Journal of Research High School

Volume 2022 (1) 94 - 103
ISSN 2688-3651 [online]

Screening P2P Loans Beyond Loan Grade: An Application of Machine Learning Algorithm XGBoost

Selina Sun[1] *

[1]Thomas Jefferson High School of Science and Technology, Alexandria, VA USA

Received October 8, 2021; Revised April 11, 2022; Accepted, April 19, 2022

Abstract

The credit risks of P2P loans fall onto the investors rather than the platforms, which provides a motivation for investors to screen loans beyond the suggestive loan grade provided by the platforms. This is made possible by P2P platforms sharing with the public the same information they use in screening. This paper studies whether screening models can be developed using machine learning algorithms to better screen P2P loans than loan grade. Based on the loans listed by one of the largest P2P platforms in the world, Lending Club, we find that machine learning algorithm XGBoost can improve default rank ordering by 10.3% in sample and 4.4% out of sample, as compared with loan grade. XGBoost also outperforms the conventional parametric Logistic regression model with a performance gain of 4.4% in sample and 1.1% out of sample.

Keywords: Marketplace Lending, Machine Learning, Logistic Regression

1. Introduction

Peer-to-Peer (P2P) lending is a practice of lending money to businesses and consumers facilitated by online lending platforms that connect borrowers and investors. Since its inception with Prosper in 2006 in the US, P2P lending has originated more than $18 billion loans by 2015 and is expected to grow further to $150 billion by 2025 (Prime Meridian, 2015).

P2P lending is also called crowdfunding, social finance, marketplace lending, or disintermediation finance, each emphasizing different perspectives of P2P lending. One characteristic that makes P2P lending different from conventional bank lending is the disintermediation, in that P2P platforms do not serve as the intermediators by holding loan originations on their balance sheets, and the credit risks of P2P loans fall directly onto the investors rather than the platforms. Therefore, the objectives of online platforms may not be necessarily aligned with investors (Vallee and Zeng, 2019). On the other side, online platforms share with the public the same information that they use to screen loans. In the US, the P2P market regulator, the U.S. Securities and Exchange Commission (SEC), demands that online platforms disclose all decisioning information to the public. Therefore, sophisticated and active investors might be able to better predict credit risks than inexperienced and passive investors who rely on the screening tools in the form of loan grade assigned by online platforms.

The objective of this paper is to study whether screening models can be developed to better capture the default risks of P2P loans than the suggestive loan grades provided by the platforms. It hypothesizes that customized screening models can predict default risks more accurately than loan grades. We take the platform grade as the baseline and evaluate the

* Corresponding Author
selinateam14@gmail.com

Advisor: Yan Zhang
yan.zhang@occ.treas.gov

benefits of utilizing the more granular raw data collected at loan application, including the borrower-provided soft information and the standard borrower credit bureau profile. Our study is based on Lending Club (LC) data. Being the second earliest P2P platform in the US and the first ever to register its offerings as securities, Lending Club is one of the largest P2P lending platforms in the world. It is the largest P2P lending platform in the US and took approximately 35%-50% of the US market share in 2015-2020 (IBISWorld, 2020). As of 2020, LC has reached a total loan volume of $50 billion since its launch in 2006 ("Top 7 P2P lending sites to lend money online from US", https://crowdfunding-platforms.com/top-7-p2p-lending-sites-to-lend-money-online-from-us). Therefore, the results based on LC could be of broader interpretation.

In terms of modeling algorithms, we study both the traditional parametric statistical models and the innovative machine learning (ML) models. Since we measure loan risks by a binary classification of loan default, the statistical model used is the Logistic regression (LR) model and the ML model is eXtreme Gradient Boosting (XGBoost, or XGB). The latter is a ML algorithm that is gaining wide popularity in credit risk modeling (Chang et al., 2018) but has not been studied as a P2P screening algorithm in an empirical study of LC loans, based on our knowledge.

