Predicting Loan Defaults in Peer-to-Peer Lending Markets

Nikesh Patel
Fall 2013
General Assembly - Data Science

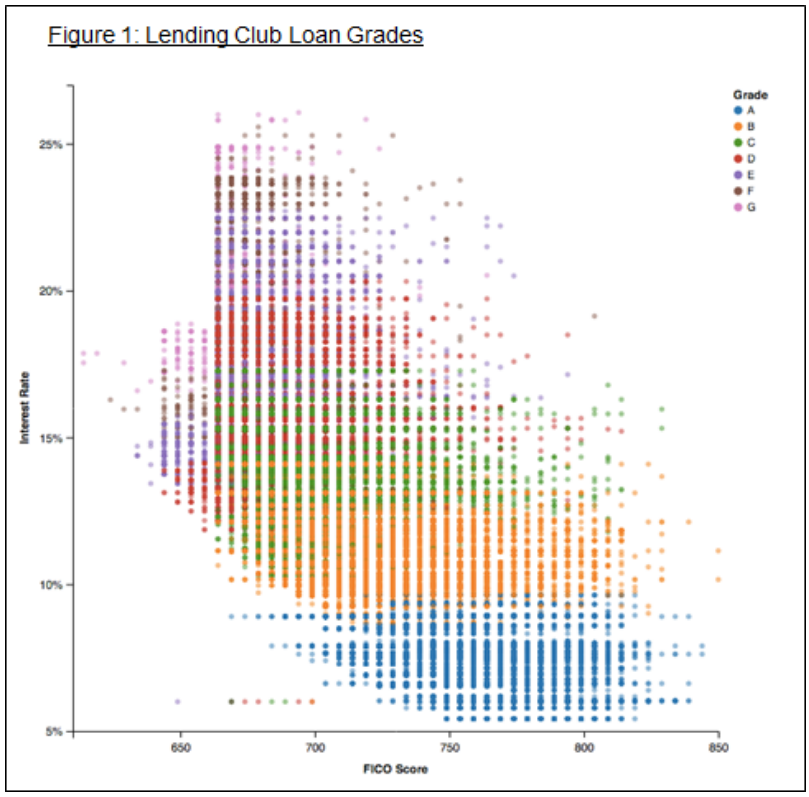
Problem Background and Hypothesis

Institutional consumer lending in the United States is a $3 trillion industry, composed of a fabric of credit providers and loan vehicles.[1] Revolving credit card debt is a significant portion of this market, with the average American card holder maintaining a balance in excess of $5200 month-over-month.[2] Although demand for consumer credit remains large, fallout from the 2007-2008 financial crisis has dampened institutional lending activity. Consumers seeking access to credit for debt refinancing or major purchases face higher borrowing standards as lenders look to avoid investments with broad exposure to the housing and job markets.[3] In an effort to capitalize on this misalignment in consumer credit demand and supply, a new breed of lenders have risen to fill the role of institutional financiers.

Peer-to-peer lenders are financial intermediaries that match private investors with individual borrowers. Unlike traditional institutional lenders that extend loans and hold them to maturity, peer-to-peer lenders "crowd fund" loans by allowing investors to buy shares of loans. The notable difference between these two models is their revenue streams: while institutional financiers profit from interest payments on loans, peer-to-peer lenders collect transaction fees on loan originations and share purchases.[4] Because of this difference, peer-to-peer lenders are not truly "lenders" at all - the private investors that purchase loan shares supply credit to the market. Peer-to-peer lenders provide two services to investors to facilitate this relationship: loan management (e.g. payment collections and disbursements) and loan screening/risk pricing. Since private investors hold the risk in P2P lending, the latter service is crucial to maintain a well-functioning market.

