PROJECT TITLE:

PREDICTIVE ANALYTICS AND RECOMMENDATION SYSTEMS IN BANKING

BATCH:

MDTE13 (DS-WD-T-B10)

NAME:

NANDHINI C

COURSE:

DATA SCIENCE

TABLE OF CONTENTS

1. Summary	Page 3
2. Introduction	Page 3
2.1 Problem Statement	Page 3
2.2 Objective	Page 3
2.3 Project Scope	Page 4
3. Data Preprocessing	Page 4
3.1 Data Overview	Page 4
3.2 Data Cleaning and Preprocessing	Page 4
4. Exploratory Data Analysis (EDA)	Page 5
5. Model Development	Page 10
5.1 Methodology	Page 10
6. Model Deployment (Streamlit Application)	Page 11
7. Screenshots	Page 12
8. Reason Behind Model Selection	Page 14
9. Conclusion	Page 14
10. Appendices	Page 15
11. References	Page 15

Summary:

This project aims to leverage predictive analytics and machine learning to address critical challenges in banking, including loan default prediction, customer segmentation, and product recommendations. A machine learning-based solution was developed, and the models were integrated into an interactive application. This approach improves customer experience, optimizes financial risk management, and enhances targeted marketing.

Introduction:

2.1 Problem Statement

The banking sector faces significant challenges, such as predicting loan defaults, understanding customer behavior, and providing personalized product recommendations. Traditional methods often fail to effectively manage these tasks due to the vast and complex nature of banking data. This project aims to address these challenges using machine learning and predictive analytics techniques.

2.2 Objective:

To build predictive models and deploy them to:

1. Predict loan defaults using supervised learning techniques.
2. Segment customers based on transaction behavior.
3. Recommend banking products to customers using recommendation systems.

2.3 Project Scope

1. Utilize historical banking datasets for analysis and model development.

2. Build and evaluate supervised, unsupervised, and recommendation models.

3. Deploy the models into an application for interactive use.

3. Data Preprocessing:

3.1 Data Overview

The datasets include historical customer information, transactional records, and product interactions. Key features include:

- Loan Default Prediction: Customer demographics, credit scores, loan details.

- Customer Segmentation: Transaction frequency, amounts, and types.

- Product Recommendations: Interaction types, product details.

3.2 Data Cleaning and Preprocessing

Categorical Features:

- Encoded categorical variables using Label Encoding and One-Hot Encoding.

- Checked for and corrected inconsistencies in entries (e.g., typos).

Numerical Features:

- Scaled features using Min-Max scaling.
- Addressed outliers using IQR and Z-score methods.
- Handled missing values with mean/median imputation or new categories for categorical data.

Unrelated Columns:

- Removed irrelevant columns, such as identifiers not used in analysis (e.g., transaction IDs).

4.Exploratory Data Analysis (EDA)

Objective:

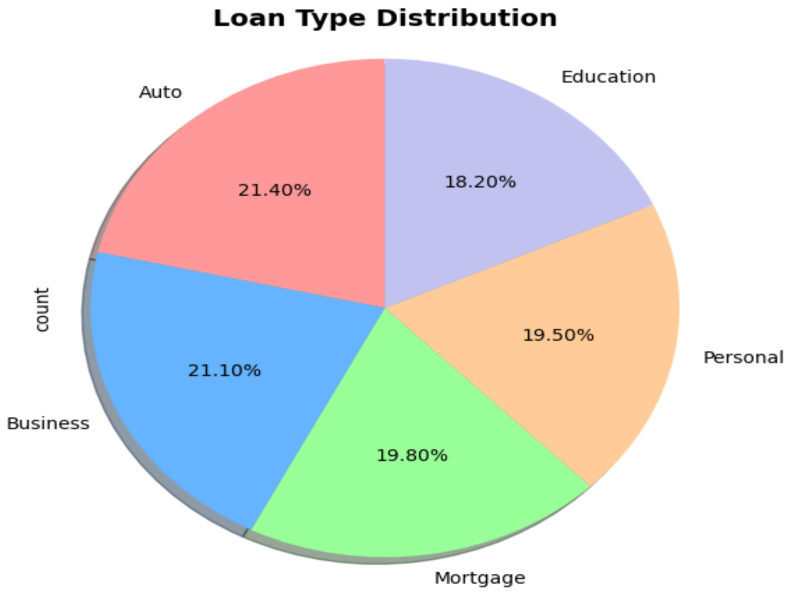
1. Understand data distributions and relationships.
2. Identify features contributing to predictions.

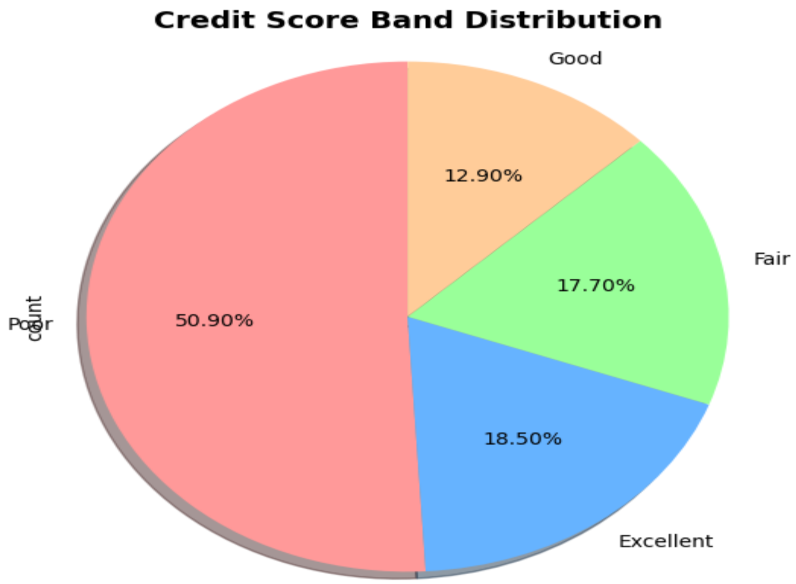
Key Insights:

