Bank Loan Analysis

Ian Tuohy

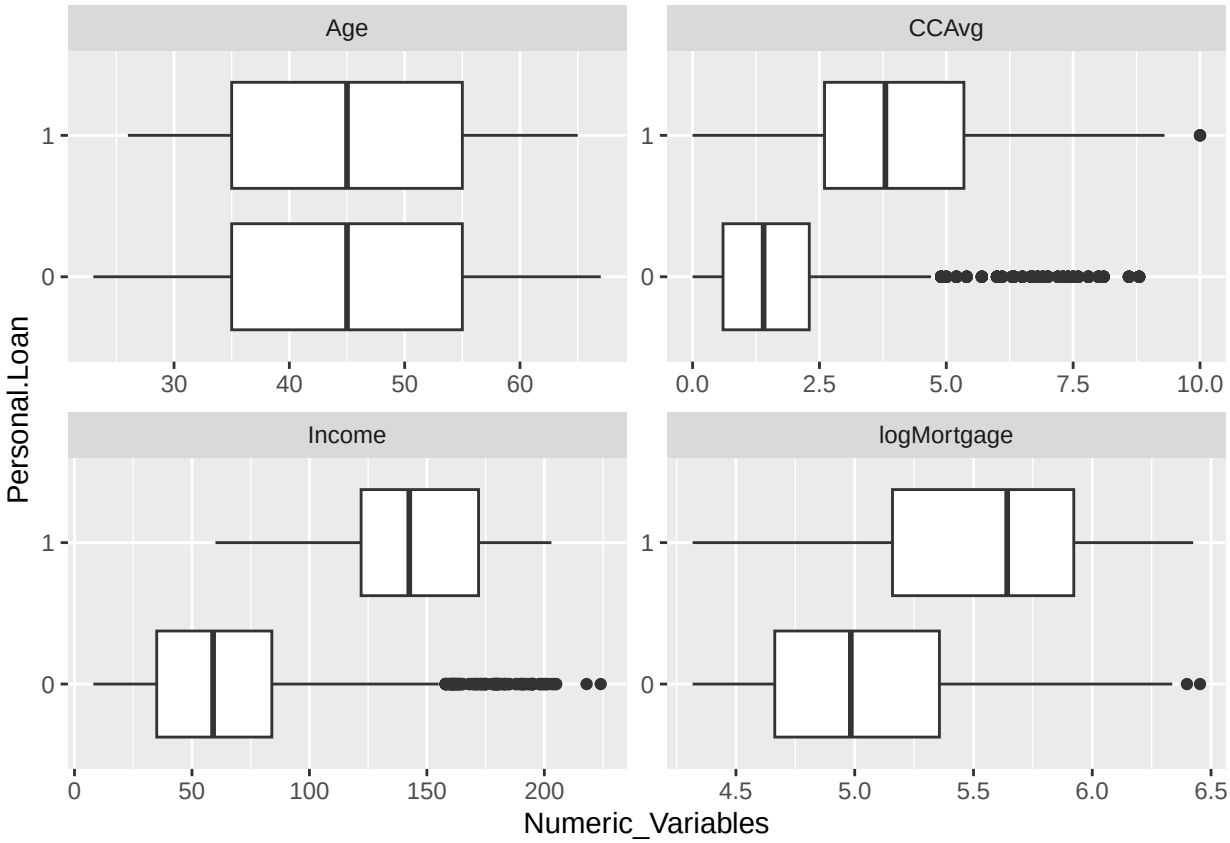
2024-2-13

Introduction

I chose this sample because I was curious about the intricacies of how banks approve different loans for different types of people. Above all, I had hoped to learn what the most important aspects of someone's banking information that a bank takes into account for approving a loan. When a bank decides to approve a loan, they want to make sure they'll make that money back, and so I'm seeing these different aspects as what makes a customer "trustworthy" to a bank. I think this is what makes this an interesting dataset.

