MODEL METHODOLOGY

Calculating Actual Return in first 12 months

Before constructing a model to improve investor performance in Prosper loans, we need to define what our metric of performance would be. We wanted to calculate expected returns after 12 months on books. The data only contained the latest month of performance. We needed to take this single month and back into the returns at 12 months on book. The return on a loan can be defined as the interest payments we expect to receive plus any fees minus the principal lost on bad loans.

First, we approximated when the customer stopped paying. If a borrower defaulted on a loan, the principal received would remain constant for all future months as the lender receives no further payments. Since term loans amortize on a fixed schedule, we can determine the month the borrower stopped making payments by comparing the actual principal received to the expected amortization pattern. This method will not be accurate for borrowers who miss a loan payment, but resume paying again before eventually defaulting. With every missed payment, the predicted month of default will be off by an additional month. However, the interest earned on these particular borrowers will be higher than those that roll straight to default. Since the amortization approach will yield the same expected principal recovered and underestimate the expected interest gaine, we prefer to take this conservative approach rather than attempt to model the additional expected interest from late payers.

Next, we calculated the expected interest received. For customers who defaulted, we can use the interest rate and term to estimate the expected interest received at the time of default. For customers who did not default, we would use the same amortization schedule for the entire 12 month period. However, for these customers we also needed to consider prepayment risk. A borrower with the liquidity to pay more than the minimum loan payment can reduce his or her overall interest expense by doing so. We need to account for the prepayment activity, or otherwise the expected interest will be overstated. Our research found that interest received in the first 12 months on good accounts is roughly 10% lower than expected. This 10% prepayment adjustment was factored into our interest calculations.

Finally, we need to add in the fees that impact an investor in Prosper loans. Lenders pay an annual servicing to Prosper equal to 1% of the outstanding balance of the loan. Lenders also pay for any collections efforts Prosper undertakes to recover the principal in the event of nonpayment by the borrower. Lenders receive all late fees paid by the borrower in the event of a missed payment. All these fees are recorded in the dataset and can be used to calculate expected return.

Modeling Expected Return

Our goal for the OptiLender strategy is to outperform Prosper’s rating grade. On average, given a risk tolerance level from an investor in the form of a Prosper rating grade distribution, our model should be able to recommend a pool of loans with comparable or lower risk that produces higher returns. This is not to say we expect to outperform Prosper’s underlying pricing model necessarily. Instead, we want to identify high return pockets within the general pool (which Prosper may or may not be aware of) and make them visible to investors.

The Prosper dataset contains over 500 attributes about the borrower that could be used at the time of funding to predict the performance of the loan. These attributes include customer demographics, employment details, credit bureau information, and any prior loan information if the customer has used Prosper before. We explored many of these variables, but ultimately settled on utilizing only 10 features in the final model, including Prosper Rating.

Attribute Name	Attribute Description
Prosper Rating	A proprietary rating developed by Prosper
Borrower State	State
Months Employed	Length of employment in months
DTI w Prosper Loan	Debt to income ratio including the monthly payment of the entire listing amount
Total Inquiries	Total Credit inquiries
Bankcard Utilization	Sum of the balances owed on open bankcards divided by the sum of the credit limits.
Revolving Balance	Total balance in USD of open revolving trades
Installment Balance	Total balance in USD of open installment trades
ALL208	Total monthly payment on all open trades reported within 6 months of the profile date
ALL701	Age, in months, of oldest trade

Table 1: Final Model Attributes

Our model is built using a Decision Tree Regressor. This approach allows us to model a continuous variable (rate of expected return), while maintaining the explainability of a tree. The decision tree structure also classifies each loan in the portfolio into a end node, which is conducive to our end goal of dividing the portfolio into high-performance and low-performance pools. The model is in fact a combination of six separate models: one tree built for each year of originations since 2010. To train each model, we took only loans that were at least 12 months old at the beginning of the origination year. For example, the 2010 model uses only 2007 and 2008 originations because the 2009 data is too young. At the start of each new origination year, we update the model with another 12 months of data and recalibrate. Prosper's exponential growth creates a natural weighting system for the age of the data, so that newer loans have a much higher weight in the model calibration but we still retain older performance for stability.

