Predictive Modeling and Investment Strategy for Loan Defaults

Phase 1: Preliminary Questions

Nima Shafikhani

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

There are several aspects that an investor could consider:

1. Investment goals / investment strategy: the objective / ultimate gain of the investment. In other words, this is what the investor would expect and want to achieve from the investment.
2. Risk tolerance: whether the investor is risk averse / risk seeking. For instance, an investor could prefer high risk - high return products, low risk - low return products, or hold mixed portfolios.
3. Asset allocation: depending on the investor's preference and risk tolerance, the investor could then decide how to allocate the investment amount into different products by considering the expected return of each asset, e.g. how much portion should be assigned to high / low risk assets.
4. Investment timeline: how the portfolio should be allocated between the 36 / 60 months assets.
5. Monitoring and adjustments: under which circumstances (e.g. thresholds / triggers) will the investment instructions change and the frequency of monitoring / reviewing investment portfolio in order to make adjustments.

(ii) Which of those decisions can you make using the available data from LendingClub and which one(s) would require additional resources?

We can make certain decisions using the LendingClub dataset, such as determining which loans to pick and how we plan to allocate the investment amount. However, if we are also investing in other lending marketplaces (i.e. those other than LendingClub), we might need to refer to additional resources (e.g. other investment options and their data). In addition, we would also have to consider our personal preferences when investing, such as how risk averse / risk tolerant we would like to be with our investments. This preference would likely be altered by our financial situations and how much we could afford to lose.

2. **(i) What is your objective when making those decisions in Q1?**

   Maximizing returns on investment and earning money would be the investor's main goal.

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

   Studying the historical trends for default rates to understand if any particular type of loan / characteristics of individuals / any other features, etc., can help make better future decisions and build a better prediction model.

   When building our model, we must also consider how we are going to split the data into training and validation sets (e.g. it might make more sense if we split the database temporally, and have a certain representation of each year in the training and validation set). Since the data is spread across many years, if we only train and validate on a certain time period, it may not be an accurate predictor of the future. The underlying assumptions (factors like GDP, public purchasing power and other economic indicators) change every year, so a model trained and evaluated on data from 2018, will probably not perform well on data from 2022, especially since there have been major changes since the COVID-19 pandemic.

   Moreover, we may possibly need separate models for the two types of loans (time spans of 36 or 60 month loans) since the characteristics of each may vary a lot from the other.

   After the model is built, we can fine-tune it based on how it performs on the validation data.

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?**

   Yes! Net Present Value (NPV) of a loan would need to be calculated to inform the decision of taking a long-term or short-term loan. Generally speaking, the sooner the loan is paid off, the better it would be in terms of securing the investment at a higher NPV since inflation decreases the value of the dollar. However, the interest rate payments would be lower since interest has less time to accrue. The sooner a loan defaults, the worse it would be! However, we are assuming that the loan will have a higher interest rate (since a low risk loan is very unlikely to default), so we *might* be able to recover our initial investment before that. Hence, it is imperative to consider NPV while making an informed decision of where and how much to invest.

   In addition, while a lot of features in the dataset may not be able to give us direct calculation results (e.g. return on investment), we might have to consider features like the above example, which are not straightforward but rather imply some underlying possibilities.

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 could allow us to pinpoint factors which could serve as flags in our decision-making process. We can train the model on those parameters and then make data driven decisions. For instance:

   - Where to invest: We can identify the characteristics of individuals who pay/default and thus make an informed decision about our portfolio
   - How much to invest: Historical data can give us perspective on how investments have fared in the past on different types of borrowers (high / low risk) and what the break-even time generally has been with each

   Besides, when using historical data, we should keep in mind that since this data is temporal, assumptions that were true in historical data may not always be valid in the future. Data distribution may shift and temporal data generally could diverge. This is why these asset prices could be considered random in a way.

   A relevant example would be the Great Recession from 2007-2009, where there was a sharp decline in economic activity and the Global GDP declined by over 5%. It would be safe to say that if we looked at LendingClub data from 2007-2009, it would not be comparable to the data that we do have.

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.)

   We may categorize from the business perspective (i.e. nature of the feature). Please see below for an extracted list of the feature examples under different categories:

Category	Column Names (Extracted)
Borrower-related	addr_state, annual_inc, application_type, earliest_cr_line, emp_length, emp_title, fico_range_high, fico_range_low, home_ownership, next_pymnt_d, pub_rec, pub_rec_bankruptcies, tax_liens, verification_status, revol_bal_joint, hardship_flag, deferral_term, hardship_amount, hardship_length, settlement_status
Loan characteristics	funded_amnt, funded_amnt_inv, grade, installment, int_rate, loan_amnt, loan_status, out_prncp, out_prncp_inv, term, desc
Payment / Credit Transactions	acc_now_delinq, chargeoff_within_12_mths,, delinq_2yrs, delinq_amnt, avg_cur_bal, max_bal_bc, num_actv_bc_tl,, num_sats, tot_coll_amt, tot_cur_bal, tot_hi_cred_lim, total_acc, total_bal_il, total_bc_limit, total_il_high_credit_limit,

   Some of these features are categorical (e.g. grade, status), and some are continuous / numerical (e.g. payment amount, late fee). Some are updated once (e.g. application type, earliest credit line), and some are updated frequently (e.g. next payment date, total payment amount).

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

   - Addr_state and homeownership could be strongly correlated, e.g., New York has expensive housing and therefore homeownership might be very low there. Another state with cheaper housing may have higher homeownership.
   - FICO could be correlated with homeownership, addr_state, loan_status. A better FICO can be associated with a non-default loan_status, and having a home as

well. FICO could also vary by state; states with richer people might have a higher FICO on average

- Grade and settlement_status would be strongly correlated, since the former is an indication of the latter.

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

As an investor, I would care about the return on investment (ROI) the most. While the features in the dataset may not let me calculate the predicted return, we would need to consider features that could help us estimate the return / changes of return-related features. In order to build such a model, we have to be careful when selecting the features to use to make sure we are not oversimplifying the problem or creating a model that cannot be generalized.

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.**

   Depending on which model we are using, it could be ok to use all the provided features. For example, if we choose to implement a neural net, then feeding all features into the model *could* lead to a better result.

   However, for a simpler model, it is not a good idea to use all of the provided features; could lead to the curse of high dimensionality. If we include all of the features, it could lead to a low bias and high variance model, and it is very likely to overfit.

   We also cannot use features without transformations since the scales and types of each are different. The categorical ones might need to be One-Hot Encoded (depending on the model we use) and the numerical ones might have to be standardized.

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

   This feature is somewhat related. For example, lower total_pymnt could indicate that the loan is still being paid. High total_pymnt could indicate that the loan is nearing completion. However, we would need more information to identify loan statuses like late or default, since total_pymnt features alone will not be enough. We would also need information about the last payment date.

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

   We can only use it if the variation in the loan status is associated with a variation in total payment historically. Since we are hypothesizing that there will not be too much of a correlation unless we factor in other features, in this case probably we should not use this to train the model.

   (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)?

Yes! The ranges of the FICO, e.g., might change over time and therefore may not allow apples-to-apples comparison. This is what we should always be aware of when using historical data, as changes may have occurred related to the borrower and also other economic factors may have changed (e.g. the macroeconomic conditions).