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. You may submit your answers as a report (slides or PDF). In addition to (or in lieu of) a report, you may also submit a notebook document (R Markdown, Jupyter, or Zeppelin, etc.). If you submit a notebook document in lieu of a report, (1) please submit the compiled version (either .html or .pdf) – code-only submissions will be rejected, and (2) treat it as report – there should be sections where you provide answers to the questions and discuss your results.

Important! If you have any questions, please don't hesitate to reach out to Michael Tepper (the hiring manager) 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 default rates by Loan Grade?
3) Is Lending Club charging an appropriate interest rate for the risk?

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 QA 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 default within the Lending Club dataset. Default is defined as LoanStatus ∈ {'Default', 'Charge Off'}. 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: Linear Regression, Logistic Regression, General Linear Model, Principal Component Analysis, Factor Analysis, K-means Clustering, Support Vector Machines, Markov Process, Hidden Markov Models, Kalman Filter, Decision Trees, 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 send it back over email.