Enhanced Banking Services Using ML and AI

Team Name: Team Prathamesh

Team Members:

- **Saloni Rai**
- **Hemanth Kumar**
- **Sevantkumar Huggi**
- **Greeshma Haridas**
- **Prathamesh Patil**

Project Index

GitHub Links:

Prathamesh - https://github.com/PRATHAMESHPATIL123654/Customer-churn-prediction-model-.git

Sevantkumar - https://github.com/Sevantkumar/Loan-Prediction

Saloni Rai – Saloni-Rai19/banking (github.com)

Greeshma Haridas – https://github.com/GreeshmaHarids/Final_Project_FeynnLab.git

Hemanth Kumar – https://github.com/Sevantkumar/Loan-Prediction

Title: Enhanced Banking Services with Machine Learning and Artificial Intelligence

Problem Statement

In the rapidly evolving financial landscape, banks face significant challenges related to customer retention, default risks, and efficient segmentation of their customer base. Traditional methods often fall short in accurately predicting customer behavior and preferences, leading to increased operational costs and decreased customer satisfaction. This project aims to leverage machine learning (ML) techniques to develop intelligent data services that enhance banking operations. By implementing models for customer churn prediction, default prediction, and customer segmentation, the project seeks to provide actionable insights that enable banks to make data-driven decisions, improve customer relationships, and optimize financial performance.

Objectives

1. **Customer Churn Prediction**: Develop a predictive model to identify customers at risk of leaving the bank. This will enable proactive retention strategies and improve overall customer loyalty.
2. **Default Prediction**: Create a model that forecasts the likelihood of customer defaults on loans, allowing banks to minimize risk and make informed lending decisions.
3. **Customer Segmentation**: Implement a segmentation model to categorize customers based on their behavior and preferences, facilitating targeted marketing strategies and personalized services.

Methodology

1. **Data Collection**: Gather relevant datasets that include customer demographics, transaction history, credit scores, and loan information.
2. **Data Preprocessing**: Clean and preprocess the data to ensure accuracy and usability for model development.
3. **Model Development**:
   - Utilize algorithms such as logistic regression, decision trees, and ensemble methods for churn and default predictions.
   - Apply clustering techniques (e.g., K-means) for effective customer segmentation.
4. **Model Evaluation**: Assess model performance using metrics like accuracy, precision, recall, and F1-score to ensure reliability.
5. **Implementation**: Integrate the developed models into the bank's existing infrastructure to provide real-time insights and recommendations.

Business Model for Predictive Analytics Models in Financial Services

Project Overview

This project involves the development of three predictive analytics models aimed at helping financial institutions optimize decision-making:

1. **Loan Amount Prediction**: Predicts the optimal loan amount a customer can take based on financial background.
2. **Customer Churn Prediction**: Identifies customers likely to leave the institution, enabling proactive retention strategies.
3. **Customer Default Prediction**: Predicts the likelihood of a customer defaulting on a loan, helping lenders manage risk.

These models provide actionable insights, reducing financial risks, improving customer retention, and optimizing lending processes.

1. Value Proposition

The predictive models deliver clear benefits to financial institutions:

- **Loan Amount Prediction**: Ensures customers are offered appropriate loan amounts, reducing rejections and optimizing loan portfolios.
- **Customer Churn Prediction**: Improves retention by identifying at-risk customers and offering targeted interventions.
- **Customer Default Prediction**: Minimizes defaults and enhances credit risk management by identifying high-risk customers early.

2. Market Opportunity

- **Financial Institutions**: Banks, credit unions, and lending companies with a need to optimize loan processes and manage customer risk.
- **Fintech Startups**: Offering these predictive services as a value-add or embedded feature in financial platforms.
- **Global Fintech Industry**: Projected to grow significantly, with machine learning solutions playing a key role in improving financial decision-making.

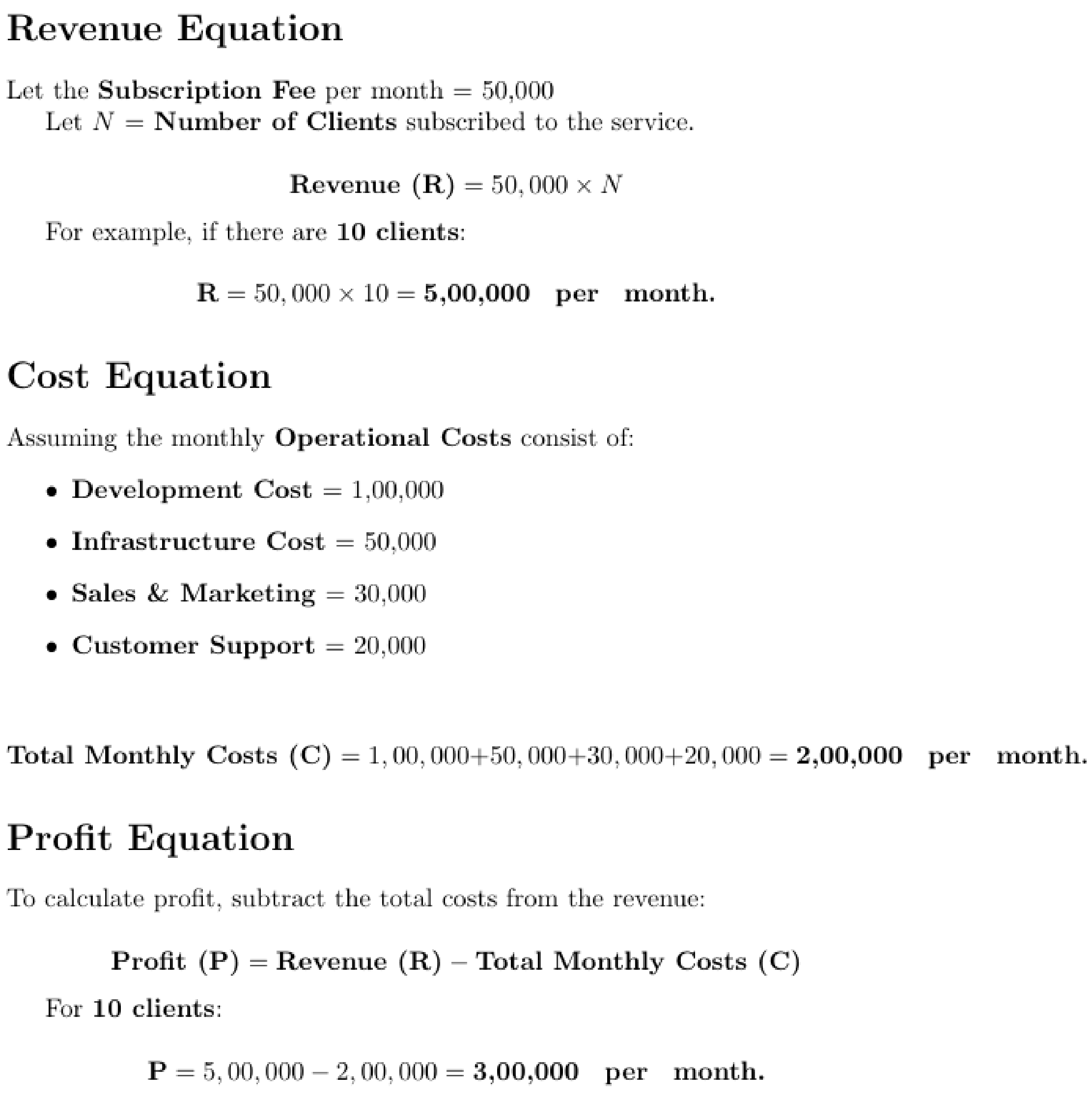
3. Revenue Model

Subscription-Based SaaS Model:

- **Subscription Fee**: ₹50,000 per month for access to the predictive platform, including all three models.
- **No Pay-Per-Use Fee**: A flat subscription fee provides unlimited access to the models for the client.

---

4. Financial Equation



Revenue Equation

Let the **Subscription Fee** per month = 50,000
Let N = **Number of Clients** subscribed to the service.

$$\textbf{Revenue (R)} = 50,000 \times N$$

For example, if there are **10 clients**:

$$\mathbf{R} = 50,000 \times 10 = \textbf{5,00,000 per month.}$$

Cost Equation

Assuming the monthly **Operational Costs** consist of:

- **Development Cost** = 1,00,000
- **Infrastructure Cost** = 50,000
- **Sales & Marketing** = 30,000
- **Customer Support** = 20,000

$$\textbf{Total Monthly Costs (C)} = 1,00,000 + 50,000 + 30,000 + 20,000 = \textbf{2,00,000 per month.}$$

Profit Equation

To calculate profit, subtract the total costs from the revenue:

$$\textbf{Profit (P)} = \textbf{Revenue (R)} - \textbf{Total Monthly Costs (C)}$$

For **10 clients**:

$$\mathbf{P} = 5,00,000 - 2,00,000 = \textbf{3,00,000 per month.}$$

---

5. Cost Structure

- **Model Development**: ₹1,00,000/month for ongoing maintenance and updates.
- **Cloud Infrastructure**: ₹50,000/month for hosting and scaling the platform.
- **Sales & Marketing**: ₹30,000/month for customer acquisition and lead generation.
- **Customer Support**: ₹20,000/month for technical and onboarding support.

