Lending Club Case Study

Amit Mohite

Neelam Jhunjhunwalla

Content

- Problem Statement
- Work Environment
- Overall Analysis Approach
- Approach to data cleaning
- Approach to Univariate
- Approach to Bivariate
- Approach to Segmented Univariate
- Summary (Insights and Conclusion)
- Recommendation and Next steps

Problem Statement

- The case study is about Lending club which wants to know the drivers behind loan default for its portfolio and risk assessment. In our case the Lending club is taking two decision on loan approval. We are interested in approved loan candidates where there are chances of default. In case of approved loan, following can happen:
    a. Loan is fully paid
    b. Loan is currently being paid
    c. Loan is charged off (applicable when defaulted)

Utilizing this data, we would understand the key factors for risk assessment

Work environment

Platform and libraries

- Python 3.11
- Pandas and NumPy for data handling
- Matplotlib and Seaborn for plots
- Scipy for statistics
- Sklearn for Isolation Forest

Development and Code management

- Visual Studio Code
- GitHub
- Microsoft/Office 365 Excel

Presentation

- Microsoft PowerPoint
- PDF

Overall Analysis Approach

For this case today we have followed the following approach:

1. Univariate analysis was done to identify descriptive analytics on the data distribution. This gave idea on data distribution
2. Analyzed the collected data for the following:
    a. Missing values
    b. outliers
    c. Single constant values used for column
3. We have removed the columns with
    a. missing values > 10% due to sparsity
    b. single values throughout
4. The outliers were eliminated using isolation forest method after testing various methods
5. Then we converted the data types to Int64 for some of the rows to support missing (NaN)
6. Analyzed the data using univariate analysis to understand the data distribution
7. The data columns were analyzed wrt leading and lagging indicator behaviour for loan default
8. Bivariate analysis and correlation matrix were formulated to see the degree of association between any two variables.
9. Established insights

Approach to data cleaning

Data dictionary review

- To understand data and critical variables
- To understand pre and post loan variables

Upfront data cleaning - Most removal of variable is done upfront based on following criteria

- Variables with high null or missing values
- Variables with constant data. E.g., member id
- Variables with data which cannot be cleaned, categorized, grouped further within current scope. E.g., title - verbatim by loan applicant

Data type fix and encoding

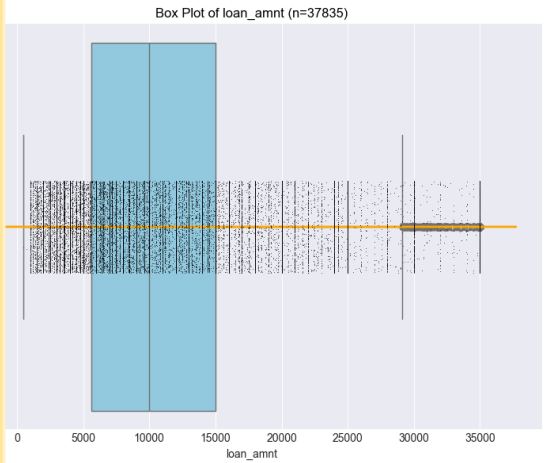
- Update default data type determined by Pandas
- Encode categorical into continuous variable for use in correlations
- Adjust data to correct interpretation. For example, Apr-68 in earliest credit line is April 1968 and not April 2024.

