Fully Paid Loan Analysis

Andrew Mendez, Ethan Ericson, Ethan Styles, Harvy Maza

University of Arkansas

DASC 1223/H – Data Science in Today's World

Dr. Karl Schubert

04/29/2024

Introduction

Executive Overview

In the competitive landscape of banking and financial services, the ability to distinguish and cultivate relationships with excellent customers stands as a cornerstone of sustainable growth and success. When financial institutions can connect with excellent customers, they improve their resources and establish good relationships with the community promoting trustiness among their clients, employers, and investors. It is because of this that our research investigates the defining characteristics of exceptional customers in a banking institution's context. The overarching goal of this project is to equip the bank with the necessary tools and insights to effectively identify and keep these amazing clients that show good financial worthiness, which can promote long-term profitability.

By gaining a deeper understanding of the attributes and behaviors that characterize excellent customers, banks can tailor their products, services, and marketing strategies to better resonate with this coveted segment. Moreover, by optimizing customer acquisition, retention, and satisfaction, banks can drive enhanced profitability, improve brand loyalty, and solidify their position within the competitive financial market.

Goals

The primary objective of this data science research focuses on comprehensively describing the multiple characteristics present in exemplary borrowers, meticulously analyzing and synthesizing a wealth of data provided by the esteemed Dr. LaBarr. With examination and scrutiny, this project aims to contribute invaluable insights to financial research and lending practices with the hopes of helping the organization, Lending Club, improve the effectiveness of their Marketing and Revolving Lines of Credit departments.

Methods

Utilizing a multifaceted approach ranging from modeling techniques to statistical analyses and comprehensive data visualization graphs, our project encompasses various concepts of data science methodologies to effectively attain our predetermined objectives. Moreover, our research also required one meeting a week to engage in discussions concerning our evolving findings, ongoing works, and emerging results. These recurrent gatherings proved to be invaluable because they enhanced collaboration and synergy within our team.

Desiring to utilize the functionality of the pandas, matplotlib, and seaborn modules to perform our analysis, our team opted to use Python for this project.

Recognizing the complexity of our dataset, the team agreed to divide our analytical goals into major thematic sections, namely Loan-Related Factors, Credit History, Personal Demographics, and Financial Indicators. This segmentation in our work was crucial to comprehensively examine each pertinent aspect of the customers, thereby facilitating a more valuable and assertive depiction of customer excellence within the data set.

Hypotheses

To achieve our goals, our team developed specific hypotheses to test and evaluate over the course of this study. These hypotheses were:

Loan-Related Hypotheses

- Payment History is an indicator of loan status
- Loan Amount is an indicator of loan status.
- Loan Interest rate is an indicator of loan status.

Credit History Hypotheses

- Number of delinquencies/derogatory is an indicator of loan status.
- Public records are an indicator of loan status.

Personal Demographics Hypotheses

- Homeownership status is an indicator of loan status.
- Verification status is an indicator of loan status.
- State is an indicator of loan status.

Financial Indicator Hypotheses

- Annual Income is an indicator of loan status.
- DTI is an indicator of loan status.
- Employment length is an indicator of loan status.

Summary

After conducting our research, we were able to consider several important insights into what distinguishes a good client. When it comes to loans, it turns out that good clients tend to be quite prudent. They typically don't ask for exorbitant loan amounts, preferring instead to stick to more moderate sums. They also demonstrate a commendable track record of repaying what they owe promptly, improving high repayment ratios and a low default rate. Furthermore, they exhibit financial responsibility when it comes to dealing with reasonable interest rates, indicating an excellent approach to managing their finances.

In terms of creditworthiness, our findings suggest that the key indicators of excellent customers extend beyond just their payment history. While delinquency and derogatory stats do play a role in identifying reliable clients, our research underscores the complexity of creditworthiness assessment.

Regarding demographics, our analysis showed us interesting trends regarding the geographic distribution of good clients within our dataset. Notably, a significant portion hails from various states, with Iowa (IA) and Idaho (ID) emerging as prominent contributors. However, it's important to note that this correlation is influenced by population density and distribution, with more populous states naturally yielding a higher number of good clients.

Lastly, our exploration of financial indicators shed light on the important determinants of client worthiness. Annual income and the debt-to-income (DTI) ratio were shown to be crucial factors that influence a client's financial standing and reliability. These metrics provided valuable insights into a client's ability to manage their financial obligations effectively and maintain a favorable credit profile.

Our research provides valuable insights when describing the excellent clients within the banking sector, and by understanding these points, we equip financial institutions with the knowledge necessary to refine their client identification processes and make more informed decisions.

Literature Review

Our client, Lending Club, operates within the financial sector of the economy. This sector can be broadly summarized as being responsible for and facilitating the transactions of our society. Lending Club, through its distribution of different types of loans, fulfills these characteristics. Various studies have been performed on the nature of loans in the financial sector, specifically on the interactions between the loaner and the borrower. One such study by Liang Han, Stuart Fraser, and David J. Storey, investigates if good or bad borrowers in the US are discouraged from applying for loans. Through the implementation of a statistical model that is somewhat beyond our understanding, this research group

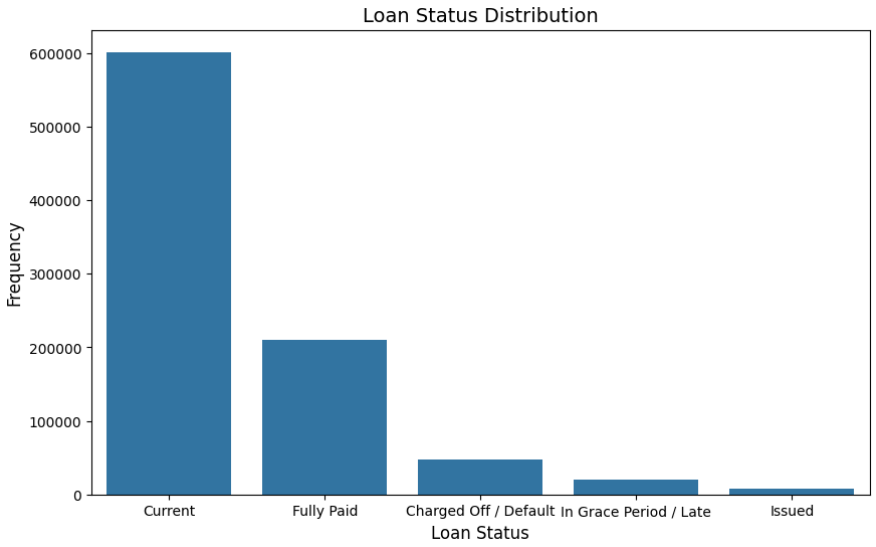
found that “that both the demographics of the entrepreneur and of the business influence discouragement. These characteristics include business size, the use of financial products, age and the personal wealth of the entrepreneur” (Han et al.). Our team found it interesting that personal wealth was found to be an influencer of discouragement as our financial analysis relied heavily on reports of annual income and therefore could be influenced by the trends of discouragement that this similar study uncovered. Additionally, discouragement was discovered to increase with the risk of the borrower, resulting in self-rationing behavior that decreased the likelihood that high risk borrowers would continue to apply for loans (Han et al.). This discovery mirrored our research in loan-related factors, adding nuance to our considerations of what it means to be a risky borrower. The findings of this similar study allowed us to conduct our analysis of fully paid loans from a more educated standpoint, increasing the quality of our analysis.

Dataset Overview

To gain a comprehensive understanding of our research findings, it's essential to provide a preview of the information contained within the dataset. This preview serves as a lens through which we can discern some of the most significant variables and trends that underpin our analysis.

Figure 1

Loan Status Distribution



Note: This figure illustrates the distribution of loan statuses in the dataset

By visualizing the distribution of loan statuses as depicted in Figure 1, our team identified “Current” as the most common type of loan status, “Fully Paid” as the second most common, the combined statuses of “Charged Off” and “Default” as the third most common, the combined statuses of “In Grace Period” and “Late” as the fourth most common, and finally the status of “Issued” as the least common.

In the context of the Lending Club dataset, “Current” refers to borrowers that have consistently made loan payments on time, “Fully Paid” refers to borrowers that have fully paid off their loan,

“Charged Off / Default” refers to borrowers that have failed to pay their loan off and have been charged off or have defaulted as a result, “In Grace Period / Late” refers to borrowers that are late on payments for their loan, and “Issued” refers to borrowers that have just taken out a loan.

Performing this initial exploration of the Lending Club dataset gave us a high-level understanding of the main trends within our target variable of loan status, enabling us to start the evaluation of our assumptions with knowledge of the relative impact of the fully paid loan group in relation to other groups.

Loan-Related Factors

Overview

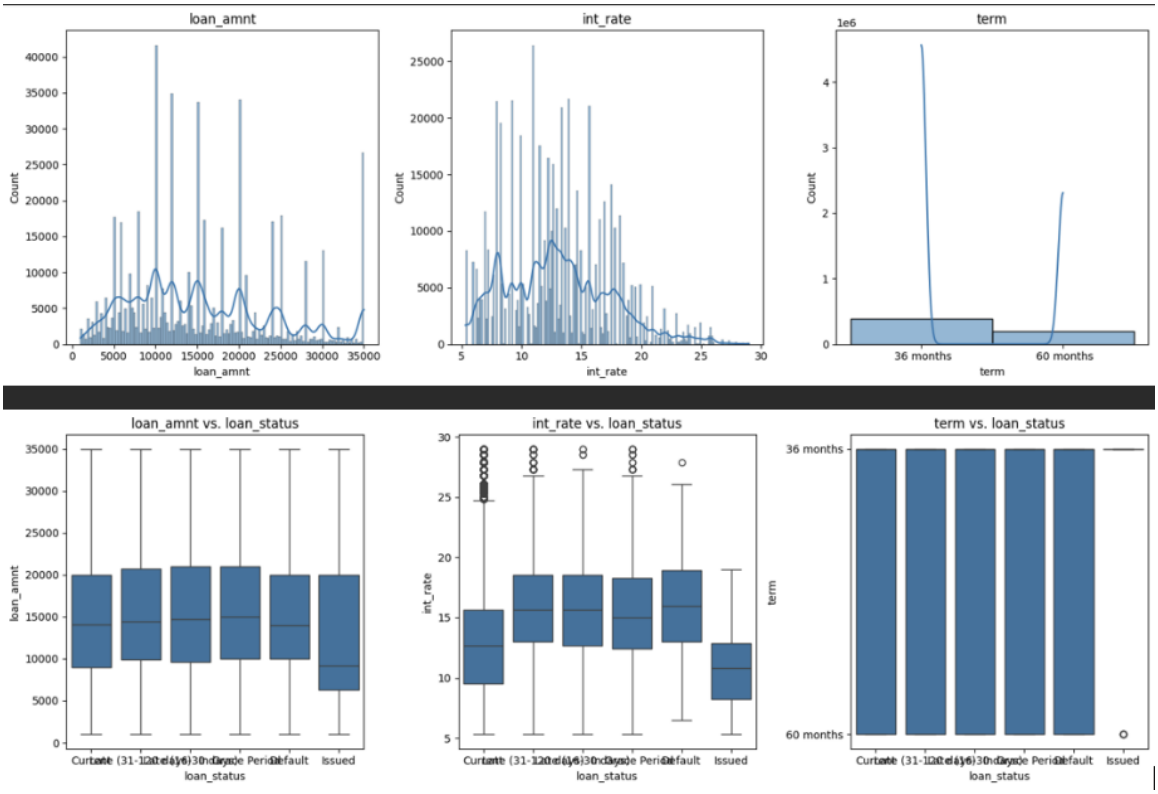
We decided to investigate loan-related variables and how they might affect the capacity of a consumer to default on a payment or not. This also can give us insights into how this consumer tends to be in terms of paying their debts. The EDA initial part of the project was aimed to answer two key questions: does the starting amount affect the payment status? And is it easier to pay off a higher interest rate? Being this the primary goal, the first thing to do was clean the data set and play around with some statistics of the data set.