After training a model using historical data, we will run every loan through its respective vintage model. The model will produce an expected rate return for the next 12 months, as well as classify the loans into leaves. The predicted rate of return does not include any information about the volatility of returns, so we have secondary process to quantify this risk. We run the training data through the models to classify the loans into leaves. We then pick the first leaf, select 100 random samples of loans within that leaf, and calculate the average return for each sample. The standard deviation of these returns is our expected volatility for that particular leaf. We repeat this process for every leaf and join the resulting risk estimates back to the loan level data. One point to note is the volatility of a pool of loans is largely dependent on the size of the sample. A particular pool that has a very high average return and very sporadic returns by loan might not be the best choice for an investor who can only afford two or three loans, but could be very valuable to an investor who can diversify this risk away by purchasing 100 loans in the pool. We vary the sample size used in the calculation of standard deviations to allow users to find pools of loans that meet their preferences or restrictions. After appending the standard deviation calculation, we summarized the results by leaf and vintage year. A subset of the data is shown below.

leaf	year	Pred_Return_Rate	Pred_Std	Act_Return_Rate
5	2010	0.064273	0.002135	0.072642
6	2010	0.105283	0.006769	0.200860
12	2010	0.034257	0.032607	0.123618
13	2010	-0.006123	0.041872	0.004297
14	2010	0.065244	0.002558	0.087413
15	2010	-0.015765	0.055524	0.080894
22	2010	0.026854	0.027749	0.069006
26	2010	0.058548	0.022455	0.096401
27	2010	0.085599	0.002379	0.137838
29	2010	0.064023	0.001890	0.075507

Table 2: Expected Risk and Return by Decision Tree Leaf

At this point we have a loan level data with expected risk and returns from the based on the leaves from the decision tree model, and the actual returns in the first 12 months. We also need the expected risk and returns assuming based on the Prosper Rating alone. Using the same train/test splits we used to build the decision trees, we calculate the average return in the training data by Rating and append it to the test data. The historical returns will be our expectation for the current vintage. We calculate the expected volatility using the same method as before, randomly selecting 100 sample portfolios from the training data and computing the standard deviation of returns. A subset of the end result is shown below.

prosper_rating	year	Prev_Return	Prev_SD	Act_Return
A	2010	0.052086	0.016691	0.053361
D	2010	0.034653	0.041008	0.141839
HR	2010	-0.034821	0.055910	0.184938
AA	2010	0.053841	0.021982	0.047579
E	2010	0.038142	0.043496	0.173507
C	2010	0.022205	0.048488	0.113642
B	2010	0.064970	0.013148	0.088318
A	2011	0.055142	0.014808	0.064211
D	2011	0.042108	0.047729	0.126505
HR	2011	-0.033197	0.059908	0.113637

Table 3: Expected Risk and Return by Prosper Rating

Backtesting Performance

To test the performance of our model, we will select leafs with comparable or better performance than the Prosper Ratings in the predicted data and measure the variance of returns in the actual data. For each Rating and Vintage combination, we would recommend loans where the predicted return is greater than the previous return of the Rating while the predicted standard deviation is lower. We then look at the actual return on those loans relative to the actual return on the Rating to determine if the investor would have been better off with our model or the default Rating In the user-facing program an investor will be able to build their own distribution of Ratings and the tool will recommend loans based on the aggregate return and standard deviation of their distribution. However, for benchmarking purposes, comparing to a single Rating is sufficient.