In order to arm investors with the information needed to purchase loans, peer-to-peer lenders first filter borrowers for those that pass a set of eligibility criteria (typically a minimum credit score and completed loan application). The default risk of borrowers is then assessed through an risk pricing algorithm that consider various metrics in addition to credit scores, such as loan size and specific credit history metrics. Based on the output of this algorithm, loans are applied a risk grade and interest rate before being made available for funding by investors.[5] Most peer-to-peer lenders also make a subset of the data they collect on borrowers publicly available to allow investors to purchase loan shares based on individual preferences and strategies.

For borrowers, peer-to-peer lending offers a valuable source of credit during a period of coolness from institutional financiers; however, this relationship is not so clear cut for private investors. Although peer-to-peer lenders are indirectly encouraged to screen out borrowers with high default risk to prevent investors from fleeing to safer competitors or alternative investment vehicles, they are fundamentally incentivized to originate more loans (rather than safe loans) due to their reliance on transaction fees. When this conflict of interest is weighed in combination with the opaqueness of the risk pricing models employed by these intermediaries, private investors can rationally question the efficacy of risk avoidance measures from peer-to-peer lenders.



The problem of verifying risk pricing algorithms employed by peer-to-peer lenders is timely and will grow in importance as new investors investigate these platforms. Using a data set of historical loans from the largest online peer-to-peer lending platform,[6] I will investigate the utility of various machine learning methods to produce a better classification model for predicting loan defaults than intermediary-supplied risk grades. I am operating under the hypothesis that in an effort to make more loans available to investors (to produce more transaction fees), peer-to-peer lenders are allowing certain high risk loans to pass through their initial screening criteria. By producing an alternative risk assessment model from the building block credit data provided, I will attempt to identify these loans so they can be avoided by investors.