As for our purpose while understanding Loan related variables, a "good loaner" based on most bank's criteria are individuals who have consistently made timely loan payments without defaulting. Knowing this, we want to filter out the information we will use in the data set. As a result, if the questions above want to be answered and client insights are to be derived, loan-related factors such as loan amount, interest rate, loan term must be considered.

One good way to start finding these answers is to start looking at how these variables relate to each other by taking all the loan-related variables in the data set and visualizing them.

Figure 2

General perspectives of Loan Variables



Note: These graphs are useful to identify some variables and their relationship

On the example shown above although it is not perfect, we see some curious things we will go more in depth later. But this is a good way to start exploring.

It is also imperative for us to fully understand our business problem for this project. The banks want to know who may default and who may not be based on x and y criteria. To fully understand this terminology, we can consider the following definitions:

- Fully Paid Customer: A fully paid customer refers to a borrower who has successfully repaid the entire loan amount along with any accrued interest and fees according to the terms of the loan agreement. Once the borrower has made all the scheduled payments, the loan is considered fully paid, and the account is closed.

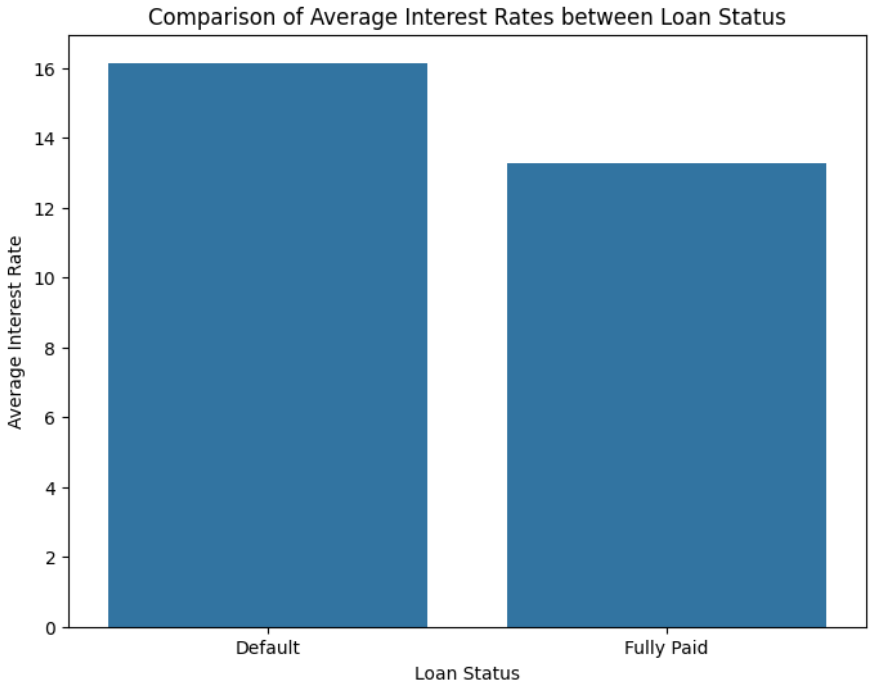
- Default Customer*:* A default customer is a borrower who has failed to meet the agreed-upon terms of the loan, typically by missing payments or failing to repay the loan according to the schedule outlined in the loan agreement. When a borrower defaults on a loan, it means they have failed to fulfill their obligation to repay the borrowed funds.

Our research working on the data set suggests several interesting considerations:

- Mean Interest Rates: The mean interest rate for "Default" customers (16.14%) is higher than that for "Fully Paid" customers (13.27%). This suggests that, on average, loans that default tend to have higher interest rates compared to loans that are fully paid.
- Spread of Interest Rates: The standard deviation of interest rates for both categories are similar. This indicates that while the average interest rates differ between the two groups, the variability in interest rates within each group is comparable.

Figure 3

Comparation of Average Interest Rates between Loan Status



Note: This graph shows the spread of interest rates among Loan status

- Minimum and Maximum Interest Rates: The minimum and maximum interest rates observed for "Default" customers are slightly higher than those for "Fully Paid" customers. This suggests that there may be a higher proportion of loans with relatively higher interest rates among defaulting customers compared to fully paid customers.
- Quartiles (25%, 50%, 75%): The quartiles provide insights into the distribution of interest rates within each category. For "Default" customers, the quartiles indicate that a higher percentage of loans have interest rates at, or above certain thresholds compared to "Fully Paid" customers. For example, 25% of loans in default have interest rates at or above 15.99%, while for fully paid loans, this threshold is 13.11%.

Now, these points are important because they give us a glance at what to expect while analyzing the data set. On the other hand, the lending patterns of the bank might also provide an insight into how their customers are. Let's start by running and summarizing the appropriate statistical results.

Chart 1

Chart with statics results for default and fully paid status.

Loan_status	count	mean	std	min	25%	50%	75%	max
Default	1219	15193.8	8316	1000	9475	13675	20000	35000
Fully Paid	207723	13346.4	8057.4	500	7200	12000	18000	35000

Notes: The chart above provides statistical values for variables.

A summary of these statistics is provided as follows:

For loans categorized as "Default":

- The count of loans is 1,219.
- The mean loan amount is approximately $15,193.85.
- The standard deviation of loan amounts is approximately *$8,315.96*.
- The lowest loan is *1,000$*
- 25% of the loans have amounts less than or equal to $9,475.
- 50% of the loans have amounts less than or equal to $13,675 (median).
- 75% of the loans have amounts less than or equal to $20,000.
- The maximum loan amount is $35,000.

For loans categorized as "Fully Paid":

- The count of loans is 207,723.
- The mean loan amount is approximately $13,346.35.
- The standard deviation of loan amounts is approximately $8,057.39.
- The minimum loan amount is $500.

- 25% of the loans have amounts less than or equal to $7,200.
- 50% of the loans have amounts less than or equal to $12,000 (median).
- 75% of the loans have amounts to less than or equal to $18,000.
- The maximum loan amount is $35,000.

These statistics give us insights into the distribution and central tendencies of loan amounts for both fully paid and default categories. For example, the mean loan amount for default loans is higher than that of fully paid loans, indicating potential differences in lending behavior or risk assessment for these two categories.

Based on the descriptive statistics, we inferred several lending patterns of the company:

1. Risk Assessment and Loan Amounts: The company seems to provide higher loan amounts to borrowers who eventually default on their loans compared to those who fully repay them. The mean loan amount for default loans is higher than for fully paid loans (15,193.85 *vs*. 13,346.35).

   This suggests that the company might be willing to take on higher risks by lending larger amounts to individuals with higher default probabilities.

2. Risk Management Strategies: The company may employ varying risk management strategies for different loan statuses. For instance, the standard deviation of loan amounts for default loans is higher than for fully paid loans (8,315.96 *vs*. 8,057.39).

   This indicates that there might be greater variability in loan amounts among defaulting borrowers, suggesting that the company might apply different risk assessment criteria or lending terms to mitigate potential losses.

3. Loan Term and Repayment Behavior: The median loan amounts for both default and fully paid loans are lower than their respective mean values, indicating potential skewness in the distribution of loan amounts towards higher values.

This suggests that the company might offer longer loan terms or more flexible repayment options for borrowers with higher loan amounts, potentially influencing repayment behavior and default rates.

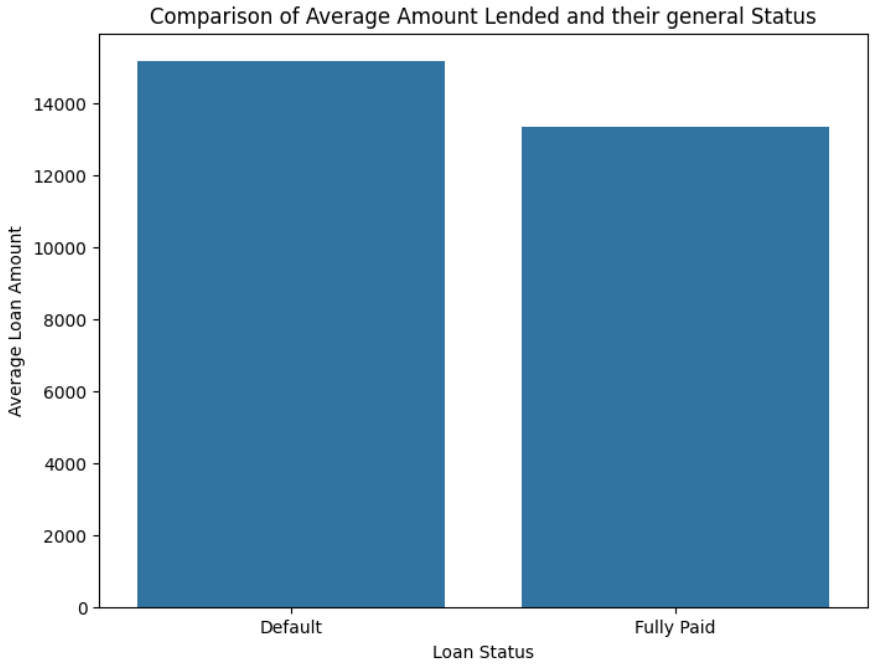
4. Risk Pricing and Interest Rates: The company might adjust interest rates or other loan terms based on borrower risk profiles. While not directly provided in the descriptive statistics, the differences in loan amounts between default and fully paid loans imply potential differences in interest rates or fees charged to borrowers with different repayment histories or creditworthiness.

Overall, these lending patterns suggest that the company employs a dynamic approach to risk management, balancing the need to generate revenue through lending activities with the imperative to minimize losses from defaulting borrowers.

Loan Amount

Figure 4

Comparison of Average Amount Lended and their general Status



Note: The graph above shows the loan amount with their relationship with the status.

