IS4116 BIS : Data Analytics Process and Interpretation

Credit Risk Assessment Analysis

20021097- P.M.B.R Vimukthi

Repo Link : Click Here

Dataset Link : Click Here

1. Introduction

Understanding the factors that influence loan defaults is essential for financial institutions to mitigate risks and refine lending strategies. This analysis explores a dataset containing information on loan applicants, including demographics, income levels, homeownership status, and loan repayment behavior, to identify key predictors of loan default. By applying data preprocessing, exploratory analysis, statistical modeling, and visualization, the project aims to develop a predictive credit risk model and provide key insights to help financial institutions make informed lending decisions.

- **Business Question**: What factors contribute most to the likelihood of a loan default, and can we build a predictive model to assess credit risk?

2. Dataset Overview

The dataset consists of **32,581** records with **12** features related to loan applicants. These features include demographic details (age, income, homeownership status), financial attributes (loan amount, interest rates), and loan repayment status (default or non-default). The data represents simulated credit bureau records, making it relevant to the **Banking and Financial Services Domain**. The primary objective is to analyze how different factors influence loan default risk, helping financial institutions refine their credit risk assessment and make data driven decisions.

3. Analytical Process

3.1 Data Cleaning and Preprocessing

Initially duplicate records were identified and removed, while outliers were handled using the IQR method to cap extreme values at appropriate thresholds. Missing values were replaced with the median due to data skewness. The categorical variables were encoded using Label Encoding (Ordinal), with One-Hot Encoding(Nominal) for advanced analysis.

3.2 Exploratory Data Analysis (EDA)

The main objective of this step is to understand the structure of the dataset and identify patterns and relationships to derive meaningful insights.

- Summary statistics (mean, median, standard deviation, etc.) were calculated for numerical variables.
- Binning was applied to create meaningful feature groups (**`Age group, Income group and Loan Amount group`**), and new features were derived, including the **`Loan-to-Income Ratio`**, which measures the loan amount relative to the borrower's income, and the **`Interest Rate-to-Loan Amount Ratio`**, which represents the cost of borrowing relative to the loan size.
- Visualizations were used to analyze data patterns and relationships. Boxplots and bar charts assisted in examining distributions and category frequencies, while heatmaps provided insights into correlations between numerical features. Scatterplots were used for correlation analysis, highlighting trends and dependencies within the dataset.

3.3 Predictive Analysis

Predictive analysis was conducted to build a reliable credit risk model. Initially, feature scaling was performed using **`StandardScaler()`** to standardize numerical variables. Following that, **`Random Forest Classifier`** was utilized to develop the prediction model, leveraging its robustness, simplicity and ability to handle complex data patterns. Model performance was evaluated using the Classification Report and Cross Validation scores and it shows **strong predictive capabilities (Fig 06)**. Additionally, feature importance was extracted and visualized to identify the most influential factors in predicting loan defaults.

4. Results and Key Findings

- Most loan applicants are 20-35 years old, with the 26-35 group being the largest. Organizations should focus on younger individuals for targeted loan offerings **(Fig 01)**.
- Low-middle(25,000 - 49,999) and middle-income(50,000 - 74,999) groups have the most loan applications, but default rates are highest in low-income groups. Higher-income(+100000) applicants default less, emphasizing income as a key risk factor **(Fig 02)**.
- The majority of applicants rent (50.9%) or have a mortgage (41.2%), which suggests the need for rental-based credit scoring models **(Fig 03)**.
- Income and loan amount show a **positive correlation**. Defaults are higher among low-income(0 - 24,999) borrowers with smaller(0 - 5000) loans, reinforcing income as a strong predictor of repayment ability. Additionally. low-income applicants pose a higher default risk, even for small loans, while high-income borrowers default less, even with larger loans **(Fig 04)**.
- **Loan grade, loan-to-income ratio, and interest rate** are the strongest default predictors. Stricter credit evaluations should be applied to borrowers with low grades, high loan-to-income ratios, and high interest rates to reduce risk **(Fig 05)**.