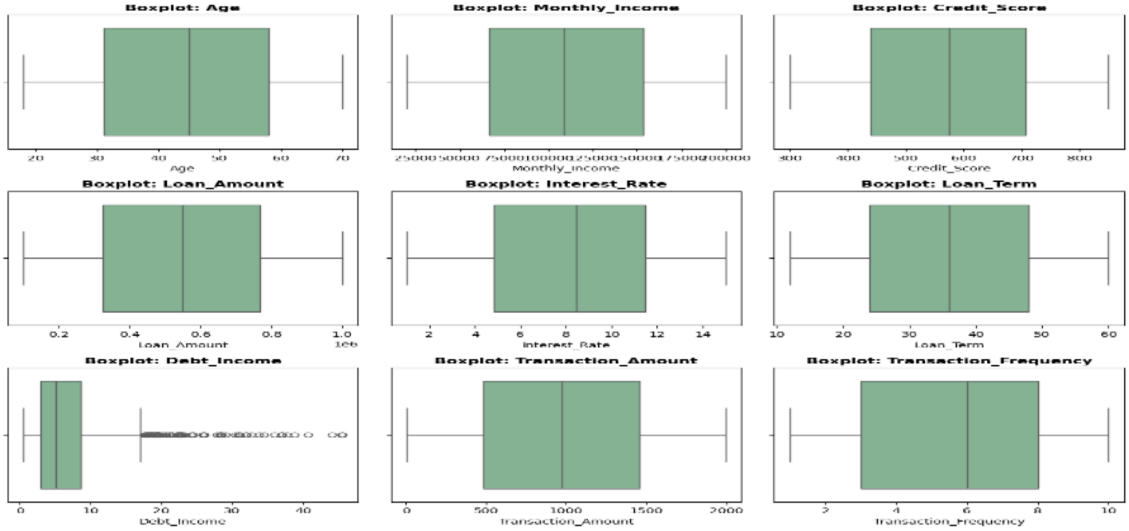
1. Customers with lower credit scores have a higher likelihood of loan defaults.
2. Transaction patterns vary significantly between customer segments.

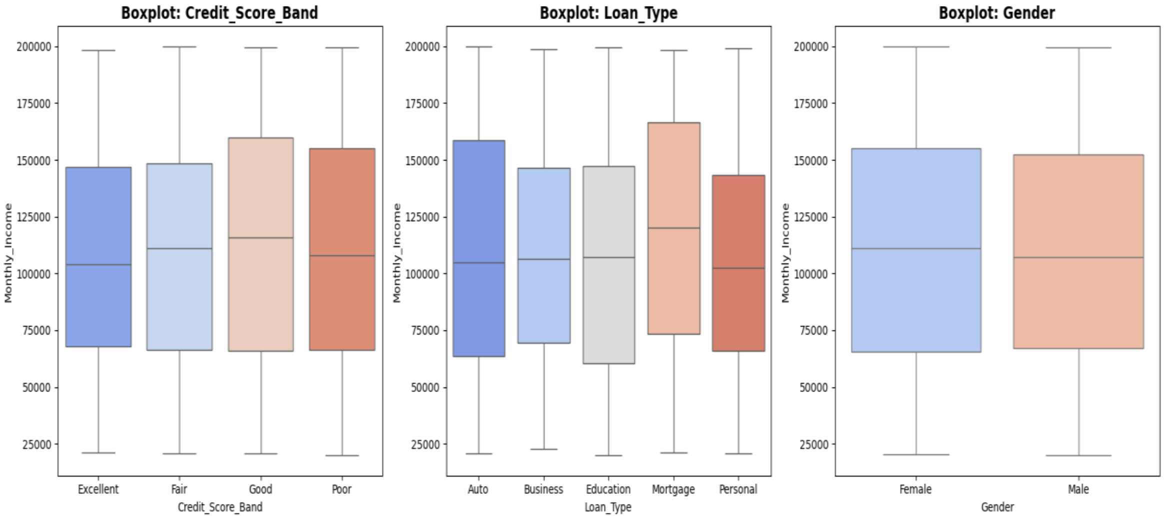
3. Certain products are more popular among specific customer demographics.

