Executive Summary

Prosper and LendingClub are two P2P lending platforms that were established at very similar times but differ slightly as a result of their interest rates, borrower fees and application requirements. The aim of this study is to replicate and compare the LendingClub data driven strategy analysis using Prosper data. The dataset for LendingClub is more diverse containing over 100 predictors when compared to the Prosper dataset.

The comparative study is divided into 6 respective parts: Introduction and Objectives, Data Ingestion and Cleaning, Data Exploration, Predictive Model Analysis, Investment Strategies and Optimization.

Study Objectives

The overall aim for a potential investor is to make as much money or "markup" as possible on their investment. To get a better understanding of which loan to invest in, the user should be able to split their decision into two parts. The first highlights the importance to decide (which loans to invest in) and the second half be used to evaluate them (how much money could they possibly earn). Both cases share the same objectives and will leverage past data to predict the potential good loans and allow investors to invest in them to generate positive returns. There are several key differences between Prosper and LendingClub:

1) LendingClub has more predictors
2) Prosper requires higher minimum loan amounts
3) LendingClub has a higher minimum APR
4) LendingClub has lower origination fee %s
5) Prosper has a higher credit score requirement of 640
6) Prosper has a higher debt-to-income ratio of 50% relative to LendingClub's 40%

Data Ingestion and Cleaning

To derive comparable insights between LendingClub and Prosper, there were multiple assumptions regarding some of the features that had to be made: principal paid considered as recoveries, loan length determined by the subtraction of last payment by issue date. A binary transformation was performed on the loan status, completed was 0 and not completed was 1 because we were more interested in determining loans that were not paid back in full. To prevent information leakage, variables that are computed during or after the term of a loan were disregarded. The correlation heatmap was did not reveal any major red flags to identify possible confounders with loan status. These include but are not limited to the principal paid and interest paid. To be specific, interest paid, and principal are determined once the loan has been provided by the investor.

Data Exploration

Since the goal of thecomparative study is to both identify the defaulting of loans and predict expected return on loans, similar to LendingClub, expected returns were calculated for prosper under 3 different scenarios: pessimistic, Optimistic and fixed-horizon The pessimistic approach

states that when the loan is paid back, the investor cannot re-invest it until the term of the loan. The optimistic approach entails that once the loan is paid back, the investor's money is returned, and the investor can immediately invest in another loan with the same return. The fixed-horizon approach involves calculating fixed-time returns for 3 different interest rates: 1,3 & 6 %. A thorough analysis of the prosper score breakdown by grade reveals that numbers are similar to that of lending club but scaled relatively higher in terms of expected returns. This is consistent with Prospers model of providing of higher returns, but also charging high amounts of service charges and fees. What we also notice is that the amount of defaulting loans by grade is consistently higher for LendingClub in comparison to Prosper. We also noticed that number of observed non-complete loans (loan status = 1) had considerably less observations in the dataset. Our analysis revealed the manual balancing would decrease the expected returns as the proportions do not reflect the underlying investment trends (There will always be more loans that complete rather than default).

Predictive Model Analysis

Stage 1: Classifying loan outcomes

A variety of models were tested using a different set of features. The first to determine the predictive power of the grade and interest rate of a loan (these are usually very strong indicators of the eventual loan status). The accuracy of models using solely grade and interest rate respectively were: 85%. This implies that these two features individually contribute to determining the loan status equally compared to when all predictors are used. This is consistent with the findings of LendingClub. Thus, to build a more robust model, the two features were dropped for future model building.

Using all the predictors except grade and interest rate, model accuracies were maximized at roughly 85%. Naïve Bayes although provided a very high accuracy was unable to account for biased data by the imbalance. Thus, for any future requirements of classification of loan status, random forest was utilized. Random Forest also proved to be the better classifier in the LendingClub case.

Stage 2: Predicting Expected Return

A variety of models were tested using a different set of features. The R^2 values of various regression models were calculated for the different return methods. The Prosper regressor model results and trend coincides like the one in the LendingClub case. In both platforms, the Random forest regressor gave the highest R^2 value across all the different types of returns when compared to the rest of the models built. It is worth noting that we utilized a few different types of regressor models than those in the LendingClub case. When testing the model stability over different time

The results indicate that there is a strong drop in the performance of the model when trained on data from 2013-2015 and tested on 2018 data. The AUC worsens to around 50 percent and the model is highly uncalibrated. This is in complete contrast to that in the LendingClub case which shows that model's performance is remarkably stable. In other words, the model performs well using older datapoints to predict for the more recent years without much effect to the performance. Thus, there is strong evidence that Prosper data had a very large change in business operation anywhere between 2013 and the end of 2018.

Investment Strategies

By leveraging both the classification and regression models to predict whether a loan will default and the estimated return, an investment strategy can be formulated to maximize an investor's average return based on the based the different mythologies to calculate the return. Four different investment strategies were tested: Random Strategy, Default-based, Simple return-based, Default and return-based strategy.

Investment Strategy	Pessimistic Return (M1) %	Optmistic Return (M2)%	M3 (1.2%)	M3 (3%)
Rand	1.5	5.7	2.5	3.4
Def	7.3	7.5	2.7	3.2
Ret	1.1	7.3	2.6	3.5
DefRet	1.2	6.3	2.7	3.3

Investment Strategy Return Percentages for Prosper

In order to depict real life scenarios, the M3 return approach seems to be a more accurate way to measure the % return when compared to LendingClub's M2 optimal return approach. Thus, the baseline for us to compare the different investment strategies would be under the M3 method. Nonetheless, the returns seem to be very close to one and another regardless of which investment strategy is being used when looking at the M3 return approach.

The magnitude of the % return is also highly dependent on the reinvestment interest rate (M3). This trend is consistent with the LendingClub case results in Table 4 because all the expected % return values are higher with a higher interest rate % regardless of the investment strategy and is very close to one and another in both cases.

Optimization

In this section, we implement three different optimization models to improve an investment strategy using Prosper. The three different optimization methods are: Directly maximize total Profit, maximize profit with budget constraint, Maximize profit with risk-return tradeoff. Results show that the optimal investment strategy for Prosper is maximizing profit using budget constraint with an expected return on 10.5% contrary to LendingClub where maximizing profit with risk return tradeoff yielded an expected return of 3.18%.

Future Extensions

It would be interesting to see how our model would perform if tied to macroeconomic external data that would represent the underlying economy's performance at the time such as oil prices or The World Bank's interest rates, to further extend the pessimistic measure's performance. Additionally, it would be interesting to conduct sentiment analysis over the investor to understand what their risk appetite exactly is as opposed to asking for the numerical input as they may lack the understanding of how the P2P lending market operates.