Exploratory Data Analysis (EDA) for Loan Default Prediction

A case study on risk analytics in banking and financial services

Presenters:

Krish Agrawal

Bharti Katal

Problem Statement

- The aim is to identify patterns indicating the likelihood of loan default.

Two key risks for the company:

- Loss of business if loan is not approved for a reliable applicant.
- Financial loss if loan is approved for a default-prone applicant.

Business Understanding

- The goal is to identify applicants who are likely to default, to minimize financial risk while maximizing business opportunities.
- Loan outcomes:
  - Fully Paid
  - Current (in progress)
  - Charged-off (defaulted)

Dataset Overview

The dataset contains information about past loan applicants:

- Demographics, income, credit history, loan details, etc.
- Target variable: Loan Status (Fully Paid, Current, Charged-off)

EDA Objectives:

- Understand the data distribution.
- Identify key variables influencing loan default.

Univariate Analysis

Key observations from univariate analysis:

- Distribution of numerical variables like income, loan amount, etc.
- Frequency of categorical variables like loan purpose, grade, etc.

Insights:

- Certain loan purposes have higher default rates.
- Grade distribution shows clear segmentation.

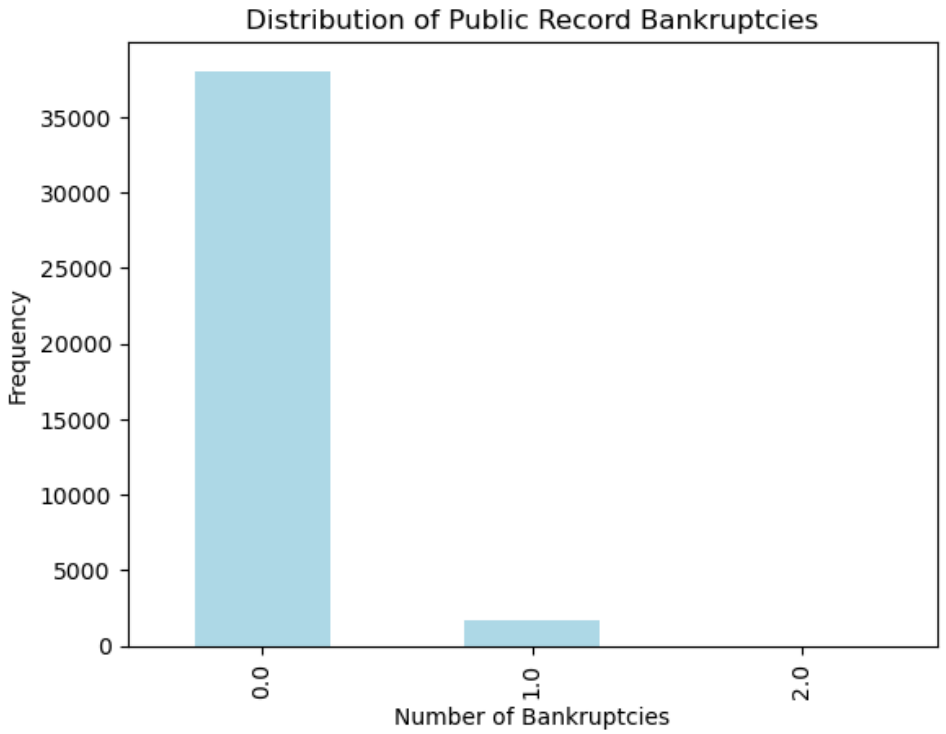
Number of Bankruptcies

- 0 Bankruptcies:

  > 35000
- 1 Bankruptcies:

  < 5000
- 2 Bankruptcies:

  Null



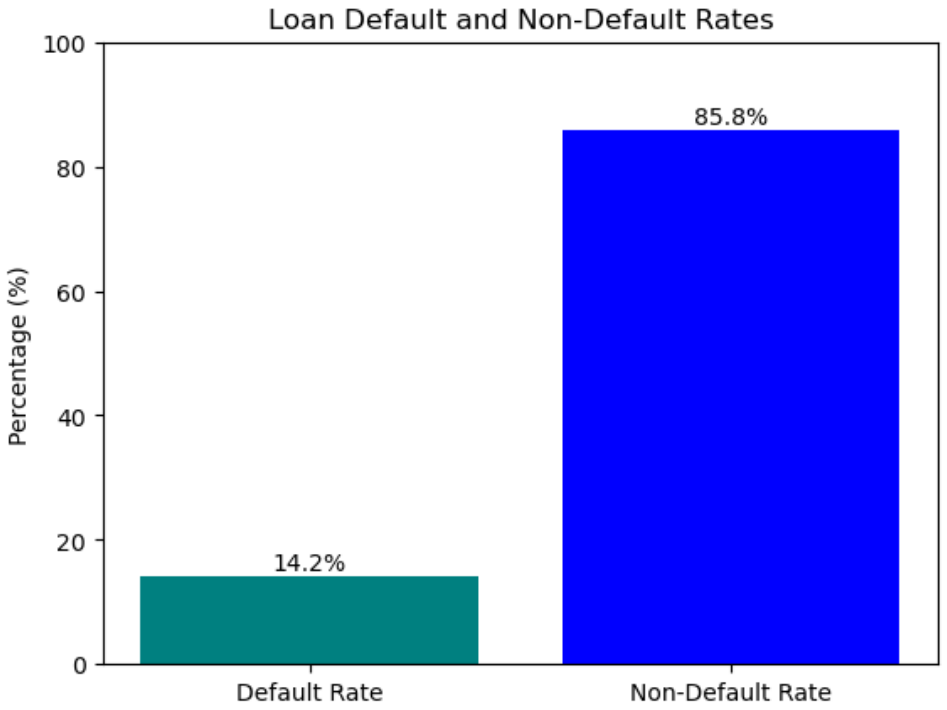
Ratio of Defaulters to Non-Defaulters

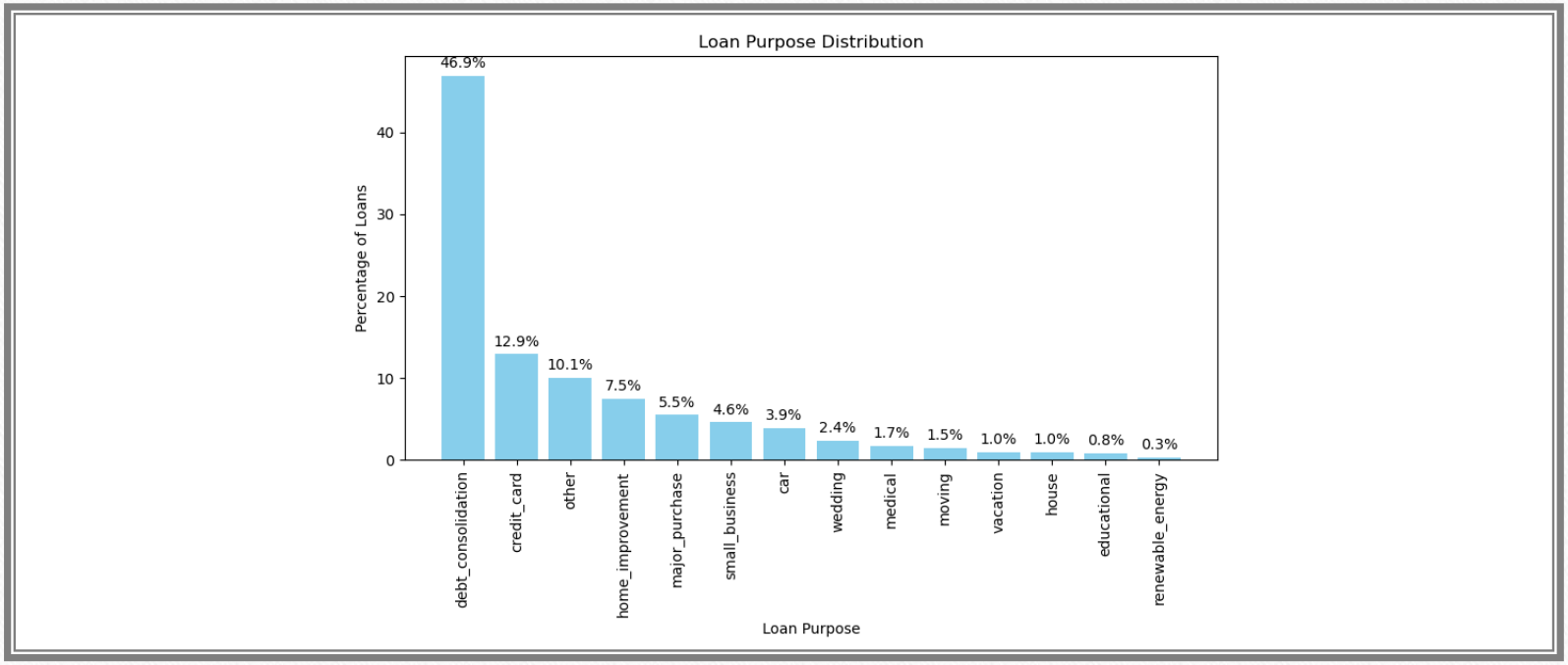
- Default Rate:

