95-828: Machine Learning for Problem Solving **Spring 2019**

Case Study Phase 2: Data Cleaning, Preparation & Exploration

Instructor: Leman Akoglu

- In this Phase of the Case Study (CS), you will perform data cleaning and preparation for model fitting, followed by exploring the data to gain some insights.
- You are expected to submit a combined *pdf* of the following via *Gradescope*:

  (1) An assignment *write-up*, including answers to the questions in this sheet, should be submitted as **.pdf**, and
  (2) An iPython *notebook*, including all your code, visualizations, and outputs, should be submitted as **.ipynb**.

  Your write-up can be hand-written, but should be clearly legible—if the TAs cannot read your solutions, you will not be given credit. Submissions can also be written in Word or LaTeX.
- Each team is required to submit only 1 report. While submitting on Gradescope, make sure you enter *andrew ids* of your team members.
- The assignment is due at **11:59 PM on Friday April 12, 2019**.
- If you have any questions, please use *Piazza* or visit course staff during office hours and recitations.
- Do not copy from other sources, share your work with others, or search for solutions on the web. Plagiarism will be penalized according to the university rules.

2.1 Data Ingestion and Preliminary Analysis

Follow the list of steps outlined below, which will walk you through reading in the data, cleaning it from outliers, selecting features to use for modeling, and exploratory analysis.

1. Read the dataset that you downloaded in Phase I into Python. You will notice the data are provided in the form of many individual files, each spanning a certain time period. Combine the different files into a single data set.

LendingClub updates some of the variables in the data set as time goes by. As such, many of the variables in the data table will contain values that were *not* available at the time the loan was issued and therefore should *not* be used to build models.

The Python `notebook` you are provided with for Phase II contains code that looks at the files over time and compares values that may have changed between them. It also ensures

that features of interest have not changed too much.

2. Remove all instances (in our case, rows in the data table) representing loans that are still current (i.e., that are not in status *Fully Paid*, *Charged-Off*, or *Default*), and all loans that were issued before January 1, 2010.

3. Visualize each of the features in the file. Are there any outliers? If yes, remove those instances.

4. Visually identify correlations between the features as well as the features and the loan status. Write down your observations.

5. For the sake of this Study, restrict yourself to the following features: `id`, `loan_amnt`, `funded_amnt`, `term`, `int_rate`, `installment`, `grade`, `emp_length`, `home_ownership`, `annual_inc`, `verification_status`, `issue_d`, `loan_status`, `purpose`, `dti`, `delinq_2yrs`, `earliest_cr_line`, `open_acc`, `pub_rec`, `fico_range_high`, `fico_range_low`, `revol_bal`, `revol_util`, `total_pymnt`, `last_pymnt_d`, `recoveries`.
   Save the resulting data set in a Python "pickle".

A key measure we will need in working out an investment strategy is the *return* on each loan, defaulted or otherwise. Next we will calculate it and add as new variable to the dataset.

Calculating the return is complicated by two factors: (1) the return should take into account defaulted loans, which usually are *partially* paid off, and (2) the return should also take into account loans that have been paid *early*.

Ideally, we would want to take into account potential future reinvestments once a loan is repaid early however this complicates things quite a bit. Instead we will introduce three different ways of calculating the return, as described below.

6a. **Method 1** (denoted $M1$–**Pessimistic**) supposes that, once the loan is paid back, the investor is forced to sit with the money without reinvesting it anywhere else. This is the worst-case scenario. Under this assumption, the *annualized return* is calculated as

$$\frac{p-f}{f} \times \frac{12}{t} \tag{2.1}$$

where

- f is the total amount invested in the loan,
- p is the total amount repaid, and
- t is the term length of the loan in months.

The downside of this method is that the assumption is hardly realistic. On the other hand, it handles defaults gracefully, by spreading the resulting loss over the term of the loan, which is reasonable since the investor was initially intending for their investment to remain "locked up" for that duration.

6b. **Method 2** (denoted $M2$–**Optimistic**) supposes that, once the loan is paid back, the investor's money is returned and the investor can immediately invest in another loan with exactly the same return. In this case, the *annualized return* is calculated as

$$\frac{p-f}{f} \times \frac{12}{m}$$

where

- m is the actual length of the loan in months; i.e., the number of months from the date the loan was issued to the date the last payment was made.

The assumption that the cash can be reinvested at the same rate may not be realistic. However, the main issue with $M2$ is that if a loan defaults early, annualizing the loss can result in a huge over-estimate of the negative return. For instance, if a loan defaults in the 1st month, the investor loses 100% of the investment. This is the maximum loss, but annualizing it would lead to a 1200% loss (!) In other words, we would be assuming the investor reinvests in an equally risky loan for the 11 remaining months of the year, each of which defaults in 1st month! Hardly realistic. Instead, you will use the following two-piece formula:

$$\begin{cases} \frac{p-f}{f} \times \frac{12}{m} & \text{if} \quad p-f>0 \\ \frac{p-f}{f} \times \frac{12}{t} & \text{if} \quad p-f \leq 0 \end{cases} \tag{2.2}$$

6c. **Method 3** (denoted $M3$) considers a fixed time horizon (e.g., T months) and assumes that any revenues paid out from a loan are immediately reinvested at a yearly rate of $i\%$, *compounded monthly*, until the T-month horizon is over (in this Study, we will consider a 5-year horizon, i.e., $T = 60$). This method is closest to what would realistically happen.

Assuming each monthly payment was of size p/m which then are immediately reinvested, we can use the sum of a geometric series to find the total return from the f initially invested:

$$\left\{ \left[\frac{p}{m} \times \left(\frac{1-(1+i)^m}{1-(1+i)} \right) \right] \times (1+i)^{T-m} - f \right\} \times \frac{1}{f} \times \frac{12}{T} \tag{2.3}$$

6. Implement all three individual variables introduced in 6a.–6c. above (as given in Eq.s (2.1), (2.2), (2.3)) and add them as new variables to the dataset.

7. As introduced in Phase I, LendingClub assigns a grade to each loan, from A through G. Answer the following questions based on analyzing the data, and populate the corresponding entries in Table 2.1.

						Avg. return	
Grade	**% of loans**	**% Default**	**avg. intrst**	$M1$	$M2$	$M3$ (2.4%)	$M3$ (6%)
A							
B							
C							
D							
E							
F							
G							

Table 2.1: Table to be filled based on Q7.

(*i*) What percentage of loans are in each grade?

(*ii*) What is the *default rate* in each grade? How do you interpret those numbers?

(*iii*) What is the average *interest rate* in each grade? How do you interpret those numbers?

(*iv*) What is the average percentage (annual) *return* per grade (as calculated using the three methods in part 6.)? (Assume two different yearly rates for $M3$: ($i = 0.002$) and ($i = 0.005$))

(*v*) Do these numbers surprise you? If you had to invest in one grade only, which loans would you invest in?

Acknowledgements

This Case Study is inspired by and based on the following:

- Provost F, Fawcett T. (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media, Inc.
- Cohen MC, Guetta CD, Jiao K, Provost F. (2018) *Data-driven investment strategies for peer-to-peer lending: a case study for teaching data science*. Big Data 6:3, 191213, DOI: 10.1089/ big.2018.0092.