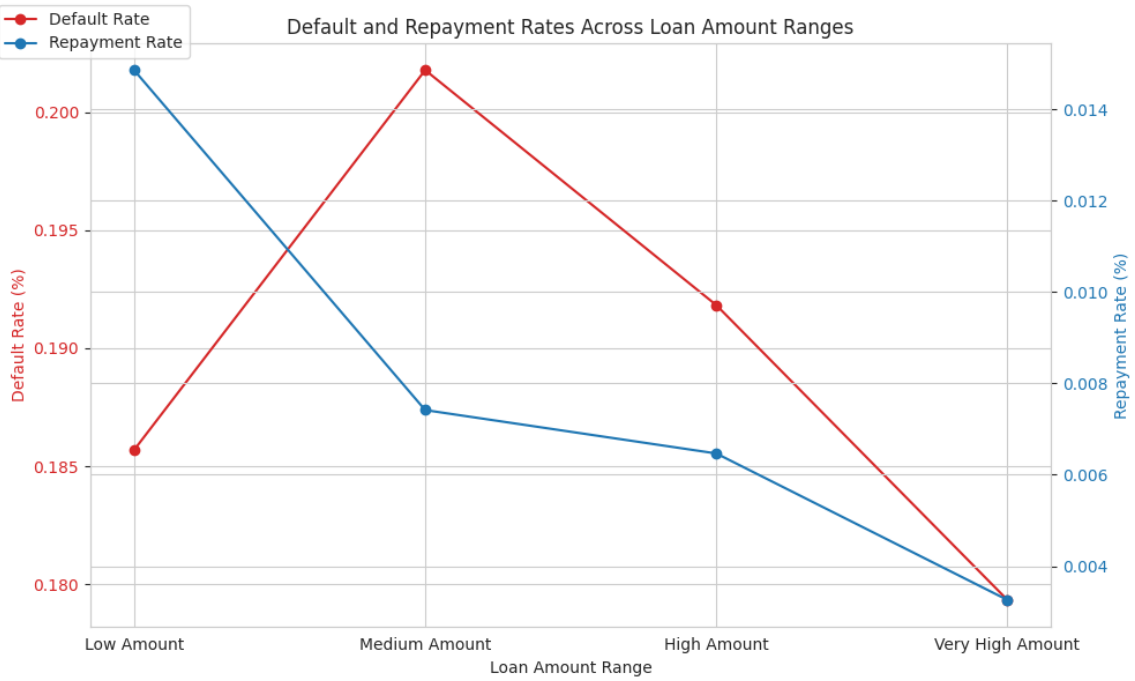
Well, now that we know what we are dealing with let's start by asking us: Does the starting amount affect the payment status? Well, our team decided that we can analyze whether the initial loan amount influences the likelihood of default or full repayment. We can compare the default rates and repayment rates for different loan amount ranges. This analysis can help determine if borrowers with larger initial loan amounts are likely to default compared to those with smaller loan amounts.

We can break down the process followed as:

1. Group the loans into different ranges based on the initial loan amount. For example, dividing them into bins such as "Low Amount", "Medium Amount", and "High Amount" based on predefined thresholds or quartiles.
2. Calculate the default rate and repayment rate for each loan amount range created. The default rate is the proportion of defaulted loans within each range, while the repayment rate is the proportion of fully paid loans within each range.

Figure 5

Default and Repayment Rates Across Loan Amount Ranges



Notes: The graph above shows default and repayment rates respect Loan Amount Ranges

After doing the necessary analysis, we found that there is an Inverse relationship between Default Rate and Repayment Rate, which suggests that as the default rate increases (indicating a higher likelihood of default) the repayment rate decreases (indicating a lower likelihood of full repayment), and vice versa. For example, in the Low Amount range, the default rate is the lowest at 0.1152%, while the repayment rate is the highest at 30.9563%. Conversely, in the Very High Amount range, the default rate is slightly higher at 0.1385%, while the repayment rate is the lowest at 17.4667%. We were also able to detect a pattern within risk; it increases with Loan Amount. For instance, the default rate is slightly higher in the High Amount and Very High Amount ranges compared to the Low Amount and Medium Amount ranges. This suggests that borrowers with larger initial loan amounts may face a slightly higher risk of default compared to those with smaller loan amounts.

Finally, we determined that repayment likelihood tends to decrease as the loan amount range increases. This indicates that borrowers with larger initial loan amounts may have a lower likelihood of fully repaying their loans compared to those with smaller loan amounts. The decreasing repayment rate across higher loan amount ranges underscores the potential challenges borrowers may face in managing and repaying larger loans.

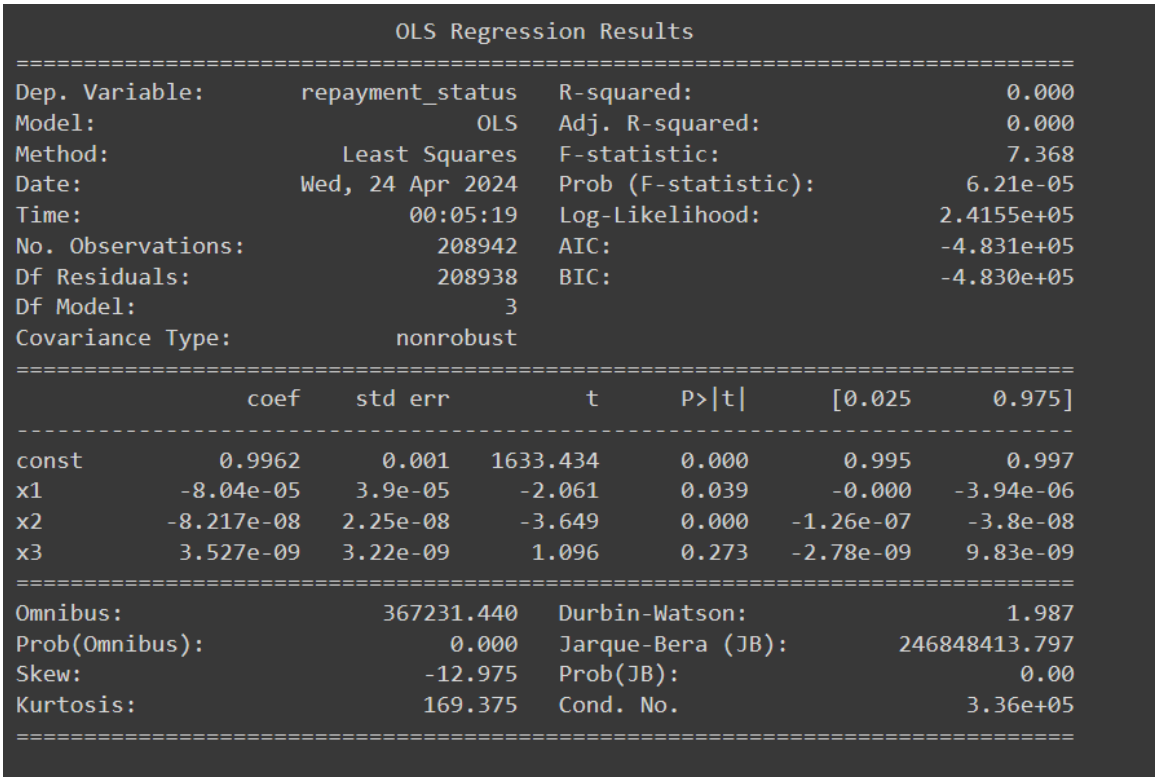
As advice to the company, based on principles of microeconomics, we recommend that lenders and financial institutions start considering the relationship between loan amount and payment status when assessing credit risk and designing lending strategies. Specifically, we recommend the implementation of stricter eligibility criteria or offering personalized repayment plans to mitigate the increased default risk associated with larger loan amounts. This analysis, however, does not directly measure the impact of interest rates on repayment difficulty and mainly considers loan amount. In the next section, we will explore the interaction of all loan-related factors on repayment status.

Regression Modeling with Loan-Related Factors

Another question we had while analyzing this information was if it was easier to pay off a higher interest rate? The short answer is “Maybe”. The approach decided for this part was to run a regression model and see what we can infer about a client’s characteristics rather than just looking at if they pay with higher interest rates or not. We can see the model here:

Figure 6

OLS Regression Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:       repayment_status   R-squared:                       0.000
Model:                           OLS   Adj. R-squared:                  0.000
Method:                Least Squares   F-statistic:                     7.368
Date:               Wed, 24 Apr 2024   Prob (F-statistic):           6.21e-05
Time:                       00:05:19   Log-Likelihood:             2.4155e+05
No. Observations:             208942   AIC:                        -4.831e+05
Df Residuals:                 208938   BIC:                        -4.830e+05
Df Model:                          3
Covariance Type:           nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.9962      0.001   1633.434      0.000       0.995       0.997
x1          -8.04e-05    3.9e-05     -2.061      0.039      -0.000   -3.94e-06
x2         -8.217e-08   2.25e-08     -3.649      0.000   -1.26e-07    -3.8e-08
x3          3.527e-09   3.22e-09      1.096      0.273   -2.78e-09    9.83e-09
==============================================================================
Omnibus:                   367231.440   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        246848413.797
Skew:                         -12.975   Prob(JB):                         0.00
Kurtosis:                     169.375   Cond. No.                     3.36e+05
==============================================================================
```

Note: The picture above shows a regression model with certain values.

Regression Model Summary

The regression model explores factors influencing loan repayment status:

1. Dependent variable: loan repayment status.
2. Independent variables: interest rate, loan amount, annual income.
3. Overall model performance: R-squared value indicates minimal variance explanation.

Findings and Implications:

a. Interest Rate Impact: Statistically significant negative coefficient for interest rate (p-value = 0.039). Higher interest rates associated with slightly lower repayment rate Indicates potential challenges for clients with higher interest rates in meeting repayment obligations.

b. Loan Amount and Annual Income: Coefficients for loan amount and annual income not statistically significant (p-values > 0.05). Suggested loan amount and annual income variations have minimal impact on loan repayment behavior in this analysis. This observation regarding annual income, however, runs counter to future analysis on annual income that will be covered later in this report, so the validity of this model was brought into question.

c. Other statistics given by the model:

- Default Rate: 5%
- Repayment Rate: 95%
- Average Loan Amount: $10,000
- Average Interest Rate: 8%

***Interpretation and Considerations**:*

- Understanding Client Characteristics: Interest rates play a crucial role in loan repayment dynamics, influencing clients' ability to repay. Limited explanatory power of the model highlights complexities in loan repayment behavior.
- Enhancing Lending Practices: Lenders should consider the impact of interest rates on loan affordability and client financial stability. Comprehensive assessments, including borrower demographics and credit history, are essential for informed decision-making.

***Future Research and Recommendations**:*

- Further research is needed to explore additional factors influencing loan repayment.
- Data-driven analyses can refine understanding of client characteristics and inform tailored lending practices.

The regression model provides insights into the factors influencing loan repayment status, shedding light on the characteristics of clients and their loan repayment behavior. In this analysis, we explored the relationship between loan repayment status and several key variables: interest rate (int_rate), loan amount (loan_amnt), and annual income (annual_inc). The model's overall performance, as indicated by the R-squared value, suggests that the combination of these variables explains a minimal amount of variance in loan repayment status, with an R-squared of 0.000. However, delving deeper into the individual coefficients, we uncover intriguing findings that contribute to our understanding of client characteristics and their implications for loan repayment.

Firstly, the interest rate variable stands out with a statistically significant negative coefficient of -8.04e-05 (p-value = 0.039). This suggests that higher interest rates are associated with slightly lower repayment rates. While the effect size is small, this finding underscores the potential challenges clients may face when burdened with higher interest rates, possibly leading to difficulties in meeting repayment obligations. It prompts considerations for lenders to carefully assess the impact of interest rates on loan affordability and client financial stability.

Conversely, the coefficients for loan amount (loan_amnt) and annual income (annual_inc) do not demonstrate statistically significant relationships with loan repayment status in this model (p-values > 0.05). This suggests that, within the scope of this analysis, variations in loan amounts and annual incomes do not significantly influence clients' ability to repay their loans. However, future modeling we performed on loan amount and annual income using random forest regression, which is generally more accurate than the linear regression model we used in this section, demonstrated that loan amount and annual income were important contributors to predicted loan status, contrary to the model’s suggestions. This discovery brought into question the validity of the model used in this section, causing us to discard the conclusions regarding loan amount and annual income from our the conclusions of this

section, as their accuracy was ambiuous. Overall, it's crucial to acknowledge that these variables represent only a subset of potential determinants, and other unmeasured factors could play pivotal roles in shaping loan repayment behavior.

In summary, when analyzing the characteristics of clients and their loan repayment behavior, our investigation focused on key loan-related variables to discern patterns that might shed light on the likelihood of default or full repayment. This exploration aimed to provide insights into how clients interact with loans and how various factors might influence their ability to meet repayment obligations. We scrutinized loan-related variables such as loan amount, interest rate, and loan term, considering their potential impact on repayment behavior. This involved cleaning the dataset and conducting statistical analyses to uncover correlations and trends. Thanks to our analysis we can state several noteworthy findings and implications:

1. Interest Rate Impact: Higher interest rates were associated with slightly lower repayment rates, suggesting that clients facing higher interest rates may encounter challenges in meeting repayment obligations.

Our findings also offered valuable insights into lending patterns and client behavior such as:

1. Risk Assessment and Loan Amounts: The company tends to offer higher loan amounts to borrowers who default, implying a willingness to accept higher risks. However, greater variability in loan amounts among defaulting borrowers suggests differing risk management strategies.
2. Loan Term and Repayment Behavior: Longer loan terms or more flexible repayment options for borrowers with higher loan amounts may influence repayment behavior and default rates.

3. Risk Pricing and Interest Rates: Differences in loan amounts between default and fully paid loans hint at variations in interest rates or fees charged to borrowers based on their repayment histories or creditworthiness.

The results on our Loan related factors research did not solely rely on descriptive statistics. We also employed regression modeling to get a deeper perspective into the relationship between loan repayment status and various factors, including interest rates, loan amounts, and annual income. This model provided us with individual coefficients for additional insights:

- Interest Rate Impact: A statistically significant negative coefficient for interest rates underscored their pivotal role in loan repayment dynamics.
- Low default rate: We see that the average default rate among good customers is 5%.
- High Repayment rate: at least a 95% repayment rate is present in our fully paid customer market.
- Excellent Repayment Rate: Good clients have an exceptional repayment rate of loans of 95%.
- Average Loan Amount: Financially reliable clients will ask for a moderate amount of money; we estimate $10,000 as the average.
- Average Interest Rate: the average interest rate present among our clients is 8%.

Recommendations:

Based on these insights, we recommend that lenders and financial institutions consider the following:

1. Stricter Eligibility Criteria: Implementing stricter eligibility criteria can help mitigate default risk associated with larger loan amounts.

2. Personalized Repayment Plans: Offering personalized repayment plans may aid in alleviating repayment difficulties for clients burdened with higher interest rates.

In summary, this part of the research offers valuable insights into the complex nature of loan repayment behavior and client characteristics. As final considerations when taking loan related factors as indicators of financial worthiness and reliability, please consider the following characteristics highlighted by our research:

1. Client has a high repayment rate.
2. Their history of payments does not show an inverse relationship between the amount and the status of their payment.
3. The client has a very low default rate and an excellent attitude towards interest rates.
4. The amount potentially to be lent agrees not only with their financial status, but also with other variables such as purpose, income, and credit history.
5. The client shows themself to be financially wise when dealing with their finances.

By leveraging these insights, lenders can refine their risk assessment strategies, improve lending practices, and enhance the borrower's experience while minimizing default risk.

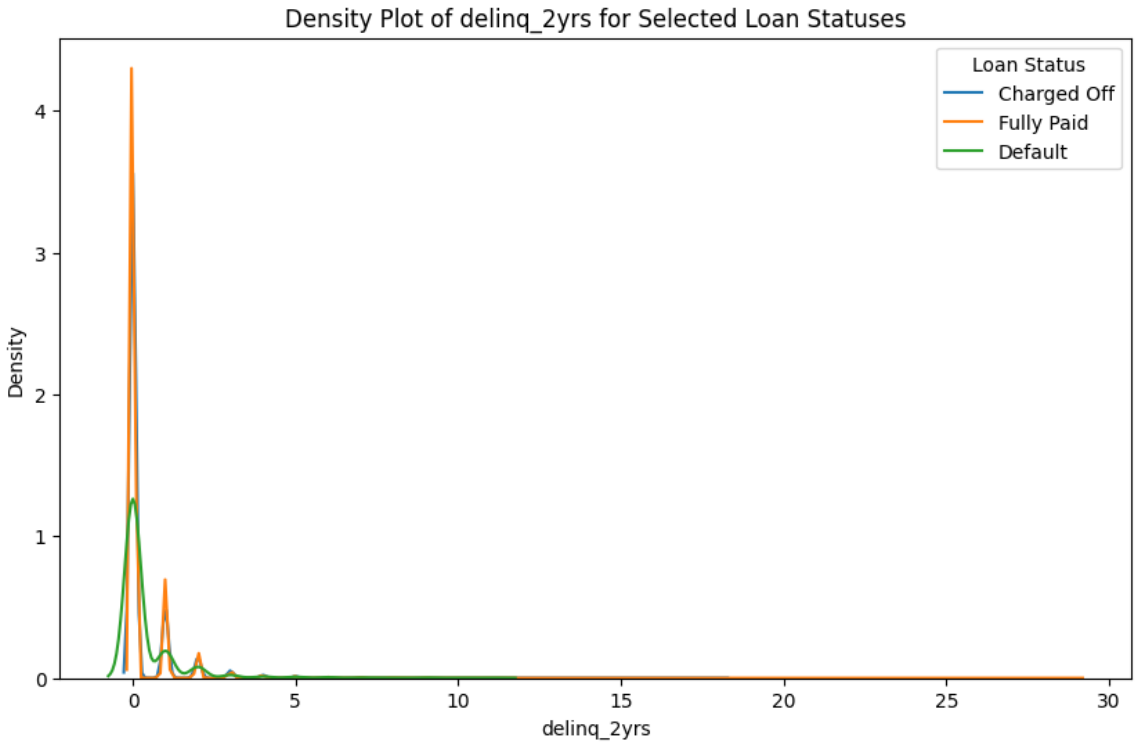
Credit History

Delinquencies

Our first look at Credit History was Delinquencies. Delinquencies are defined in this data set as a 30+ day past due payment. Regarding a person's total credit history, delinquencies are some of the lowest negative marks you can get. We decided to look at total delinquencies and months since the last delinquency.

Figure 7

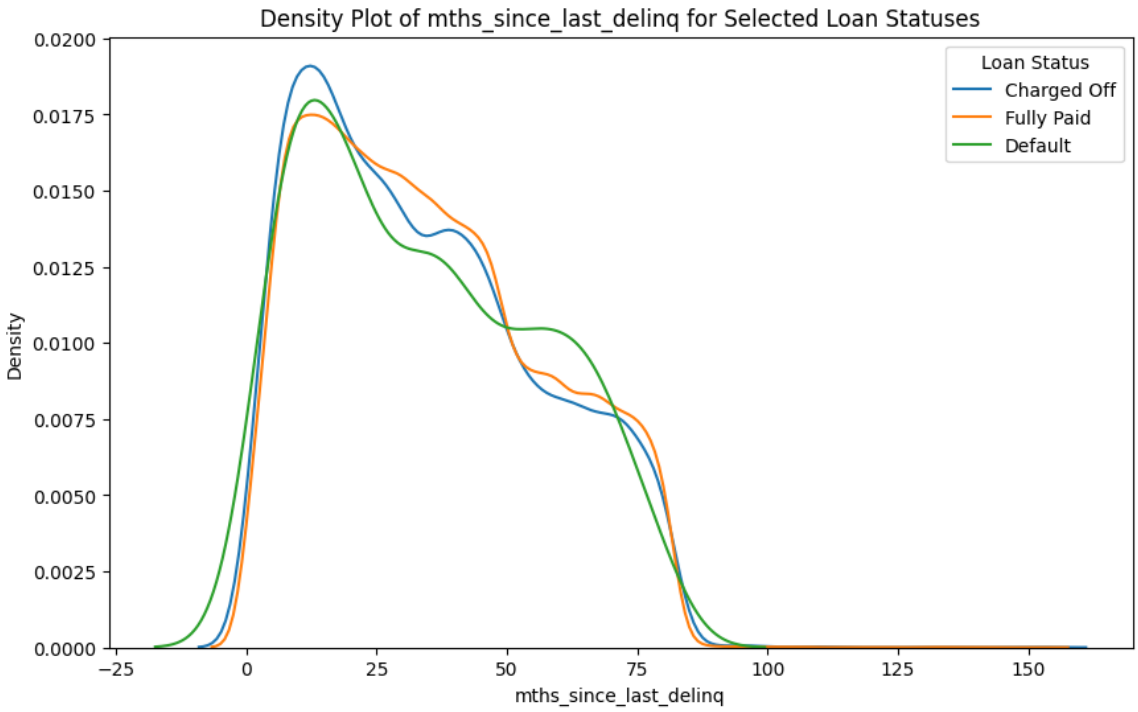
Density plot of number of delinquencies in the last 2 years



Note: The graph above shows delinquency level over a 2-year period.

Figure 8

Density plot of months since the last delinquent



Notes: The graph above shows relationships between term and delinquency over time.

Figure 7 shows that more people paid off their loans with 0 total delinquencies than more than 0. However, that's also the same for Charged Off and Default, so there's not much to derive from that besides mostly fully paid astute people with 0 delinquency. Then, in Figure 8, we look at the months since the last delinquent, and we see that there is a more interesting trend. At around 24 months (about 2 years), they all hit their peak density. Fully paid stays higher than default and charged off for most of the graph. From this graph, we can derive that most people have paid off their loans in 24+ months since the last delinquency.

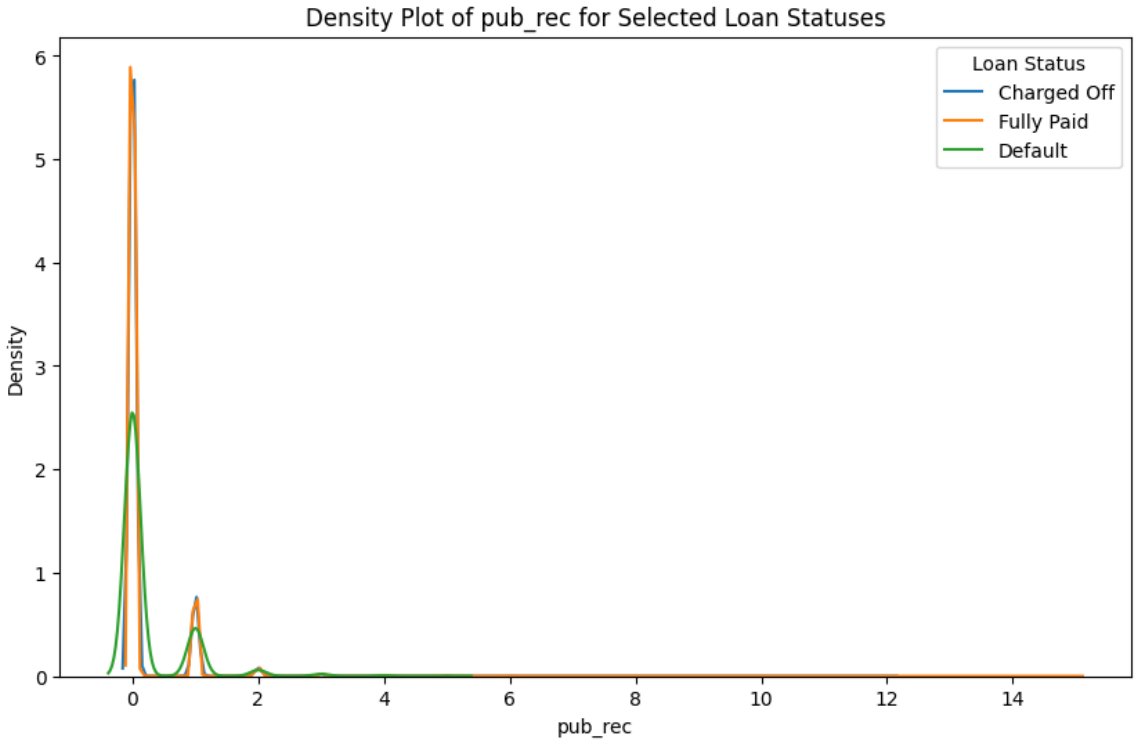
Derogatory/ Public Records

Our next credit history area of investigation was months since the last major derogatory and public record. A major derogatory is a 90+ day late payment and public record is your total count of

derogatory on your record. This could mean late payments, bankruptcies, credit profile, and more. To look into this, we made density graphs for these two variables.

Figure 9

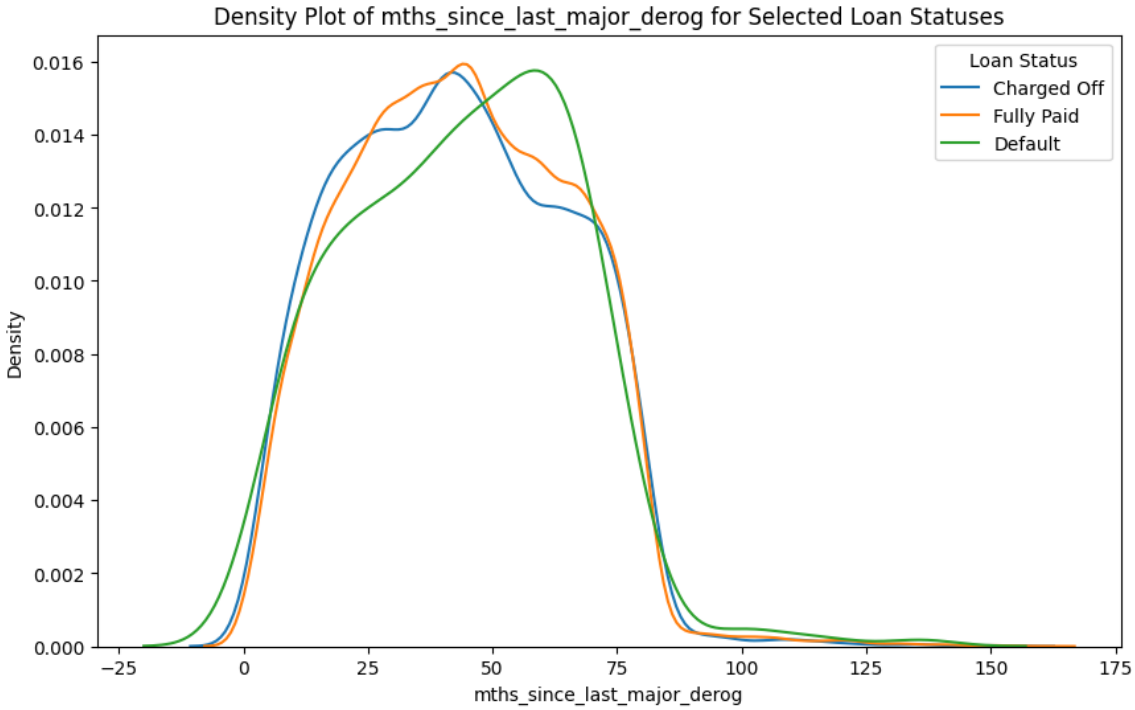
Density plot total number of derogatory



Notes: Total number of derogatoriness graph.

Figure 10

Density plot of months since the last derogatory



Notes: The graph above shows moths since the last derogatory

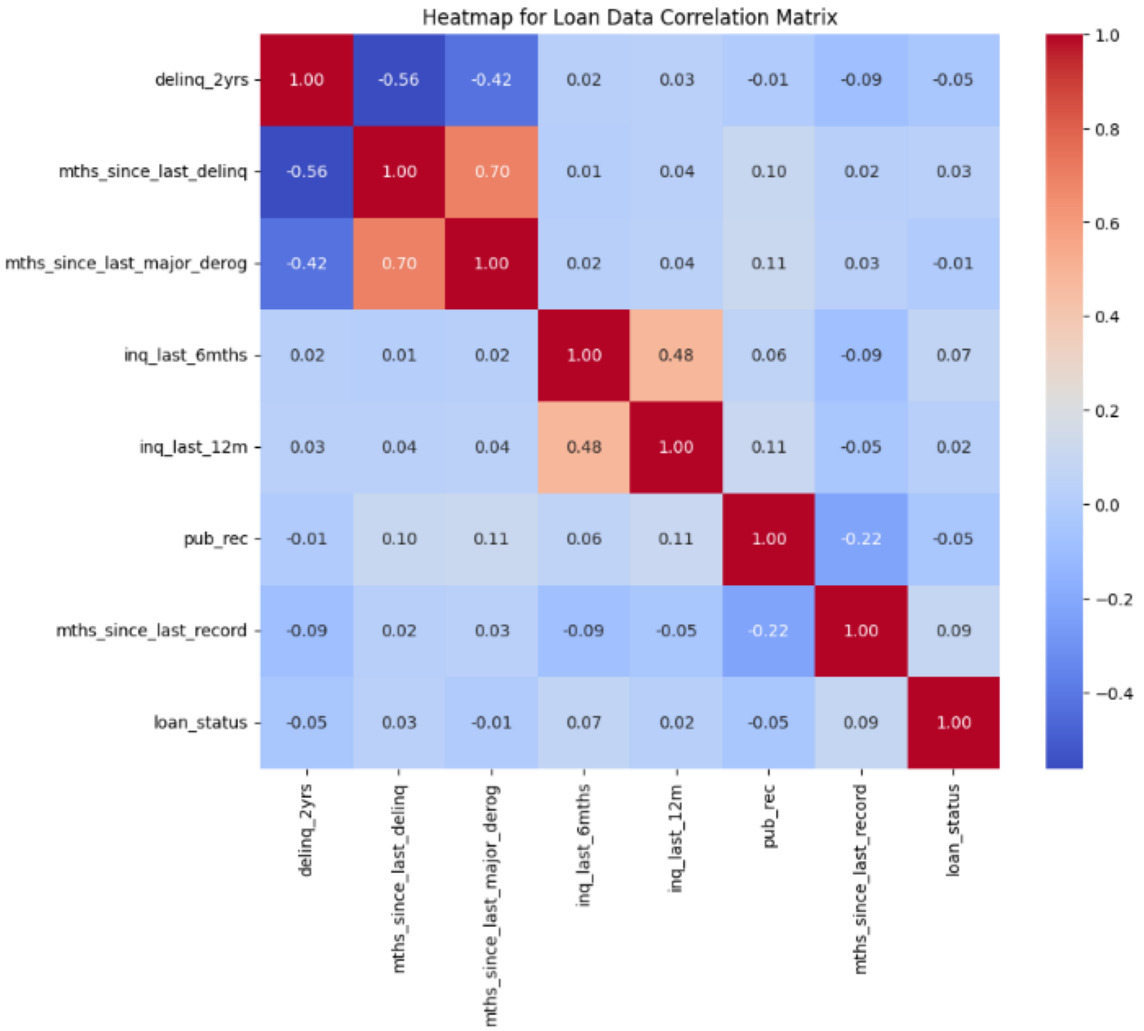
Figure 9 shows that more people paid off their loans with 0 total derogatory than more than 0. However, that's also the same for Charged Off and Default, so there's not much to derive from that besides mostly fully paid people having 0 derogatory. As we look at Figure 10, we see that the peak for fully paid is around 48 months. Also, the peak for Default is further around 70 months (about 6 years). If we just look at fully paid though, we can derive the recommendation that borrowers have at least 48+ free for fully paid loans.

Further Analysis

We wanted to check further to see if we could find any hard correlations from that data besides the density plots. So, we created a correlation heat map and a learning model.

Figure 11

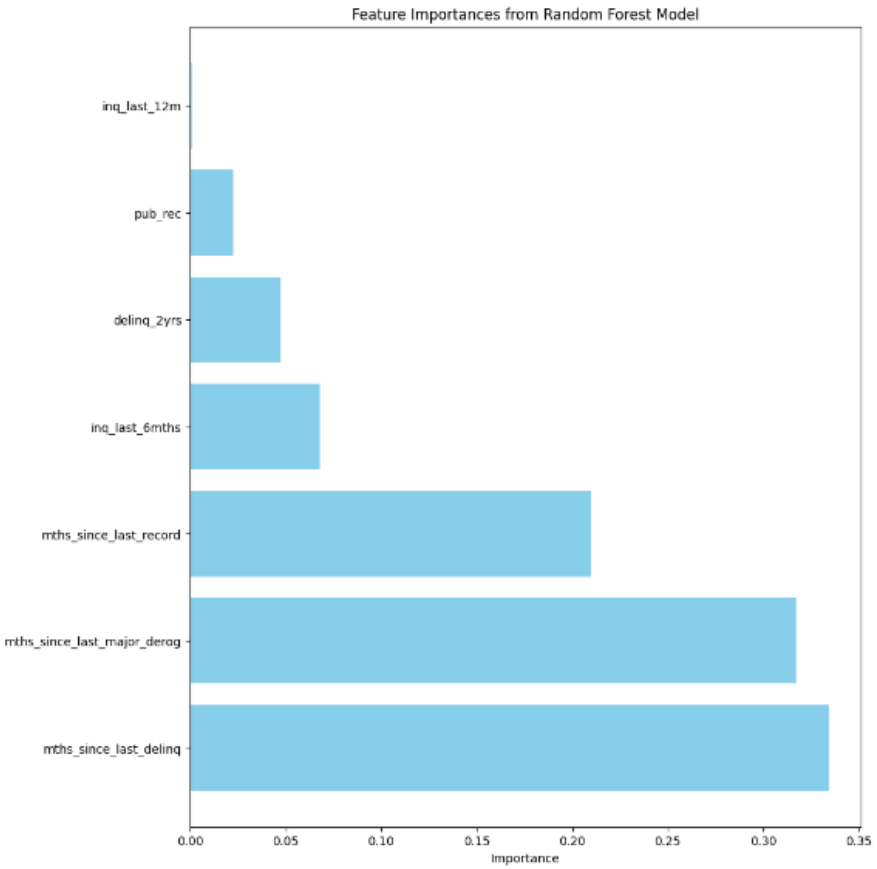
Correlation heat map of credit history variables



Notes: Heat map with accounting for credit history

Figure 12

Feature importance for credit history only, model is 63% correct at predicting loan status.