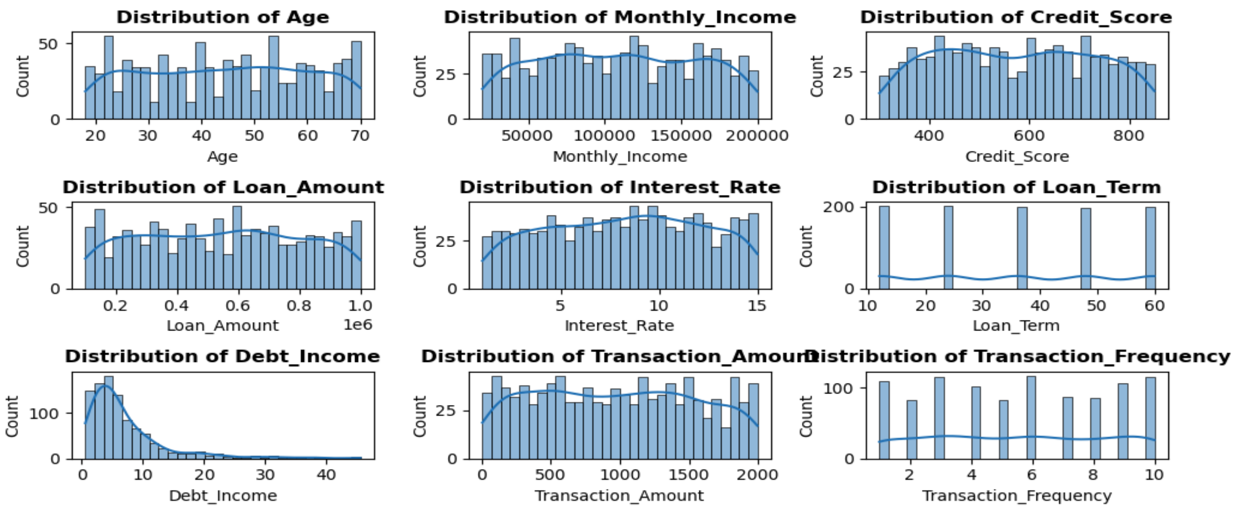
EDA PLOTS:

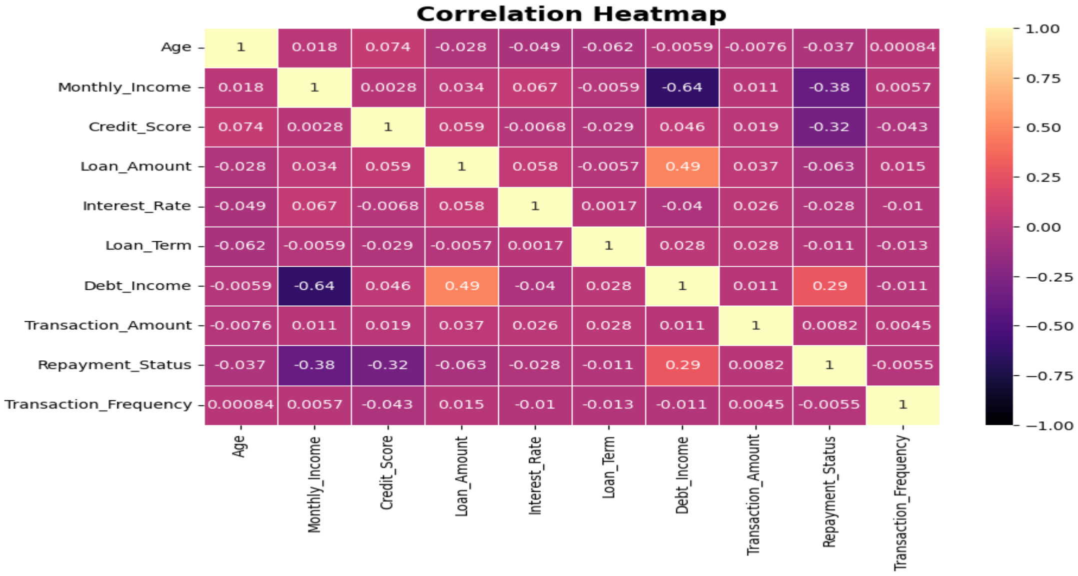


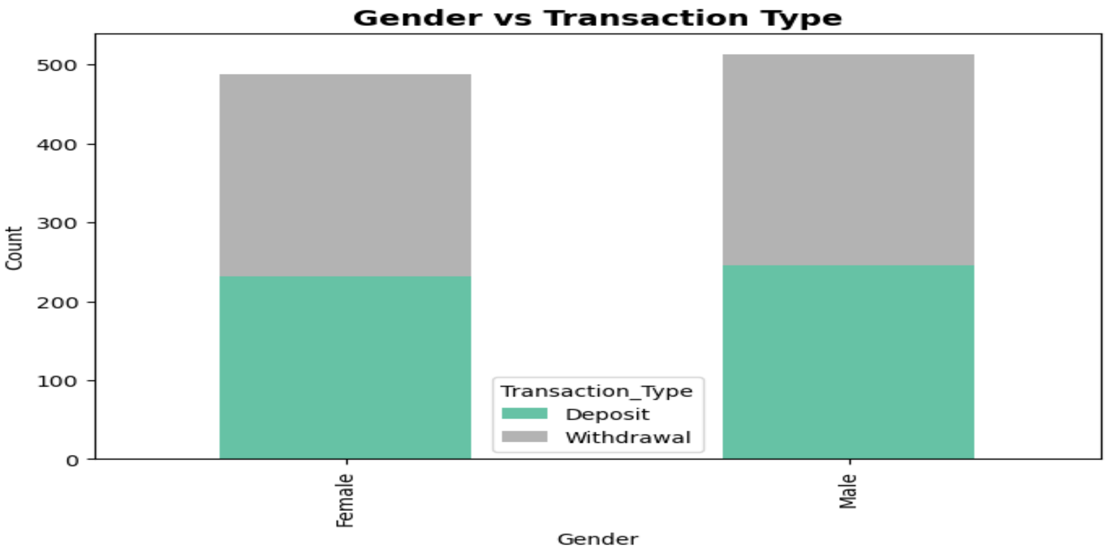


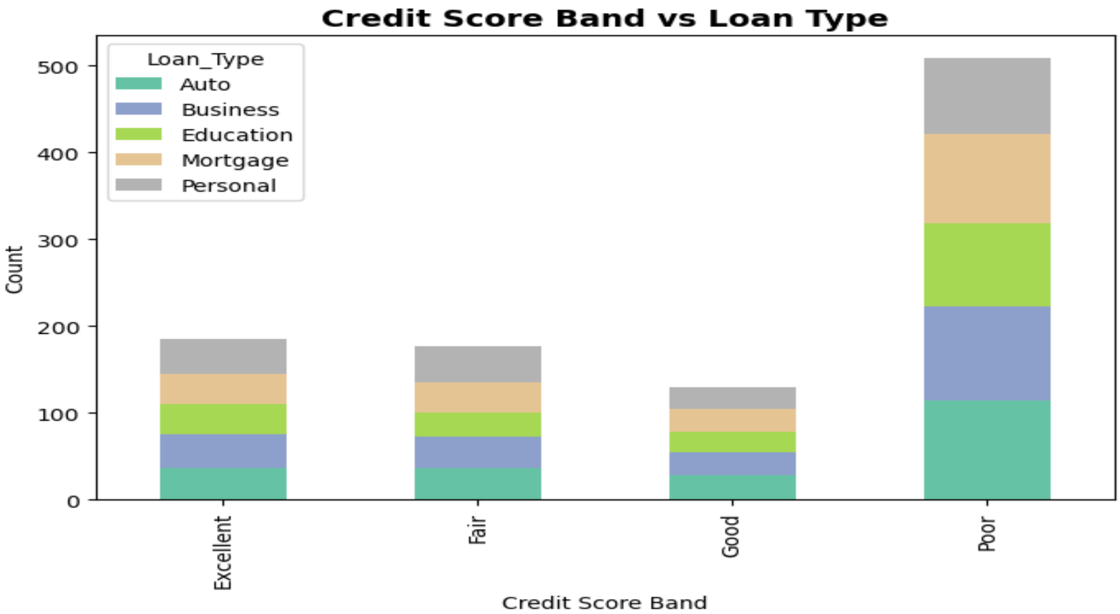

Boxplot: Age
Boxplot: Monthly_Income
Boxplot: Credit_Score
Boxplot: Loan_Amount
Boxplot: Interest_Rate
Boxplot: Loan_Term
Boxplot: Debt_Income
Boxplot: Transaction_Amount
Boxplot: Transaction_Frequency

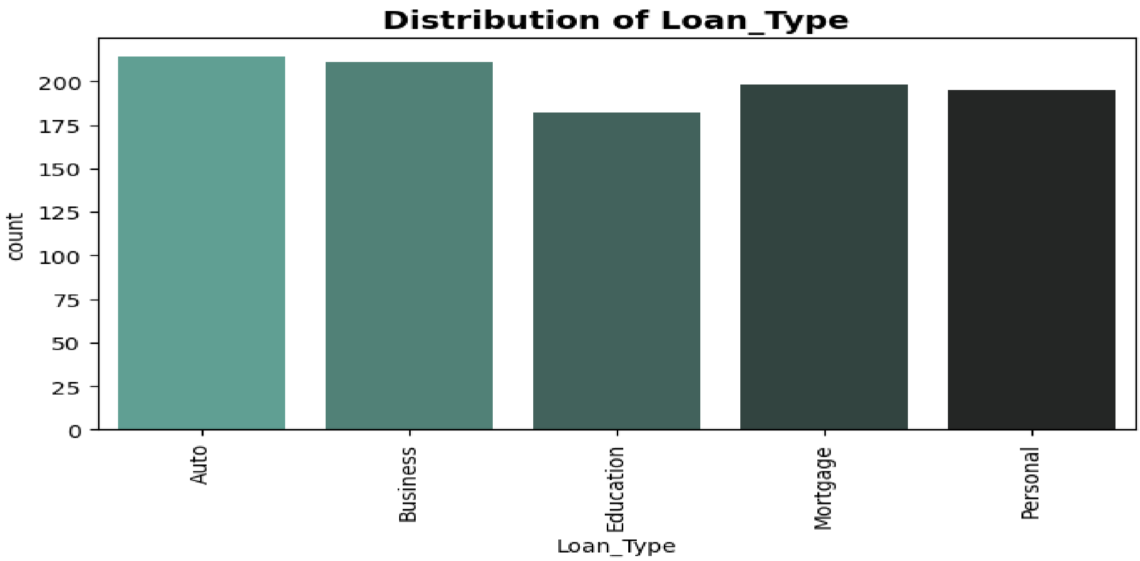

Boxplot: Credit_Score_Band
Boxplot: Loan_Type
Boxplot: Gender
Monthly_Income
Excellent
Fair
Good
Poor
Credit_Score_Band
Auto
Business
Education
Mortgage
Personal
Loan_Type
Female
Male
Gender