We find that our models perform better in predicting default risks than platform grades. Given the LC loans used in the paper, the XGB model achieves a 10.3% increase in rank ordering defaults in sample and a 4.4% increase out of sample. The XGB model is also able to classify defaults more accurately than the Logistic regression model with rank ordering gain of 4.4% in sample and 1.1% out of sample. While loan grade is found to be reasonably predictive, other loan and borrower variables can be utilized to screen loans further, including payment-to-income (PTI) ratio, debt-to-income (DTI) ratio, Fair Isaac Corporation (FICO) score, and number of accounts opened in the past 24 months.

The paper contributes to the strand of emerging literature studying the screening ability of lenders in marketplace lending. One sub-group of this literature identifies information that can be used to approximate platform grade or even further to screen loans on top of platform grade. Using Prosper data, Iyer et al. (2015) show that investors can gain by leveraging the number of current delinquencies, debt-to-income ratio, amount delinquent, and number of credit inquiries in the last six months in addition to the FICO-based grade. Based on Lending Club data, Emekter et al. (2015) identify DTI, FICO, and revolving line utilization as contributing to loan screening alongside loan grade. Serrano-Cinca et al. (2015) suggest that besides loan grade, loan purpose, annual income, homeownership, credit history, and borrower's indebtedness are relevant to risks of LC loans. Another sub-group of the literature tests the use of financial technology in the form of ML algorithms in screening. Malekipirbazari and Aksakalli (2015) compare Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Logistic regression in predicting loan defaults and find RF performs the best. Chang et al. (2015) conclude that Naïve Bayes performs better than SVM and Logistic regression in LC loan default prediction. This paper evaluates the ML algorithm of XGB, which has not been documented in existing literature for screening P2P loans. Our analysis is based on a LC dataset with a much more recent snapshot and therefore includes more loans. We assess model screening ability not only in sample but also out of sample.

The rest of the paper is organized as follows. Section 2 discusses the LC data and defines the analysis samples. Section 3 presents model specifications and performance measures. Section 4 reports the regression results. Section 5 concludes.

2. Data

The study uses a Lending Club dataset that was downloaded on July 3, 2021 from the Kaggle website (https://www.kaggle.com/wordsforthewise/lending-club). It contains all the loans accepted by Lending Club during the period from June 2007 (the beginning of its loan issuing) to December 2018. In aggregate, there are 2.26 million loans with a total loan amount of $34 billion. Each loan is reported as one data record with a rich set of data fields (151 variables in total) including information collected at the application (loan application submission and borrower credit pull), LC assigned risk grade and pricing, and loan performance read by the end of 2019Q1 (loan status, hardship program, settlement program, etc.).

LC loans are in the form of unsecured personal loans with fixed term and installment (i.e., monthly payment). The loan amount is in the range of $500 to $40,000, and the loan term includes 3-year and 5-year. The 5-year term loans were not rolled out until May 2010. Out of the two product terms offered by LC in the data, roughly 71% are of 3-year term. To obtain a more homogenous sample, we focus on the 3-year loans. Most (95%) of the loan applications are filed by individuals, so we exclude joint accounts to simplify the analysis without too much loss of generosity. As this paper evaluates loan credit risks by following up loan performance within two years of the origination date, we also drop loans that are originated after February 2017. They are right censored by the data read snapshot of 2019Q1 and therefore do not have the full 2-year performance window imposed. The LC loan volume has been growing at an accelerating pace and the originations from March 2017 through December 2019 take up 36% of the total loan volume. Imposing a shorter 1-year performance window allows us to include more loans into the analysis, but the performance window is too short to have a good read of loan performance. On the other side, if we go with the full 3-year performance window, even more loans will be excluded from our analysis, and the gain is marginal – not many loan defaults occur in the third year.

After the exclusion of 5-year termed loans, joint loans, and loans originated after February 2017, we have an analysis data sample of 982,245 loans. The application fields are well populated, however, some of the credit bureau fields have missing values. We impute the missing values with the median of the variable, so as not to distort the variable distribution. For variables with high missing rates, we also construct corresponding missing variable indicators to account for possible implications of the missing patterns to loan risks.