Notes: Predicting Loan status

Looking at Figure 11 and Figure 12 we can see that just the credit history is not enough to predict the loan status. We do see though that in relative retrospect to just credit history the most important indictors are the months since last delinquency and derogatory.

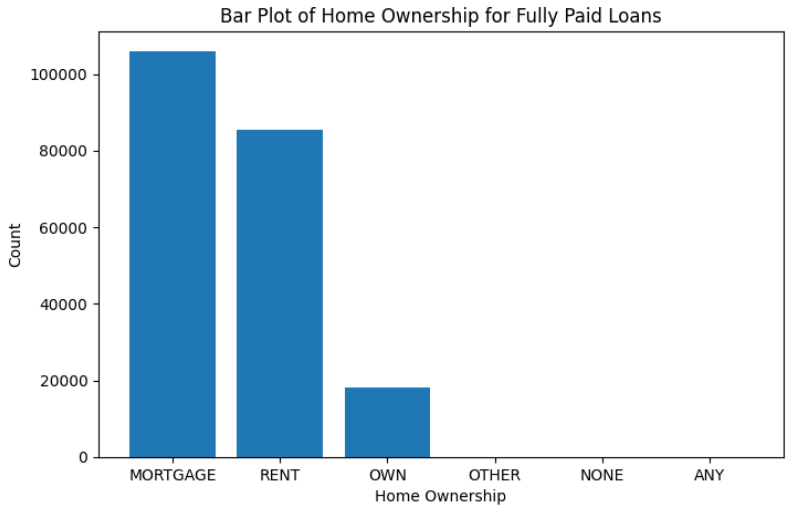
Personal Demographics

Home Ownership

Before getting a deeper look at our observations, we initialized our analysis by importing necessary packages, primarily pandas and matplotlib. While geopandas was part of our code, it remained unused throughout our analysis. Initially, we constructed a data frame using the “read_csv” function from pandas, utilizing a file path leading to the dataset. Subsequently, we filtered this data frame to exclusively include Fully Paid Loans, aligning with our designated topic. Our approach then entailed generating two graphs for each topic to facilitate comparison. To achieve this, we computed value counts for home ownership and verification status in both the original data frame (encompassing all loans) and the filtered data frame (comprising only fully paid loans). Following this, we utilized matplotlib to craft bar plots for these four distinct data frames.

Figure 13

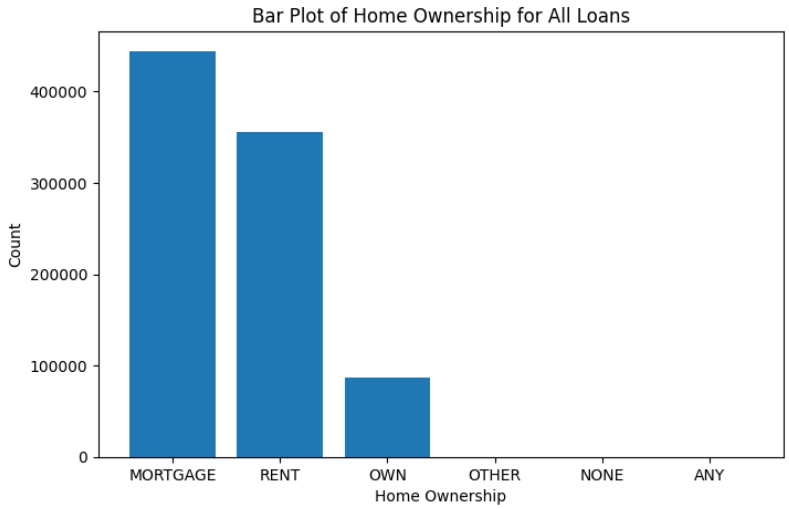
Bar Plot of Home Ownership for Fully Paid Loans



Note. This figure demonstrates the distribution of different Home Ownership for fully paid loans.

Figure 14

Bar Plot of Home Ownership for All Loans



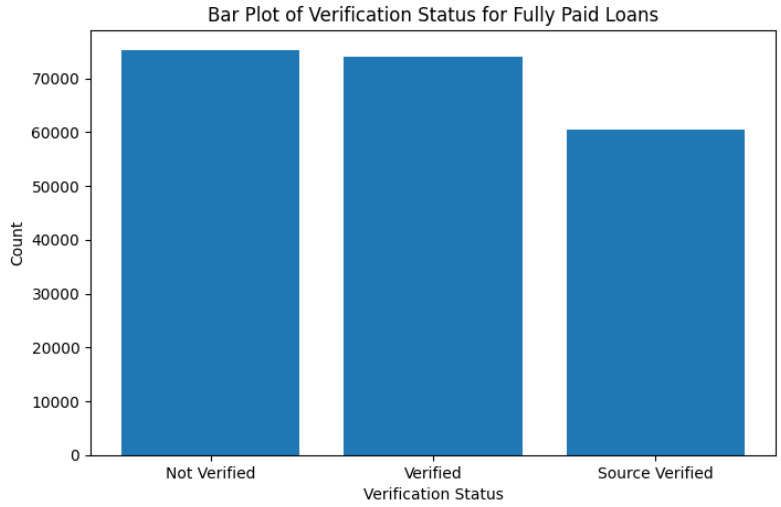
Note. This figure demonstrates the distribution of different Home Ownership for all loan types.

As can be seen in the graphs the distribution of home ownership is practically the same for just fully paid loans and all loans, with most mortgaging their homes, followed by renting, and a small number of homeowners. To us this indicated that home ownership is not a good indicator of the ability to fully pay off a loan as there are no distinct features showing a difference in the ownership of those who fully pay their loans from everyone else. The same comparison was made for verification status.

Verification Status

Figure 15

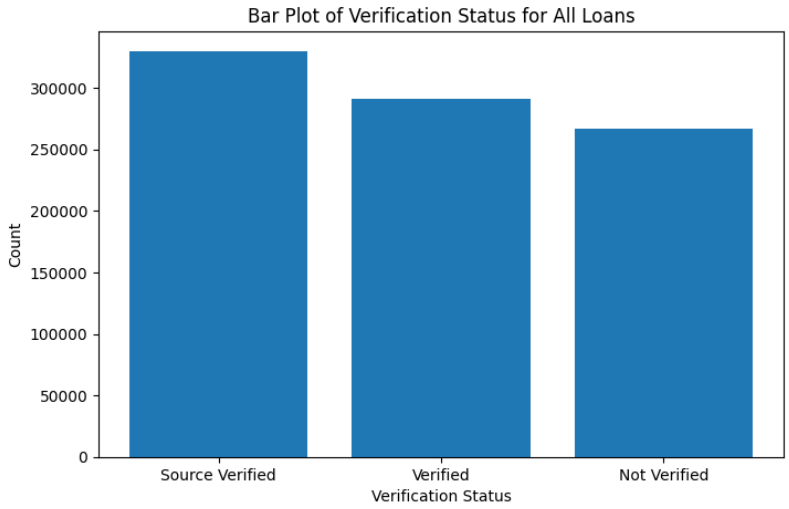
Bar Plot of Verification Status for Fully Paid Loans



Note. This figure demonstrates the distribution of different verification statuses for fully paid loans.

Figure 16

Bar Plot of Verification Status for All Loans



Note. This figure demonstrates the distribution of different verification statuses for all types of loans.

Interestingly a significant difference can be seen in the different verification statuses. For all loans the source verified is the highest, followed by verified, and finally not verified loans. Which makes

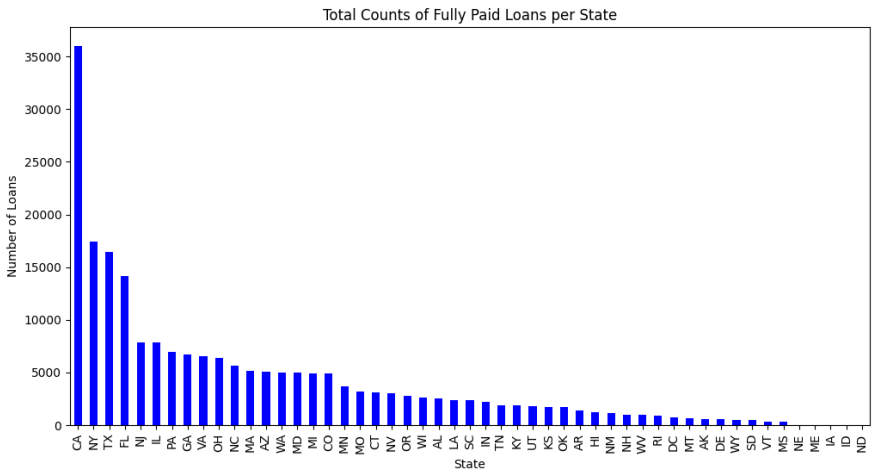
sense as you would want to verify everything you can before giving a loan. But interestingly, this is reversed for fully paid loans. While verified loans are not verified closely, not verified still makes up most fully paid off loans, suggesting that they are more likely when there is no verification. This does seem strange, however, suggesting unknown influences.

States

States were our next topic of interest in personal demographics. We created a data frame of value counts for state just as we had the previous two variables (having one for fully paid and another for everything together). We also created a new variable that was a ratio of the fully paid loans compared to all loans by state by dividing the two data frames together. This produced the following bar graphs.

Figure 17

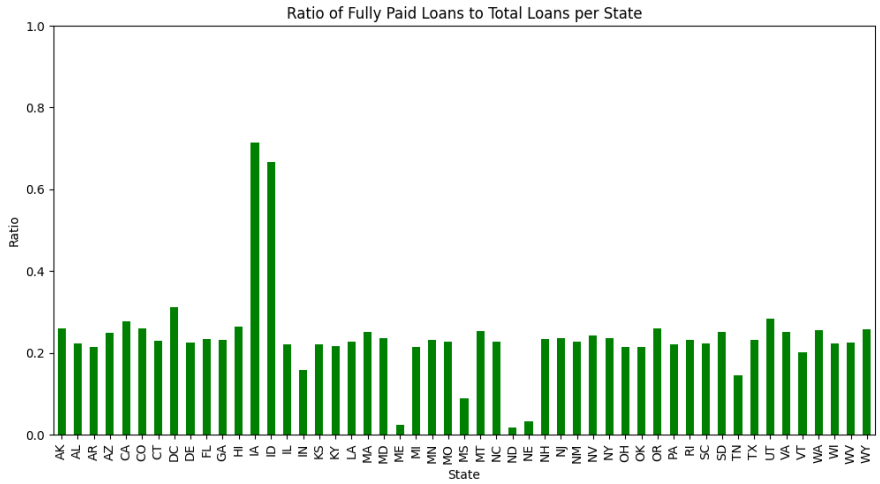
Total Counts of Fully Paid Loans per State.



Note. This figure demonstrates the number of fully paid loans in each state.

Figure 18

Ratio of Fully Paid Loans to Total Loans per State.



Note. This figure demonstrates how much fully paid loans make up a state's total loans.

The first bar graph shows that the most fully paid loans come from California, New York, Texas, and Florida. Not a big surprise given the populations these four states boast. The second graph we made to try and address population to an extent. In this graph we can see that in Iowa and Idaho fully paid loans make up over 60% of all the loans in those states. But looking back at the total counts graph there really aren't that many loans in those two to begin with. We concluded that state could be a good indicator given you take both population and fully paid ratio into account.

Financial Indicators

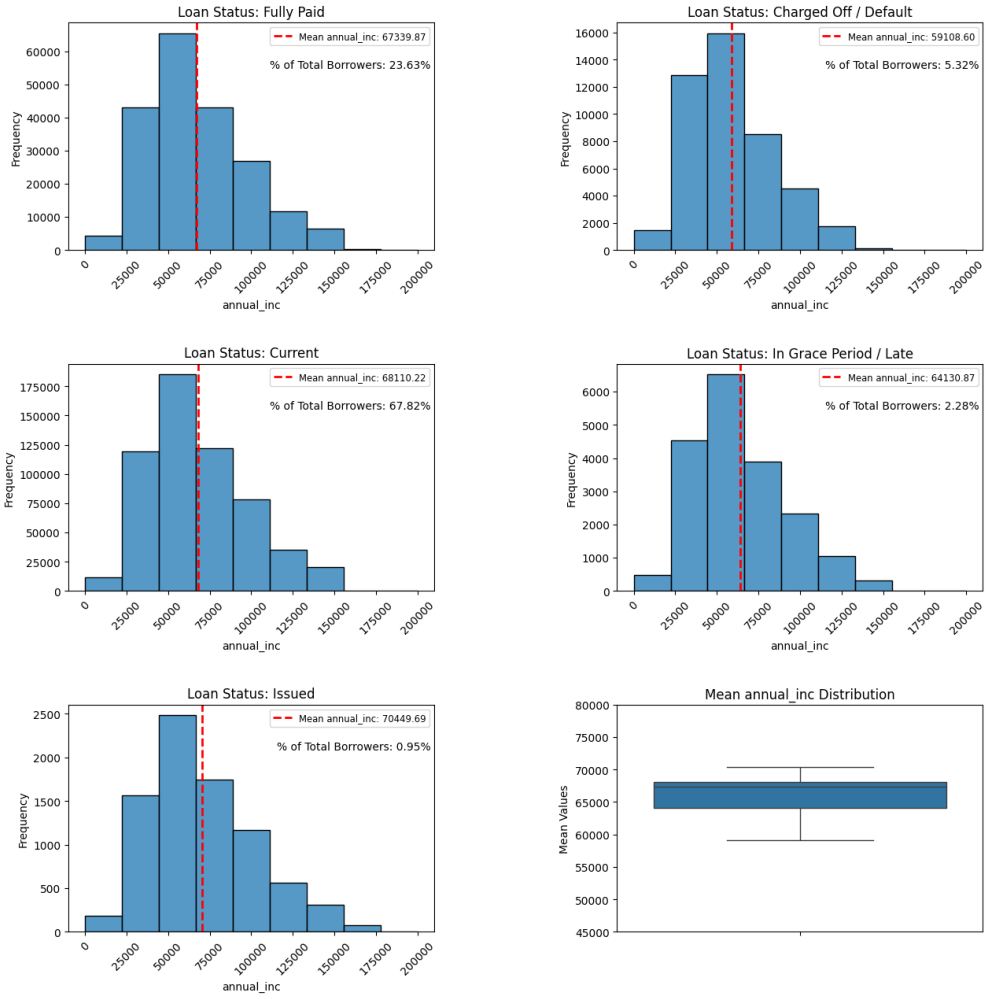
To assess the impact of the financial indicator's annual income, debt to income ratio, and employment length on a borrower’s ability to fully pay off their loan, we performed exploratory data analysis on each of these variables, segmenting their distribution by loan status.

Annual Income

Annual income is the self-reported annual income provided by the borrower during registration (Lending Club, 2024b). It is commonly used by lenders as a financial indicator of the liquidity of a borrower, which according to our hypotheses means that it should also indicate if a borrower is likely to fully pay off their loan or not. To validate the truth of this assumption, we created a series of subplots containing the distribution of annual income for a particular loan status and compared each plot’s summary statistics to identify key differences, or lack thereof, between the groups.

Figure 19

Annual Income Distribution by Loan Status



Note: This figure illustrates the distribution of annual incomes across all loan statuses in the dataset

After creating Figure 19, our team compared the mean annual income of each of the distributions to understand how annual income changes as loan status changes. Contrasting the mean annual incomes for the Fully Paid group and the Charged Off / Default group, we found that there existed an around $8,000 dollar disparity. This finding gave credence to our hypothesis that annual income is a predictor of a borrower's ability to fully pay off their loan.

Comparing the Current and In Grace Period / Late groups' mean annual income, we saw around a $4,000 disparity, again indicating that annual income is tied to differences in loan statuses. Because this disparity in income exists both between borrowers that fully pay off their loans and borrowers that

do not, and between borrowers that pay on time and borrowers that do not, we validated our initial conclusion that annual income is a financial indicator of a borrower's ability to pay off their loan.