---

6. Customer Segments

- **Banks and Lending Institutions**: Interested in reducing churn and defaults, and optimizing loan portfolios.
- **Fintech Startups**: Looking for integrated solutions to enhance their offerings to customers.
- **Credit Unions and Microfinance Institutions**: Benefiting from AI-driven risk management tools.

---

7. Key Activities

- **Model Development and Maintenance**: Continuously improving the accuracy and performance of the models.
- **Data Collection and Updates**: Regularly updating models with new financial data for better prediction accuracy.
- **Client Acquisition**: Engaging in direct sales and marketing efforts to bring in new clients.
- **Customer Support**: Providing 24/7 technical assistance and onboarding support to ensure customer satisfaction.

---

8. Key Resources

- **Data Scientists and Developers**: For model development, testing, and improvement.
- **Cloud Infrastructure**: To host the models and manage the prediction requests.
- **Sales and Marketing Teams**: To engage and acquire potential clients.

---

9. Key Partners

- **Cloud Providers**: AWS, Google Cloud, or Microsoft Azure for scalable and reliable infrastructure.
- **Financial Data Providers**: Collaborating with third-party data aggregators to improve model accuracy.
- **Marketing Agencies**: To help drive awareness and lead generation for potential clients.

---

10. Customer Relationships

- **Dedicated Account Managers**: For large clients, providing personalized service and technical support.
- **Automated Platform**: A user-friendly interface where clients can access the models and generate predictions with minimal friction.
- **24/7 Support**: Offering real-time assistance through email, phone, and live chat.

---

13. Scalability

- **Geographical Expansion**: Targeting financial institutions in new regions as the service matures.
- **Additional Features**: Expanding the platform to include new predictive models, such as fraud detection or personalized customer offers.
- **Automated Integration**: Allowing financial institutions to easily integrate the models with their existing systems through APIs.

Conclusion

The predictive models for loan amount, churn, and default status offer a compelling business model for financial institutions. By adopting a subscription-based revenue model, we can generate stable and predictable income while providing valuable, data-driven insights that help institutions make better financial decisions. The scalability, clear cost structure, and market demand make this a viable long-term opportunity.

Step 2 – Implementation

Customer Churn Prediction Model Report

Prepared by: Prathamesh Patil

Github link – https://github.com/PRATHAMESHPATIL123654/Customer-churn-prediction-model-.git

---

1. Introduction

In the competitive landscape of modern business, customer retention is of paramount importance. This report outlines the development of a customer churn prediction model designed to identify customers at risk of leaving. The analysis focuses on various features that influence customer behavior, enabling businesses to implement effective retention strategies.

---

2. Feature Overview

The dataset utilized in this analysis includes several key features that provide insights into customer demographics and behaviors:

- **Cust No.**: A unique identifier for each customer, removed from the analysis as it does not contribute to predictive modeling.
- **First Name and Surname**: Personal identifiers excluded to maintain customer privacy.
- **Credit Score**: A critical indicator of financial reliability, impacting customer retention.
- **Geography**: The geographical location of customers, which may reveal regional trends in churn.
- **Gender and Age**: Demographic variables that could affect customer behavior and churn rates.
- **Tenure**: The duration of customer engagement, with longer tenure generally correlating with reduced churn risk.
- **Balance**: Customer account balance, reflecting engagement and satisfaction levels.
- **Num Of Policies**: The number of policies held by a customer, indicating trust and reliance on the service.
- **Credit Card**: A binary feature indicating credit card ownership, potentially influencing loyalty.
- **Active Member**: Indicates whether a customer is actively using the service, which can directly affect churn.
- **Salary**: The financial status of customers, influencing their ability to maintain accounts.
- **Exited**: The target variable indicating customer departure, essential for modeling.

3. Data Preprocessing

Data preprocessing is vital for ensuring the quality and reliability of predictive models. The initial step involved dropping non-essential features, specifically the customer identification number.

Missing values were addressed through label encoding for categorical features and the KNN imputer for numerical ones, ensuring a complete dataset for analysis.

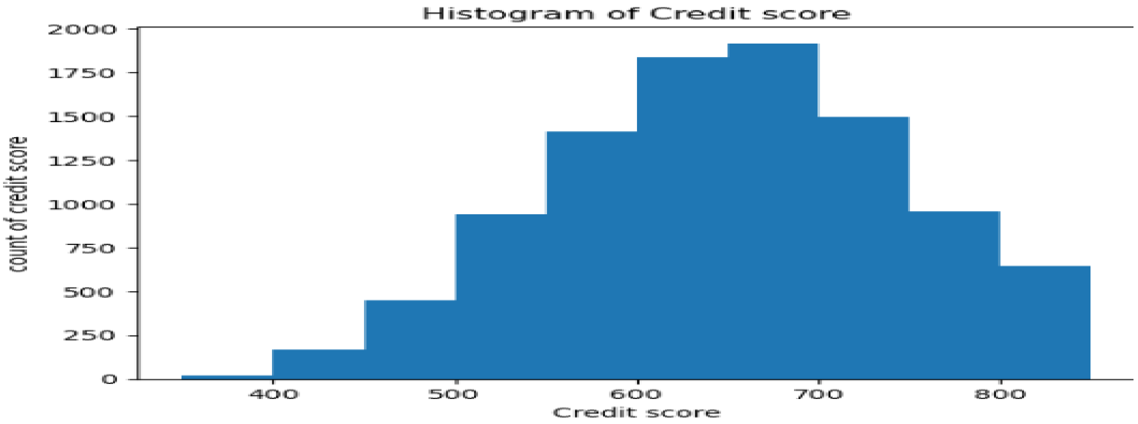
A summary function was created to compute descriptive statistics for each feature, including measures such as mean, median, standard deviation, and outlier detection. Identifying outliers was critical, as they can significantly impact the model's performance.

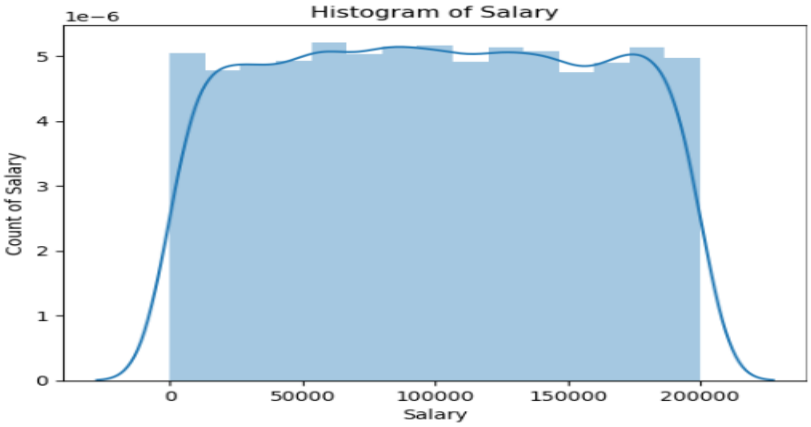
Outlier treatment strategies were employed, utilizing quartile and standard deviation methods to clean the dataset and enhance the robustness of the analysis.

