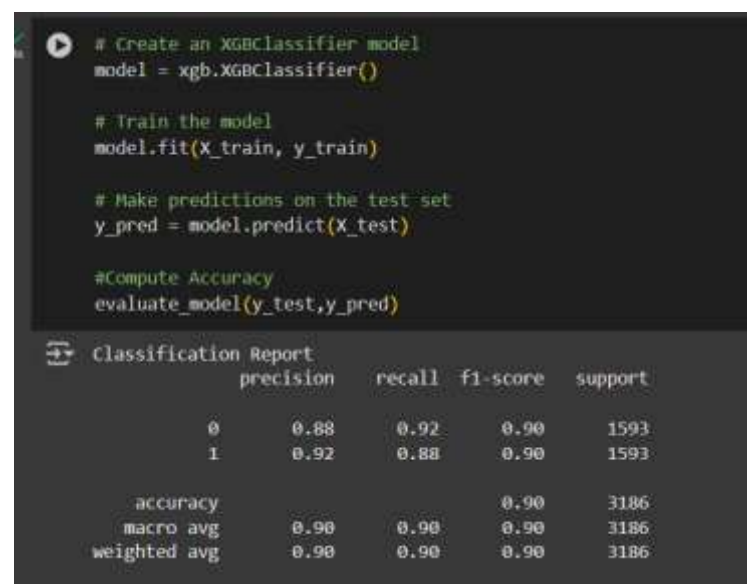
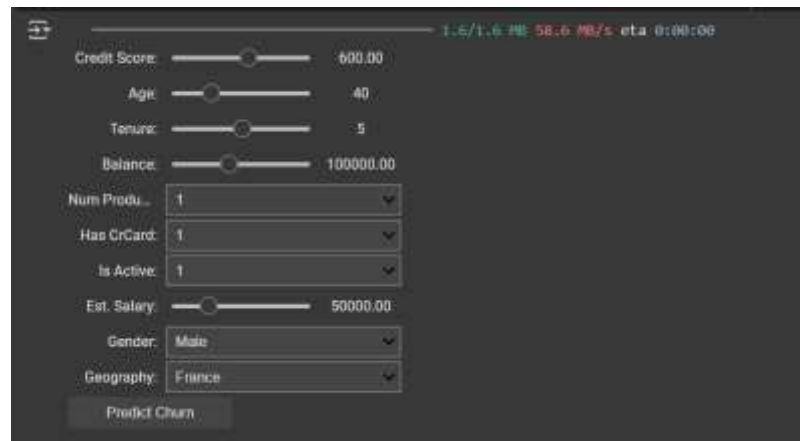


