Lending Club Case Study

Samuel Sushanth Kolli

Manoj Shah

PGMLAI C62

Contents

- **Problem Statement**
- **Data Description**
- **Data Understanding**
- **Data Cleaning & Pre-processing**
- **Univariate Analysis**
- **Bivariate Analysis**
- **Multivariate Analysis**
- **Correlation Analysis**
- **Suggestions**
- **References & Useful Links**

Problem Statement

The problem statement can be divided into the following sections:

Introduction

- This assignment aims to provide an understanding of how real business problems are solved using Exploratory Data Analysis (EDA).
- The case study focuses on risk analytics in banking and financial services, and how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

- The company is a consumer finance company that specializes in lending various types of loans to urban customers.
- When a loan application is received, the company must decide whether to approve or reject the loan based on the applicant's profile.

Two types of risks are associated with the loan approval decision:

- If the applicant is likely to repay the loan, not approving the loan results in a loss of business for the company.
- If the applicant is not likely to repay the loan (i.e., they are likely to default), approving the loan may lead to a financial loss for the company.

Business Objectives

- The company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.
- The largest source of financial loss for the company is lending loans to "risky" applicants, known as credit loss.
- The aim is to identify these risky loan applicants using EDA, which can help the company reduce credit loss by denying loans, reducing loan amounts, or lending at higher interest rates to risky applicants.
- The company wants to understand the driving factors (driver variables) behind loan default, which are strong indicators of default.

Problem Statement

The problem statement can be divided into the following sections:

Data Understanding

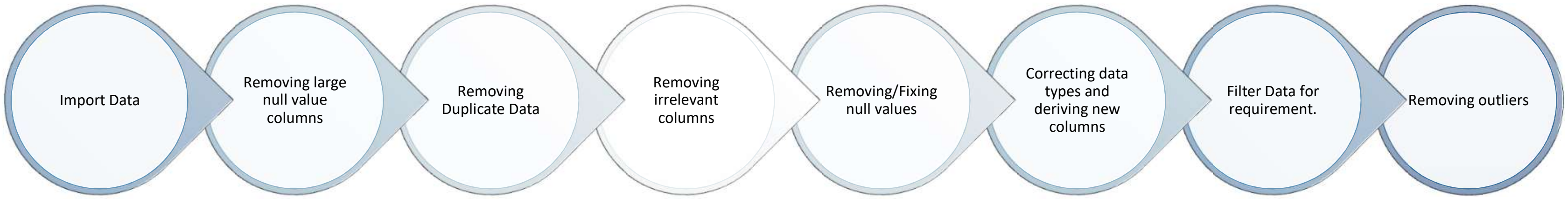
- The dataset provided contains complete loan data for all loans issued through the time period 2007 to 2011.
- A data dictionary describing the meaning of the variables is also provided.

Results Expected

- Write all code in a well-commented Python file, mentioning insights and observations from the analysis.
- Present the overall approach of the analysis in a presentation, including the problem statement, analysis approach, results of univariate, bivariate analysis, etc., in business terms, and visualizations summarizing the most important results.
- Submit an Ipython notebook clearly explaining the thought process behind the analysis, code, and relevant plots.
- Submit a GitHub repository link containing the files and a README.md file describing the project briefly.

Data Description

- The dataset provided contains information about past loan applicants and whether they 'defaulted' or not. It includes various attributes related to the loan and the applicant.
- Lending Club, with over 39,717 records and 111 columns, providing rich information about borrowers' past credit history and loan details. This extensive dataset allowed your team to conduct thorough analysis, identifying relationships and assessing their impact on borrowers' ability to fulfill loan agreements successfully. With such a wealth of data, you were well-equipped to explore various factors influencing loan outcomes.



Import Data

Removing large null value columns

Removing Duplicate Data

Removing irrelevant columns

Removing/Fixing null values

Correcting data types and deriving new columns

Filter Data for requirement.

Removing outliers

Data Understanding

The dataset contains information about loan applicants, including their loan status (fully paid, current, charged-off) and various attributes related to the loan and the applicant. The goal is to identify patterns that indicate if a person is likely to default on the loan, which can be used for decision-making, such as denying the loan, reducing the loan amount, or lending at a higher interest rate to risky applicants.

Dataset Attributes:

Primary Attribute: Loan Status

- Three distinct values: Fully-Paid, Charged-Off, Current
- Fully-Paid: Customers who have successfully repaid their loans.
- Charged-Off: Customers who have defaulted on their loans.
- Current: Loans in progress, no conclusive evidence regarding future defaults.

Decision Matrix:

Loan Acceptance Outcome:

- FullyPaid: Applicants who have repaid both principal and interest.
- Current: Applicants making installments, not categorized as defaulted.
- Charged-off: Applicants who failed to make timely installments, resulting in default.

Loan Rejection:

- Cases where the loan application is declined are not included in the dataset as there's no transactional history available for these applicants.

