#### **Loan Default and Customer Segmentation**

| Author | Date | Version |
| --- | --- | --- |
| Nivedha. D | 22/12/2024 | V1.0 |

[1. Introduction](#_jb8h0udxb6gr)

[2. Dataset Description](#_xgv6fa86nwas)

[Customer Data](#_fnso0p6ws6pw)

[Transaction Data](#_p9y4gd6st5w)

[3. Exploratory Data Analysis (EDA)](#_ox55qoiqse2e)

[Customer Data](#_zbrqawtyivvf)

[Transaction Data](#_1tqrob1k2bx1)

[4. Methodology](#_7q9u12nfj7pv)

[Data Preprocessing](#_9g5xixo1a0xe)

[Modeling](#_lrm314odv395)

[Clustering](#_fsce86f5y522)

[5. Modeling and Evaluation](#_m8hlxxca9jm2)

[Loan Default Prediction](#_4sd33sr49vx1)

[Customer Segmentation](#_ttm4642e1os5)

[6. Customer Segmentation](#_pr2r2tti90c)

[7. Conclusion](#_8z87291urvlk)

[Recommendations](#_i67xrgw9ujbf)

#### 

### 

### **1. Introduction**

This project focuses on addressing two critical objectives:

1. Predicting loan default probabilities using machine learning techniques.
2. Segmenting customers based on transaction behavior for targeted business strategies.

By combining predictive modeling and clustering, this project aims to improve risk management and enhance customer engagement for financial institutions.

### **2. Dataset Description**

#### **Customer Data**

* **Source**: Synthetic dataset generated using Faker.
* **Key Attributes**:
  + customer\_id: Unique identifier for each customer.
  + age: Age of the customer.
  + annual\_income: Annual income in USD.
  + loan\_amount: Total loan amount.
  + credit\_score: Customer's credit score.
  + repayment\_status (renamed to default): Indicates if the customer defaulted (1) or not (0).

#### **Transaction Data**

* **Attributes**:
  + transaction\_id: Unique identifier for each transaction.
  + customer\_id: Associated customer.
  + transaction\_type: Categorical feature (e.g., purchase, withdrawal).
  + transaction\_amount: Transaction value.
  + transaction\_date: Date of the transaction.

### **3. Exploratory Data Analysis (EDA)**

#### **Customer Data**

* **Summary Statistics**:
  + Checked for null values and ensured data completeness.
  + Columns like credit\_score, annual\_income, and loan\_amount were visualized against default status.
* **Visualizations**:
  + Histogram of credit\_score vs. default using seaborn.
  + Scatterplots to explore relationships among income, age, and loan defaults.

#### **Transaction Data**

* **Statistics**:
  + Timeframe: Transactions span from the earliest to the latest recorded dates.
  + Types: Unique transaction types identified for clustering.
* **Grouping**:
  + Aggregated by customer\_id for average transaction amount and count.

### **4. Methodology**

#### **Data Preprocessing**

1. Removed unnecessary columns (e.g., name, customer\_id).
2. Applied label encoding to categorical fields (e.g., default).
3. Standardized continuous features (e.g., credit\_score, loan\_amount) using StandardScaler.
4. Engineered features from transaction data (e.g., average transaction amount).

#### **Modeling**

* **Logistic Regression**:
  + Split data into train-test sets (80-20 split).
  + Evaluated model using metrics such as accuracy and ROC-AUC.

#### **Clustering**

* **K-Means Algorithm**:
  + Determined optimal clusters using the elbow method.
  + Evaluated clusters with Silhouette and Davies-Bouldin scores.

### **5. Modeling and Evaluation**

#### **Loan Default Prediction**

* **Model**: Logistic Regression.
* **Performance**:
  + Predicted probabilities for loan defaults.
  + Key metrics: Precision, recall, and F1-score.

#### **Customer Segmentation**

* **Clusters**:
  + Segmented customers into groups based on transaction behavior.
  + Performed PCA for dimensionality reduction and visualized clusters.
* **Insights**:
  + Average transaction amount and frequency varied significantly across clusters.

### **6. Customer Segmentation**

1. **Features for Clustering**:
   * Average transaction amount per customer.
   * Number of transactions.
   * Frequency of transaction types.
2. **Business Insights**:
   * High-value customers with frequent transactions identified.
   * Customers with high default risks were profiled for targeted interventions.

### **7. Conclusion**

This project successfully demonstrates:

1. **Loan Default Prediction**: A robust logistic regression model was built to identify high-risk customers.
2. **Customer Segmentation**: K-Means clustering revealed actionable insights for marketing and risk management strategies.

#### **Recommendations**

* Integrate the loan default prediction model into existing loan approval processes.
* Use segmentation insights for targeted campaigns and personalized customer offers.