  14.2%

- Non-Default Rate:

  85.8%





Distribution of customers based on loan purpose

Bivariate Analysis

Key observations from bivariate analysis:

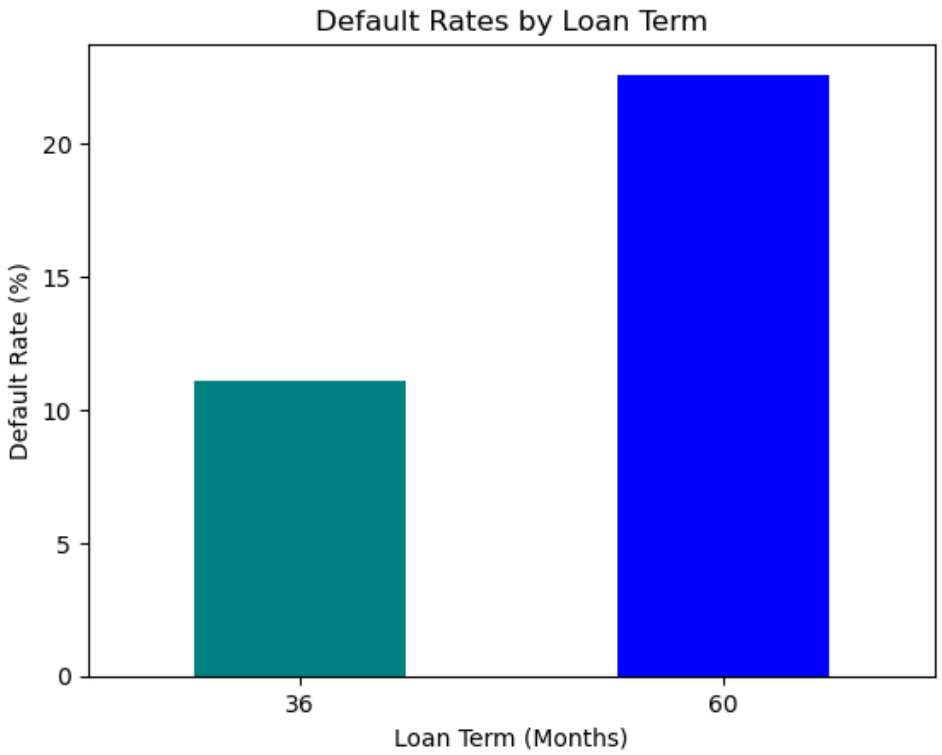
- Correlation between loan amount and default likelihood.
- Relationship between income levels and loan grades.

Insights:

- High loan amounts and low grades correlate with higher defaults.
- Income levels influence repayment behavior.

