CASE STUDY – PHASE 1

Name	Andrew ID
Bharat Madan	bmadan
Dhwani Panjwani	dpanjwan
Sahil Ahuja	sahilahu
Sujai Chaudhary	sujaic

1. (i) 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) A. One of the most important decisions you make as an investor is how much money to invest. For example, if you have saved $100,000, you wouldn't want to invest all of it, and even if you did, you won't put all of the money in the same investment. It would be better to diversify your portfolio to mitigate the risks associated with it.

B. Decide the type of investment based on your risk appetite. In the context of this case study, we would base our decision on the borrower's "grading scheme" and credit history.

C. We must also decide the time duration for which we want to invest the money - 36 months or 60 months.

D. Incase a loan defaults, we would want to know how would the money come back to us- whether it will be insured, payed back by a third party, etc

(ii) For part A,B and C we can make the decisions using the data from the Lending Club whereas for part D, additional resources such as type of insurance and which party would be liable in case of default.

2.(i) 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?

Our objective from the decisions of the first question is to maximize profits while also managing investment risk. In order to distinguish better decisions from worse ones, we can use data to analyze the performance of our investments. For example, we can evaluate the return on investment (ROI) for each borrower we invest in, and compare that to the level of risk associated with that investment. If we find that some investments are providing higher ROI while also being less risky than others, we can consider those to be better decisions. On the other hand, if we invest in borrowers with high rates of interest but consistently see a higher rate of default or loss, those would be considered worse decisions. Ultimately, by using data analysis to evaluate the performance of our investments, we can make more informed decisions that maximize profits and minimize risk.

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?

It is important to note that different loans may default at different times, and some may be repaid early. These factors can greatly impact downstream analysis and decision-making.

For example, if a borrower repays a loan early, we may miss out on potential interest that could have been earned if the loan had continued to its full term length. Conversely, if a borrower defaults near the end of the loan term, we may have already earned most of the interest we would have earned if the loan had been repaid in full, but we may still lose out on a portion of the principal.

Therefore, the time period of the loan is a crucial factor that can affect our returns. It is important to consider the loan term length and potential scenarios that may occur during that time when making investment decisions. By doing so, we can better understand the potential risks and rewards of each investment, and make more informed decisions that maximize our returns while minimizing our risk.

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?

Yes, historical data will be helpful for our purpose. Suppose we want to build a model which predicts if a loan will be fully repaid or not, our model can look at the borrower details of historical loans and compare that with the borrower details of a loan which is still being funded and predict the chances of full loan repayment.

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? (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.)

There are 151 features for loans that were issued on LendingClub between April 2008 and September 2019. These features provide insights into different stakeholders involved in the lending process for a given loan. There are some features that provide information about the borrower, especially their background. Then there are features that inform us about the financial records of the borrower, these are specifically important as they will tell us about their current financial health and how have they fared till now for their various loans. There are some features provided by LendingClub themselves that would give us the specifications about the loan like grade, int_rate, term, etc. In the end, there are features that define whether the borrower has been previously or is right now on any hardship and settlement plans.

According to LendingClub, “With a debt settlement program, you negotiate debt forgiveness. This means your creditor agrees to accept less than the amount you owe—usually about 50 to 80 percent less.”

According to LendingClub, “Hardship plans are commonly offered to borrowers in the lending industry because they allow borrowers time to adjust to a life event (like a medical emergency, temporary job loss, unexpected car or home repairs, death in the family, or other events)”.

These features can be categorized based on Source of Measurement (Self-Reported, Credit Report, LendingClub Record, Hardship and Settlement third-party vendor), type (numerical, categorical, text), and temporal characteristics(fixed with time or dynamic.

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

Features likely to be strongly correlated –

- Annual_inc and Annual_inc_joint: These features both describe the borrower's income and are likely to be strongly correlated if the joint applicant's income is similar to the primary borrower's income.
- Last_fico_range_high and Last_fico_range_low: These features both describe the borrower's credit score and are likely to be strongly correlated, as a higher credit score in one will generally result in a higher credit score in the other.
- Dti and Dti_joint: These features both describe the borrower's debt-to-income ratio and are likely to be strongly correlated if the joint applicant's debts and income are similar to the primary borrower's debts and income.
- Percent_bc_gt_75 and All_util: These features both describe the borrower's credit utilization and are likely to be strongly correlated, as a higher credit utilization in one may result in a higher credit utilization in the other.
- Grade and Int_rate: These features are both related to the interest rate on the loan and are likely to be strongly correlated, as higher-risk loans (higher grade) will generally have higher interest rates.

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

As an investor, I would like to maximize my returns and minimize the risk. So, I must find loans where there is a good trade-off between these two, as I could earn more by investing in high-interest loans but there would always be a risk of the borrower defaulting. On the other hand, investing in safe loans will not only give low returns but if paid before the stipulated time I would lose valuable interest earnings. I think the following features would help me in making an informed decision –

Annual_inc, Delinq_2yrs, Desc, Emp_length, Last_fico_range_high, Last_fico_range_low, Grade, Int_rate, Term, Hardship_flag, Pub_rec_bankruptcies, Verification_Status, Term

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)?

a. Using all the provided features won't be a good idea because, features like zip_code, url, etc. don't seem to impact our investment decisions in any way. Also, the feature grade seems to be highly correlated with the features related to the borrower like annual income, monthly FICO score, etc. adding all these variables together will make the interpretability of the model a challenge and it will be difficult to measure the impact of each feature.

b. Again, using the features as is won't be a good idea because the dataset contains features of mixed types and many of the numeric features are in different scales. Hence, it will be useful to apply pre-processing techniques like encoding and scaling to broaden the scope of models we can fit to our dataset.

i) Yes, the feature total_pymnt is related to the loan status. If the total payment is equal to the principal+interest then the loan status is fully paid. If there is some outstanding amount of principal or interest, the loan status will be current or late depending on when the loan term ends.

ii) For future loan investments we can’t use total pymnt as a variable in our model. When checking if a loan is worth investing in, we won’t have total pymnt feature available to us because the loan term has’nt started yet and it’s still in funding stage.

iii) Taking the example of fico range low, it is important that the value of this variable is that of when the loan was issued because, when we are using the model to decide if a loan is worth investing in or not, we will only have access to the current fico score of the borrower and not the future score. If the value of fico range low is not that of the loan issue date we will end up making incorrect predictions and this will impact our downstream modelling as well.