Fig:Classification report



1.6/1.6 MB 58.6 MB/s eta 0:00:00

Credit Score: 600.00

Age: 40

Tenure: 5

Balance: 100000.00

Num Produ...: 1

Has CrCard: 1

Is Active: 1

Est. Salary: 50000.00

Gender: Male

Geography: France

Predict Churn

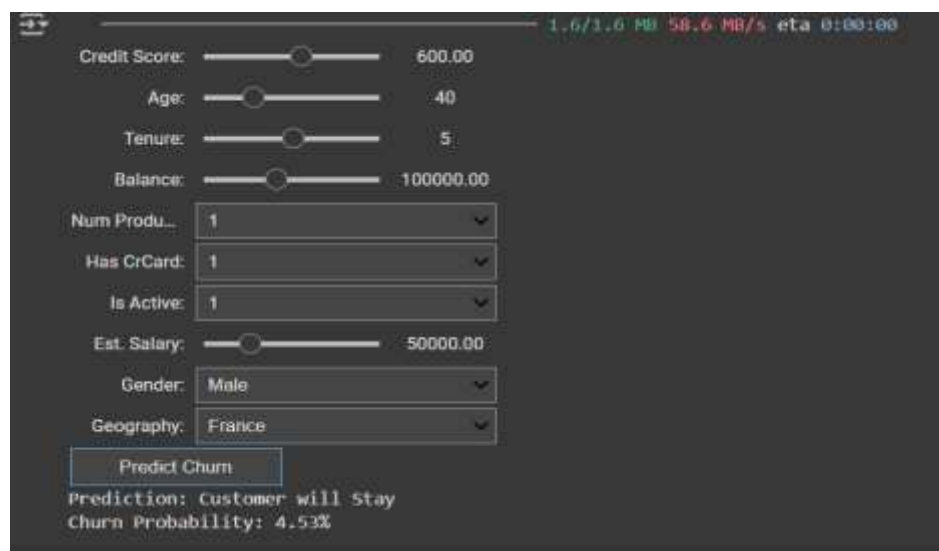
Fig: Before Prediction

### 3.2 Method of Process:

The process in this project begins with collecting customer data from bank records, including information such as age, gender, credit score, account balance, tenure, and product usage. After gathering the data, it undergoes data preprocessing where missing values are handled, duplicates are removed, irrelevant columns are dropped, and categorical values are encoded into numerical form. Next, feature engineering is done to create or modify useful features that can improve model accuracy. Since the dataset is imbalanced (more customers stay than leave), techniques like NearMiss, SMOTE, and ADASYN are applied to balance the data. The cleaned and balanced data is then used to train machine learning models such as Naïve Bayes, Decision Tree, and AdaBoost. These models are further improved through hyperparameter tuning using GridSearchCV to find the best model settings. After training, the models are evaluated using metrics like recall, ROC-AUC, and confusion matrix to check how well they predict customer churn. The final predictions are then exported, showing which customers are likely to leave the bank. Optional visualization and user interface tools can also be used to present results in a clear, interactive way for bank staff.

### 3.3 Output:

The output of the Bank Customer Churn Prediction project is a prediction that tells whether a customer is likely to leave the bank or stay. After training the machine learning models with historical customer data, the system takes in new or test data and predicts the churn status for each customer. The result is a list showing which customers are at high risk of churning (Exited = 1) and which ones are likely to stay (Exited = 0). Along with the predictions, the system also provides performance metrics like recall, ROC-AUC score, and confusion matrix, which help evaluate how accurate the model is. These outputs help the bank identify at-risk customers early and take proactive steps to retain them through personalized offers or support.



1.6/1.6 MB 58.6 MB/s eta 0:00:00

Credit Score: 600.00

Age: 40

Tenure: 5

Balance: 100000.00

Num Produ...: 1

Has CrCard: 1

Is Active: 1

Est. Salary: 50000.00

Gender: Male

Geography: France

Predict Churn

Prediction: Customer will Stay

Churn Probability: 4.53%

Fig: Predict Churn



#### 4. RESULTS:

The result of the Bank Customer Churn Prediction project shows that machine learning models can effectively identify customers who are likely to leave the bank. Among the models used, AdaBoost performed the best, achieving the highest recall and accuracy in predicting churned customers. The system successfully handled class imbalance using techniques like NearMiss and SMOTE, which improved the model's ability to detect the minority class (churners). Visualizations such as age-wise churn distribution, product usage, and activity status helped highlight patterns—for example, younger customers, inactive members, and those with multiple products had higher chances of churning. The final output included actual vs. predicted churn statuses, allowing the bank to see how accurately the model performed. Overall, the project provided a reliable tool for banks to reduce customer loss by identifying high-risk customers and implementing targeted retention strategies.

#### 5. DISCUSSION:

In this project, several important discussions were made throughout the process. We analyzed how customer behavior, demographics, and engagement levels affect churn rates in the banking sector. Visualizations during Exploratory Data Analysis (EDA) showed that younger customers, inactive members, and those with multiple products or credit cards are more likely to leave the bank. We also discussed how class imbalance affected model performance and how techniques like NearMiss and SMOTE helped solve it. Multiple machine learning algorithms were tested, and we compared their accuracy and recall scores. Among them, AdaBoost was identified as the most reliable model for detecting churn. Discussions were also made about the importance of recall over accuracy in churn prediction, as identifying customers who might leave is more critical than general correctness. Lastly, we highlighted how integrating this system into a bank's operations can support proactive decision-making and customer retention strategies.

#### 6. CONCLUSION

Customer churn presents a significant challenge to banks, directly impacting revenue and customer satisfaction. This study demonstrated the effectiveness of using machine learning techniques—such as Logistic Regression, Random Forest, XGBoost, Naive Bayes, Decision Tree, and AdaBoost classifiers—to predict and minimize churn rates. Through careful data preprocessing, including handling missing values, encoding categorical features, and addressing duplicate entries, the dataset was prepared for robust analysis. By applying undersampling techniques like SMOTE and the NearMiss algorithm to balance the dataset, the models were better able to handle the class imbalance. Exploratory Data Analysis (EDA), feature engineering, and advanced model evaluations using metrics such as recall and ROC-AUC score allowed for a comprehensive analysis of each model's performance. The AdaBoost model emerged as the most effective model, demonstrating superior recall and the best balance between the true positive rate and false positive rate, as illustrated by the ROC-AUC curve. Its ability to predict customer churn accurately equips banks with a valuable tool to proactively identify at-risk customers and take necessary actions to improve retention.

