Ross Brown Happy Money Data Science Assessment

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:

A1) Monthly total loan volume in dollars:

Apr-2015	$539401075
Aug-2015	$555331400
Dec-2015	$667910550
Feb-2015	$366908525
Jan-2015	$533132575
Jul-2015	$696238600
Jun-2015	$429777175
Mar-2015	$390003275
May-2015	$483189475
Nov-2015	$567247325
Oct-2015	$738221400
Sep-2015	$450246800

A1) Monthly average loan size?

$15,260.90

A2) What are the default rates by Loan Grade?

A	5%
B	10%
C	16%
D	24%
E	29%
F	38%
G	43%

A3) Is Lending Club charging an appropriate interest rate for the risk?

The answer to this question, I believe, is based on Lending Club's profit from loans that are paid back, after accounting for the cost of loans that default, and administrative and marketing costs. Gross revenue is determined by the interest rate, which can be calculated. But the data is based on loans originating in 2015, and some of the loans have five-year payback periods. Therefore, the ultimate outcome of these loans, in terms of revenue realized or interest income not collected due to default, is yet to be determined. Although there is other data in the data set that could be used to estimate gross revenue and lost money due to default at a given point in time (such principal and interest received to date), this calculation is also not conclusive, due to what may happen during the remaining two years of the loan terms.

I don't have sufficient knowledge of the financial industry to answer this question, if it can be answered with the data provided. Presumably if I worked for Happy Money, I would acquire that knowledge.

I made a calculation that does not take into consideration the factors described above. That calculation was:

Gross revenue = Loan amount x (term / 12 x interest rate)

This calculation also does not accurately reflect how compounding interest works, I believe, but I am handicapped by the limitations of my knowledge as noted.

The outcome of the calculation showed the following:

Amount of revenue from loans not defaulted:	2.700119e+09
Amount of revenue not realized from loans that defaulted:	6.500872e+08

This would seem to indicate that Lending Club is not charging enough interest.

Section B

Data is often messy, please review and QA the Lending Club dataset and summarize your thoughts on any structural issues:

B1a) Is there missing data? Is the missing data random or structured? Are some attributes missing more than others?

Yes, there is missing data. Some seems to be structured, some appears to be random.

I prepared several visualizations to detect structured missing data, including matrixes, bar charts, and missing data correlation heatmaps. The initial matrixes and bar charts showed missing data, including several examples of the same missing data for the same cases, i.e., structured missing data. However, with the large number of variables for these initial visualizations, the matrices do not label the variables.

I removed all variables with 50% or more missing data, and reviewed bar charts and matrices again, as well as a missing data correlation heatmap. The matrices and bar charts indicated that a handful of the remaining variables had random missing data. The following variables were among those that had the most missing data: Employment length, months since recent inquiry, mo_sin_old_il_acct, and num_tl_120dpd_2. In the case of some variables, there are multiple variables that seem to be related indicators of the same factor (i.e., lateness in payments, recentness of new accounts being opened), including the variables with missing data. For many of these sets of variables relating to the same factor, often where one variable had missing data, the other related variable was complete. Therefore, I was not concerned about the few variables that had missing data after those with 50% or more missing data were removed. (All cases with missing data were removed before the machine learning analysis.)

The missing data correlation heatmap flags variable pairs with negative correlations if one appears and the other does not, and positive correlations "if one variable appears the other definitely also does." Given the strength and direction of the correlations on the heatmap, I did not see any patterns that were cause for concern.

B1b) Are there any glaringly erroneous data values?

I prepared box plots that flag suspected outliers. These are shown in separate Jupyter notebooks on the repo. GitHub doesn't render Jupyter notebooks, so please click on the theta symbol at the top right of the notebooks to view the interactive visualizations.

I made several box plots, but was not able to upload them to github due to the size of the files. However, there are two notebooks in the repo, one showing an example without suspected outliers, and another showing box plots for features with suspected outliers.

For all of these box plots, the data can be seen by hovering over the points on the plots. To view the data in areas of the plots where there is too much data to discern individual data points, selecting the area with your mouse will zoom in on it, and the individual data points are then visible. (Double click to reset the zoom.) Below are the variables for which there were many such suspected outliers. (I did not create box plots for all variables, just those on a scale that may have outliers, or for which outliers could be expected, or were likely to be problematic.)

Variables with suspected outliers:

- annual income
- mo_sin_old_rev_tl_op
- mo_sin_rcnt_rev_tl_op
- mo_sin_rcnt_tl
- num_accts_ever_120_pd
- num_actv_bc_tl
- num_actv_rev_tl
- num_bc_sats
- num_tl_90g_dpd_24m
- num_sats

Without domain knowledge, I'm reluctant to say with confidence that any of these suspected outliers are erroneous values. Additional knowledge of these features and expected values would allow me to identify values that are possibly erroneous. Also, there may have been erroneous data values for the variables with 50% or more missing data, but I removed those variables.

2) Using any format and any modeling technique that you prefer, please create a model to predict default within the Lending Club dataset. Show any work that you would deem important in evaluating this process.

I used random forests and was able to predict default with 85% accuracy. Code is in the Jupyter notebook in the repo.

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.

a) Linear regression makes predictions using a linear function of the input features. The predicted response is the weighted sum of the input features; the prediction is a line for a single feature, a plane when using two features, or a hyperplane in higher dimensions.

b) In a data science context, running a linear regression analysis allows us to quantify the effect different factors have on an outcome of interest. Running a linear regression analysis on home sales data, for example, quantifies the effect the number of bedrooms and the location have on the price of a home, thereby allowing us to estimate future home sale prices.