The "Issued" subplot was used to understand what the mean annual income is of the group of borrowers that have just taken out a loan. Using our conclusion that annual income is a predictor of a borrower's ability to fully pay off their loan, we were able to make predictions about the incoming group of Issued status borrowers' ability to pay off their loans. Due to the issued group's mean annual income of $70,449, which is higher than both the Charged Off / Default and In Grace Period / Late groups' mean annual incomes, we assumed that the majority of these new borrowers would be able to fully pay off their loan, or at least pay their loans on time, increasing the likelihood that they will fully pay off their loans in the future.

On each subplot, our team generated a value representing the percentage of total borrowers represented by that loan status. This measurement describes the same trend that our group measured in our Dataset Overview analysis referenced at the start of this paper. Our team included this measurement to help us put into context the implications of the group's mean annual income. For instance, since the Issued group only contains 0.95% of borrowers, we must take our observation of the mean annual income with a grain of salt, as there are not a lot of data points to average over compared to the Current or Fully Paid groups.

The subplot "Mean annual_inc Distribution" summarizes all our analysis regarding annual income by graphing the distribution of mean annual incomes as a box plot. The range of this boxplot indicates how varied the mean annual incomes over all loan statuses; thus, a large range indicates that annual income most likely is a predictor of which loan status a borrower will be grouped into, specifically providing our team with insight into the effectiveness of annual income in predicting if a borrower will

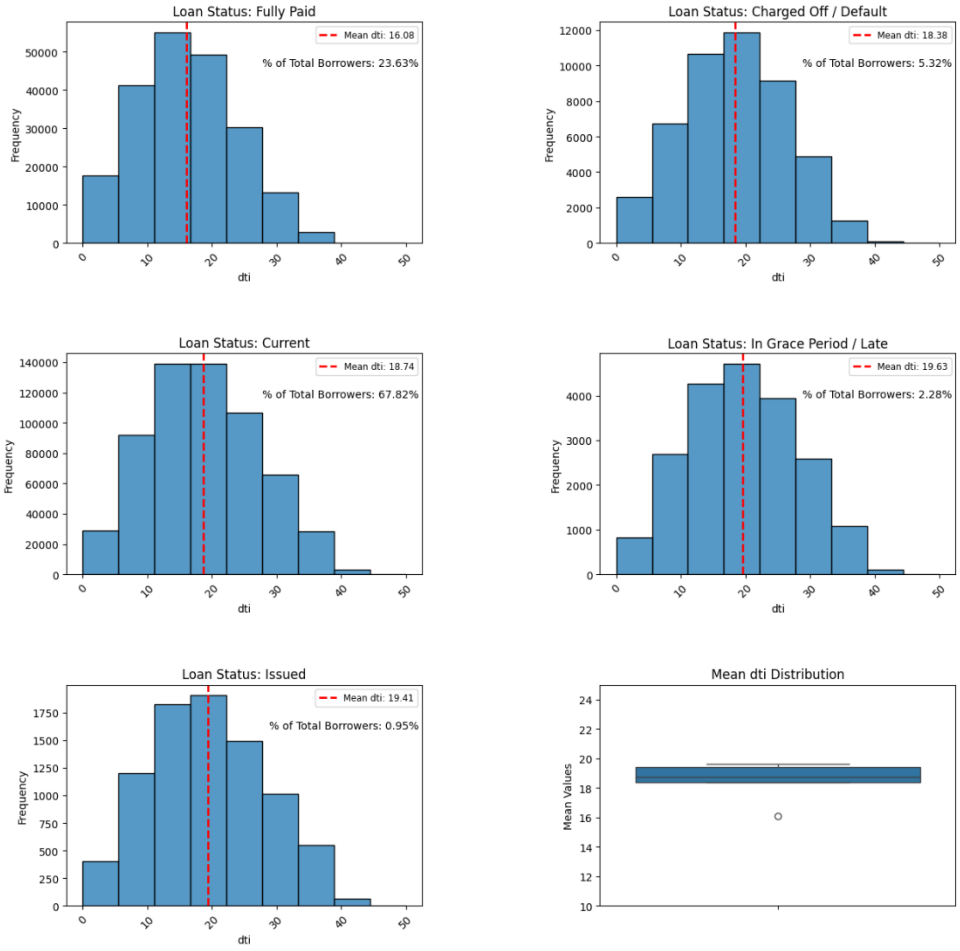
fully pay off their loan. After analyzing this plot and taking note of its moderately high range of $11,000, we re-confirmed our initial conclusion regarding annual income.

Debt to Income Ratio

Debt to Income Ratio (DTI) describes the percentage of a borrower's monthly income that borrower's monthly debts take up (Lending Club, 2024b). Like annual income, DTI is used by lenders to evaluate the financial liquidity of a potential borrower; therefore, according to our hypothesis, DTI should play a role in predicting if a borrower will be able to fully pay off their loans. Following the same data exploration process we employed for borrower's annual income, our team split the DTI distribution over multiple subplots corresponding to different loan statuses and displayed summary statistics for each plot.

Figure 20

DTI Distribution by Loan Status



Note: This figure illustrates the distribution of DTI ratios across all loan statuses in the dataset

From Figure 20, our team derived several important insights. To begin with, there exists a disparity of around 2% between the mean DTI of the Fully Paid group and the Charged Off / Default group, and a disparity of around 1% between the mean DTI of the Current group and the In Grace Period / Late group. These differences may seem small, but since we are dealing with percentages of monthly income, these values correspond to significant amounts of cash allocated to paying debts. So, even a difference of 2% or 1% is an indicator that DTI does play a role in determining if a borrower will be able to pay their loan off.

Investigating the Issued group, our team used our conclusions regarding DTI to estimate that due to the group's high DTI relative to the Fully Paid group, the incoming wave of borrowers is less likely to fully pay off their loans.

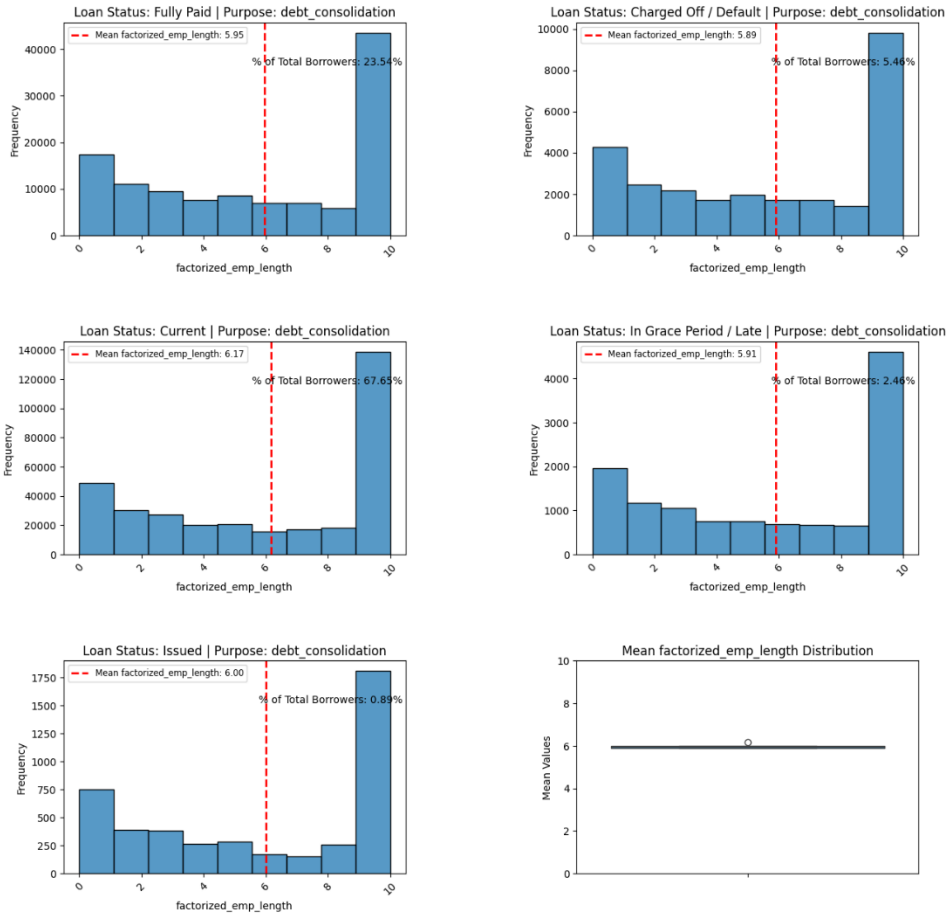
Connecting our team's analysis to the "Mean DIT Distribution" boxplot, we observed a total range of around 2% among the mean DTI ratios of the different loan statuses. As discussed earlier, even this small percent is indicative of DTI's influence on a borrower's ability to fully pay off their loan; thus, our initial hypothesis that DTI impacts borrower's repayment behavior was validated by our investigation in this area.

Employment Length

Employment length describes the number of consecutive years the borrower has been employed, with values ranging from zero to ten, where zero represents zero years of employment and ten represents ten or more years of employment (Lending Club, 2024b). Unlike annual income and DTI, employment is not used by lenders to estimate liquidity, but instead is used to measure the financial stability of a borrower. According to our hypothesis, the financial stability gained from being employed longer should result in a borrower being more likely to pay off their loan. Once again, our team split the distribution for employment length up into subplots corresponding to the collection of loan statuses.

Figure 21

Employment Length Distribution by Loan Status



Note: This figure illustrates the distribution of employment lengths across all loan statuses in the dataset

Figure 21 displays a surprising trend: all the mean employment lengths are very close together, regardless of the distribution. Unlike DTI, this lack of a large disparity cannot be due to the nature of the units of our financial variable since employment length is measured in years, which is not a small unit of measurement like percentage is. Because of these observations, our team concluded that employment length is not an indicator of if a borrower will pay off their loan fully.

This conclusion is validated by the exceedingly small range seen in the “Mean factorized_emp_length Distribution” plot. Since a broad range in this plot indicates that the variable, we are analyzing is an indicator of a borrower’s ability to pay off their loan, the inverse must also be true; a small range indicates that the variable we are analyzing is not an indicator of a borrower’s ability to pay

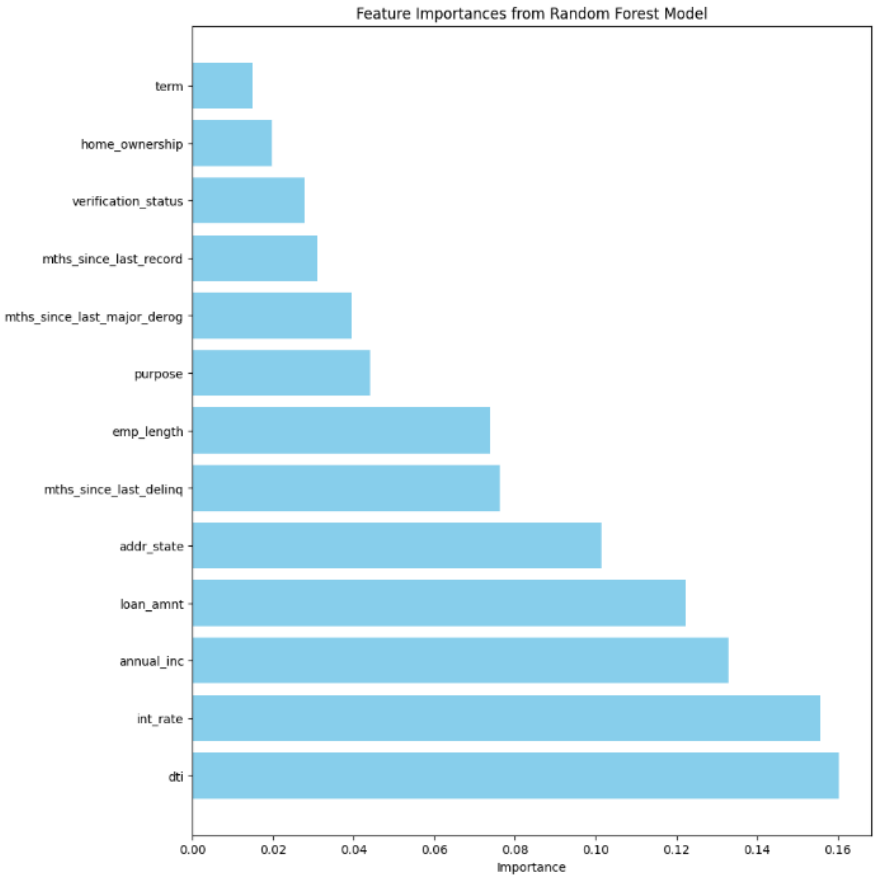
off their loans. Therefore, by both our team's comparison of different loan status plots and our investigation of the distribution of mean employment lengths, we concluded that our hypothesis regarding employment length was invalid.

Random Forest Regression

After performing our EDA and model development by section, we wanted to get a look into how our empirically derived results compared to those output by a random forest regression model we trained on a portion of the variables of the dataset.

Figure 22

Feature importance for general variables, model is 82% correct at predicting loan status.



Notes: Graph Overview of the model