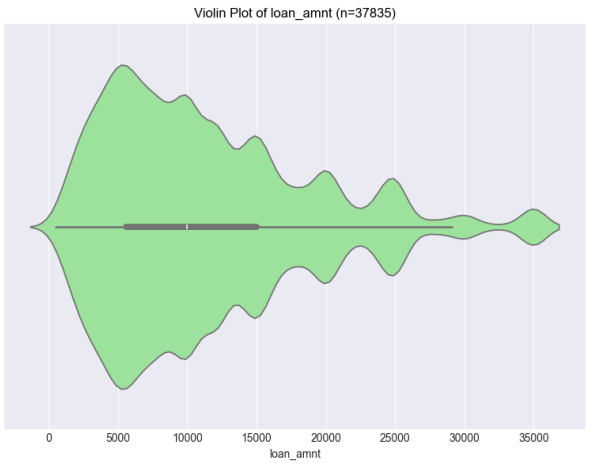
Outlier removal and capping

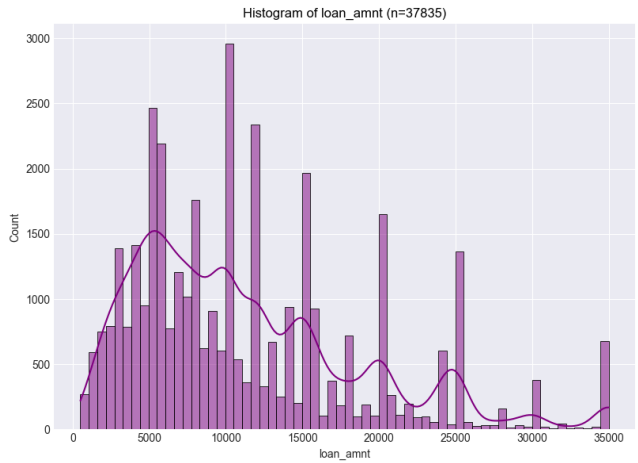
- Selection of outlier method by testing each method on critical data points. We concluded on `Isolation Forest` as it affected the least records and created same impact on data.
- Z score was next best method and IQR performed the worst.
- Due to low nature of outlier, both capping and removal was possible, and we concluded on capping.

Note: Data cleaning outcome evolved iteratively, and final code may not reflect iterations.

Approach to Univariate Analysis







Box Plot with KDE and Strip Plot Overlay (Left) - Visualizes the distribution of quartiles, median, and outliers, overlaid with a kernel density estimate (KDE) and individual data points.

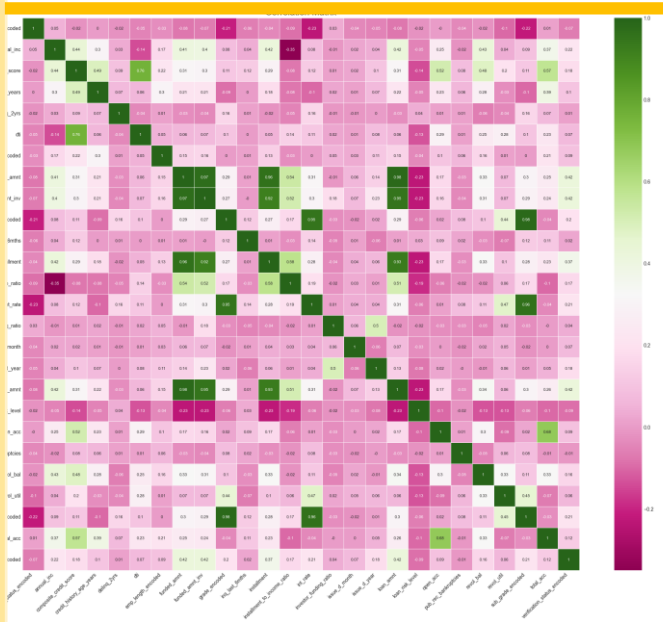
Violin Plot (Center) – To visualize shape of density and data peaks

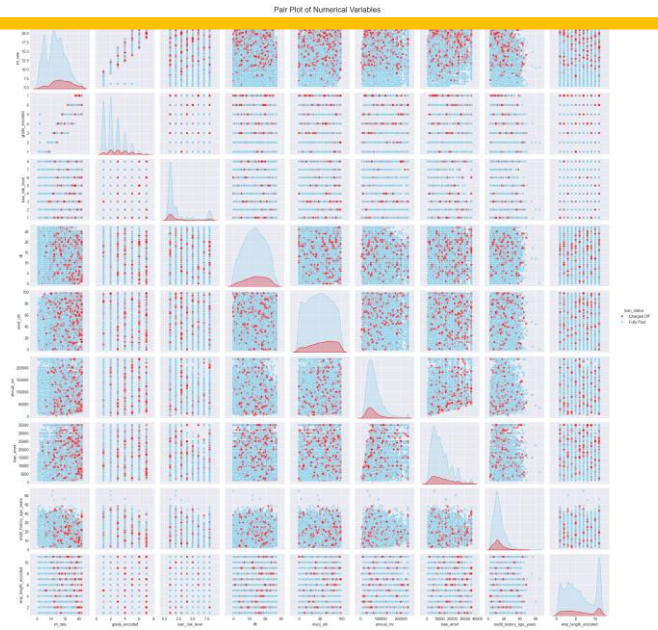
Histogram with KDE Line (Right) - Frequency distribution with bars representing count of observations, overlaid with a kernel density estimate (KDE) line.

Simple cross tab is used for categorical variables. Many of the categorical variables were converted to continuous during encoding and analyzed using plots.

We used function and loop to create summary of many variables at same time. We end up creating more plots than needed but allows faster plot creation and re-runs. Elimination of unwanted plots done during analysis and summarization.

Approach to Bivariate Analysis





Correlation matrix with heatmap for easy visualization of correlations.