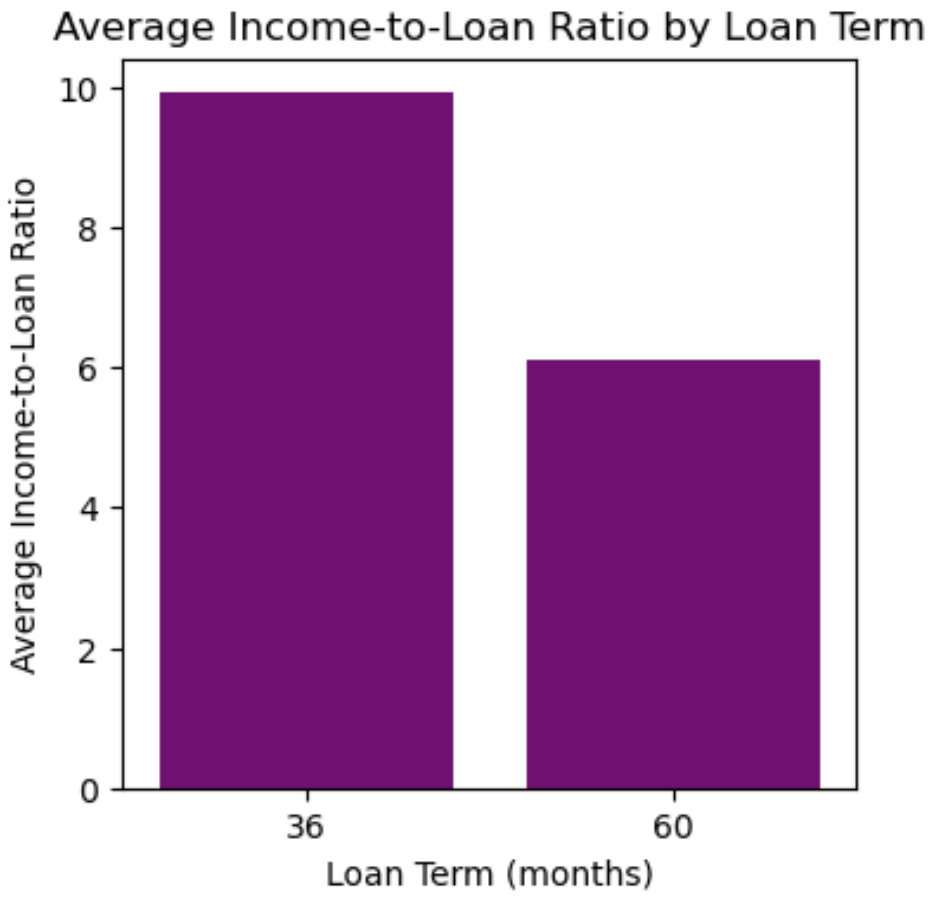
Default Rate Based on Loan Term

- 36 Months: ~11%
- 60 Months: ~22%

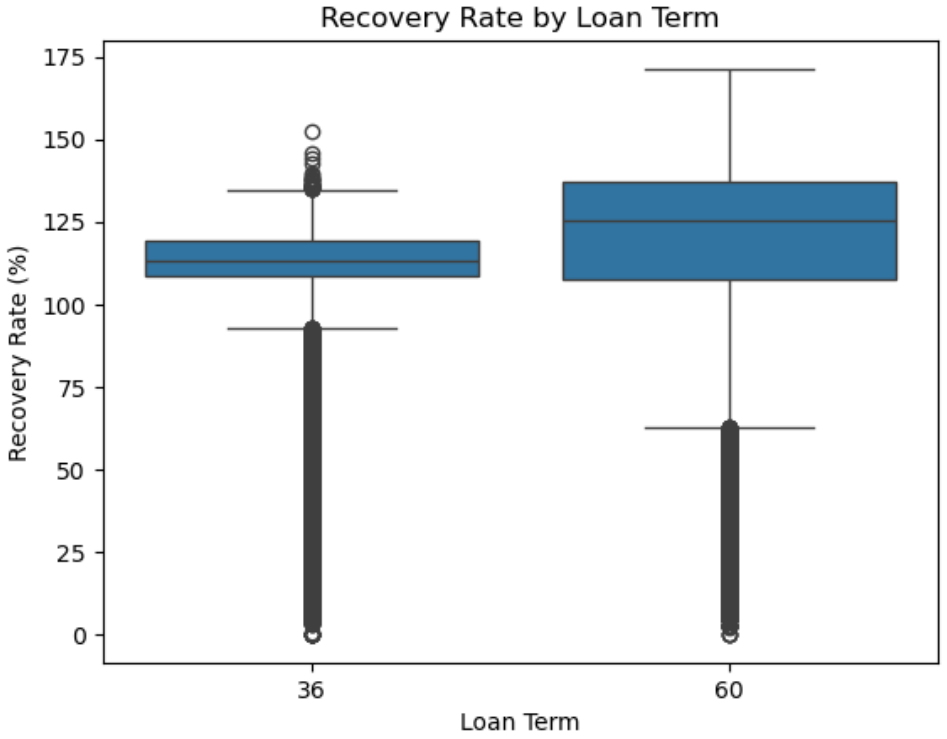