4. Exploratory Data Analysis (EDA)

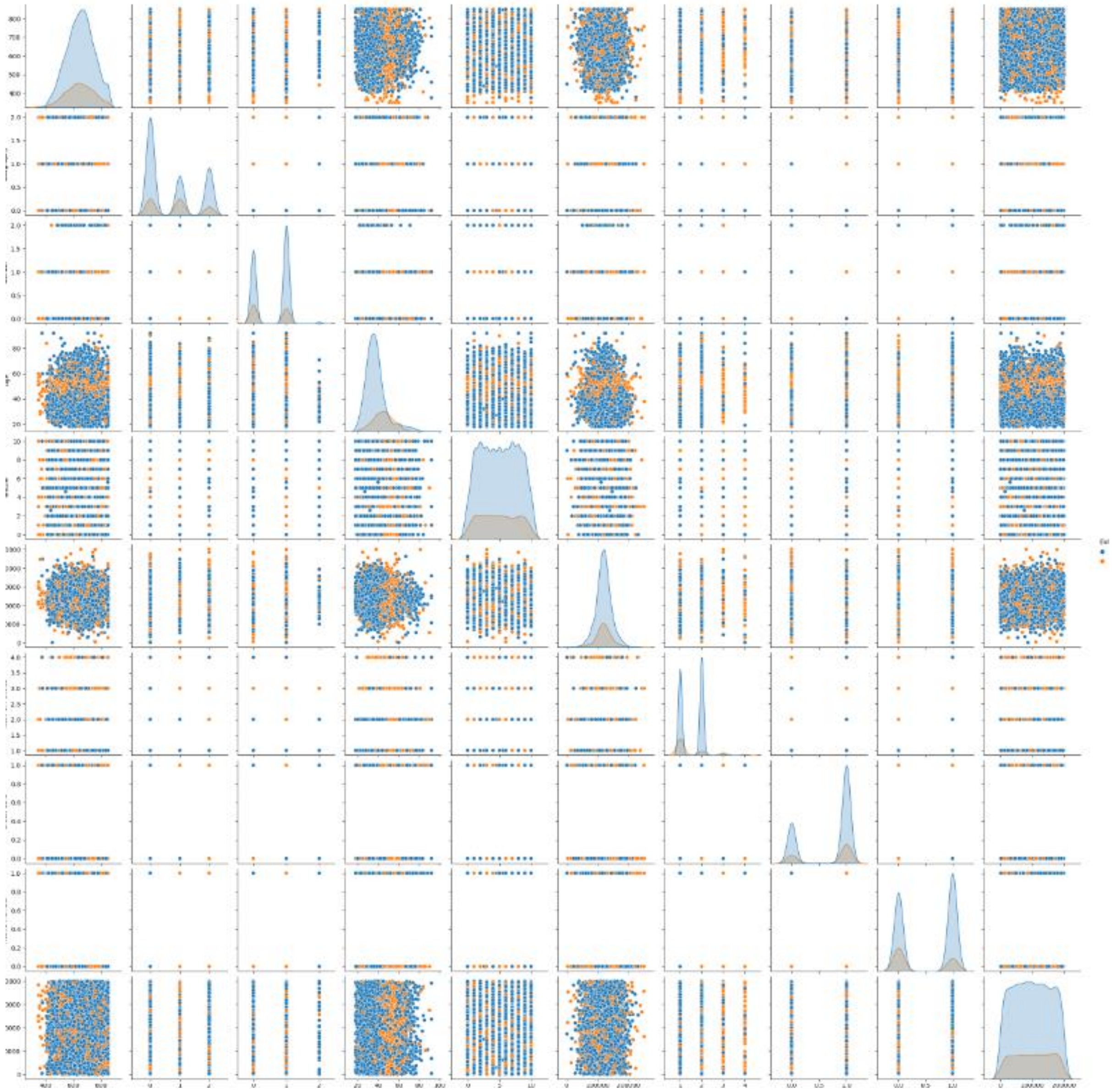
Exploratory Data Analysis was conducted to identify patterns and relationships within the data. Histograms were generated to visualize the distributions of credit scores and salaries, providing insights into customer demographics and financial health.

Additionally, categorical variables were examined through pie charts to understand the gender distribution among customers. These visualizations aided in identifying relevant features for the predictive model.




Histogram of Salary
1e−6
5
4
3
2
1
0
Count of Salary
0
50000
100000
150000
200000
Salary

```
<seaborn.axisgrid.PairGrid at 0x2873a75db90>
```



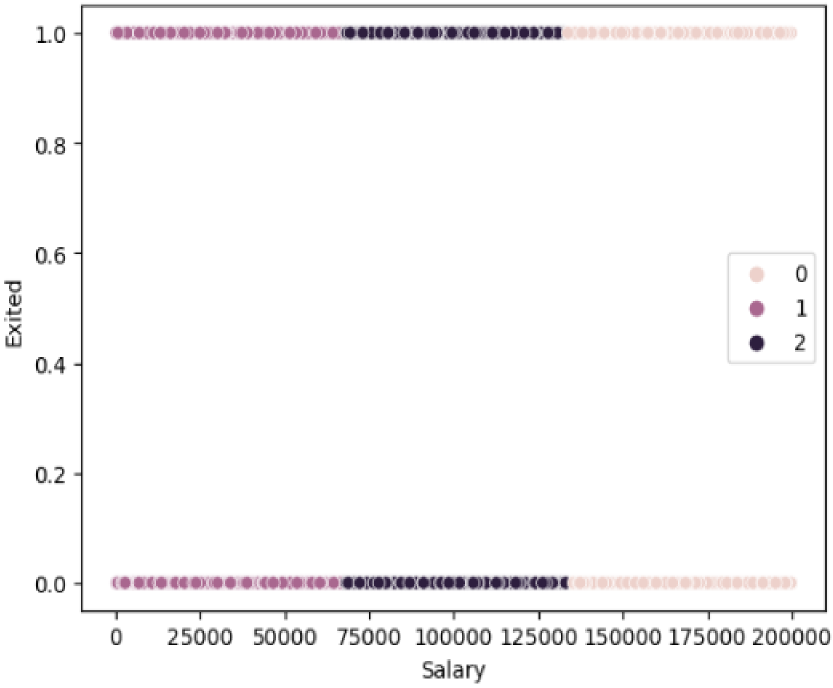
5. Model Building

A variety of machine learning algorithms were utilized to develop the customer churn prediction model, assessing each model based on accuracy, precision, recall, and F1 score. This comprehensive evaluation ensured the identification of the most effective algorithms.

Stratified K-fold cross-validation was implemented to enhance the robustness of the models, allowing for multiple training and testing iterations to reduce the risk of overfitting and improve generalizability.

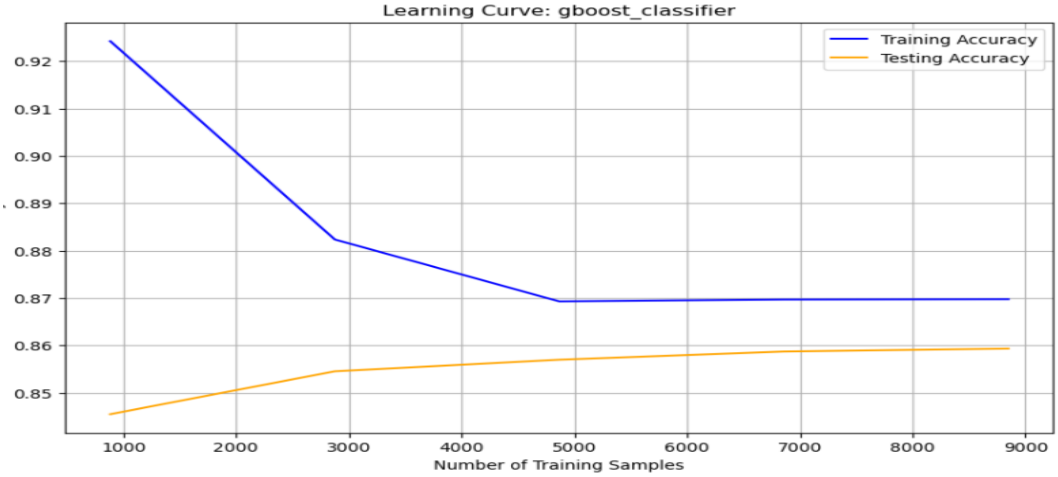
6. Clustering Analysis

K-means clustering was applied to explore customer segmentation further. This analysis identified distinct customer groups based on behaviors and demographics, informing targeted strategies for reducing churn. Visualization of these clusters through scatter plots provided additional insights into customer profiles.



7. Learning Curve Generation

Learning curves were generated for the built models, illustrating how performance varied with the size of the training dataset. These visualizations provided insights into model performance and informed decisions regarding the need for additional data.



8. Conclusion

The development of the customer churn prediction model involved comprehensive data preprocessing, feature exploration through EDA, and rigorous testing of various machine learning models. The insights derived from this analysis contribute to understanding customer behavior and developing strategies for enhancing customer retention, ultimately benefiting business operations.

Loan Eligibility Prediction

Prepared by: Sevantkumar S Huggi

Github - : https://github.com/Sevantkumar/Loan-Prediction

---

1. Introduction

This project focuses on building an automated system to predict loan approvals based on key factors such as applicant income, credit history, and loan amount. Traditionally, loan approval has been a manual process, but by using data and machine learning, we can make the process faster, more consistent, and efficient. This helps banks reduce the risk of loan defaults, improve decision-making, and provide a better customer experience by speeding up the loan approval process. The project aims to benefit both lenders and borrowers by making loan evaluations more accurate and scalable.

---

2. Analysis of the Code

2.1 Dataset

The dataset used in this project contains information about loan applicants and their respective loan approval status. It includes a variety of demographic, financial, and categorical features that help in predicting whether a loan application will be approved or not. Below are the columns in the dataset:

- **Loan-ID:** Unique identifier for each loan application.
- **Gender:** The gender of the applicant (Male/Female).
- **Married:** Marital status of the applicant (Yes/No).
- **Dependents:** Number of dependents the applicant has.
- **Education:** Educational qualification of the applicant (Graduate/Not Graduate).
- **Self-Employed:** Employment status of the applicant (Self-employed or not).
- **ApplicantIncome:** Monthly income of the applicant.
- **CoapplicantIncome:** Monthly income of the coapplicant (if any).
- **LoanAmount:** Total loan amount requested by the applicant.
- **Loan-Amount-Term:** The term of the loan (in months).
- **Credit-History:** The credit history of the applicant (1 for good, 0 for bad).
- **Property-Area:** The location of the applicant's property (Urban/Semiurban/Rural).
- **Loan-Status:** The target variable indicating whether the loan was approved (Y) or not (N).

