MLPS Case Study Phase 1

Zirui Zheng ziruizhe
David Lee dlee3
Keyu Zhang keyuzhan

1. *As an investor, what are the decisions you would need to make? (ii) Which of those decisions can you make using the available data from LendingClub and which one(s) would require additional resources?*

   i) As an investor, we need to first decide if it is wise to invest in Peer-to-Peer lending. Then, if we decide to invest, we need to figure out which loans to invest in, and how many notes to invest for each loan.

   ii) From the LendingClub data, we can decide which loans to invest. However, we need more information, such as stock/bond return, to decide on whether to invest in P2P. Also, we need more information like individual preferences between returns and risks to decide how many notes to invest for each loan.

2. *What is your objective when making those decisions in Q1? (ii) Explain how you would be able to distinguish "better" decisions from "worse" ones using the data?*

   i) Our objective is to balancing between maximizing return and minimizing risks. We can create an objective(utility) function that consists of both returns and risks. It will be a concave function that has a global max, which functions as our goal: maximum utility.

   ii) We will compare the data and pick the loans that have higher utility. As mentioned in Q1, based on different risk-return preferences among different customers, the objective function would include such a parameter.

3. *Note that loans are temporal entities (36 or 60 months-long term). Different loans could default at different times; some will default soon after approval, some much later. Some, on the other hand, might be repaid early, before their term ends. Would these facts affect your downstream analysis and decision-making? How/Why?*

   i) Yes, they would affect our downstream analysis and decision-making because some loans might be only partially repaid if they default, or some might be repaid early. We could recoup some losses from loans that were partially paid back but also lose some interest from loans that were paid back early.
   These facts would change our decision making because they would make the realized return lower than the nominal, expected return beforehand.

4. *Based on the discussions thus far, do you think historical data would be helpful? In which ways could you use such data to help make the decisions of your interest?*

   i) Yes. Historical data could help us estimate the prepayment risk, default rist, and adjust the expected return, so that we will have a better estimate on each loan provided.

5. *Next you will take a look at the data. (i) Write down a high-level description of the different features—that is, the variables describing the loans. How would you categorize these features? (Note that there may be multiple ways of categorizing the features; think in terms of the source of the measurements, the type, and temporal characteristics.) (ii) Just based on the feature descriptions, give an example to features that are likely to be (strongly) correlated. (iii) Which do you think are most valuable to an investor like yourself?*

   i) We describe the data features as follows:

   a) Features that describe the initial loan:
      - General info: “id”, “desc”, “url”, etc.
      - Interest rate: “int_rate”, etc.
      - Amount: “loan_amnt”, “funded_amnt”, etc.
      - Time: “term”, etc.
      - Status: “verification_status”, “loan_status”, etc.

   b) Features that describe the loaner:
      - Employment: “emp_length”, “annual_inc”, etc.
      - Home: “home_ownership”, “addr_state”, etc.
      - Hardship: “hardship_flag”, “hardship_type”, etc.

   c) Derived features:
      - FICO score: “fico_range_low”, “fico_range_high”, etc.
      - Loan grade: “grade”, “sub_grade”, etc.

   d) Features that describe the loan repayment:
      - Delinquencies: “delinq_2yrs”, “acc_now_delinq”, etc.
      - Payment: “total_pymnt”, “last_paymnt_amnt”, etc.
      - Debt settlement: “debt_settlement_flag”, “settlement_amount”, etc.

   ii) The derived features(FICO score, Loan Grade) seem to be highly correlated to the features describing the loaner, because they’re estimated by those features.

   iii) The most valuable features to an investor are the ones that describe the loaners and their loan repayment activities. Comparing to loan amount or interest rate, we care much more about if a loaner is able to repay all the debt.

6. *Next we will question whether or not it is a good idea to (a) use all of the provided features and (b) use them as is in our downstream modeling.*
   - *i) Consider the feature total pymnt (payments received to date). Do you think this feature is related to the loan status? Why?*
   - *ii) When investing in future loans, could you train a model that uses total pymnt as a variable? Why (not)?*
   - *iii) It is unclear whether the values of the variables in the dataset are current as of the date the loan was issued, or as of the date the data were provided. (For example, suppose we download the data in Dec 2017, and consider the feature fico range low for a loan that was issued in Jan 2015. It is unclear whether the score listed was the score in Jan 2015, or the score in Dec 2017.) Would this matter for your downstream modeling? Why (not)?*

i) Yes. If the total payments are much less than what the payments should be during the time period then the loan status is more likely to be late, default or charged off.

ii) No. We won't do that because we've already had the "loan_status" variable, which is highly correlated to the "total_pymnt" variable.

iii) Yes. It would matter since as the investor we would want the value of the features to be on the date the loan was issued. We cannot see ahead as the investor when we are deciding to invest or not and thus must use the value of the features on the date of the loan.

To answer this question, we don't need to use all of the provided features in general. A lot of features are highly correlated, so we don't need to use all of them.