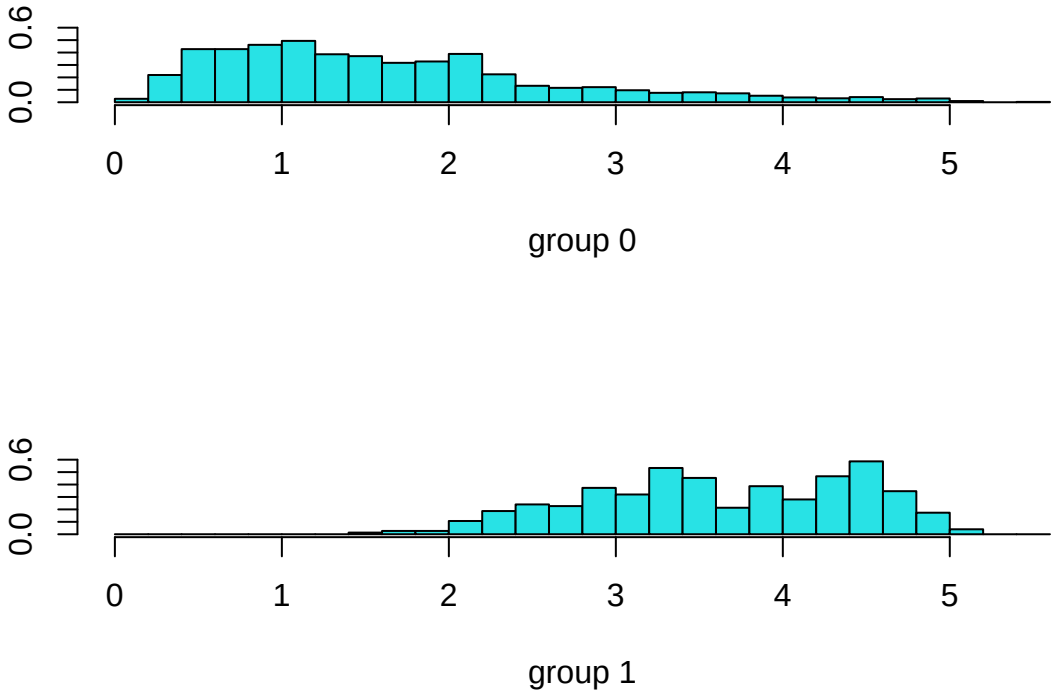
Data Cleaning

In order to clean this data, there wasn't much work to be done. The data came from kaggle mostly cleaned. The biggest issue I had run into was making sure that any of the variables I kept in the dataset were variables that I was able to explain. Any variables that weren't as easy to explain or didn't have accurate and detailed descriptions from kaggle I had removed. Lastly, there were a few variables that required updating from numeric to factor variable types. The most important variables in this dataset that I've found were Income and Mortgage. These are both variables that show (in thousands) either the amount of income someone makes, or the amount left on their mortgage, respectively.



Data Visualizations

In the above visualizations, I wanted to create boxplots of each of the numeric variables plotted against Personal.Loan, our response variable. I was curious if I could come to any broad conclusions from just this simple graphing, before making any models. It's easy to see in our visualizations that the Income variable seems to be an interesting candidate, as it has two very distinct and different boxplots depending on whether or not someone was approved. Alongside this, Mortgage was an interesting variable. I decided to use the log of Mortgage, and it too seems to have two quite different resulting boxplots.



Data Visualizations

In the above visualization, I decided to make an LDA of our income variable. I decided to do this because of its aforementioned boxplot, and how distinctly different it seemed. After further analysis, it seems like Income is our best predictive variable for personal loan approval. The interesting part about this variable is that given the nature of personal loans, it seems like the people with the lower income should be the ones most in need of personal loans. Despite this, banks tend to approve people with a higher income more frequently. Going back to my introduction, this is one of the points that I had wanted to learn. It seems as though banks have determined people with higher incomes more "trustworthy" and more likely to pay off their personal loan, thus being approved more often than people with lower incomes.

Conclusion

If I had more time, I would've run a few different models ontop of the LDA analysis that I had done. I think that this is a really interesting dataset, and I especially enjoy using data to get a sneak peek into how different industries utilize data in order to make decisions. Previously, when I had used this dataset, I had decided to make a decision tree as well as an Elastic-NET model. These were both models that were beneficial with this particular dataset as they're relatively robust when it comes to variables that aren't as important for predicting the end result. Despite using modeling techniques that deal with this well, it is possibly the biggest challenge in analysis. Another challenge with the analysis is

Github Repository

This package is being hosted on my github, under ituohy/DemoPackage.