2.2 Libraries

The following libraries were utilized in this project:

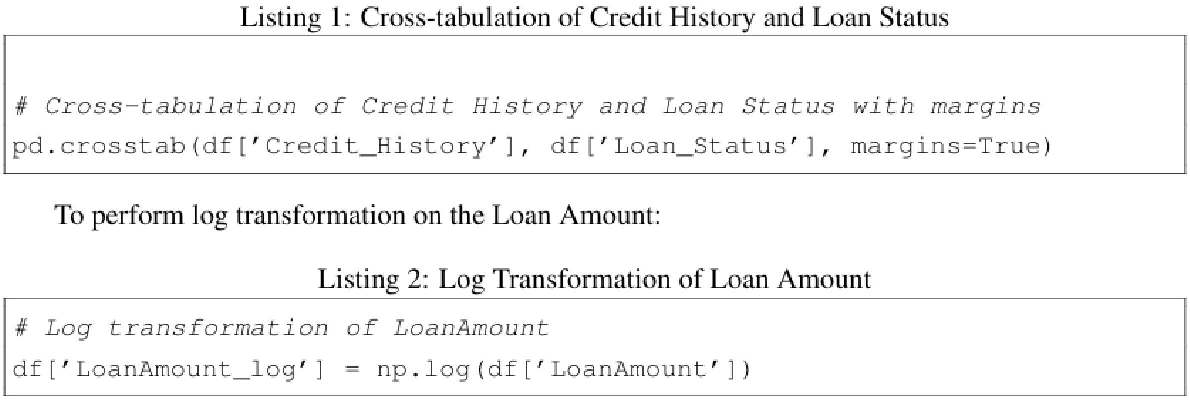
- **pandas**
- **numpy**
- **matplotlib.pyplot**
- **sklearn**

2.3 Computing Variables

The computations aimed to provide better insights and normalize the data for more efficient processing.

Log Transformation of Loan Amount

Cross-tabulation of Credit History and Loan Status was performed, along with a log transformation of the loan amount to improve the data distribution.



Listing 1: Cross-tabulation of Credit History and Loan Status

```
# Cross-tabulation of Credit History and Loan Status with margins
pd.crosstab(df['Credit_History'], df['Loan_Status'], margins=True)
```

To perform log transformation on the Loan Amount:

Listing 2: Log Transformation of Loan Amount

```
# Log transformation of LoanAmount
df['LoanAmount_log'] = np.log(df['LoanAmount'])
```

2.4 Data Preprocessing

This involves handling missing values, encoding categorical variables, and normalizing the data to ensure optimal model performance. The following steps were performed:

1. **Handling Missing Values:**
   - Categorical features had their missing values filled using the mode.
   - Numerical features were filled with the mean value of their respective columns.
2. **Feature Engineering:**
   - New features, such as **Total Income** (sum of ApplicantIncome and CoapplicantIncome), were created.
   - Logarithmic transformations were applied to relevant features to reduce skewness.
3. **Encoding Categorical Variables:**

- Categorical variables were encoded using LabelEncoder to convert text labels into numeric values.

2.5 Naive Bayes Algorithm

The Naive Bayes algorithm was employed to predict whether a loan application would be approved or not. This classification algorithm operates on the principle of calculating probabilities based on input data and assumes that all features are independent.

2.6 Visualization

Data visualization was conducted using the Matplotlib library. Boxplots were created to identify outliers in applicant income and loan amounts, while histograms displayed the distribution of income and loan amounts. These visualizations helped reveal key data patterns and informed preprocessing steps.

3. Conclusion

This project implements a Naive Bayes classifier to predict loan approval status based on various financial and demographic factors. The process included:

- Data cleaning and preprocessing.
- Feature engineering.
- Model training and evaluation.
- Applying the model to a test dataset for predictions.

The approach can be extended with other algorithms and techniques to further improve prediction accuracy.

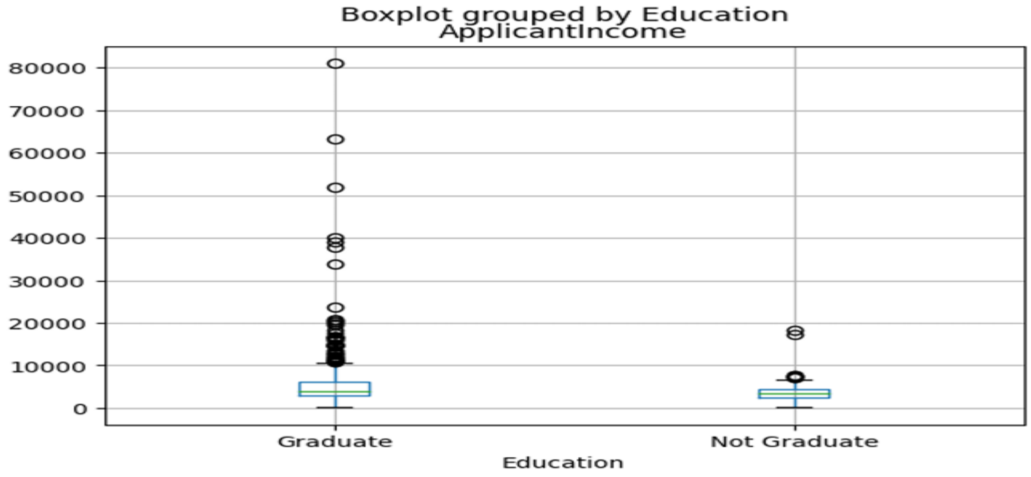


Figure 1: Box plot

Bank Customer Segmentation Project

Name : Saloni Rai
GitHub Link: Saloni-Rai19/banking (github.com)

1. Introduction

This project focuses on customer segmentation in the banking sector, utilizing a dataset that includes various attributes related to customers' demographic and financial information. The goal is to analyze how customers can be segmented based on their likelihood of loan approvals and financial behaviors. This understanding aids financial institutions in enhancing their offerings and targeting the right audience.

The dataset includes the following key features:

Feature	**Description**
Age	The age of the client.
Job	The type of job the client has (e.g., admin, technician).
Marital	The marital status of the client (e.g., single, married).
Education	Highest level of education attained (e.g., primary, tertiary).
Default	Whether the client has credit in default (yes or no).
Balance	Average yearly balance in the client’s bank account.
Housing	Whether the client has a housing loan (yes or no).
Loan	Whether the client has a personal loan (yes or no).
Contact	Type of communication contact (e.g., cellular, telephone).
Day	Last contact day of the month.
Month	Last contact month of the year.
Duration	Duration of the last contact (in seconds).
Campaign	Number of contacts performed during this campaign.
Pdays	Days since last contact from a previous campaign.
Previous	Number of contacts before this campaign.

Feature	Description
Poutcome	Outcome of the previous marketing campaign (e.g., success).
Deposit	Whether the client subscribed to a term deposit (yes or no).

---

2. Exploratory Data Analysis (EDA)

During the EDA phase, several key insights were drawn:

- **Loan Behavior by Marital Status**: Married individuals are more likely to take out loans, while divorced clients are the least likely.
- **Education Impact**: Clients with secondary education tend to take more loans than those with tertiary education.
- **Defaulters vs. Non-Defaulters**: The dataset is skewed toward non-defaulters, with 10,994 non-defaulters compared to only 168 defaulters, suggesting potential model bias.
- **Personal Loans**: A majority of clients do not take out personal loans.
- **Term Deposits**: Most clients do not subscribe to term deposits, although longer contact durations correlate with increased deposits.

Due to high occurrences of unknown values in the **poutcome** column, this feature was removed to enhance data accuracy.