Income to Loan Ratio Based on Loan Term

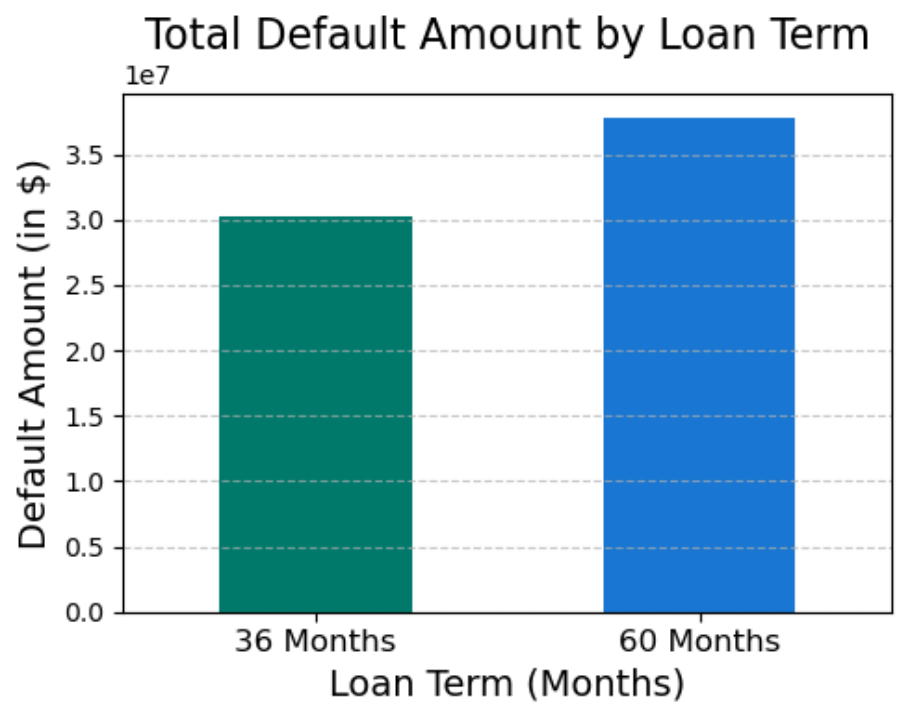
- 36 Months : 60 Months
- Ratio: ~ 5 : 3