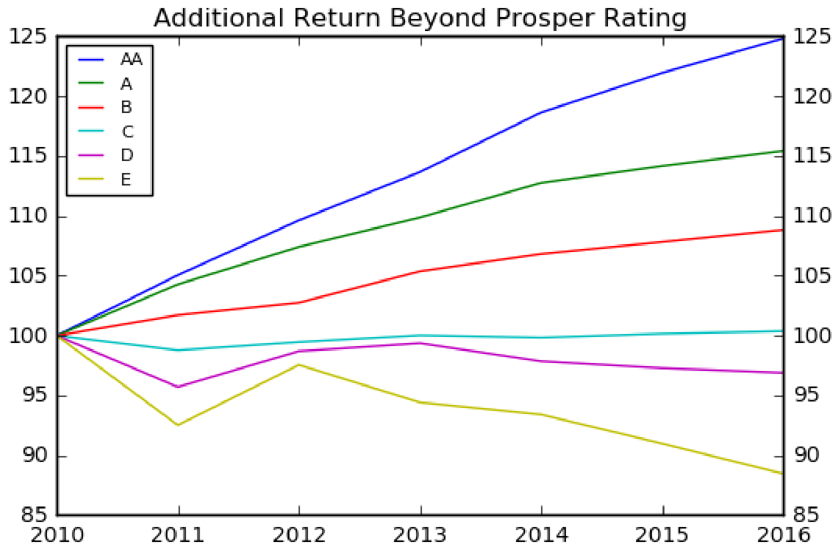

Figure 1: Variance of Model Return to Prosper Rating Return

The model works very well for lower risk segments. Investors that have a risk profile of Rating "B" or higher would be better off investing using our recommendations. Over the 6 year backtesting window, investors could expect at least a 10% lift in expected returns. Unfortunately, the model does not perform very well for high risk segments. Prosper's "D" and "E" ratings produce 20-30% more return over 6 years than our model recommendations. This does not mean that our recommendations would have lost money, but rather that the investors paid a heavy opportunity cost. Prosper Ratings of "C" perform comparably to our model output over the backtest period, but this is not a reassuring message to an investor who abandoned the security of the default rating system for higher returns.

There are a number of reasons for why the model would succeed on good ratings and fail on high risk pools. It's likely that the features most predictive of inability to pay back a loan are very different for low and high risk customer. For example, a customer with good credit could be financially knowledgeable and therefore managing a large number of credit products is not a risk factor for him or her, while a customer with poor credit may be more likely to become overwhelmed if there are too many credit cards and loans to pay every month. We might have seen higher performance if we had constructed two separate models: one specifically tailored to high risk borrowers and the other focused on low risk. Another reason the model underperforms could be the stability of performance in the rating grades. AA and A rated customers tend to have more stable performance, and barring any serious economic changes their returns should be roughly the same year after year. D and E rated customers could be more sensitive to minor impacts. In this case, predictions for higher risk pools could benefit from monthly model re-calibration with less weight applied to older historical data.

A third potential source of model error is the Prosper Rating itself. Prosper is a relatively new company in the grand scheme of the lending industry and its models have not been around long enough to become stable. Add to this that higher risk pools are fairly unstable anyways, and it's highly likely that Prosper has built many different models through the years to define its Rating Grade. We use the historical data to set the return expectation by Rating for the upcoming year, but the types of loans that composed a D rating in 2010 might be entirely different from those in 2011. In an environment where the underlying definition of a rating changes frequently, including that feature in our model could be detrimental to results. Other possible reasons include a lack of account level examples in the lower ratings to calibrate the model, and the fact that these accounts are more likely default in the first year leading to higher volatility of returns.

Model Weaknesses and Assumptions

The Prosper dataset we used in our analysis had a number of limitations. The most significant limitation is the performance data only shows the most recent month, not a time series since origination. We are able to see the final outcome of the loan, but not the path it took to get there. Overcoming these limitations required making assumptions about the unknown factors.

