Happy Money®

Happy Money Data Science Assessment

Data Scientist

Dataset Details:

The Lending Club dataset is a collection of installment loan records, including credit bureau data (e.g., FICO, revolving balances, etc.) and loan performance data (e.g., loan status).

The data, and data dictionary, can be downloaded by following this link: Lending Club Data

This will be the dataset used to answer the questions below. Please answer all questions in both sections, keeping in mind that quality is FAR more important than quantity.

Important! If you have any questions at all, please don't hesitate to reach out to Meng Zhao at mzhao@happymoney.com and Michael Tepper at mtepper@happymoney.com.

Questions:

Section A - KPI Reporting

As part of the data team, a key aspect of our work is determining the key performance indicators (KPIs) from a large set of unfamiliar data. We need you to determine the KPIs that can provide guidance for the following needs:

1) What is the monthly total loan volume in dollars and what is the monthly average loan size?
2) What are the charge-off rates by Loan Grade[1]?
3) Are there any statistically significant differences in charge-off rate by Loan Grade? Please explain how you made this determination.
4) Is Lending Club charging an appropriate interest rate for the risk?

[1] Use the following condition to determine the charge off status: If loan_status = {'Charged Off', 'Default'}, then the loan has charged off, otherwise it has not.

Happy Money®

Section B - Modeling

Prior to creating a model, it is important to inspect the quality of the dataset:

1) Data is often messy, please review and clean the Lending Club dataset and summarize your thoughts on any structural issues:
    a) Is there missing data? Is the missing data random or structured? Are some attributes missing more than others?
    b) Are there any glaringly erroneous data values?

Let's build a model:

2) Using any format and any modeling technique that you prefer, please create a model to predict charge off[2] within the Lending Club dataset. Show any work that you would deem important in evaluating this process and discuss some of the key features selected.

Section C

3) Please choose **one** of the topics below and **concisely** explain it to:
    a) Someone with significant mathematical experience.
    b) Someone with little mathematical experience.
    c) Topics: Logistic Regression, Principal Component Analysis, Factor Analysis, Markov Process, Hidden Markov Models, Gradient Boosted Trees, Survival Modeling, Kernel Density Estimation, **or** the Curse of Dimensionality.

Important! Please include all code used to generate any analysis or plots in a document of your choice and upload it to the link included in the email.

[2] Use the following condition to determine the charge off status: If loan_status = {'Charged Off', 'Default'}, then the loan has charged off, otherwise it has not.