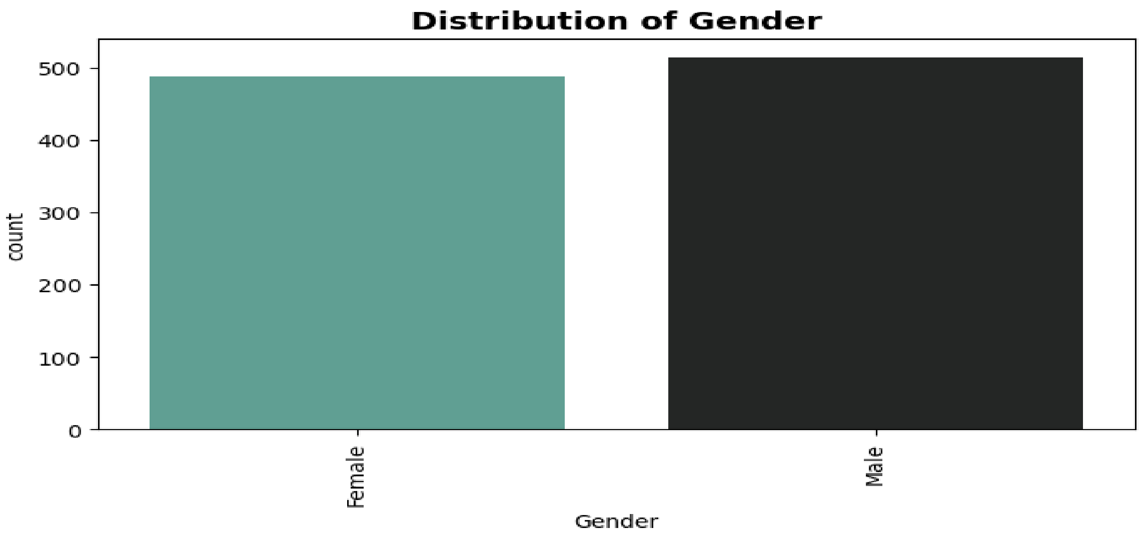

Distribution of Age
Distribution of Monthly_Income
Distribution of Credit_Score
Distribution of Loan_Amount
Distribution of Interest_Rate
Distribution of Loan_Term
Distribution of Debt_Income
Distribution of Transaction_Amount
Distribution of Transaction_Frequency
Count

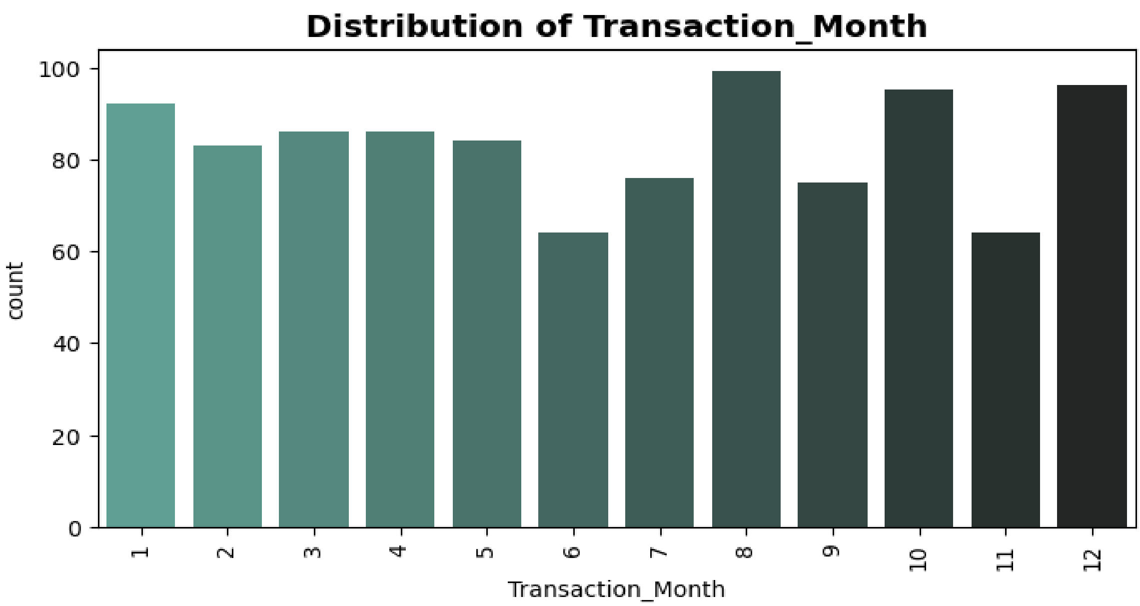

Correlation Heatmap
Age: 1, 0.018, 0.074, -0.028, -0.049, -0.062, -0.0059, -0.0076, -0.037, 0.00084
Monthly_Income: 0.018, 1, 0.0028, 0.034, 0.067, -0.0059, -0.64, 0.011, -0.38, 0.0057
Credit_Score: 0.074, 0.0028, 1, 0.059, -0.0068, -0.029, 0.046, 0.019, -0.32, -0.043
Loan_Amount: -0.028, 0.034, 0.059, 1, 0.058, -0.0057, 0.49, 0.037, -0.063, 0.015
Interest_Rate: -0.049, 0.067, -0.0068, 0.058, 1, 0.0017, -0.04, 0.026, -0.028, -0.01
Loan_Term: -0.062, -0.0059, -0.029, -0.0057, 0.0017, 1, 0.028, 0.028, -0.011, -0.013
Debt_Income: -0.0059, -0.64, 0.046, 0.49, -0.04, 0.028, 1, 0.011, 0.29, -0.011
Transaction_Amount: -0.0076, 0.011, 0.019, 0.037, 0.026, 0.028, 0.011, 1, 0.0082, 0.0045
Repayment_Status: -0.037, -0.38, -0.32, -0.063, -0.028, -0.011, 0.29, 0.0082, 1, -0.0055
Transaction_Frequency: 0.00084, 0.0057, -0.043, 0.015, -0.01, -0.013, -0.011, 0.0045, -0.0055, 1
Age, Monthly_Income, Credit_Score, Loan_Amount, Interest_Rate, Loan_Term, Debt_Income, Transaction_Amount, Repayment_Status, Transaction_Frequency
1.00, 0.75, 0.50, 0.25, 0.00, −0.25, −0.50, −0.75, −1.00