A risk to the revenue stream of lenders is prepayment by the borrower. Prepayment lowers the total interest rate received as the borrower pays back the remaining principal before further interest accrues. While the argument could be made that lender now has the full principal back to reinvest, this only holds true if the limiting factor of growth for the lender was access to capital. It also ignores the fixed cost of originating a loan, which becomes a larger percent of total costs if the loan is prepaid. Furthermore, good customers are much more likely to repay a loan than bad customers, so a lender runs the risk of seeing its interest income dry up while it maintains a riskier portfolio than originally intended. For any given loan, we can calculate the expected interest received based on a standard loan amortization pattern and compare it to the actual interest to calculate the cost of prepayment. Aggregating this calculation gives us some insights into the prepayment loss of the portfolio. However, due to the lack of time series data, we cannot determine when prepayment occurs.

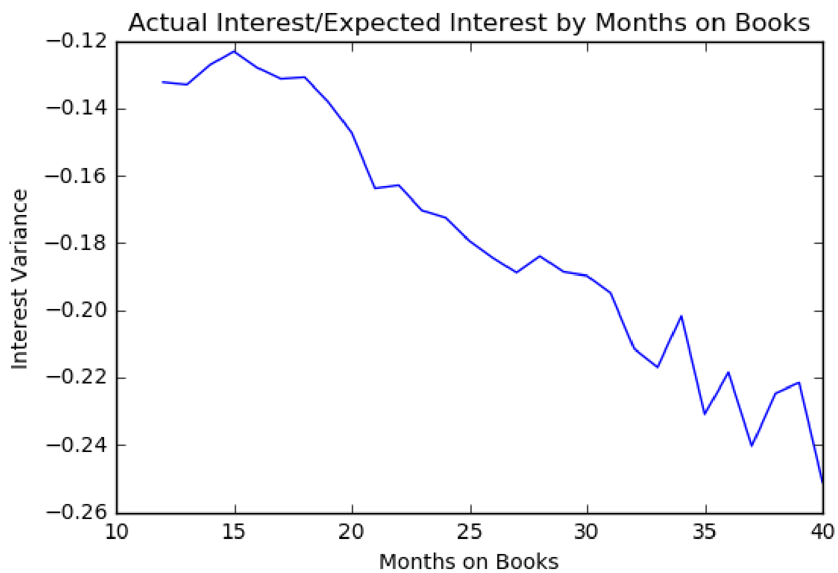

Figure 2: Prepayment Analysis

We compared the expected interest to actual interest and aggregated the results by months on books. What we found was a linear trend of variance in expected and actual interest. Accounts less than 12 months old typically had ~10% less interest income than expected, so when we built our model we assumed a 10% prepayment reduction on expected interest income in the first year. Since the data has no performance history, the accounts less than 1 year old will also be the newer accounts, so it's possible that the prepayment rate is drastically different across origination periods, in which case our model will be using improperly calculated returns. Also, the application of a flat 10% prepayment adjustment is simplistic. Some segments of the portfolio will have high prepayment risk while others will be very low, and predicted returns will be skewed accordingly.

Another obstacle data is the loan performance data is stored in a separate table from the origination data and there is no direct link between the two tables. We were able to reverse engineer a join using common attributes in both tables with a join rate above 50%. The success of the join is contingent on a single loan from the performance data joining with a single listing on the origination data with no duplicates. As a result, the success rate is dependent on the number of applications Prosper receives. The more apps on a particular day, the higher the likelihood of two accounts sharing all the same attributes we use to join. We had better luck with joins on earlier vintages and worse luck in later vintages, as the number of applications has grown exponentially. Still, as the success rate of joins is dependent on the number of apps and the likelihood of two customers applying with the same attributes, we can think of our populations as pseudo-random. We assume the B-rated accounts we joined would be no different than the B-rated accounts we did not, as long as our join population is large enough to contain a representative sample. However, there is the possibility that there is some factor that makes borrowers more likely to be in the non-join population. We could be receiving a disproportionately low amount of the most common types of loans, or we could be excluding applications resulting from a targeted ad campaign run by Prosper. If this is the case, and our join and non-join populations are significantly different, than we run the risk that the model will not scale to the entire population of loans in implementation.