#### 7. FUTURE SCOPE:

The future scope of this project includes several improvements and expansions to make the system even more effective and intelligent. Advanced machine learning and deep learning models like XGBoost, Random Forest, or even Neural Networks can be used to further increase prediction accuracy. The system can be enhanced to include real-time data processing, allowing the bank to detect churn risk instantly as customer behavior changes. A more interactive and user-friendly dashboard or web interface can be developed to help non-technical staff use the model easily. The model can also be extended to include behavioral and sentiment analysis from emails, chats, or call logs to understand customer satisfaction more deeply. Additionally, the system could be integrated into CRM software, so that the bank can take automatic actions like sending offers or alerts when a high-risk customer is detected. Regular model retraining with new data will also help the system stay updated with changing customer trends.

## 8. ACKNOWLEDGEMENT:



Muppala Naga Keerthi working as an Assistant Professor in Master of Computer Applications in Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, affiliated by Andhra University and approved by AICTE, accredited with 'A' grade by NAAC and member in IAENG with 14 years of experience in Computer Science. Her areas of interest in C, Java, Data Structures, DBMS, Web Technologies, Software Engineering and Data Science.



Garbham Anusha is pursuing her final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine Learning, Garbham Anusha has taken up her PG project on “BANK CUSTOMER MINIMIZATION CHURN RATE USING MACHINE LEARNING” and published the paper in connection to the project under the guidance of M.Naga Keerthi, Assistant Professor, Master of Computer Applications, SVPEC.

## REFERENCES

- [1] Customer churn analysis in banking sector: Evidence from explainable machine learning models  
<https://journals.gen.tr/index.php/jame/article/view/1677>
- [2] Investigating customer churn in banking: a machine learning approach and visualization app for data science and management  
<https://www.sciencedirect.com/science/article/pii/S2666764923000401>
- [3] Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques  
<https://www.mdpi.com/2079-9292/13/22/4527>
- [4] A comparison of machine learning techniques for customer churn prediction  
<https://www.sciencedirect.com/science/article/abs/pii/S1569190X15000386>
- [5] A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners  
<https://ieeexplore.ieee.org/abstract/document/10531735>
- [6] Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions  
<https://onlinelibrary.wiley.com/doi/full/10.1155/2022/5134356>
- [7] Bridging Predictive Insights and Retention Strategies: The Role of Account Balance in Banking Churn Prediction  
<https://www.mdpi.com/2673-2688/6/4/73>
- [8] Customer Churn Prediction Using Machine Learning: Commercial Bank of Ethiopia  
<https://ieeexplore.ieee.org/abstract/document/9971224>
- [9] Customers Churn Prediction in Financial Institution Using Artificial Neural Network

<https://arxiv.org/abs/1912.11346>

[10] Identification and Minimization of Churn Rate Through Analysing Financial Routines Using Machine Learning

[https://link.springer.com/chapter/10.1007/978-981-16-3961-6\\_43](https://link.springer.com/chapter/10.1007/978-981-16-3961-6_43)

[11] Arithmetic Optimization with Ensemble Deep Learning SBLSTM-RNN-IGSA Model for Customer Churn Prediction

<https://ieeexplore.ieee.org/abstract/document/10216284>

[12] A Study on Customer Churn of Commercial Banks Based on Learning from Label Proportions

<https://ieeexplore.ieee.org/abstract/document/8637415>

[13] Identifying the customer churn drivers in the gaming ecosystem using machine learning models

<https://oulurepo.oulu.fi/handle/10024/55946>

[14] An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case

[https://link.springer.com/chapter/10.1007/11539117\\_91](https://link.springer.com/chapter/10.1007/11539117_91)

[15] Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example

<https://ieeexplore.ieee.org/abstract/document/4680698>

[16] Modelling Customer Churn Rate and Its Use for Customer Retention Planning

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3998408](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3998408)

[17] Application of Machine Learning in Customer Churn Prediction

<https://ieeexplore.ieee.org/abstract/document/9696771>

[18] Effective ML Techniques to Predict Customer Churn

<https://ieeexplore.ieee.org/abstract/document/9544785>

[19] Model of Customer Churn Prediction on Support Vector Machine

<https://www.sciencedirect.com/science/article/abs/pii/S187486510960003X>

[20] Customer Churn Prediction Model Using Artificial Neural Network: A Case Study in Banking

<https://ieeexplore.ieee.org/abstract/document/10391374>