Unlocking Financial Insight: Data-Driven strategies for mitigating Loan default risk

Data Science MATH 1130: Final Project Report

By: Nehemiah Abdi,

Luca Vivona,

Minchan Chae,

Beamlak Melku,

XinFeng Teng

December 1, 2023

Overall Research Topic: Exploring Data patterns for smart lending decisions: Uncovering key information from consumer data to prevent loan default and maximize loan payments.

Subtopic 1 - Family status

Background/Introduction of problem:

One of the main questions we are trying to answer is whether there is a relationship between family status and loan payments towards loan providing companies. We hypothesize that people without any children are greatly represented in clients with On-Time payments and the people with children and providing for people are greatly represented in clients with late-Time payments.

Data Analysis Steps:

In this detailed credit EDA analysis, informed by extensive datasets from Venkatasubramanian Sundar Mahadevan, encompassing 'applications_data,' 'columns_description,' and 'previous_applications,' a thorough exploration of loan repayment nuances in relation to the borrower's family status was conducted. Graphing the value counts of 'NAME_FAMILY_STATUS,' which includes married, single/not married, civil marriage, separated, widow, and unknown, revealed that married individuals overwhelmingly represent the majority of borrowers. While exploring the distribution, the analysis utilized a bar graph to compare two distinct groups within the application data: those with children and those without. 'NAME_FAMILY_STATUS' was set as the dependent variable, and 'CNT_CHILDREN' as the independent variable, showcasing a high level of outliers in the number of children within each family status. Notably, individuals without children, particularly Widow, Unknown, and Single/Not Married, were found to be significantly overrepresented in this group.

To further understand the situation, data for plotting was derived from the frequency distribution of the number of children in a dataset named 'with_children.' Logarithmic transformations were applied for better visualization of the number of children, addressing the high level of outliers. Subsequent to data cleaning to address possible outliers, a chi-squared test was employed to rigorously examine payment difficulties. Hypotheses were framed under Case Hypothesis I and II, indicating a significant difference in the distribution of difficulties with installment payments between individuals with and without children, supporting the observation that families with children face more challenges in payments. Exploring the correlation between having children and loan approval, histograms were employed to provide a visual summary of the data's overall shape. Individuals without children exhibited lower approval rates, raising concerns about potential financial risks, particularly evident in cases where a person with children misses a payment. Application type analysis, comparing Approved, refused, canceled, and unused offers, revealed significant differences based on the presence of children. People without children tended to receive more approvals and fewer refusals, cancellations, and unused offers, contributing to a nuanced understanding of the intricate relationship between income, employment stability, family dynamics, and loan repayment behavior. This comprehensive approach, blending data science techniques and Case Hypothesis framework, validated initial hypotheses and provided insights into the unique challenges faced by families with children in loan repayment.

Results / Conclusions

Upon careful examination of various categories, it becomes apparent that individuals without children tend to receive a considerably low amount of approval with respect to their counterparts of people who have children. This observation raises the possibility of an underlying financial risk, especially as we show within our first histogram there is a higher likelihood in cases where a person with children misses a payment, as the dynamics are reversed in comparison.

Drawbacks of the Analysis Performed and Any Concerns:

Problem With Our Analysis:

Sample Size Discrepancy - The sizes of the two groups (individuals with and without children) may not be balanced. A significant difference in sample sizes can impact the statistical power of the chi-squared test and affect the reliability of the results.

Confounding Variables - Other relevant factors that could influence payment difficulties may not have been considered. For example, income, employment status, or other demographic variables might confound the relationship between having children and payment difficulties.

Causation vs. Correlation - Statistical association does not imply causation. While you may observe a significant difference in distributions, it doesn't necessarily mean that having children causes payment difficulties. There could be other external factors like economic downturns, losing of ones Job, Divorce, etc influencing the observed patterns.

Subtopic 2- Employment, Income conditions

Background/Introduction of the Problem:

In addressing the complications of loan repayment, we dive into the relationship between income levels, employment stability, and their impact on customer ability to make the ideal loan payment. Our hypothesis insists that during economic downturns, individuals without current employment or have a low income are more likely to face challenges in making timely payments on pre-existing loans. This exploration is key for financial institutions seeking to understand how income conditions and employment stability intersect, providing crucial insights for risk mitigation and refined lending strategies.

