Problem Statement

To begin our coverage of the results of our analysis, let's revisit our problem statement. One of the most lucrative - and riskiest - banking activities is lending. This truth is so evident that most private citizens trust their future financial well-being to this activity in the form of investments. Banks, however, rely on lending not only as a long-term background activity for success, but as a daily reality of business. Therefore, one of the greatest dangers to any given bank is a default on a loan, and one of the most important missions to those same banks should be better identification of risk in lending.

The aim of this project was exactly that - identifying any and all factors that have a significant correlation to defaulting on loans, ranking them, and putting together a realistic roadmap on which loans to avoid, and how to mitigate risk on ambiguous loans that are worth the chance. These insights should be put into practice immediately, as each loan underwritten on shaky risk analysis is a potential money sink for any given bank, and we should be able to accurately predict loans which have defaulted through risk factors they share.

Data

In the course of conducting this analysis, we used data from an anonymous financial institution. This set had information on nearly 230,000 loans, each of which had a variety of features for analysis in addition to information on whether the loan defaulted or was paid. The features available in this dataset are below:

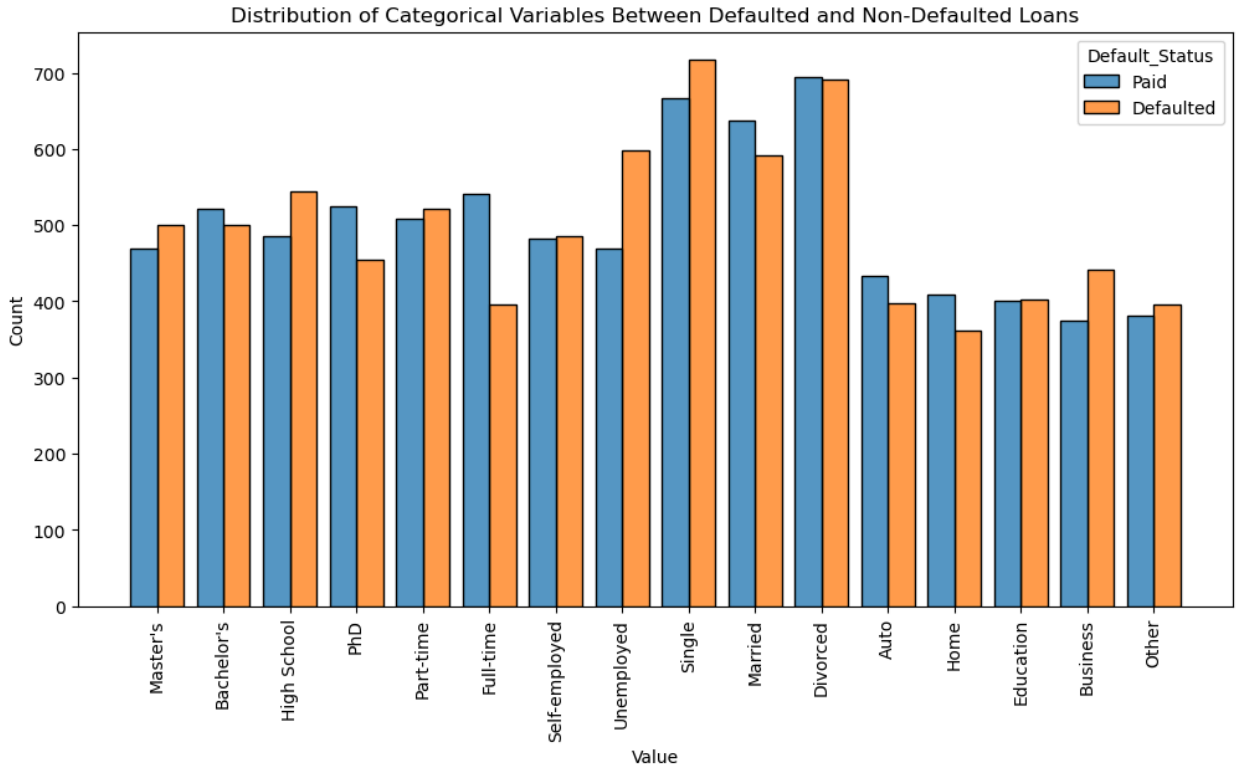
- Age
- Income
- Amount of loan
- Credit score
- Employment status
- Type of employment
- Months employed in current role
- Number of credit lines
- Interest of loan
- Loan term
- Debt to Income Ratio
- Ratio of loan amount to income
- Level of education
- Marital status
- Whether applicant has a mortgage
- Whether applicant has dependents
- Whether applicant has a co-signer
- Purpose of the loan

Analysis

Our initial aim was to compare what we knew about loan recipients who defaulted, when compared to those who did not. Our first step was to separate the data regarding recipients who defaulted from those who did not, and perform a summary analysis of statistical and correlative significance. While there were several differences between the two, none of these differences were substantial enough to give us actionable insights into our risks. For example, the average age on defaulted loans was 36.5, 8 years younger than the average age of 44.5 of those who

did not. The median income on defaulted loans was a little under $20,000 less ($66, 566) than the median income for those who did not default ($84,237), although interestingly we found that applicants with no higher education than a Bachelor's degree were less likely to default than those who had completed a Master's, while also being less likely to default than those with only a high school diploma. Figure 1 below displays a histogram of various features and their ratio of paid vs. defaulted loans.