---

3. Feature Extraction

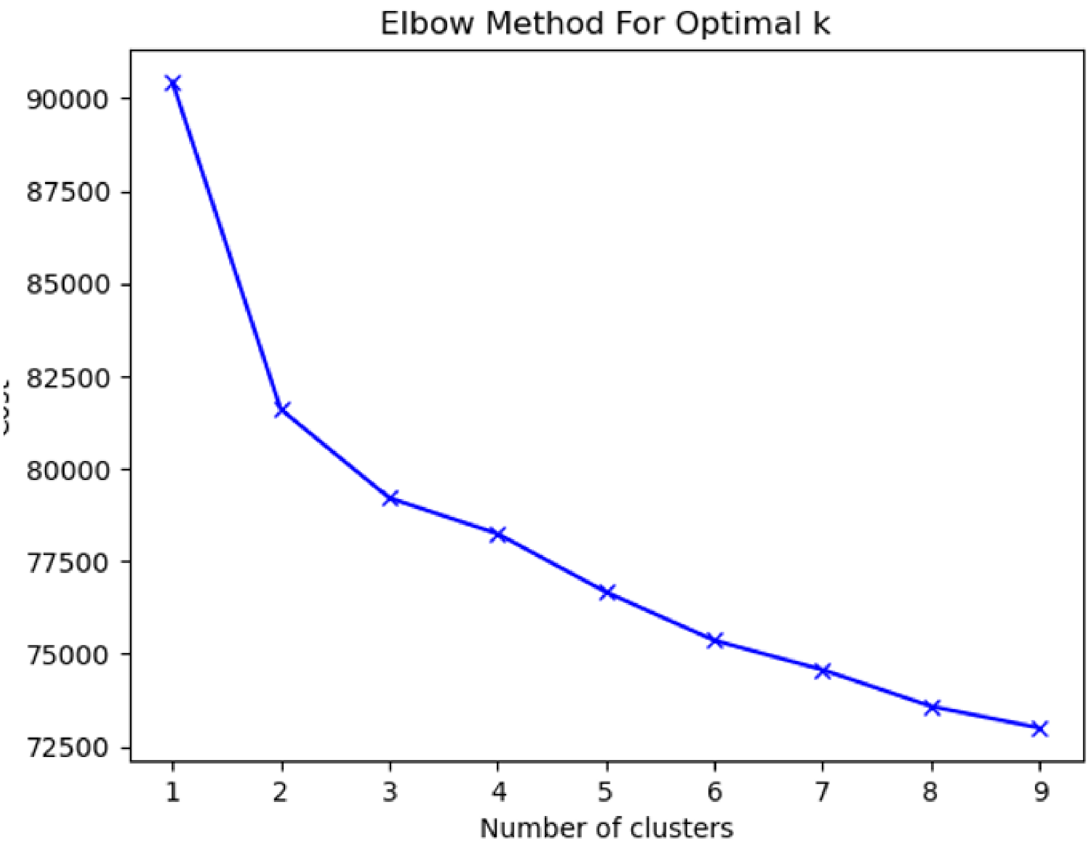
Two techniques were employed for feature extraction: **kModes** and **Multiple Correspondence Analysis (MCA)**.

3.1. kModes Algorithm

The **kModes** algorithm is tailored for clustering categorical data. The following steps were undertaken:

1. **Initialization**: Randomly select initial modes (centroids) from the dataset.
2. **Assignment**: Assign each data point to the closest mode based on a dissimilarity measure.
3. **Update**: Update the modes of the clusters by finding the most frequent values for each attribute.
4. **Iteration**: Repeat the assignment and update steps until the modes stabilize.

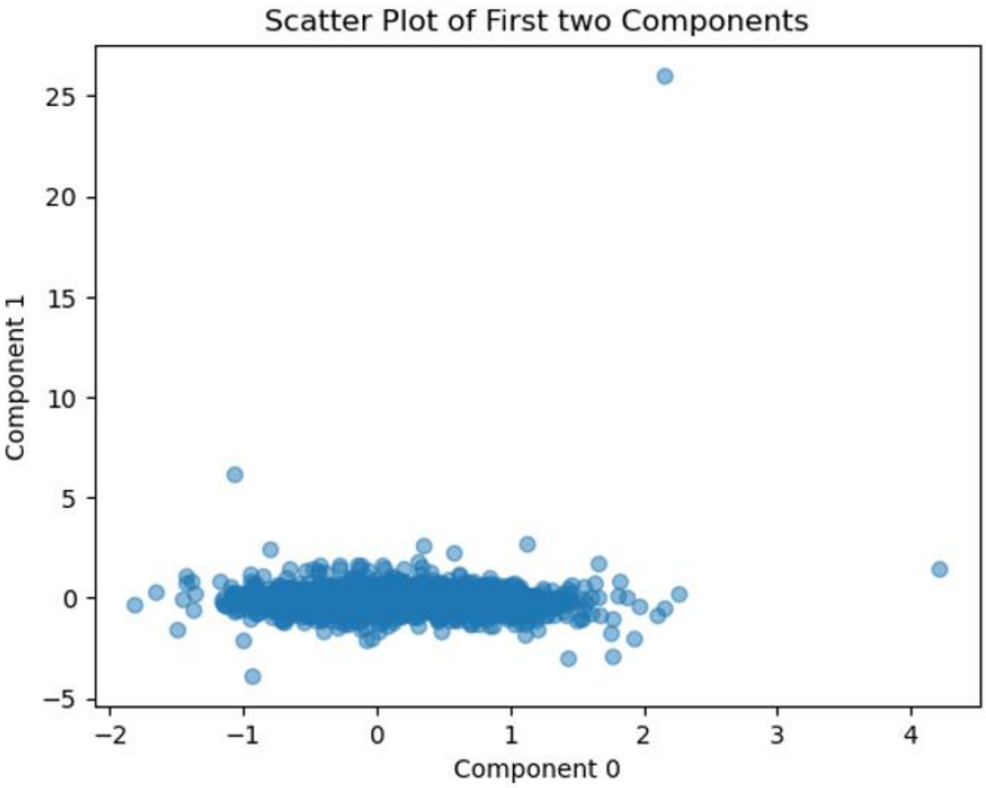
Using the Elbow Method, the optimal number of clusters was determined to be three. Below are the key characteristics of each cluster:

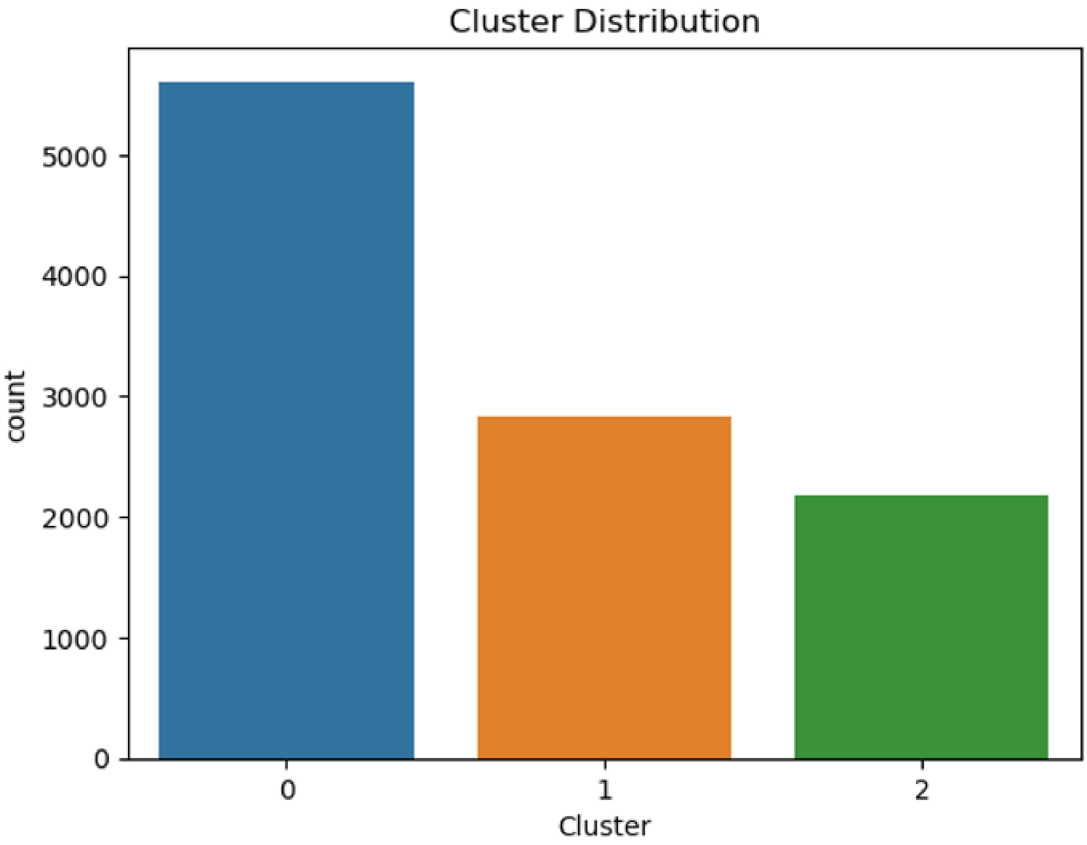


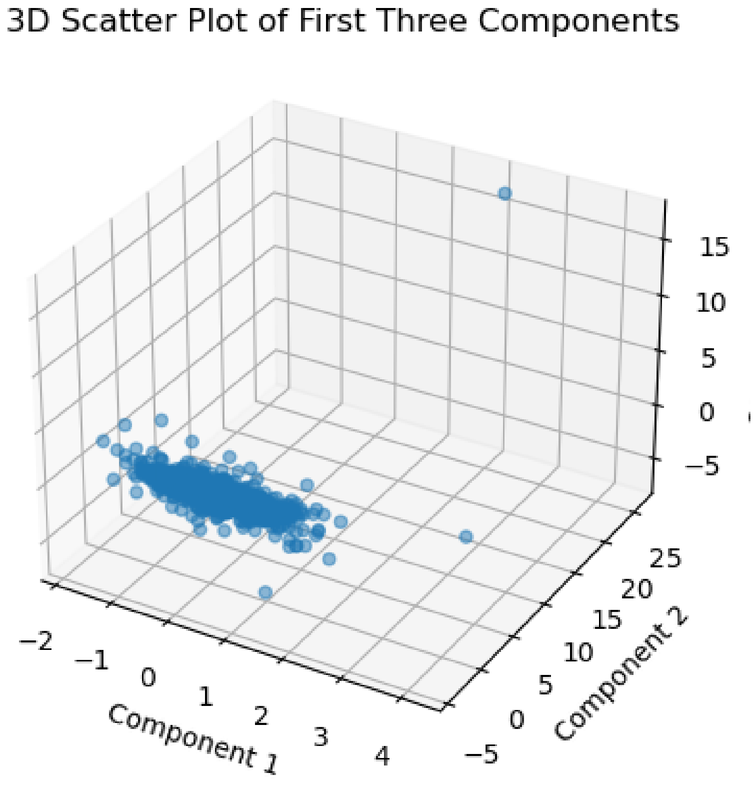
Feature	Cluster 0	Cluster 1	Cluster 2
Age	32	34	30
Job	Blue-collar	Management	Management
Marital	Married	Married	Single
Education	Secondary	Tertiary	Tertiary
Default	No	No	No
Balance	0	0	0
Housing	Yes	No	No
Loan	No	No	No
Contact	Cellular	Cellular	Cellular
Deposit	No	No	Yes