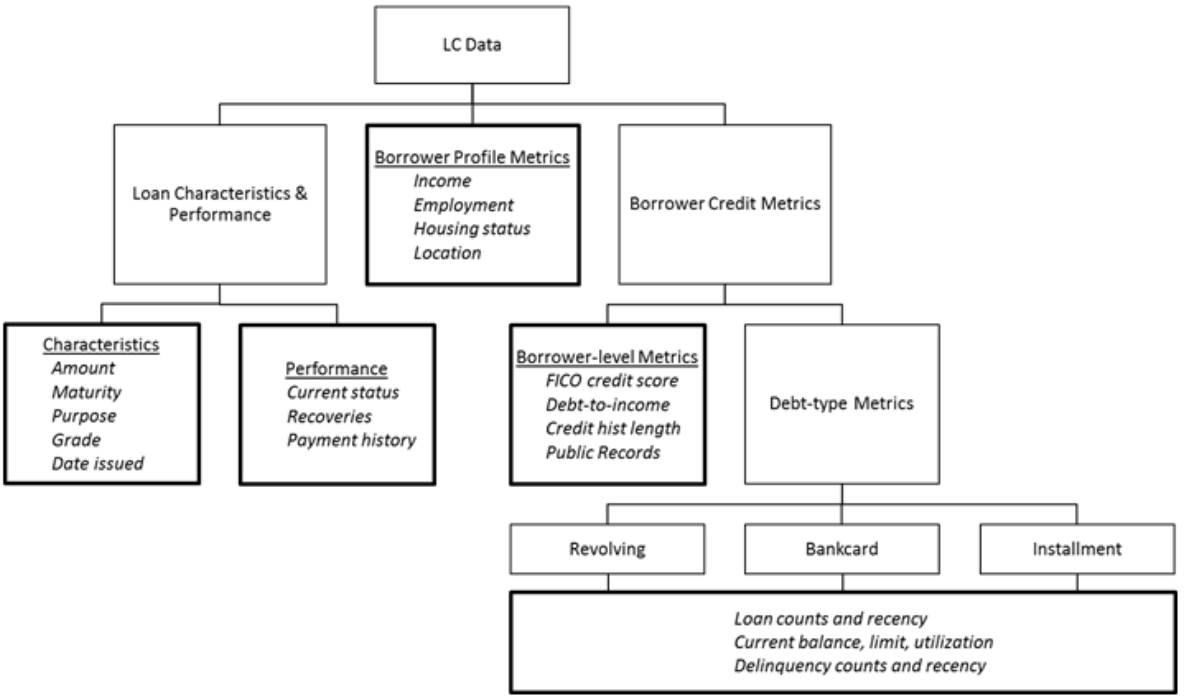
Data Overview

Lending Club (LC) is the largest peer-to-peer lending platform by loans issued and revenue. According to the company, the average LC borrower has a FICO credit score above 700 and self-reported annual income over $70,000, with mean loan values around $13,500.[7] The data set used in this research is the population of all current and historical LC loans, dating back to the company's first loan offerings in 2007. 207,757 loans are represented in total;[8] the data includes all loans that are fully paid (e.g. principal + interest return to investors), current, late (tiered by length of delinquency), in default, and charged off (debt written off as a loss). The data was downloaded from LC's Statistics website, where it is available in two compressed CSV files (*LoanStats3a* and *LoanStats3b*).[9] A dictionary of the data features is also provided on this site in Excel format. LC publishes other data files that were not used for this research, including a data set of rejected loans and loans that are currently available for purchase.

The features provided in the data can be grouped into three categories: loan characteristics and performance metrics, borrower profile metrics, and borrower credit metrics. The loan amount, duration, purpose, LC assigned risk grade and applied interest rate are features included in the loan characteristics category. This group also includes the date the loan was issued, payment history, and loan status (e.g. current, late, default, etc). Features within the borrower profile group include employment details and income of the borrower, current housing status, and location. The last category - credit metrics - contains the majority of the data set's features. Credit metric features can be subdivided into two types: borrower-level, and debt type-level. Borrower-level credit features include the FICO credit score, the debt-to-income ratio (the percent of pre-tax monthly income used for debt payments), length of credit history, and counts and recency of public records such as

bankruptcies and tax liens. Debt type-level features are sets of credit metrics broken out by the type of debt: revolving (principally credit cards), bankcard (principally debit cards), and installment (such as car loans or student loans). Features available for each debt type include number and recency of accounts, total balance and credit limit, utilization (e.g. balance/credit limit ratio), and counts and recency of delinquencies.

Figure 2: Feature Type Hierarchy



Data Cleaning and Feature Selection

The initial processing step in preparing the data for analysis was identifying the class feature. The *loan status* feature contains the current standing of each loan - since this analysis is concerned with predicted loans likely to default, loans that are currently making payments or are late were excluded, leaving only loans with statuses of fully paid, in default or charged off remaining in the data set. The difference between default and charged off is nuanced - after a loan is over 120 days late, LC classifies a loan as in default; however, it in unclear how long loans hold this status before being charged off as a loss. According to LC, only 14% of loans in default have a partial recovery of principal after 9 months;[10] for this analysis, all loans in default were classified as "bad loans" in addition to charged off loans. Correspondingly, fully paid loans were assigned a "good loan" class - together, these classifications compose the boolean *bad loan* feature at the center of this research analysis.

The restriction of the LC data set to fully paid, charged off and defaulted loans posed an immediate challenge on sample size. Lending Club issues loans with 36-month and 60-month maturities; accordingly, only loans extended before November 2010 or November 2008 had the opportunity to mature by the date the data was accessed, depending on their duration. Given that Lending Club began operations in 2007, the impact to the data set was substantial - of the initial 207,757 loan observations, over 80% were excluded due to not reaching maturity, prepayment, or a default. The size of the final data subset used for this analysis was 39,649 loan observations; in absolute terms, this is still an adequate sample size for this analysis. Unfortunately, this initial restriction also repercussions for the feature selection process.