Scatter/pair plot, to visualize relationships between pairs of variables in a dataset through a grid of scatterplots, helping identify correlations, patterns and clusters, thus providing a comprehensive overview of variables.

Again, function and loops are used to faster plot creation. We have also limited variable of interest

~85% of the records are for Fully Paid loans. This makes Bivariate less vocal about impact on loan status and should be handled before correlations but not done in current scope.

Approach to Segmented Univariate

Using same plots and cross tabs as univariate analysis and data is sliced by loan status – Charged off and Fully Paid.

X and Y axis of the respective plots are matched by pre-calculating min and max limits.

This section also use function and loop to create plots faster. Analysis is limited to variables of interest from earlier steps.

Summary – (Insights and Conclusion)

Univariate Analysis (continuous variables):

1. Common loan amounts are between $5000-$10000, with a notable preference of $10000. Funded amount shows a similar pattern with concentration around $10000 with most interest rate 10% to 15$. The installments range around between $150- $400 with major installments around $300. The debt-to-income ratio is between 10-20 and median around 15.

Univariate Analysis (discrete variables):

1. term - Most loans have a term of 36 months (73%), with the remaining 27% having a term of 60 months.

2. home_ownership - Most borrowers either rent (48%) or have a mortgage (44%), with a smaller portion owning their home (8%).

3. verification_status - Verification status is distributed with 43% not verified, 32% verified, and 25% source verified.

4. purpose - Debt consolidation is the most common purpose (47%), followed by credit card (13%) and other (10%). Same data is also summarized as loan_risk_level earlier

5. addr_state - Most loans (18%) are from single state CA i.e. State of California

Bivariate Analysis:

1. Installment and Loan Amounts: There is a high positive correlation between installment and loan_amnt (0.93), funded_amnt (0.95), and funded_amnt_inv (0.97), indicating that higher loan amounts correspond to higher monthly installments. These are obvious correlations and as expected

2. loan_status_encoded shows a moderate positive correlation with grade_encoded (0.2), indicating that higher loan statuses (fully paid) are linked to better grades

Summary – (Insights and Conclusion)

Scatter plot/ pair plot:

- There is a clear distinction in int_rate distributions between Fully Paid and Charged Off loans. Loans with higher interest rates are more likely to be charged off, indicating that a higher interest rate might be a significant risk factor.
- The loan grade shows a notable pattern where lower grades (which are higher risk) are more associated with Charged Off loans.
- Higher dti values slightly tend to be associated with charged-off loans, suggesting that higher debt-to-income ratios could be a contributing factor to loan defaults.

Overall, the features that seem to have the biggest impact on the likelihood of a loan being charged off are:

1. Interest Rate (int_rate): Higher rates are strongly associated with charged-off loans.
2. Grade (grade_encoded): Lower grades (as per grade_encoded) correlate with higher risk. Grade A being the best and G being the worst.
3. Debt-to-Income Ratio (dti): Higher ratios are slightly more associated with defaults.
4. Annual Income (annual_inc): Lower incomes show a higher risk of default.
5. Credit History Age (credit_history_age_years): Shorter credit histories are more prone to defaults.

Correlation by loan status segment

a) By Charged off:

- **Installment and Loan Amount (0.93)**: Higher loan amounts correspond to higher installments.
- **Grade and Interest Rate (-0.95)**: Lower grades are associated with higher interest rates, indicating higher risk.

b) Fully Paid

Installment and Loan Amount (0.93): Higher loan amounts result in higher installments**.**

Grade and Interest Rate (-0.95): Lower grades correlate with higher interest rates, indicating **higher risk loans**

Recommendations and Next steps

Recommendations:

1. Interest Rate Adjustments:

- Implement stricter criteria for loans with higher interest rates due to their strong association with defaults. Adjust the interest rate model to better reflect default risk.
- A further deep dive is needed to understand if high interest rate justify risker loans

2. Loan Grade Improvements:

- Loan grade is good indictor of default but is not always accurate. Enhance the loan grading system to more accurately capture default risk. Tighten criteria for lower grades and provide incentives for higher-grade loans.

3. Enhanced Data Collection and Analysis:

- Improve data collection methods to capture more granular information on borrower behavior and financial health. Regularly update the analysis to reflect changing trends and patterns in loan defaults.

Future Investigations:

1. High Interest Rates and High-Risk Loans:

- Further deep dive is needed to understand if high rewards from higher interest rates justify the increased risk of defaults.

2. Deep dive into loan Grade

- Current analysis did not explore relationship of grade with rest of the parameters. This analysis is needed to make recommendation to better loan application grading

Thank You