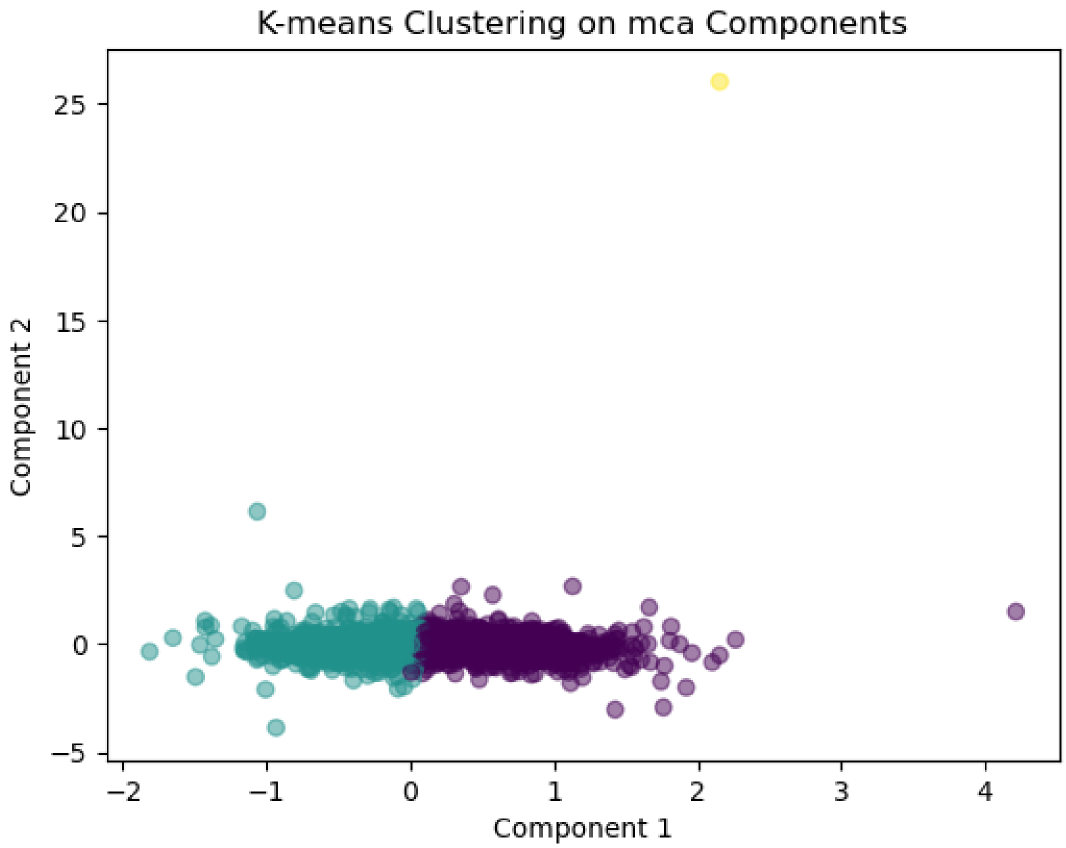
3.2. MCA Analysis

Multiple Correspondence Analysis (MCA) was applied to explore relationships between multiple categorical variables, resulting in scatter plots that visualized the clustering of clients based on their attributes.




Cluster Distribution
count
5000
4000
3000
2000
1000
0
0
1
2
Cluster


3D Scatter Plot of First Three Components
15
10
5
0
−5
25
20
15
10
5
0
−5
Component 2
−2
−1
0
1
2
3
4
Component 1



4. Conclusion

From the analysis, the following conclusions can be drawn:

- **Marital Status**: Married clients are more likely to take housing loans, while single clients tend to avoid both personal and housing loans.
- **Term Deposits**: Single clients are more likely to subscribe to term deposits, while married clients do not typically do so.
- **Loan Duration**: Clients with higher balances prefer longer loan durations.

This segmentation allows banks to tailor their products to specific client groups, enhancing marketing and customer service strategies.

Bank Data Analysis for Fraud Detection

By: Greeshma Haridas

Github - https://github.com/GreeshmaHarids/Final_Project_FeynnLab.git

---

Abstract

The banking industry faces significant challenges related to fraud, especially with credit card transactions. To combat this issue, we propose a solution leveraging machine learning models for real-time fraud detection. This report details the development of a fraud detection prototype based on a large dataset of credit card transactions. The goal is to enhance the detection process, minimize financial losses, and improve customer trust in banking institutions.

---

1. Problem Statement

Fraudulent transactions cost banks millions each year and harm customer trust. The primary objective of this project is to identify fraudulent transactions from real-time banking data, specifically focusing on credit card fraud. By developing a machine learning-based solution, we aim to identify patterns and anomalies that signal potential fraud, thereby reducing the risk of unauthorized transactions.

---

2. Market/Customer/Business Need Assessment

The banking industry heavily relies on customer trust. Fraudulent activities, such as unauthorized transactions, lead to substantial financial losses and damage banks' reputations. Implementing a machine learning-based fraud detection system can significantly mitigate these risks. This solution will provide:

- Real-time detection of suspicious activities
- Automated alerts to banking staff
- Increased security and customer confidence

---

3. Target Specifications and Characterization

- **Real-Time Fraud Detection**: Utilize machine learning models to process large transaction datasets and flag potentially fraudulent activities within seconds.
- **Operational Efficiency**: Automating fraud detection reduces the burden on bank employees and increases detection accuracy.
- **Customer Satisfaction**: Faster and more accurate fraud detection enhances customer security, improving retention rates.

4. Data Overview and Preprocessing

The dataset contains various features, including transaction amounts, customer demographics, and transaction timestamps. Initial steps included:

- Handling missing values and data types
- Removing duplicates
- Converting categorical variables to appropriate data types
- Feature engineering to derive insights from transaction timestamps and customer ages

Key Findings from Exploratory Data Analysis (EDA):

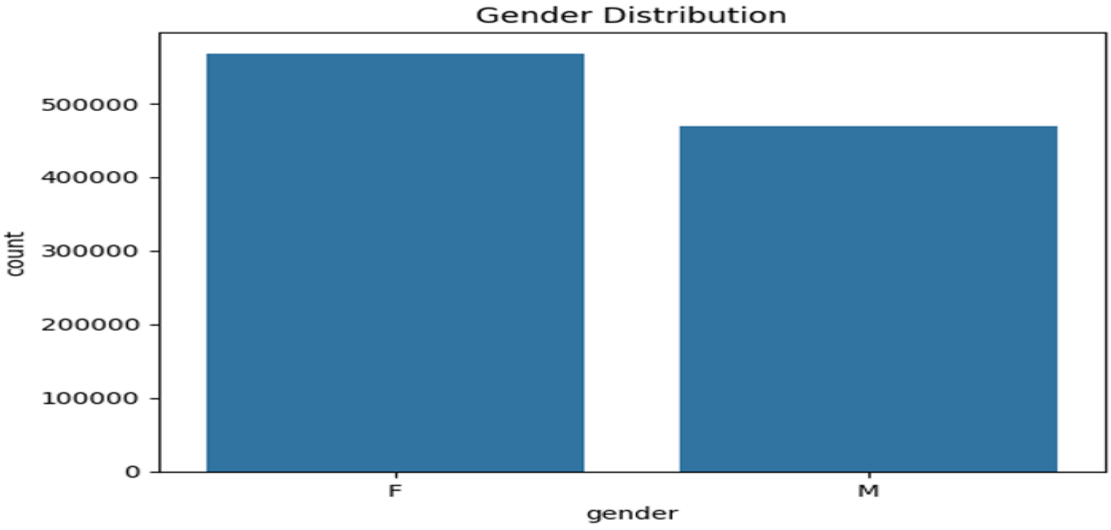
- **Transaction Amount Distribution**: Transaction amounts exhibit a skewed distribution, which was addressed through log transformation.
- **Time-Based Patterns**: Transaction patterns vary by time of day and day of the week, with peak activity observed on weekends and during late afternoons/evenings.