Figure 1



The most encouraging early metric in determining risk appeared to be the ratio between the amount of the loan and the annual salary of its applicant. It is important to distinguish this metric from the Debt to Income ratio, which measures total debt to the annual salary of the applicant, and appears to be much less useful. Below are figures which demonstrate the predictive power of the income to amount ratio and salary (Figure 2) to the simpler comparison between income and loan amount (Figure 3)

Figure 2

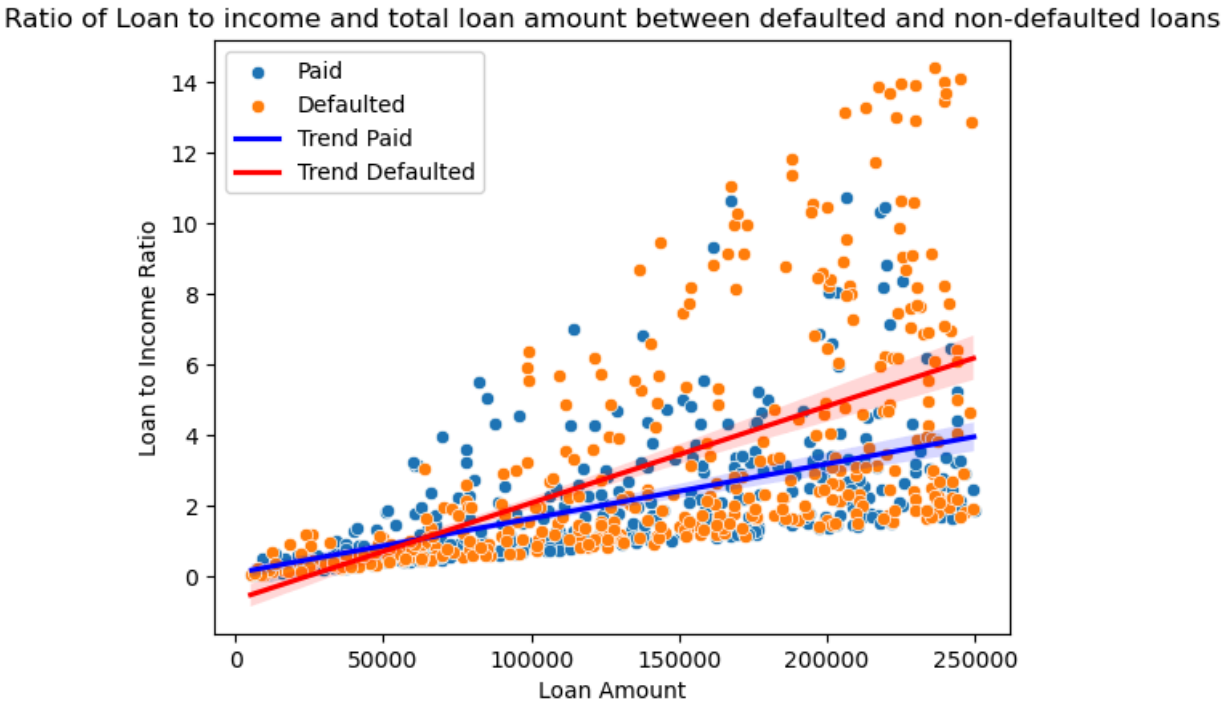
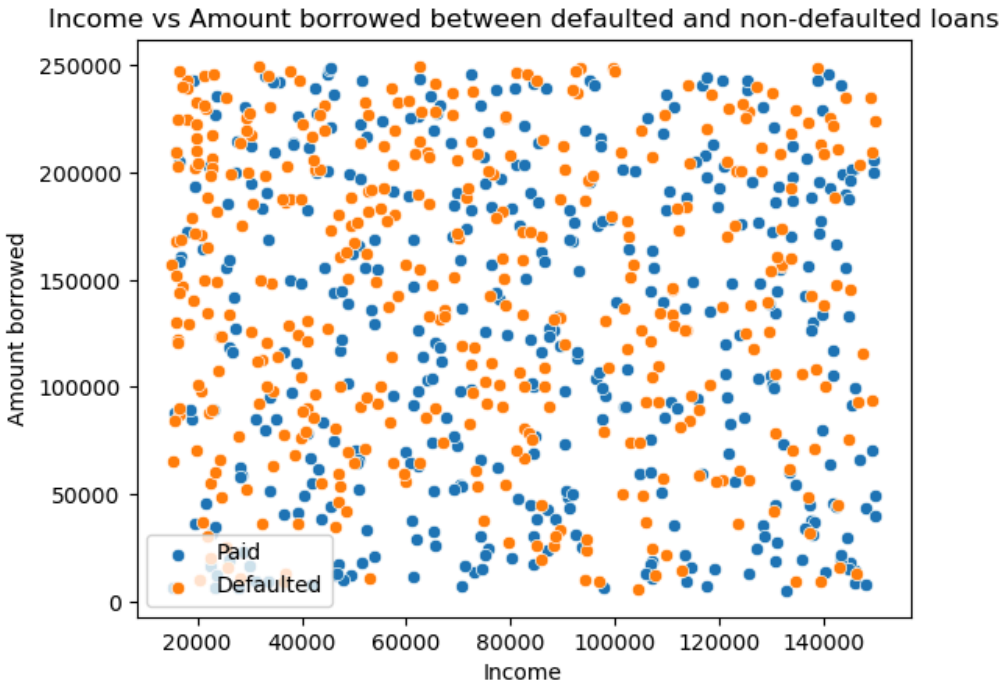