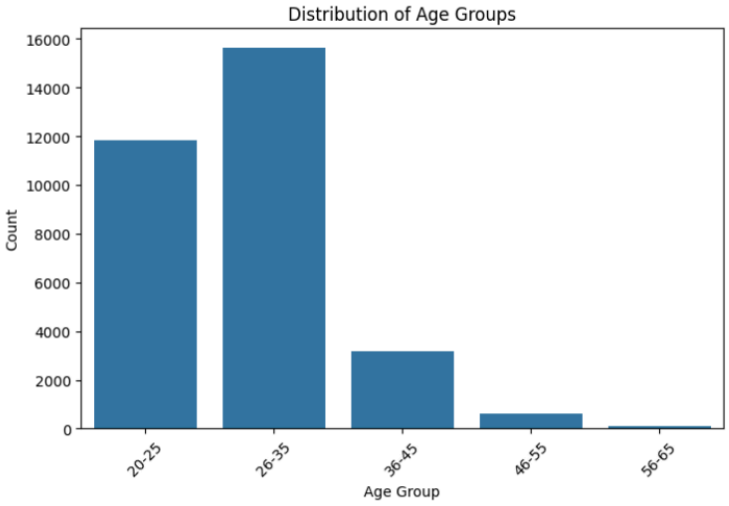


Fig 01 - Distribution of Age groups

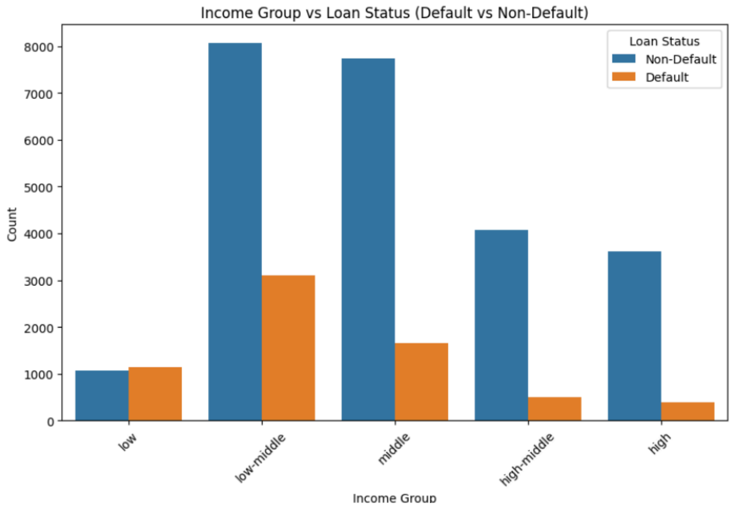


Fig 02 - Distribution of Income groups

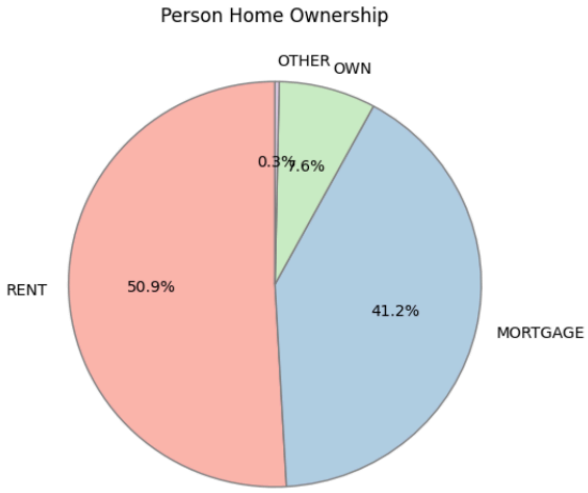


Fig 03 - Distribution of Person Home Ownership

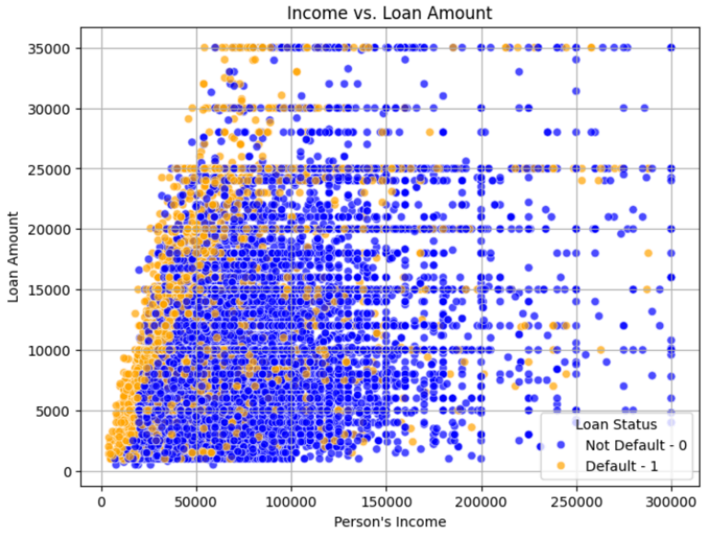


Fig 04 - Income & Loan Amount Correlation

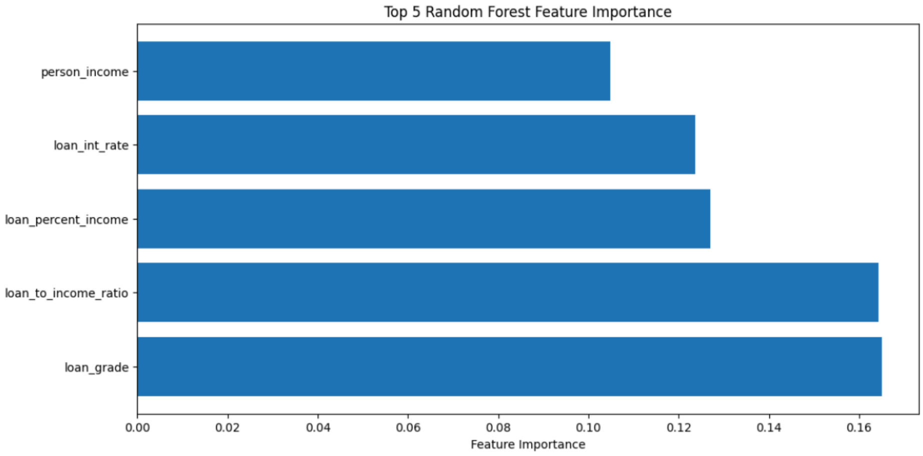
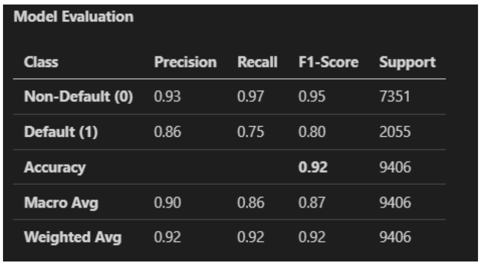


Fig 05 - Distribution of Age groups



Model Evaluation

Class	Precision	Recall	F1-Score	Support
Non-Default (0)	0.93	0.97	0.95	7351
Default (1)	0.86	0.75	0.80	2055
Accuracy			0.92	9406
Macro Avg	0.90	0.86	0.87	9406
Weighted Avg	0.92	0.92	0.92	9406

Fig 06 - Model Evaluation