When borrowers submit their application to the LC online platform, they need to disclose loan amount (loan_amnt), loan purpose (purpose), annual income (annual_inc), home ownership (home_ownership), and employment length (emp_length). In real time, the platform processes the applications and makes underwriting and pricing decisions using its screening models. The borrower credit bureau information is pulled and leveraged in the credit decisioning. The LC dataset discloses borrowers' FICO score in the data in the form of a range with upper bound and lower bound (fico_range_high and fico_range_low). The DTI ratio (dti) is calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. Based on the date when the borrower's earliest reported credit line was opened (earliest_cr_line_date), we can infer how long the borrower's credit history is (credit_history). LC also makes public the detailed credit profile of borrowers in the form of credit bureau variables, such as the number of trades opened in past 24 months (acc_open_past_24mths), the number of inquiries in the past 6 months (inq_last_6mths), and the percent of bankcards with utilization greater than 75% (percent_bc_gt_75). For approved loans, the installment is calculated and reported given the loan amount, term, and interest rate. We create a new variable, PTI ratio (pti), as the installment amount over the borrower's self-reported monthly income.

The loan status variable reports loan performance in the following nine status: 1) "Charged Off", 2) "Current", 3) "Default", 4) "Does not meet the credit policy. Status: Charged Off", 5) "Does not meet the credit policy. Status: Fully Paid", 6) "Fully Paid", 7) "In Grace Period", 8) "Late (16-30 days)", and 9) "Late (31-120 days)". The hardship program field flags borrowers who encounter hardship and receive assistance, and the settlement field identifies troubled borrowers who settle debts with lenders. This paper measures loan credit risks by a user defined technical "default" variable that takes the value of 1 if a loan is more than 30 days late or worse within 2 years since loan origination, and 0 otherwise. Specifically, a loan technically defaults if it reaches the status of 1) "Charged off", 3) "Default", 4) "Does not meet the credit policy. Status: Charged Off", 9) "Late (31-120 days)", or receives hardship assistance, and debt settlement.

Table 1 reports the definition and summary statistics of the data fields used in our analysis. Panel A lists the loan characteristic variables. On average, the default rate and interest rate are very close at 11.9% and 12.0%, respectively. The mean loan amount funded is $12.6K. The applied loan amount and funded loan amount are very similar, as shown by the small loan amount difference (amnt_diff), suggesting that the requested loans are pretty much funded fully.

Table 1 Panel B provides information on the application variables. The annual income is reported to be $74.3K on average. The data report detailed employment history in 12 categories of not reported, < 1 year, 1 year through 9 years with the increment of 1 year, and 10+ years. We translate the categorical employment length into a continuous employment length by coding "not reported" as 0, "< 1 year" as 0.5, "10+ years" as 10, and other employment length categories as they are. The average employment length is 5.5 years. Loan purpose has the following categories: Car, Credit card, Debt consolidation, Educational, Home improvement, House, Major purchase, Medical, Moving, Renewable energy, Small business, Vacation, Wedding, and Other, among which the top three categories are Debt consolidation, Credit card, and Home improvement, taking up 53.8%, 23.9%, and 6.6% of the total loans, respectively. Home ownership shows that 45.0% of the borrowers own homes with mortgage, 11.7% own homes paid off, and 43.2% rent. Types of income verification status include not verified, source verified, and verified, each of which takes up about a third of the total. LC verifies income if it determines that the borrower's requested loan amount is too high relative to the self-reported income. The income verification involves checking W-2 or paystub (source verified) or reaching out to the employers (verified). A loan can be listed as factional or as whole, the latter is to facilitate institutional investment. Factional loans consist 45.6% of the total and the remaining 54.4% are whole loans. The whole listing did not start until October 2012 and has since increased continuously to the level of close to 90% by the end of 2018.