Gender vs Transaction Type
Count
0, 100, 200, 300, 400, 500
Transaction_Type
Deposit
Withdrawal
Female
Male
Gender


Credit Score Band vs Loan Type
Count
0, 100, 200, 300, 400, 500
Loan_Type
Auto
Business
Education
Mortgage
Personal
Excellent
Fair
Good
Poor
Credit Score Band


Distribution of Loan_Type
count
0
25
50
75
100
125
150
175
200
Auto
Business
Education
Mortgage
Personal
Loan_Type


Distribution of Gender
count
0
100
200
300
400
500
Female
Male
Gender


Distribution of Transaction_Month
count
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
Transaction_Month

5. Model Development

1. Splitting the data into training (80%) and testing (20%) sets.

2. Evaluating models using cross-validation and hyperparameter tuning.

5.1 Methodology

Loan Default Prediction:

- Algorithms: Logistic Regression, Random Forest, Gradient Boosting, Decision Tree.
- Data split into 70% training and 30% testing sets.
- Evaluation metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Customer Segmentation:

- Algorithms: K-Means and Hierarchical Clustering.
- Features standardized before clustering.
- Evaluation metrics: Silhouette Score and Cluster Visualization.

Recommendation System:

- Algorithms: Collaborative Filtering and Matrix Factorization.

- Evaluation metrics: Precision, Recall, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG).

Model Performance:

- Gradient Boosting provided the best accuracy for loan default prediction.
- K-Means effectively segmented customers into meaningful groups.
- Collaborative Filtering achieved high precision in product recommendations.

6. Model Deployment

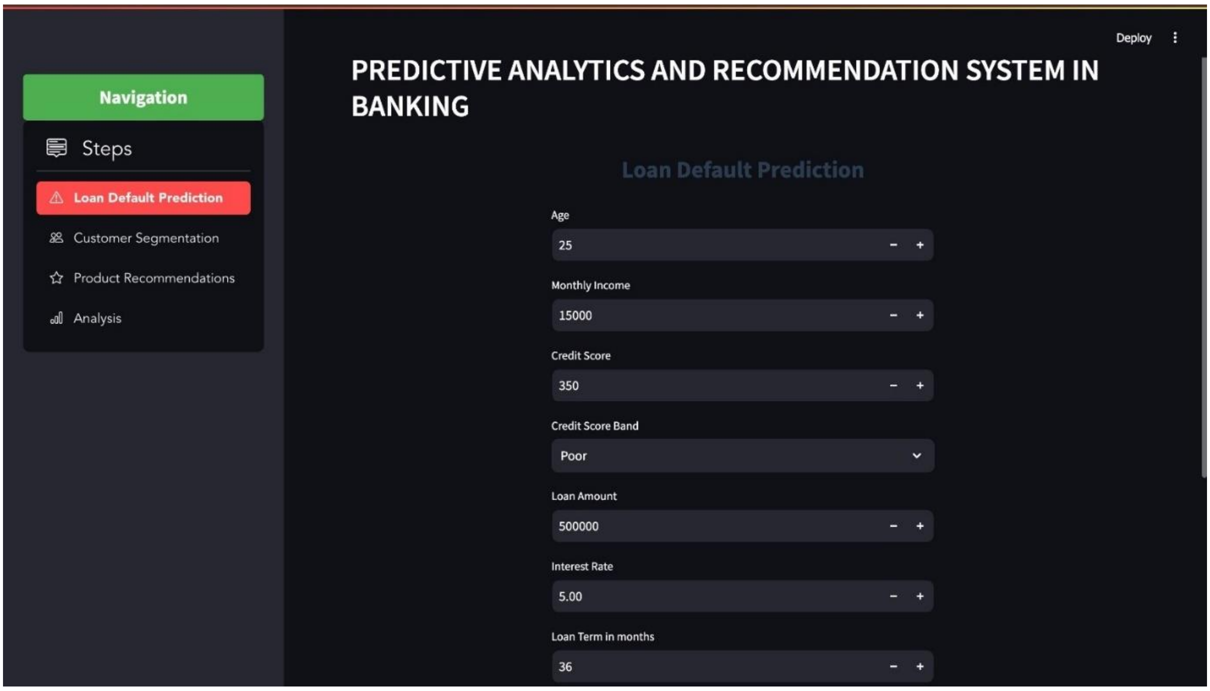
Overview of Application:

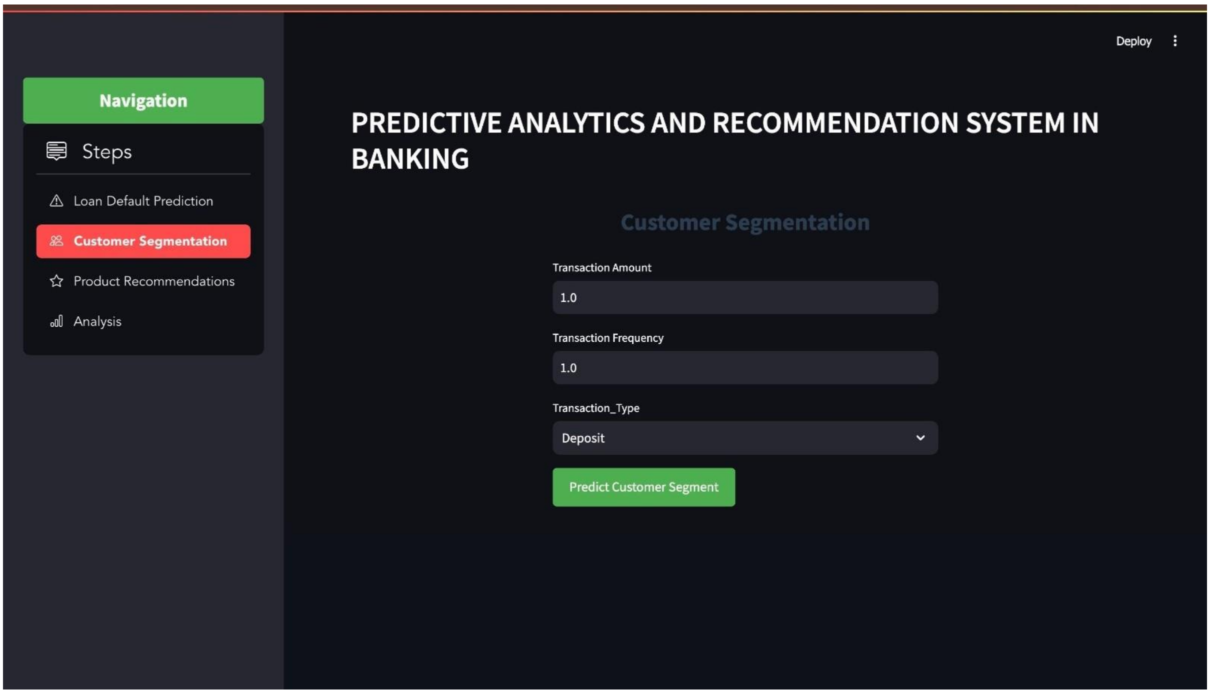
- Deployed models using Streamlit.
- Features:
  - Input forms for customer or transaction details.
  - Real-time loan default predictions, customer segmentation, and product recommendations.

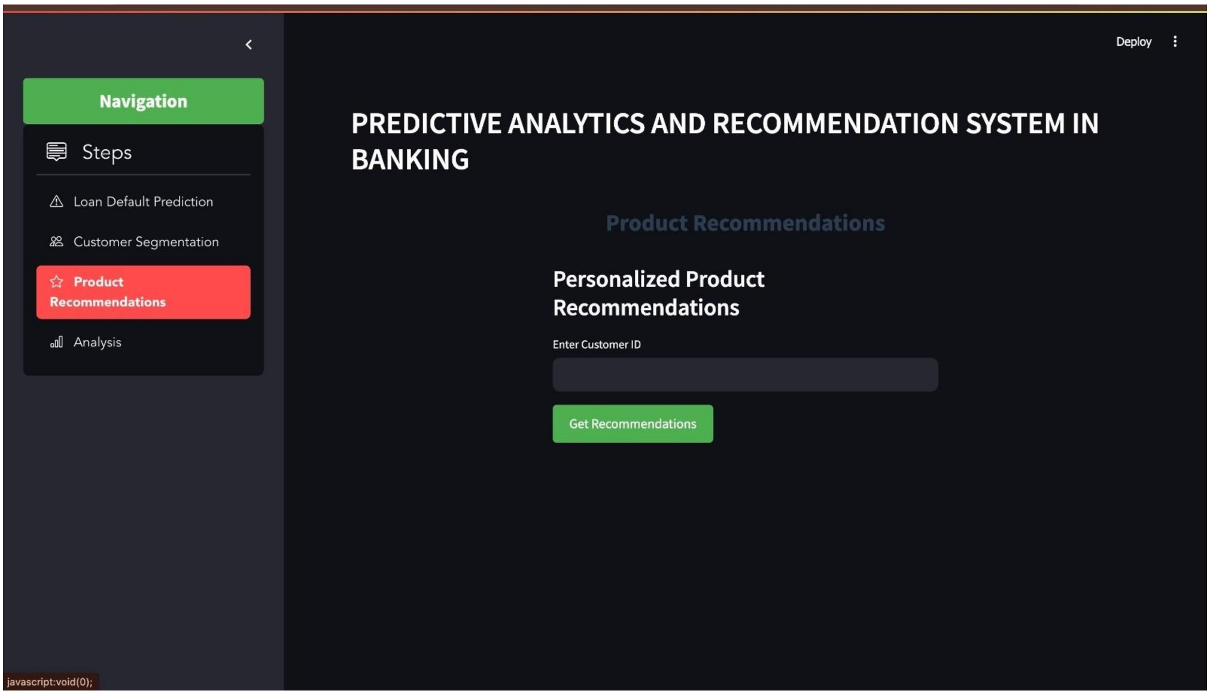
Deployment Process:

1. Saved trained models using joblib.
2. Integrated models into a Streamlit app.
3. Tested deployment for user interaction and scalability.