The feature importance of each of the variables used in this model roughly corroborated our findings. Our findings regarding the following variables were validated by this random forest model:

- Interest Rate
- Month since last delinquency
- Months since last record
- State
- Annual Income
- DTI

However, the following variables had more varying levels of influence on loan status than we had concluded as the result of previous models or EDA:

- Loan Amount – More than expected.
- Home Ownership – More than expected.
- Employment Length – More than expected.
- Term – Less than expected.
- Months since last derogatory – Less than expected.
- Verification Status – Less than expected.
- Purpose – Was not the focus of our analysis.

The differences between the random forest model's results and ours can be explained by the complexity of the problem we were looking to solve with our analysis. There are a lot of variables at play in determining loan status, much more than a simple EDA can describe by itself, so it is to be expected that the results from our EDA would not be entirely accurate. However, our EDA did correctly identify many of the most significant influencers of loan status, demonstrating that the way we conducted our study, and the findings we ascertained from it, are valid to a degree.

Conclusion

Overview

When working with data from banking and financial services, the importance of distinguishing and fostering professional relationships with exceptional customers emerges as a fundamental pillar for sustainable growth and success. As financial institutions constantly connect with these good clients, they augment their resources and create strong relationships within the community, creating trust among clients, employees, and investors alike. It is against this backdrop that our research focused on delineating the defining characteristics of exceptional customers within the banking market, aiming to furnish the bank with the requisite tools and insights to effectively identify and retain these esteemed clients, thereby fostering long-term profitability. By delving deeper into the attributes and behaviors of these customers, banks stand a good chance to customize their products, services, and marketing strategies better. Furthermore, through optimized customer acquisition, retention, and satisfaction, financial institutions can enhance their profitability, fortify brand loyalty, and consolidate their position within the fiercely competitive financial landscape.

Goals and Recommendations

Our primary objectives that we sought to accomplish with this study were to understand the qualities that make a “good” borrower and to apply our findings to the Marketing and Revolving Lines of Credit departments within our client, Lending Club’s, business. In respect to the Marketing department, the characteristics of a borrowers that are more likely to fully pay off their loans can be used as a template for Lending Club’s new target market, generating an incoming wave of borrowers that are more likely to fully pay off their loans. Additionally, the Revolving Lines of Credit department can use the characteristics of a good borrower as a benchmark to compare current borrowers against. If a current

borrower is failing to meet the qualities of a good borrower, their risk of defaulting increases. The Revolving Lines of Credit department can identify these at-risk individuals and increase their likelihood of success through renegotiation of loan terms.

Results

Upon conclusion of our research, our findings illuminate several critical insights into distinguishing the good clients within the banking sector. Notably, exemplary clients exhibit prudence in their loan requests, opting for moderate loan amounts and demonstrating commendable repayment records, thereby minimizing default rates. In addition, our analysis of creditworthiness shows us that this assessment extends beyond payment history to encompass various indicators. Moreover, our exploration of demographic trends sheds light on the geographic distribution of exemplary clients within our dataset, with notable contributions from states such as Iowa (IA) and Idaho (ID). Lastly, our examination of financial indicators highlights the significance of metrics such as annual income and debt-to-income ratio in gauging client worthiness and financial reliability.

Some of our hypotheses were not validated by our study, including the following:

Credit History Hypotheses

- Public records are an indicator of loan status.

Personal Demographics Hypotheses

- Homeownership status is an indicator of loan status.

Financial Indicator Hypotheses

- Employment length is an indicator of loan status.

While these hypotheses seemed intuitive to our team at the beginning of the study, the results of our analysis demonstrate that the variables that the hypotheses center around are not reliable indicators of loan status.

Having identified which of our assumption were correct and which were not, we generated the following example figures to represent the quantifiable characteristics of a good borrower:

Loan Indicators

- Payment History: High Repayment Rate (95%)
- Loan Interest: Is fine with a reasonable Interest rate (29 - 40%)
- Loan Amount: Ask for a moderate Loan Amount (avg. $10,000)

Credit Indicators

- Months since the last Derogatory: 48+ months
- Months since the last Delinquency: 24+ months

Demographic Indicators

- State: Varies by Population
- Less Verified the better

Financial Indicators

- Annual Income: $60,000 or more
- Debt to Income: Within 1% of 16% DTI – This can change based on annual income.

In synthesizing these findings, our research furnishes Lending Club with invaluable insights to refine client identification processes and make informed decisions, ultimately enhancing their ability to

cultivate and sustain relationships with exceptional clients, thereby fortifying long-term profitability and resilience within the dynamic financial landscape.

Bibliography

Lending Club. (2024a). LendingClub_Data LendingClub_Data.csv [File Acces], Dr. Aric LaBarr. April 15, 2024.

Lending Club. (2024b). LendingClub Data Dictionary [File Access]. Dr. Aric LaBarr. April 15, 2024.

All images, statistics, metrics, models, graphs, analysis, insights, and calculations used in this research are based and created according to Lending club (2024a) data, the information on this research is for academic purposes only.

Han, Liang, et al. “Are Good or Bad Borrowers Discouraged from Applying for Loans? Evidence from US Small Business Credit Markets.” Journal of Banking & Finance, vol. 33, no. 2, Feb. 2009, pp. 415–424, https://doi.org/10.1016/j.jbankfin.2008.08.014.