ANALYSIS AND VISUALIZATION OF LENDING CLUB LOAN DATA

DATA CLEANING, OUTLIER DETECTION, AND CORRELATION ANALYSIS

AUTHORS

MEGHA KHANT & ANUSHMITA MUKHERJEE

BUSINESS UNDERSTANDING

- Introduction:
  - The company specializes in lending various types of loans to urban customers.
  - Decision to approve loans based on applicant's profile to minimize risk.
- Risks:
  - Loss of business if a good applicant is rejected.
  - Financial loss if a bad applicant is approved.
- Objective:
  - Identify patterns that indicate a person’s likelihood to default.
  - Use this information for making informed lending decisions.

DATA UNDERSTANDING

- Dataset Overview:
  - Complete loan data for loans issued from 2007 to 2011.
  - Contains information on loan status: fully paid, current, and charged-off.
- Data Dictionary:
  - Provides descriptions of variables in the dataset.
  - Link to Data Dictionary

DATA PREPARATION

- Steps Taken:
  - Handling missing values.
  - Removing irrelevant columns.
  - Converting data types for analysis.
- Outcome:
  - Cleaned and prepared dataset ready for analysis.

UNIVARIATE ANALYSIS

- **Objective:** Understand individual variable distributions.
- **Key Findings:**
  - Higher Default Rates in Lower Loan Ranges and Outliers in High Loan Amounts
  - Many defaults occur between 12.5% to 15% interest rates, with the median in this range.
  - Higher instalments may increase default risk.
  - High Positive Skew in Revolving Balances:
  - Broad Distribution of Revolving Utilization Rates
- **Visualizations:**
  - Histograms, box and bar plots for key variables.

BIVARIATE ANALYSIS

- **Objective:** Explore relationships between variables.
- **Key Findings:**
  - Larger loan amounts correlate with higher default rates, indicating increased risk with larger loans.
  - More open accounts and higher revolving balances/utilization ratios are linked to higher default rates.
  - Higher received interest and recovery fees are associated with default loans, suggesting defaults may occur after substantial payments and recovery efforts.
- **Visualizations:**
  - Pair plots showing key relationships.

MULTIVARIATE ANALYSIS

- **Objective:** Understand interactions between multiple variables.
- **Key Findings:**
  - Loan amount, funded amount, and instalment show almost perfect correlation
  - Interest rate and instalment, as well as revolving balance and utilization, show moderate positive correlations
  - Variables such as total payment with total payment invested, and recoveries with collection recovery fee, are almost perfectly correlated
- **Visualizations:**
  - Heatmaps plots.

KEY INSIGHTS

- Driving Factors Behind Loan Default:
  - High loan amounts and interest rates correlate with higher default rates.
  - Certain loan purposes have higher default rates.
- Business Implications:
  - Adjusting interest rates based on risk.
  - Refining loan approval criteria.

RECOMMENDATIONS

- Target Larger Loans and High Utilization Rates:
  - Since larger loans and higher revolving balances/utilization are correlated with defaults, focus on stricter evaluation and monitoring for these borrowers to mitigate risk.
- Monitor Payment Patterns:
  - Given the strong correlation between total payments and recoveries, closely track borrower payment behaviours and intervene early for signs of financial distress.
- Loan Approval Criteria:
  - Improve criteria for loan approval by incorporating insights from correlations, such as considering the interplay between interest rates, installments, and borrower income, to better predict and prevent defaults.

CONCLUSION

- After doing the EDA on the data set, it indicates that loan defaults are significantly influenced by high loan amounts, high revolving balances, and utilization ratios, along with higher interest rates and instalment payments.
- Borrowers with higher incomes and higher emp_title able to pay back the loan better, leading to lower default rates.
- Also, home_ownership and grad are also important factor in loan defaulting.
- Defaulting borrowers are leading to higher recovery costs.