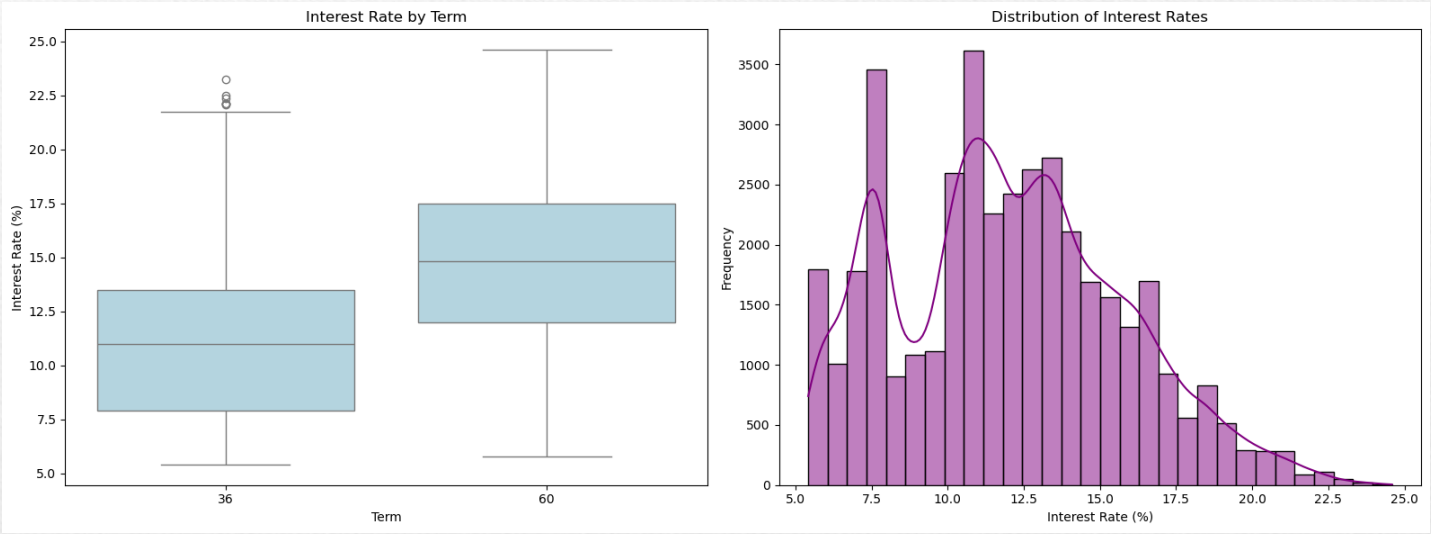
Recovery Rate Based on Loan Term



Total Default Amount Based on Loan Term



Interest Rate Based on Term(Duration) and Distribution of Interest Rates



Multivariate Analysis

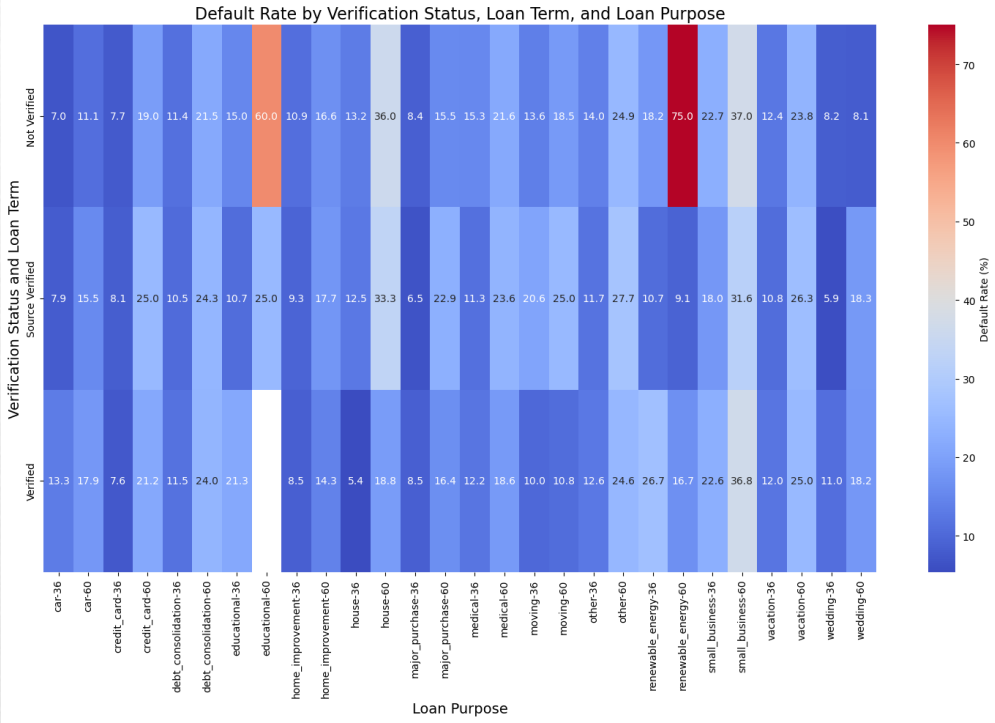
Key observations from multivariate analysis:

- The default rate varies significantly based on the combination of Verification Status, Loan Term, and Loan Purpose.
- Verified loans generally show a lower default rate compared to Not Verified or Source Verified across most purposes.
- Loan purposes such as "small business" and "educational" exhibit higher default rates, especially for longer-term loans (60 months).
- Shorter-term loans (36 months) have comparatively lower default rates across most purposes and verification statuses.

Insights:

- Applicants with unverified or source-verified statuses tend to be riskier, indicating the need for stricter verification processes.
- Longer-term loans (60 months) are more likely to default, particularly for high-risk purposes like "small business."
- Tailored loan policies based on purpose, term, and verification status can help mitigate risk and reduce default rates.

Default Rates based on Verification Status, Loan Term and Loan Purpose



Insights and Business Implications

Actionable Insights:

- Adjust lending criteria based on high-risk loan purposes.
- Implement stricter checks for low-grade loans.
- Offer lower amounts or higher interest rates to risky applicants.

Business Implications:

- Minimized financial loss.
- Increased profitability through better risk management.