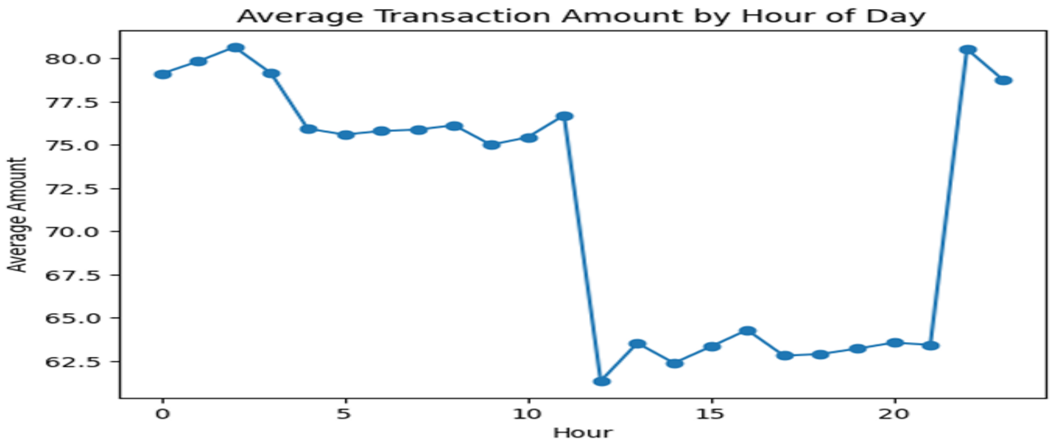
Activity Insights:

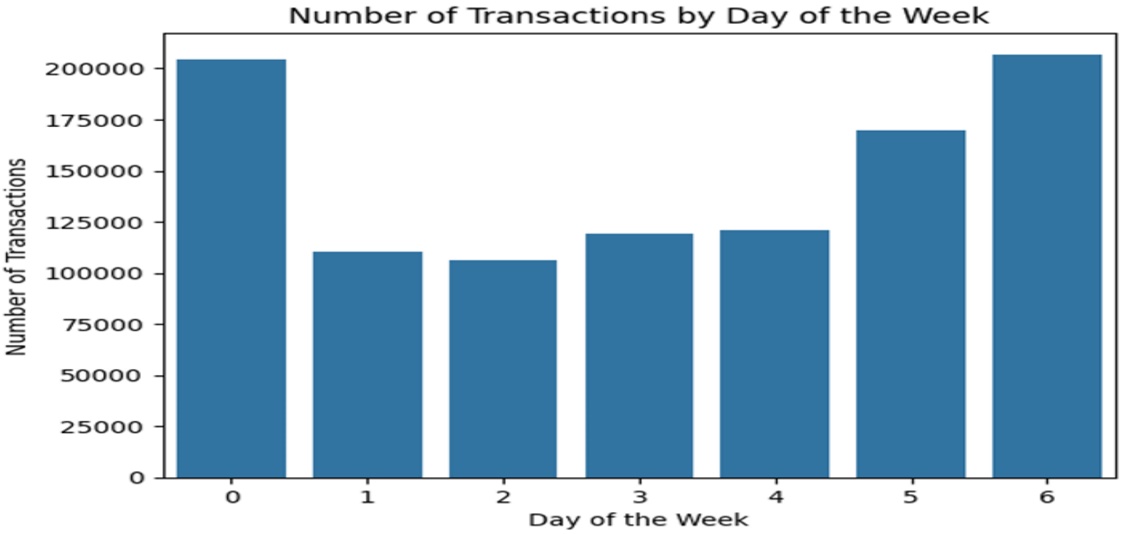
- High average transaction amounts occur at midnight and 10 PM, with a decline after 9 AM. The lowest amounts occur between 10 AM and 3 PM, suggesting a midday lull, followed by increased spending from 4 PM onwards.
- Transaction volumes peak on Sundays and Saturdays, with over 200,000 transactions on both days. Weekdays show fewer transactions, with Fridays experiencing a rise.

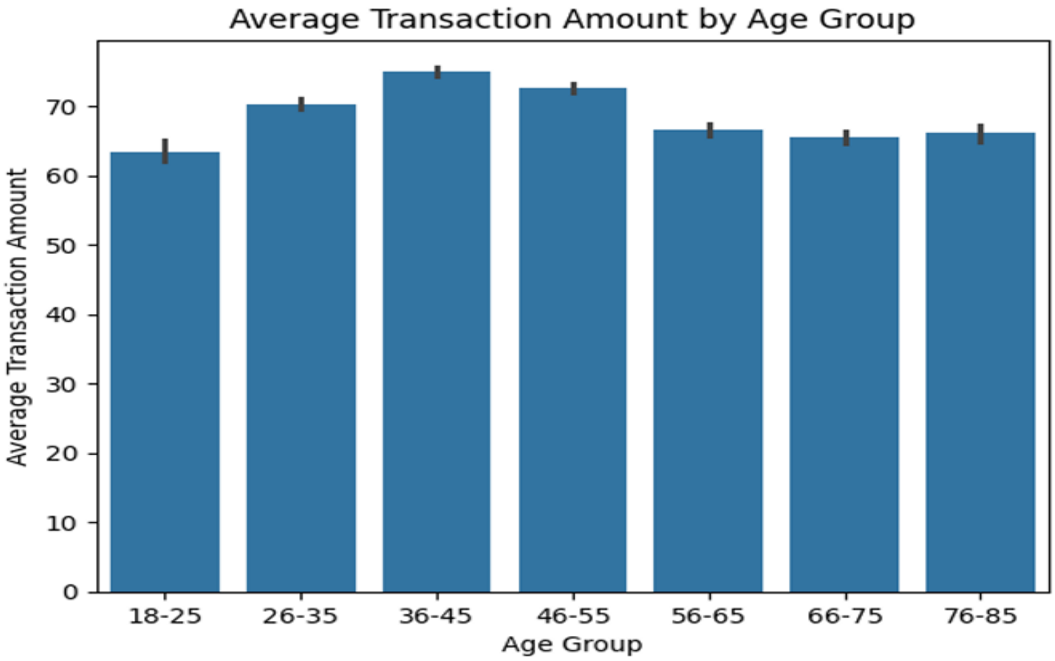
Age-Based Insights:

- Individuals aged 36-45 likely represent peak earning years, enabling higher spending.
- The 46-55 age group also spends considerably, while the 26-35 age group may have lower average transactions due to different financial priorities.




Average Transaction Amount by Hour of Day
Average Amount
80.0
77.5
75.0
72.5
70.0
67.5
65.0
62.5
0
5
10
15
20
Hour

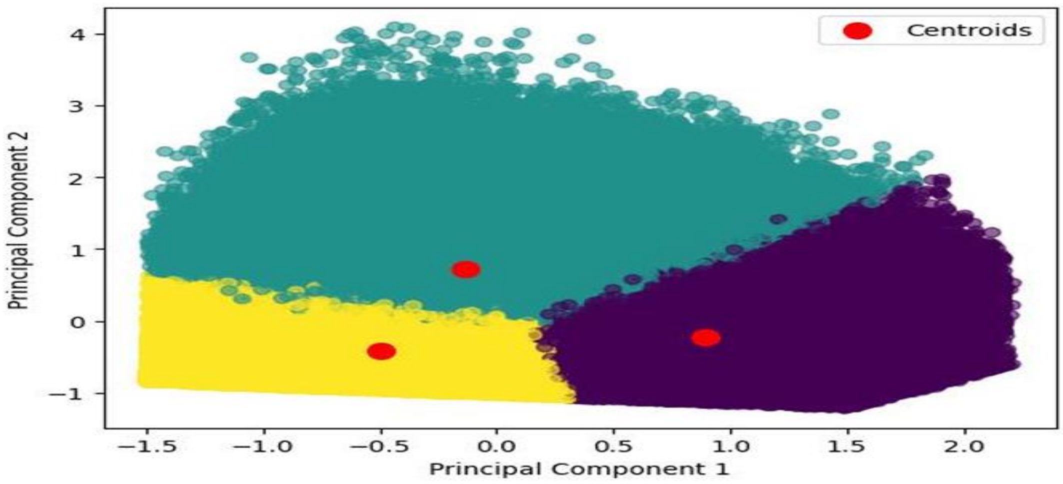

Number of Transactions by Day of the Week
Number of Transactions
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
4
5
6
Day of the Week


Average Transaction Amount by Age Group
Average Transaction Amount
70
60
50
40
30
20
10
0
18-25
26-35
36-45
46-55
56-65
66-75
76-85
Age Group

5. Feature Engineering and Selection

Significant features derived include:

- Customer age, calculated from the date of birth
- Transaction hour, day of the week, and month
- Log-transformed transaction amounts to address skewness
- Clustering analysis revealed distinct spending behaviors across demographics. KMeans clustering identified patterns based on age and transaction amounts.