Data Understanding

Annual Income (annual_inc): Reflects customer's annual income, influencing loan approval likelihood.

Home Ownership (home_ownership): Indicates whether the customer owns a home, impacting collateral availability.

Employment Length (emp_length): Represents customer's employment tenure, indicating financial stability.

Debt to Income (dti): Measures the portion of monthly income used for debt payments, influencing loan approval.

State (addr_state): Denotes customer's location, useful for demographic analysis and identifying trends in default rates.

Loan Characteristics:

Loan Amount (loan_amt): Amount requested by the borrower.

Grade (grade): Rating based on creditworthiness, indicating loan risk.

Public Records (public_rec): Derogatory public records impacting loan risk.

Public Records Bankruptcy (public_rec_bankruptcy): Number of bankruptcy records affecting loan success rate.

Data Cleaning and Manipulation

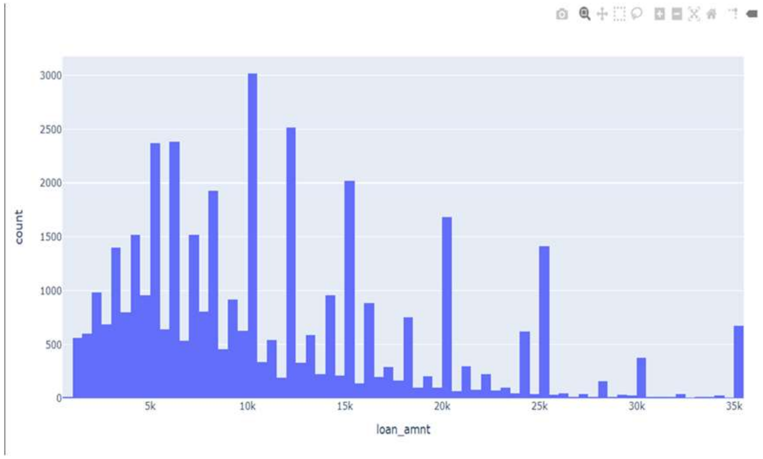
This section will cover the steps taken to clean and preprocess the data, such as handling missing values, removing outliers, and converting data to a suitable format for analysis.

1. **Missing Values in Annual Income (annual_inc):**
    1. Filled missing values in annual_inc with the mode value of annual_inc corresponding to the emp_length field.
    2. Assuming missing emp_length values (1015) represent business owners, added their employment duration with the mode value of emp_length, which is 10+ years.
2. **Employment Length Mapping:**
    1. Mapped employment length to the respective number of years in integers.
3. **Handling Missing Values in Home Ownership (home_ownership):**
    1. Imputed 'NONE' values as 'OTHER' for home_ownership.
4. **Standardizing Verification Status:**
    1. Replaced 'Source Verified' values with 'Verified' since both indicate verified income sources.
5. **Handling Null Values in Public Record Bankruptcies (pub_rec_bankruptcies):**
    1. Dropped 660 rows with null values in pub_rec_bankruptcies as they couldn't be imputed.
6. **Data Analysis:**
    1. Created a 'default' column based on the 'loan_status' column.
    2. Created 'issue_year' and 'issue_month' columns.
    3. Plotted Histogram and Boxplot of the 'annual_inc' column.
    4. Divided values into 'low_inc', 'mid_inc', and 'high_inc' categories based on 'annual_inc'

Univariate Analysis

Univariate analysis will be performed on the individual variables to understand their distributions, central tendencies, and other characteristics. This analysis will help identify potential driver variables that may influence loan default.

1. **Low Annual Salaries:** Exercise caution when lending to individuals earning less than $40,000 annually. Implement rigorous income verification and assess repayment capacity thoroughly for applicants in this income bracket.
2. **Interest Rates:** A significant portion of defaulted loans had interest rates between 13% to 17%. Consider offering loans at lower interest rates when possible to reduce the risk of default.
3. **Higher Loan Amounts:** Evaluate applicants seeking loan amounts of $15,000 and above carefully. Ensure strong credit history and repayment capability for larger loans.
4. **Funded Amounts:** Align funded amounts with the borrower's financial capacity. Conduct thorough credit assessments for larger loan requests to mitigate risk.
5. **Debt-to-Income Ratios:** Implement strict debt-to-income ratio requirements to prevent lending to individuals with unsustainable levels of debt relative to their income.
6. **Monthly Installments:** Monitor and assess applicants with monthly installment amounts between $160 to $440 closely to mitigate the risk of loan defaults.