Over the course of Lending Club's lifespan, the credit underwriting policy used to screen loans for funding has evolved. Newer iterations of the policy appear to require more granular credit history data from prospective borrowers, resulting in additional features being made available for analysis. Many of the debt-type credit metrics, especially for bankcard and installment debt, are available due to these policy changes. A negative consequence of these iterations is inconsistency

among the features available for loans originating under different credit policies, as Lending Club has not backfilled data as new features are defined. This consequence was particularly problematic as the majority of the loan observations in the restricted subset originated under older credit policies, thus missing newer features. Less than 40% of the loans contained data for any feature added under a post-2010 credit policy, and far fewer contained more than one new feature. Although I investigated using regression techniques to fill in some of the missing features, I ultimately dismissed this solution due to high cost (in tuning time and number of features) and unclear benefits (ultimate predictive significance of the imputed features). I elected to prefer sample size over data granularity, and reduced my feature space to only features available under pre-2010 credit policies. Although the penalty of this decision was substantial, reducing the feature space from 67 initial features to 25, over 96% of the original subset sample size was preserved.

Another challenge that became apparent during analysis of the LC data set was multicollinearity. Prior to restricting the feature space to only features available under pre-2010 credit policies, I calculated correlation coefficients for all available features using *Pandas*' Pearson correlation method. 18 feature pairs had correlation coefficients over 0.8, indicated a strong linear relationship between the variables, and over 30 had coefficients above 0.7. In situations where multicollinearity is present in a data set with a large feature space, principal component analysis (PCA) can be leveraged to reduce the feature space to its uncorrelated components. While PCA has useful applications for data sets with similar characteristics to the LC data, I elected not to conduct PCA after investigating some of the most highly correlated features using Lending Club's data dictionary. After researching credit reporting metrics and their underlying calculations, clear relationships could be deduced from nearly all of the highly correlated features. For example, the number of public records on file for a borrower was highly correlated with the number of times they declared bankruptcy. This relationship makes sense, because bankruptcies are included in public records; I used this information to produce a non-bankruptcy public record feature, thus eliminating the correlation issue. A less obvious example was the correlation of a borrower's revolving balance and their combined credit card limit. The correlation of these variables usually manifested for borrowers looking to refinance credit card debt to lower rates (this information is available through the loan purpose feature). If a borrower has a high credit card balance, they are likely to be "maxed out" at the credit limit allowed by their card providers, thus explaining the high correlation between the balance and credit limit features. For this example I elected to use the revolving credit utilization ratio to capture the relationship of these features.

The LC data also included a series of text features which I vectorized in order to incorporate into this analysis. These features were the stated purpose of the loan, borrower's housing status, and borrower's state residency. Values in the loan purpose feature fell into two primary categories - debt *refinancing* and discretionary spending *financing*. Purposes within the first group include debt consolidation and credit card refinancing, while purposes in the latter may include small business financing or wedding expenses. I initially considered vectorizing the purpose feature into buckets representing loans that functionally refinanced existing debt versus loans for initial financing of some purchase or activity. This representation would be ideal, as borrowers making rational refinancing decisions improve their financial stability by reducing monthly payments, while loans used for initial financing are incremental to existing debt obligations. I ultimately abandoned this strategy, because this classification decision could not be easily made - a home improvement loan may be refinancing a high installment loan instead of initially financing a discretionary spending project. Future analysis of the LC data could benefit from attempting to model this financing distinction. Another text field worth investigating in the future is the description features. The description feature is an optional text field that borrowers may fill in to describe why they are seeking a loan or make a case for why they are a responsible borrower. I did not include this feature in this analysis, though I investigated whether borrowers who included a description were more or less likely to default and found no relationship. A word frequency analysis of this field is worth investigating in future research of the LC data.

Classification Methods and Results

A series of different classification techniques were trialled to predicted loan defaults/charge offs. The same workflow was repeated for each technique: fit the model with all available features, tune parameters to prevent overfitting, and drop insignificant features. Prior to fine tuning any single technique, I chose to fit a "kitchen sink" model with each method before deep diving on individual classification methods. Five classification techniques were trialled in total - logistic regression, LinearSVC, naive Bayes, random forests, and AdaBoost with decision trees. All models were fitted with a 75/25 "train-test" sample split and employed a 15-fold cross validation technique for scoring. Models were scored with the cross validation AUC metric, with special attention paid to how false positive levels increased with improvements in true positive classification.