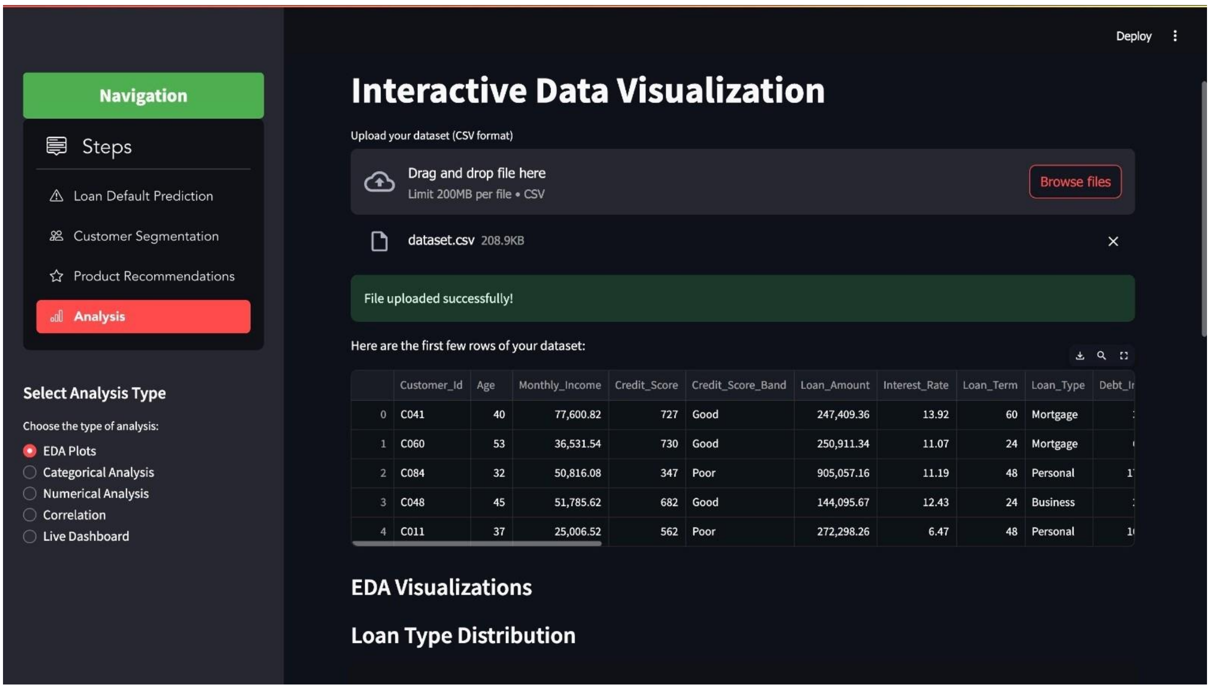
Screenshots:







Deploy
Navigation
Steps
Loan Default Prediction
Customer Segmentation
Product Recommendations
Analysis
PREDICTIVE ANALYTICS AND RECOMMENDATION SYSTEM IN BANKING
Product Recommendations
Personalized Product Recommendations
Enter Customer ID
Get Recommendations
javascript:void(0);



Deploy
Navigation
Steps
Loan Default Prediction
Customer Segmentation
Product Recommendations
Analysis
Select Analysis Type
Choose the type of analysis:
EDA Plots
Categorical Analysis
Numerical Analysis
Correlation
Live Dashboard
Interactive Data Visualization
Upload your dataset (CSV format)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
dataset.csv 208.9KB
File uploaded successfully!
Here are the first few rows of your dataset:
Customer_Id Age Monthly_Income Credit_Score Credit_Score_Band Loan_Amount Interest_Rate Loan_Term Loan_Type
0 C041 40 77,600.82 727 Good 247,409.36 13.92 60 Mortgage
1 C060 53 36,531.54 730 Good 250,911.34 11.07 24 Mortgage
2 C084 32 50,816.08 347 Poor 905,057.16 11.19 48 Personal
3 C048 45 51,785.62 682 Good 144,095.67 12.43 24 Business
4 C011 37 25,006.52 562 Poor 272,298.26 6.47 48 Personal
EDA Visualizations
Loan Type Distribution

Reason Behind Model Selection

Loan Default Prediction:

- Gradient Boosting outperformed other models in terms of accuracy and robustness.

Customer Segmentation:

- K-Means provided intuitive and actionable clusters.

Recommendation System:

- Collaborative Filtering excelled in suggesting relevant products.

Conclusion

Project Impact:

1. Improved loan risk management through accurate predictions.
2. Enhanced marketing strategies via customer segmentation.
3. Increased customer satisfaction with personalized recommendations.

Future Work:

1. Incorporate additional datasets for improved model training.

2. Implement real-time feedback loops for model refinement.

3. Optimize application scalability for larger user bases.

Appendices

Model Performance Metrics:

- Loan Default Prediction: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

- Customer Segmentation: Silhouette Score, Davies-Bouldin Index.

- Recommendation System: Precision, Recall, MAP, NDCG.

References

1. Python libraries: NumPy, Pandas, Scikit-learn.

2. Visualization tools: Matplotlib, Seaborn.

3. Deployment: Streamlit.

4. Banking datasets from open repositories.