6. Model Development

The **XGBoost** algorithm was selected for its high performance in classification tasks. The model training process involved:

- Splitting the data into training and testing sets
- Training the model on selected features

Key Performance Metrics:

- **Accuracy**: 1.00
- **ROC AUC Score**: 0.98

Feature importance analysis indicated that transaction amount, customer age, and transaction hour are critical predictors of fraudulent activity.

7. Results and Insights

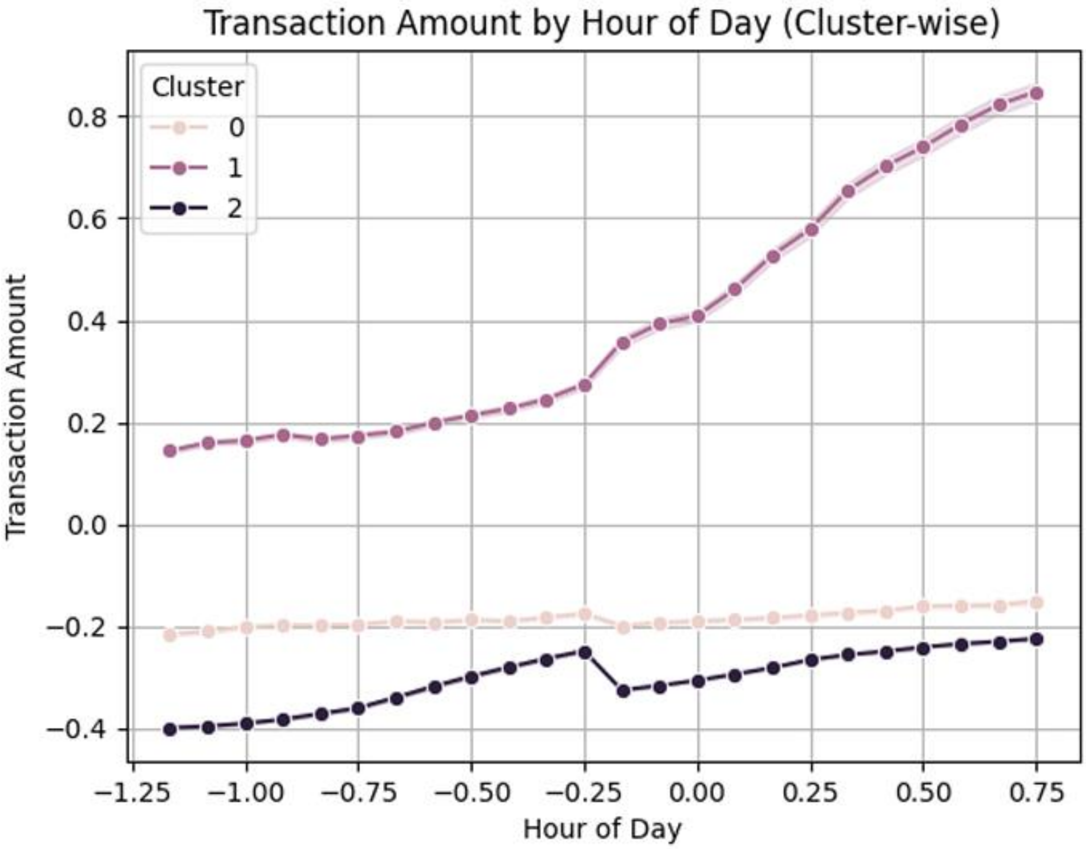
- The confusion matrix and classification report provide detailed insights into the model's predictive performance.
- Cluster analysis reveals varying spending behaviors:
    - **Cluster 0**: Older individuals with conservative spending habits
    - **Cluster 1**: Younger demographics with higher transaction amounts
    - **Cluster 2**: Mid-age group with moderate spending patterns

Spending Behavior:

Cluster 1 exhibits higher transaction amounts, indicating a demographic with greater disposable income. Clusters 0 and 2 show lower amounts, suggesting more conservative spending habits.

Age Distribution:

Cluster 1 contains a younger demographic compared to Clusters 0 and 2, which may include older individuals with different spending priorities.



Default Customer Prediction model

Name: Hemanth Kumar
Github link - https://github.com/HemanthKumarBodduboina/Default-prediction-model-.git

---

Problem Statement:

Financial institutions must accurately predict loan defaulters to mitigate risks and optimize their lending strategies. Traditional methods often overlook intricate borrower behaviors and relationships between financial metrics. To address this, I developed a model using **XGBoost**, a powerful machine learning algorithm capable of handling complex classification tasks, with the aim of improving defaulter prediction.

Objective:

The objective of this project was to build a highly effective machine learning model that could predict loan defaulters by leveraging XGBoost and advanced data analysis. The ultimate goal is to provide financial institutions with actionable insights that can lead to reduced default rates and better risk management.

Features Used:

- **LoanID:** Unique identifier for each loan.
- **Age:** Borrower's age.
- **Income:** Annual income of the borrower.
- **LoanAmount:** Amount of the loan.
- **CreditScore:** Borrower's credit score at the time of loan application.
- **MonthsEmployed:** Number of months the borrower has been employed.
- **NumCreditLines:** Number of active credit lines the borrower has.
- **InterestRate:** Interest rate applied to the loan.
- **LoanTerm:** Duration of the loan in months.
- **DTIRatio:** Debt-to-Income ratio.
- **Education:** Education level of the borrower.
- **EmploymentType:** Type of employment (Salaried, Self-employed, etc.).
- **MaritalStatus:** Marital status of the borrower.
- **HasMortgage:** Indicator of whether the borrower has a mortgage.
- **HasDependents:** Indicator of whether the borrower has dependents.

- **LoanPurpose:** Purpose of the loan (Personal, Business, Education, etc.).
- **HasCoSigner:** Whether the loan has a cosigner.
- **Default:** Target variable indicating whether the borrower defaulted (1 for default, 0 for no default).

Methodology:

1. Data Acquisition and Preprocessing:

- **Data Collection:** The dataset used for this project contained diverse borrower-related features such as **Age**, **Income**, **LoanAmount**, **CreditScore**, and others. This provided a well-rounded view of borrower financial health and creditworthiness.
- **Data Cleaning:** A crucial step in this project was ensuring the dataset's quality. Missing values were imputed using median or mode values, and outliers in features like **Income** and **LoanAmount** were handled carefully to prevent skewed results.
- **Data Transformation:** Categorical features like **MaritalStatus**, **EmploymentType**, and **LoanPurpose** were one-hot encoded to make them suitable for model training.
- **Data Splitting:** The dataset was split into **80% training** and **20% testing** sets to evaluate model generalization.

2. Feature Engineering:

- **New Feature Creation:** Derived additional features like:
  - **Loan-to-Value Ratio (LTV):** LoanAmount relative to borrower's Income.
  - **Employment Stability:** Indicating whether a borrower was employed for over 24 months.
- **Feature Interaction:** Explored feature interactions between **CreditScore** and **DTIRatio**, finding that low credit scores combined with high DTI significantly increased default risk.

3. Model Selection and Training:

- **Algorithm Choice:** I selected **XGBoost** for its superior performance with imbalanced data and its ability to model complex non-linear relationships. Its regularization mechanisms also helped prevent overfitting.
- **Hyperparameter Tuning:** Performed grid search and cross-validation to optimize:
  - **Learning Rate**
  - **Max Depth**
  - **Subsample Ratio**
  - **Number of Estimators**

4. Model Evaluation:

- **Metrics Used:**
  - **Accuracy:** Overall correctness of the predictions.
  - **Precision & Recall:** Important to balance the detection of defaulters while minimizing false positives.
  - **F1-score:** A balanced measure of precision and recall.
  - **ROC AUC:** Provided insight into the model's discriminatory power.
- **Confusion Matrix:** This helped to visualize false positives and false negatives in the predictions.

Results:

- The **XGBoost** model achieved:
  - **Accuracy:** 84.72%
  - **F1-score:** 0.835

This demonstrates a significant improvement over traditional models and confirms the model's ability to predict loan defaults effectively.

Key Insights:

- Features like **CreditScore**, **DTIRatio**, and **LoanAmount** had the greatest influence in predicting defaults. Borrowers with low credit scores and high DTI ratios were found to be at greater risk of default.
- Surprisingly, the presence of a **CoSigner** was less influential than expected, suggesting that cosigners may not always mitigate default risk.

Conclusion:

This project successfully demonstrated the effectiveness of using **XGBoost** in predicting loan defaulters. The model's ability to handle high-dimensional data and non-linear relationships proved valuable for financial institutions looking to mitigate risk and make more informed lending decisions.

Recommendations:

- **Continuous Monitoring:** Periodic retraining with new data will help maintain model accuracy as market conditions change.
- **Model Expansion:** Future iterations should explore ensemble models or deep learning techniques for improved performance.