Logistic regression and LinearSVC were the first two techniques used to classify loans in the LC data set. On first pass, both models struggled to classify a single loan as likely to default. Defaults and charge offs represent approximately 20% of loans in the data, scaling from 8.7% for grade A loans to 41.5% for grade G. Bucketing continuous variables into boolean features

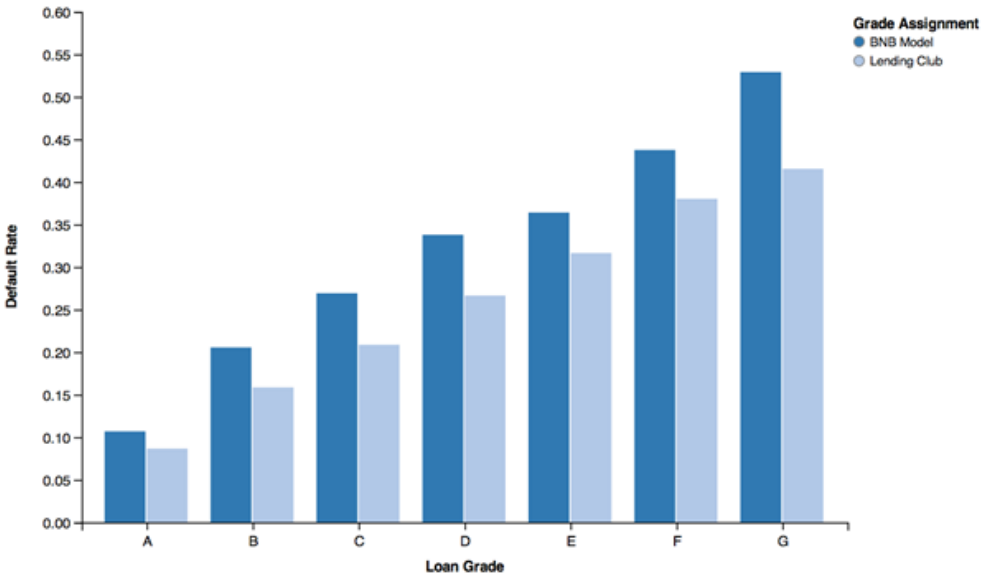
improved both models, and adjusting the penalty parameter increased the accuracy of the LinearSVC model. Both techniques achieve similar maximum AUC scores of between 0.53 and 0.55; however, overwhelming false negative rates for detecting bad loans undermined the utility of both models. Initial findings from analysis of coefficients from the logit model informed feature selection. Loan purpose features were important signals in the model. The small business purpose feature had a strong positive coefficient, while the credit card feature had a significant negative coefficient. This relationship partially validates an earlier assumption that loans used for initial financing (e.g. small business loans) were riskier than loans for refinancing debt (e.g. debt consolidation loans), because the latter reduces net monthly payments while the former is incremental to current debt. Loan maturity was also a significant driver of default risk, with longer maturities equating to greater default risk. The top and bottom quintiles for annual income, revolving debt utilization and FICO score were also significant predictors of default risk. Surprisingly, home status was less significant than other features for predicting default. Although I hypothesized that home owners would be riskier than renters due to their inability to relocate for lower rents as well as the implication of their choice of a Lending Club loan over a home equity loan, renters proved marginally more risky. In absolute terms, both features did not contribute heavily to model predictions.

Performance from decision tree-based models paralleled the logit and LinearSVC techniques after tuning and controlling for overfitting; in absolute terms, all techniques struggled to predict loans defaults. Two ensemble techniques with decision trees were trialled - a random forest model, and the AdaBoost meta-classifier. To control for overfitting, the random forest classifier restricted tree depth to a range between 10 and 20 nodes. Insignificant variables were pruned after analyzing p-values following initial model fittings. During these trials, I decided to remove state-level location features from all models due to few loan observations in smaller states and concerns about the historical bias. While some states with adequate observation counts appeared to have strong predictive relevance (for example, loan origination in Nevada increased default likelihood), I was uncomfortable generalizing this measure to all future loans due to potential correlation with the 2008/2009 housing crisis and recession. Among the ensemble methods, the AdaBoost meta-classifier model had the highest AUC score (0.54) despite minimal measures to avoid overfitting, although both models performed poorly in absolute terms. Like the logit and LinearSVC model, the ensemble methods functioned poorly at identifying bad loans, with overall recall scores failing to breach 10% in both in- and out-of-sample testing.