Figure 3



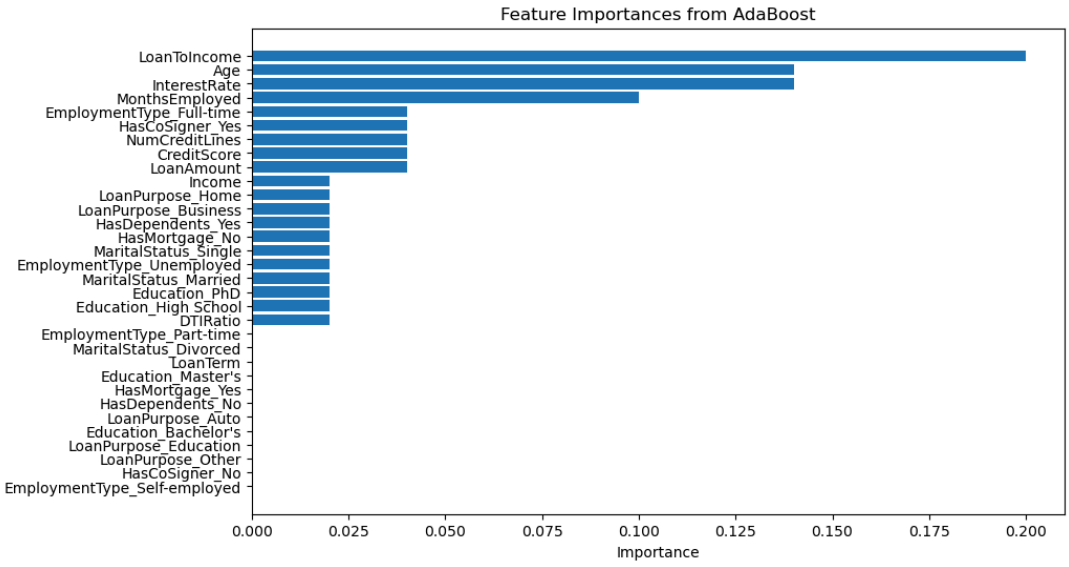
In addition to scatterplots and histograms, we also investigated simple correlations in our exploratory analysis. Unfortunately all the correlations to one another were rather weak, and we

were not able to include categorical variables in these heatmaps alongside our continuous variables. Unsupervised clustering methods were also unable to distinguish any clusters of variables which may be useful at a glance. These heatmaps will be shown in Appendix A.

Moving from our rather lackluster exploratory data analysis, we scaled and encoded our data accordingly. Without this necessary step, it would be impossible to glean insights from wildly different values such as would be seen in the relationship between age or different ratios to salary or loan amounts. This also allowed us to properly train models with encoded categorical data.

From here, we move forward into selecting a classification model. We quickly found that a Random Forest or ADA boosted model would work best, and that the ADA model performed slightly better. We saw a very slight superiority in the ADA model when compared to random forest, with a a ROC AUC score of .75 compared to an optimized Random Forest model's score of .73, with the ADA model's recall score being .71 for non-defaults and .65 for defaults. These are strong scores, with both recall scores being significantly better than a coin flip. When looking into feature importances, we saw a few features being utilized much more than others. Specifically, the ADA model uses the Loan to Income ratio most significantly, followed by age, interest rate, and months employed following it, with months employed lagging the previous two. It is important to note that the interest rate is calculated by the bank and is reflective of risk, so it isn't a useful measure for banks to vet out applicants. This chart is below in Figure 4.