Data Analysis Steps:

In this detailed credit EDA analysis, informed by the extensive datasets from Venkatasubramanian Sundar Mahadevan, including 'applications_data', 'columns_description', and 'previous_applications', we employed a series of specific coding and visualization techniques to delve deeply into the nuances of loan repayment patterns in relation to employment and income levels. We began by extracting critical data such as 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', and 'DAYS_ID_PUBLISH', transforming these from days to years, a vital step for making the age and employment-related data more comprehensible. This transformation was achieved through coding that converted negative values to positive and subsequently into years, providing a clearer perspective on the demographics of the loan applicants. We then focused on analyzing the 'Occupation_Type' and 'AMT_Income_Total' data to determine the varying income levels across different occupations. This analysis revealed distinct patterns, such as low-income workers like low-skilled laborers and cleaning staff earning below $50,000 annually, in contrast to higher-income roles like managers and high-skill tech staff, earning above $75,000.

We further categorized the individuals into 'Low Income' and 'High Income' groups based on a $40,000 yearly salary threshold, a strategic decision that allowed us to focus on clear income delineations without a middle ground, which in turn facilitated a more targeted analysis. Using scatter plots, we explored the correlation between income and years employed, revealing a significant link between high income, longer employment duration, and fewer payment difficulties. This indicated that while long-term employment is beneficial, high income is a more decisive factor in loan repayment capabilities. We also employed bar plots to compare the frequency of late installments and loan approval rates among different income groups and occupations. These visualizations highlighted that high-skill tech staff, despite high

incomes, displayed a higher propensity for late payments, a pattern not observed in other high-income roles.

Furthermore, we created box plots to analyze customers with and without payment difficulties, comparing their income levels and credit amounts. This analysis was split between low and high-income groups, revealing that low-income individuals with payment difficulties often had lower credit amounts, suggesting that their financial challenges were not merely a function of high loan amounts but also related to their income levels. We also investigated the specific occupation types within these groups, using bar plots to display the frequency of previous loan application approvals. This detailed examination showed that high-skill tech staff, despite being high earners, faced more payment difficulties compared to other occupations, an insight that could be pivotal for financial institutions in evaluating loan applications.

In summary, through a series of calculated data transformations, visualizations, and statistical analyses, we were able to extract profound insights from the credit EDA datasets. Our approach highlighted the critical interplay between income, employment duration, and loan repayment behavior, providing a nuanced understanding of the factors influencing loan repayment risks. This analysis not only underscores the importance of considering income and occupation type in loan approval processes but also reveals the complex dynamics within different income and occupational groups, offering valuable perspectives for financial risk assessment and strategy development.

Results/Conclusions:

The analysis of the relationship between income levels, employment stability, and loan repayment ability reveals key insights: High-income individuals, particularly in stable, long-term employment, generally show better loan repayment behavior. However, low-income workers, regardless of their employment duration, face more challenges in timely loan repayments. Interestingly, high-skilled technology staff, despite their high incomes, show a tendency for late payments, indicating that factors beyond income and employment duration might influence loan repayment patterns. This study suggests that while income is a crucial factor in assessing loan repayment risks, financial institutions should also consider unique occupational trends and outliers in their lending strategies.

Drawbacks of the Analysis Performed and Any Concerns:

Problem With Our Analysis:

Interpreting Missed Payments as Early Warning Signs: Our analysis presupposes that missing a payment is a reliable early warning sign, but this assumption may oversimplify the issue. Various external factors, beyond an individual's control, could contribute to missed payments, challenging the accuracy of this indicator.

Fixed Parameters in Analysis: Our analysis relies on fixed parameters, potentially overlooking dynamic factors such as changes in living expenses, access to credit, and individual financial management skills. The complexity of these variables could confound the relationships we are attempting to assess.

Limited Sample Representativeness: It's important to acknowledge that our sample, derived from a single bank's dataset, may not be fully representative of the broader population. Generalizing findings based on this limited scope might not accurately reflect the diverse financial behaviors present in the wider community.

Granularity Sacrificed in Threshold Definition: Establishing a threshold for defining high income, such as using the mean, introduces a trade-off by sacrificing granularity in

the information. Aggregating beyond this threshold may obscure nuanced variations in income levels, limiting the depth of our insights.