The most performant model for predicting loan defaults in the LC data I tested was a naive Bayes classifier. Although many of the features in the LC data set are continuous or frequency variables, which are modeled well under a Multinomial naive Bayes classifier, many binary variables are also present in the feature space. Accordingly, I elected to bin continuous features into boolean fields and use a Bernoulli naive Bayes (BNB) model. Although high false negative frequency remained a weakness in the classifier, the BNB model had substantial gains in true positive prediction over all previous techniques, achieving an out-of-sample recall score above 25%. AUC for the BNB model was 0.58 - comparatively better than previous techniques, but overall an indicator of the poor predictive accuracy of the model. Although this result is disappointing, it is not truly surprising - if Lending Club could produce a model that could accurately predict if loans were likely to default, it would not approve those loans for funding to being with. The LC data set is composed of loans that have already passed Lending Club's screening process; here, loans are assigned a risk grade - a proxy for their likelihood of default - *precisely because* Lending Club cannot predict if they will default. Through this lens, a more fair comparison of the efficacy of the BNB model is to compare the default rate of grades assigned by Lending Club with grades assigned by the BNB model.

Lending Club has 7 loan grade - A through G, with A being assigned to the 'safest' loans based on LC's risk pricing techniques. To compare the estimation power of the BNB model for determining default risk with the loan grades assigned by Lending Club, I assumed that Lending Club grades were uniformly distributed across the range of default risk probabilities (0%-100% likely to default). Based on the uniform distribution assumption, each grade represented one-seventh of the range, with each grade occupying consecutive "sevenths" (e.g. A is 0-14% default probability, B is 14%-28%…) as predicted by the BNB model. After predicting probabilities for all loan observations in the sample, I used this distribution assumption to assign grades to each loan. To compare the performance of the BNB model to Lending Club's risk pricing technique, I calculated the default rate among the pool of loans assigned to each grade, and plotted the results for all grades.

Figure 3: BNB Model Grading Performance



For 'safe' loan grades (e.g. A-C), the BNB model was worse than the internal Lending Club model for predicting defaults, demonstrated by the higher default rate measured compared to the default rate in the pool of 'safe' loans as classified by Lending Club. Although a higher default rate in safer grades is undesirable, it is an expectation in riskier grades (F and G). For these riskier grades, the BNB model "outperformed" the Lending Club-assigned grade by capturing more defaults within the BNB model-assigned grade. This comparison is not truly "apples to apples", since the distribution of grades along the range of risk probabilities as modeled by Lending Club may not be uniform. Lending Club may also purposefully misclassify loans by spreading risky loans throughout the grading distribution to maintain a market for riskier credit (which likely has a larger applicant market due to increasing credit standards from traditional lenders). Even with these caveats noted, the BNB model appears to perform well for its original purpose - predicting loans that are likely to default. This optimization is evident in the comparison of the efficacy of the BNB model's grading system to Lending Club's - while the BNB model performs poorly for low risk grades, its outperforms for the high risk grades that it was designed to predict.

