Loan Repayment Challenge

1 Introduction

At MoneyLion we are constantly working to assess risk of our applicants more accurately. Being successful in this objective enables us to better price customers and mitigate losses on our portfolio of loans. The following challenge asks you to work with a data set of loan repayment. It is intentionally meant to be open ended. The point is not to arrive at a predetermined answer or search for the lowest possible standard error. Rather, the hope is that it will force you to ask relevant questions about the data, do some preliminary exploration, perform the necessary manipulations or aggregations, generate visualizations, and reach conclusions or insights. The most important thing to remember is that we are evaluating your thought process and ideas! The more you explain your thinking, in a clear and succinct manner, the better. If you get stuck, describe what additional information or data you might look to collect, and trying a different idea is highly encouraged.

2 Data

You are provided with 3 files: loan.csv, payment.csv and clarity_underwriting_variables.csv. The files are comma separated, with the column names in the first row. The detailed description of each column is available in the MoneyLion Data Scientist Assessment Data Dictionary document. This can be found in the dictionaries.zip file.

3 Rules

You may use any language, packages, or external libraries for the challenge, though Python or R are preferred. An IPython notebook might be the best way to show your code and write your comments/thoughts to follow along. DO NOT use any help from other people, sources, online forums, etc., your submission should be solely your ideas and work. There is no hard limit on the amount of time that can be spent on this challenge. Please disclose how much time you spent on the challenge when you've completed it. For your submission, zip up your source code (you can leave out external packages or libraries) and any text files and email it back to us.

4 Guidelines

We want to see a model built for predicting the loan risk or quality (loan repayment) on a given applicant. But before diving into this task, start with more simple analyses as it may inform your model building process. Take a look at some distributions of some of the variates. Maybe the loan repayment value can be modified in order to make it easier to model. Perhaps there are some idiosyncrasies for geographies, or maybe different bands of clearfraudscore partition loan repayment nicely. Some visualizations may provide insights. Consider the problem in a business context, what would you want to predict generally? Loans with what type of performance?