**Case Study - Phase 3**

**Yuhao Zhu**

**1. (i) As an investor, what are the decisions you would need to make?**

a. Which loan should we choose as an asset worth to invest?

b. What is the ratings for each loan?

c. For how long we should hold a loan asset before selling to others? (If the loans are tradable)

d. How to construct the portfolio? (i.e. the proportions of each loan we hold)

e. What is the desired position? (i.e. the amount of money we allocate)

f. What is the risk preference and requirements? (Such as VaR, Durations, Convexities, Sensitivity rates, etc.)

g. What is the profit target under risk? (Considering Sharpe Ratio, Alphas, Risk-free Arbitrages, etc.) h. Other decisions on investment strategies according to personal preferences.

**(ii) Which of those decisions can you make using the available data from LendingClub and which one(s) would require additional resources?**  
As we only know about the data of loans, we would choose a. to make decision, that is, to  
determine whether a loan can be chosen to invest or not based on some criteria. Decision b. is a  
variation of decision a. that change the binary classification problem into a multi-classification  
problem. Other decisions would require additional resources.

**2. (i) What is your objective when making those decisions in Q1?**   
The objective is to maximize the Rate On Investment (ROI) with certain level of risks.

**(ii) Explain how you would be able to distinguish \better" decisions from \worse" ones using the data?**

We will use the Sharpe ratio to determine whether a loan is a worthy asset or not, where and are the HTM (Hold-To-Maturity) rate of return and risk, and the is the risk-free rate of return, which is usually the return of 1-year treasury bond. (It is plausible to assume that all loans on LendingClub will be held to maturity.)

**3. Would these facts affect your downstream analysis and decision-making? How/Why?**

These facts would affect our analysis since the HTM return of a loan is calculated by the discount cash flow over time. Since different loans could default at different times, those loans could be classified into different investment decisions or different ratings.

**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?**

Historical data would be helpful since we will construct machine learning models using historical  
data to classify the loans.

**5.(i) Write down a high-level description of the different features that is, the variables describing the loans. How would you categorize these features?**

The table has 151 features in all, features can be categorized to several types such as numerical (‘loan\_amnt’, ‘int\_rate’, ‘installment’ …), categorical (‘home\_owner’, ‘loan\_state’…), ordinal (‘grade’…); some of the features are more essential to describe loans, loan\_amnt, int\_rate, grade, annul\_inc. Some features are changing with time changed, such as ‘grade’, ‘annual income’. Some features are stable, such as loan\_status, zipcode.

**(ii) Just based on the feature descriptions, give an example to features that are likely to be (strongly) correlated.**

Grade and interest rate may be strongly correlated, the higher the grade, the lower the interest rate, it’s a tradeoff between risk and profit. Term and grade are also correlated; long term loans are always graded lower.

**(iii) Which do you think are most valuable to an investor like yourself?**

I think interest rate, grade and term are most valuable to me. Because I need to invest in loan that is suitable for my planned time, as well as an interest rate that I’m satisfied with, and a grade with risk I can take.

**6. (i) Consider the feature total pymnt (payments received to date). Do you think this feature is related to the loan status? Why?**

I think this feature is related with the loan status. If the total\_pymt is close to the total amount of loan, then it indicates that the loan is close to finish or the lender is close to pay all the money back. I think if the loan is default, it is more likely that the total payment is small.

**(ii) When investing in future loans, could you train a model that uses total pymnt as a variable? Why (not)?**

I do not think so. When investing the in future loans, there is no available data for total\_pymt since the loan has not start yet. So, I could not train the model using total\_pymt as the variable.

**(iii) Would this matter for your downstream modeling? Why (not)?**

Yes. For downstream modeling, it is important to find the indicator evaluating whether the loan is safe or not and it should be the up-to-dated evaluation. Suppose a lender is very trustable in 2015, however, he/she ran some crisis and could not pay the loan on time from 2016. The outdated 2015 data would overestimate his/her performance in 2016 or 2017.