Figure 4

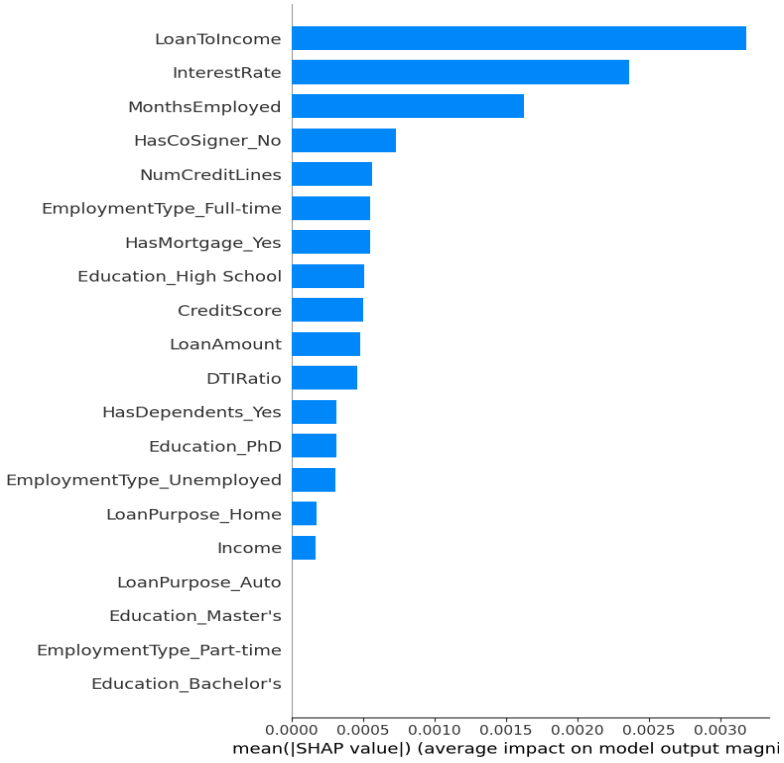


Beyond the feature importance, we took a further step and conducted a Shapely Additive Explanations analysis (SHAP) to look into a more robust explanation into the model output. The difference between the SHAP analysis and a basic feature importance is that the feature importances works by assigning a score to each feature to attempt to calculate their importance

within the model, whereas a SHAP analysis leverages game theory outside of the model itself to gain a broader and more general weighting of each of the features. In other words, our initial feature importance chart explains the way a certain model in a certain context weighs each feature, whereas a SHAP analysis provides a more general view of the actual importances of each of the features.

It is also important to note that in our SHAP analysis we have removed age as a factor, as age cannot be used in home lending as a factor or approval of a loan. More on this in our conclusion. We still see the top three features in the same order, with Loan to Income ratio as the most important factor, Interest rate second, and months employed third. We do however get a more robust look into the small differences among our subsequent factor, seeing that loan amount is marginally less important than having a mortgage, but is more important than having dependents. While these are not useful as primary features in decision making, they are interesting to use as a potential area for future study in a more precise model, or to see if there are additional features within the feature that may be more useful as indicators than the original was.

Figure 5



We set out initially to both create an optimized model which will be useful in identifying risk factors for defaults on the loan, as well as to identify the individual factors most tied to the risk for default. It is also important in this analysis to recognize legislation around fair lending practices, as we cannot use features that are protected by law, however useful, to deny lending.

We found throughout our study that the ratio of the loan amount to the applicant's income is the most useful predictor of the 18 we used, and by a significant margin. After that, we found that age and interest rate were the next two most important, followed by the months employed in their current role.

Of those four, only two are universally useful in the approval process. Interest rate is not useful, as that is determined by the lendor as a result of perceived risk, rather than existing apart from the risk itself. Age is useful in many situations but the Fair Housing Act of 1968 prohibits its use as grounds for denying a home loan. Marital status is also barred from use in consideration for loans under the Equal Credit Opportunity Act (ECOA), but fortunately it is not useful when compared to other features.

In conclusion, we have discovered that a ratio should be calculated for each loan between the amount of the loan and the annual income of the applicant. Along with this, there should be a secondary consideration of the months the applicant has been employed in their current role, though this should not necessarily rule out those with a strong ratio. We have created and tuned a model which has a high recall for both defaulters and non-defaulters in our current data, with recall being important as it is resistant against false positives. Further study should be done to find additional features which are not protected by fair lending laws which are also useful, and those features should include both new factors in themselves as well as ratios between continuous factors,