Lending Tree Final Project

Group Members: Vikram Maduskar, Robert Persichitte, Pavan Reddy, and Catherine Williams

1. Motivation:
2. Problem Statement:
3. Introduction and Description of Data:
4. EDA:
5. Models:
6. Results and Conclusion:

# Project Objective

Create a model to choose Lending Club investments that maximize expected return.

The goal of the analysis was to choose a model that maximized the expected rate of return on investments. We considered the predictors that would be available to an investor at the time of purchase and which of them would be useful for our prediction. We considered which variables might be a good indicator of lifetime return. We also considered the risks and limitations of our predictions.

# Predictor Variables

Determining effective predictors for the Lending Club data set is challenging for a few reasons. The data set has many variables (150+) and one would need separate appropriate response variables (discussed in the next section) from the full set of features based on study goals.

But the sheer number of remaining variables results in a number of additional complications that need to be accounted for:

* Predictors available at time of loan underwriting
* Descriptive vs predictive data
* Sparse data
* Categorical data
* String data
* Collinear features

## Predictors available at underwriting:

Before building a model, we needed to understand the variables are actually known at the time of loan decision vs being collected throughout the life cycle of the loan? Clearly this would impact the fidelity of our model. We have flagged which variables would likely be available to loan underwriters but will further review to make sure we can utilize all relevant factors.

## Descriptive variables:

There may be multiple variables which are more of a descriptive nature vs having predictive value. We have reviewed the data set for these types of variables (URL for example) but will take another pass in the coming weeks.

## Sparse data:

Much of the data is sparse (no value i.e. NaN) and we need an appropriate strategy to either fill in a value, remove that observation, or remove that variable. We note that removing observations without completely available data would limit our data set to 8000+ rows (from a total of 1.8M+ observations). We will examine if there is an effective fill strategy but for now we are removing the sparsest variables from the data set which leaves us around 800K+ observations to work with.

## Categorical data:

The data set seems to have categorical data of potential predictive which we have transformed using one-hot encoding. These variables include: 'application\_type', 'term', 'verification\_status', and 'home\_ownership'.

## String data:

The data set also includes string data that could be useful if transformed into appropriate formats including factors like length of employment, interest rates, and revolving credit.

## Collinear features:

Finally, there are a significant number of features (34) which exhibit high correlation. One presumes it should be possible reduce total feature count by 17.

# Response Variable

We needed to choose a variable to assess the expected return of an investment. We ultimately chose the total amount of repayments to assess the expected return. To understand why that is the best predictor, I will first explain some of the alternatives we considered and the limitations in all the possible variables to assess. We researched as predictors loans that were current or paid in full, loans that were charged, and the total amount of repayments.

The predictors for loans that are current or paid in full and loans that were charged off can all be assessed with a simple logistic regression. Having an accurate idea of the chances of a loan being successfully repaid is essential to both predictors. First, looking at loans that are current gives us the advantage of seeing more recent observations—namely loans that have been issued within the last 60 months and are currently in repayment. Having more recent data would overcome some of the challenges of assessing the effect of broader economic trend on a borrower’s ability to pay. Looking at loans that have been charged off would give us an idea of which loans are riskiest. Knowing that information, we could select investments that have the lowest probability of being charged off. It is the same basic principle as looking for loans that are current but looking for loans that have become totally worthless. There is a category of loans somewhere in the middle which is loans that are either in a hardship payment plan or simply late on payments. We will have to define our categorical boundaries to include or exclude these values. That categorization is challenging because loans with late payment–and late fees–may end up being a better expected return.

The biggest limitation of building a model that predicts the likelihood of repayment is the impact on total return. Because riskier loans have higher interest rates, it may be better to select the risky loans, and accept losses from defaults hoping that the high interest rate will offset any losses. A model predicting repayment does not help us determine the appropriate balance between risk and interest rate.

To account for the problems mentioned above, we chose as our response variable . In other words, we are looking at the total return compared to the amount invested. A $25 loan that was charged off after received $5 in payments would have a value of 0.2. A $25 loan that was paid in full including $10 of interest would have a value of 1.4. This encompasses total return of an individual loan. By encompassing total return of individual loans, we can predict the expected value of a group of loans.

This variable still has limitations. First, we must limit our analysis to loans that will have no future expected payments. In other words, we can only look at loans that have been paid in full or charged off. This means we will have to look at loans that have fully run their course which can take up 60 months or more if the loan was deferred at any point. Second, we do not have information about when the payments were made. A loan could have had no payments for 90 days and then paid principle, interest, and late fees. In other cases, the borrower may choose to pay the loan in full before the maturity date. Because we cannot calculate these nuances, we will only consider the return over the lifetime of the loan.

# Limitations

Our initial concern is that the data set that is available from Lending Club includes loans as early as 2010. Given that the economic trends of our country have not stayed consistent since that time, it is hard to guarantee that predictions made from this data would hold true to today’s economic situations. While we would ideally want to solve for this, our model includes no economic trend data. While we could possibly control for economic trends and time periods of growth and recession in our models, that is beyond the scope of this analysis.

As a potential next step for our analysis we would incorporate broader economic indicators to understand their impact on the expected rate of return on the loans. This would be especially difficult given the complexity of predicting economic trends. While we have retrospective data on borrowers and lenders, there is an advantage to knowing what happened to economic markets while the loan was in repayment. We would not have that information when deciding which loans are suitable for investment. This is a common problem across the entire financial sector. So, while this is not unique to us, it is important to always caution a model’s interpretation with knowledge that because financial data is yet to be perfectly predicted.

Because of this limitation, the actual return of the loans our model chooses may differ from the expected return we calculate. They may differ if the broader economic situation is not as stable as it has been during the period we selected our training data from.

# Next Steps

The next stage of our study will include higher order features since our current R2 via Linear Regression is relatively low (0.162 on our test set). We will continue to explore models that have greater predictive value.