Business Application and Continuing Research

Although the predictive accuracy of the modeling techniques used in this analysis to predict defaults in aggregate was poor, the research process yielded valuable insight into variables that drive credit risk. While historical information about a borrower's ability to pay off debt is a signal of their propensity to make timely future debt payments, what ultimately drives loan default are changes in an borrower's financial stability. The predictive significance of the loan purpose features in the LC data set support this reasoning. Although two borrowers may have similar credit histories, the knowledge that one borrower is using a loan to fund a small business explicitly puts that borrower at greater risk for default. A potential strategy to gain insight into the borrower characteristics beyond credit history that support credit risk is to create financial stability models for borrowers. Under this strategy, a borrower's credit history is analyzed with their pre-loan financial situation to calculate a benchmark of their absolute ability to meet debt obligations. This calculation is then completed for the borrower's post-loan financial situation, potentially though a Monte Carlo simulation in order to incorporate forecasted variables. Default risk is then represented by a relationship between the absolute level of the post-loan calculation and the delta between the pre- and post-loan calculations (in order to include the positive or negative "stability shock" of the loan to the borrower's financial stability).

Another opportunity to explore for future work with this data set would be reincorporating the borrower location features. While fitting at the state-level leaves a model open to bias and overfitting, location data can help incorporate economic health features like county unemployment levels and city-level housing market indicators into the feature space. Many public and private entities publish economic data by location, notably the Bureau of Labor Statistics and real estate companies like

Zillow. Location features can also be employed to "de-anonymize" the data set to discover new information about borrowers. Location and employment features are valuable hook variables for de-anonymization, and could be fused with a social network to identify a borrower's education level and consumer interests. In addition to expanding the feature space, increasing sample size may also improve model performance. One route to enable sample expansion is to extend the data set to other peer-to-peer lending platforms like Prosper, a Lending Club competitor. The sample can also be expanded simply by waiting - as Lending Club extends more loans, and loans that originated after 2010 mature, the sample will grow. The usable feature space will also grow with time, since more loans originating under post-2010 credit policies will mature.

From an application perspective, the clearest benefit of this research is to optimizing Lending Club loan portfolios. Optimization can be achieved through two strategies - avoiding high risk loans outright, and identifying local loan "mispricing" (e.g. safe borrower's paying high interest rates). Another potential use case for extending this research is to investigate risk pricing and returns on the second market for peer-to-peer loans. Loans with late outstanding payments trade at steep discounts to par on secondary markets, despite that fact that some of these loans recover and reach maturity. Predicting which loans among all late loans are most likely to recover could be a lucrative next step for research on the LC data set, although transaction fees would likely need to be incorporated into any potential model.

---

1. Federal Reserve Consumer Credit Data: http://www.federalreserve.gov/releases/g19/Current/ ↩
2. TransUnion 2013 Q3 Credit Report: http://newsroom.transunion.com/press-releases/transunion-q3-report-demonstrates-consumers-manag-1070204 ↩
3. WSJ - Credit Crunch Moves Beyond Mortgages: http://online.wsj.com/news/articles/SB118773982869404682 ↩
4. This is a simplification for illustrative purposes; institutional lenders can package and resell loans as collateralized debt obligations (CDOs), thus acting functionally like peer-to-peer lenders. The chief difference here is that CDOs are synthetic loans composed of many different debt vehicles, whereas peer-to-peer loans are backed by individual borrowers. ↩
5. Lending Club's loan assessment process: https://www.lendingclub.com/public/how-we-set-interest-rates.action ↩
6. Bloomberg - LendingClub Said to Reach $2.3 Billion Valuation in DST Funding: http://www.businessweek.com/news/2013-11-13/lendingclub-said-to-reach-2-dot-3-billion-valuation-in-dst-funding ↩
7. Lending Club Investing Portal: https://www.lendingclub.com/public/steady-returns.action ↩
8. Lending Club updates its data files with new loans daily; therefore, more loan data may have become available since the conclusion of this research. The data set used here was accessed 11/4/2013. Also of note is that the number of loans in the data set exceeds the number quoted on LC's website. This is likely due to the exclusion of loans issued under LC's previous credit screening policy. LC reported figures are available here: https://www.lendingclub.com/info/demand-and-credit-profile.action ↩
9. Lending Club data download site: https://www.lendingclub.com/info/download-data.action ↩
10. Loan Status Migration for loan in default: https://www.lendingclub.com/info/statistics-performance.action ↩