Bivariate Analysis

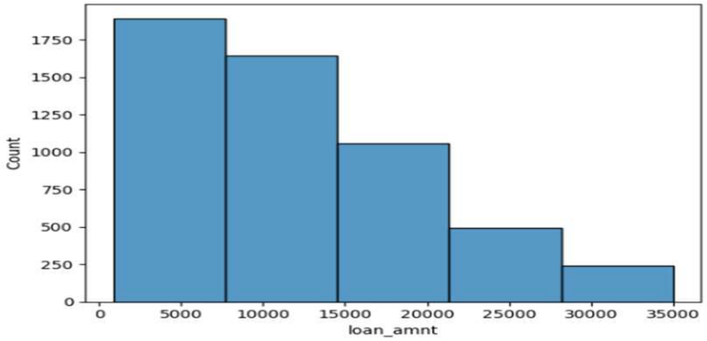
Bivariate analysis will be conducted to explore the relationships between pairs of variables. This analysis will help identify combinations of variables that may be strong indicators of loan default.

Findings:

- A majority of defaulted loan applicants received amounts of $15,000 or higher.
- Applicants with high Debt-to-Income (DTI) ratios were prevalent among those who defaulted.
- Defaulted loans often had interest rates ranging from 13% to 17%.
- Most defaulted loan applicants reported annual incomes below $40,000.

Inferences:

- High Loan Amounts: Applicants receiving $15,000 or more are prone to defaulting. The company could mitigate this by conducting more thorough assessments and potentially capping loan amounts for higher-risk applicants.
- DTI and Interest Rates: High DTI ratios and interest rates between 13% to 17% correlate with defaults. The company should review its interest rate determination process and consider adjusting rates based on DTI ratios.



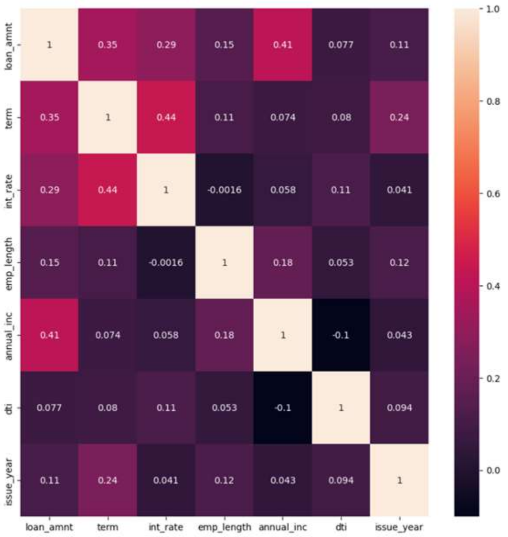
Multivariate Analysis

- Multivariate analysis techniques, such as clustering or dimensionality reduction, may be employed to uncover patterns and relationships among multiple variables simultaneously.
- Multivariate analysis involves analyzing data with more than two variables simultaneously.
- Unlike univariate (one variable) and bivariate (two variables) analysis, it examines relationships among multiple variables simultaneously.
- Widely used across fields like economics, social sciences, biology, marketing, and environmental science.
- Can include various types of variables, such as categorical, numerical, or a combination of both.

Correlation analysis

Correlation analysis will be performed to measure the strength and direction of the linear relationship between variables. This analysis will help identify variables that are strongly correlated with loan default.

1. **Loan Amount, Funded Amount, Funded Amount invested, and Installment:**
    1. Highly correlated with each other.
2. **Annual Income and Debt-to-Income (DTI):**
    1. Annual Income is negatively correlated with DTI, indicating that as Annual Income increases, DTI tends to decrease.



Suggestions

Based on the insights gained from the EDA, suggestions and recommendations will be provided to address the business objectives and minimize the risk of lending to risky applicants.

Major Driving Factors to Predict Default Risk:

- Debt-to-Income (DTI) Ratio
- Loan Grades
- Verification Status
- Annual Income
- Public Recorded Bankruptcies

Recommendations:

- Review the interest rate determination process and consider adjusting rates based on Debt-to-Income (DTI) ratios to align with the borrower's ability to repay.
- Carefully evaluate debt consolidation loan applicants, consider interest rate adjustments or offer financial counseling services.
- Consider housing stability during the underwriting process to assess the applicant's ability to repay the loan.
- Review the verification process to ensure effective assessment of applicant creditworthiness and make improvements if necessary
- Loan Risk Factors are Loan amounts of $30,000 or higher to Applicants with annual income below $25,000 and interest rate above 15% and Loan grades E, F, and G and Loans with a 60-month term
- Loan Acceptance Criteria are Loans for weddings, major purchases, cars, and credit card consolidation and Calculated interest rate less than 7.5% and Loan grades A and B and Applicants who own a house and Loans with a 36-month term

References & Useful Links

- This section will include references and useful links related to the topic of risk analytics in banking and financial services, as well as any other relevant resources used during the analysis.

- **GitHub Repository Link: https://github.com/ /lending-club-case-studyTechnology /**

Package	**Version**	**Documentation**
Python	3.11.4	https://www.python.org/
Matplotlib	3.7.1	https://matplotlib.org/
Numpy	1.24.3	https://numpy.org/
Pandas	1.5.3	https://pandas.pydata.org/
Seaborn	0.12.2	https://seaborn.pydata.org/