As for borrower creditworthiness as reported in Panel C of Table 1, the mean DTI ratio is 17.8. LC imposes a maximum DTI restriction for loan approval. Most of the time the DTI cap is set at 39.99%, but in the first half of 2016, it rose to a concerning level of 50%. The lowest FICO score on average is 694. Similar to DTI, LC requires borrower FICO to be above a certain level to list a loan. With a few fluctuations in the earlier years, the FICO threshold has been 660 since November 2008. The credit history, which measures the number of years since the borrower initially establishes a credit profile, is 16.2 years on average. LC requires at least 3 years of credit history to approve a loan. Out of the many bureau variables, this paper focuses on the top 15 that are shown to be highly related to loan risks. They are selected by an XGB model with number of iterations = 500, maximum tree depth = 3, learning rate = 0.05, minimum loss split = 0.05, minimum child weight = 5, lambda = 1, alpha = 0, subsample ratio of columns = 0.8, and subsample = 1. The section "Research Methodology" provides a detailed explanation of XGB and its hyperparameters. Other variable selection methods like stepwise regression lead to very similar variable selections.

Table 1: Variable definition and summary statistics

Panel A: Loan characteristics

Variable	Definition	Mean	SD
default	loans become 30+ DPD or worse within 2 years of loan issuance	11.9	
int_rate	loan interest rate	12.0	4.0
funded_amnt	loan amount funded	12,551	8,079
amnt_ diff	difference between the listed loan amount and funded loan amount	4	175

Panel B: Application information

Variable	Definition	Mean	SD
annual_inc	self-reported annual income provided by the borrower during registration	74,348	69,908
pti	ratio of the monthly payment of the loan (installment) to borrower self-reported monthly income	7.7	4.3
emp_ length	employment length in years	5.5	3.8

purpose: debt consolidation	loan purpose is to consolidate debts	53.8%	
purpose: credit card	loan purpose is to pay off credit card balances	23.9%	
purpose: home improvement	loan purpose is for home improvement	6.6%	
home: mortgage	home is owned with a mortgage	45.0%	
home: rent	home is rented	43.2%	
home: own	home is owned without a mortgage	11.7%	
verification: not verified	income is not verified	32.5%	
verification: source verified	income source is verified	38.2%	
verification: verified	income is verified	29.3%	
initial listing status: fractional	loan listed as fractional	45.6%	
initial listing status: whole	loan listed as whole	54.4%	

Panel C: Borrower creditworthiness

Variable	Definition	Mean	SD
dti	ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income	17.8	8.4
fico_range_low	lower boundary of the borrowers FICO range at loan origination	695	31
credit_history	credit bureau history in years	16.2	7.7
acc_open_past_24mths	number of trades opened in past 24 months	4.5	3.1
avg_cur_bal	average current balance of all accounts	12,212	15,593
bc_open_to_buy	total open to buy on revolving bankcards	9,397	14,490
bc_util	ratio of total current balance to high credit (or credit limit) for all bankcard accounts	60.7	27.1
delinq_2yrs	number of 30+ days past-due incidences of delinquency in the borrowers' credit file for the past 2 years	0.3	0.9
inq_last_6mths	number of inquiries in past 6 months (excluding auto and mortgage inquiries)	0.7	1.0
mo_sin_old_rev_tl_op	months since oldest revolving account opened	180.3	93.9
mort_acc	number of mortgage accounts	1.5	1.9
mths_since_recent_bc	months since most recent bankcard account opened	23.7	30.6
mths_since_recent_inq	months since most recent inquiry	6.8	5.5
num_tl_op_past_12m	number of accounts opened in past 12 months	2.1	1.8
percent_bc_gt_75	percentage of all bankcard accounts > 75% of limit	45.2	35.2
revol_util	revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit	52.2	24.1
tot_hi_cred_lim	total high credit or credit limit	158,580	172,866
total_bc_limit	total bankcard high credit or credit limit	20,469	20,912

Notes: The statistics are generated based on a sample of Lending Club loans of single applicant 3-year loans originated during June 2007 through February 2017 with performance snapshot of 2019Q1, total 982,245 loans.

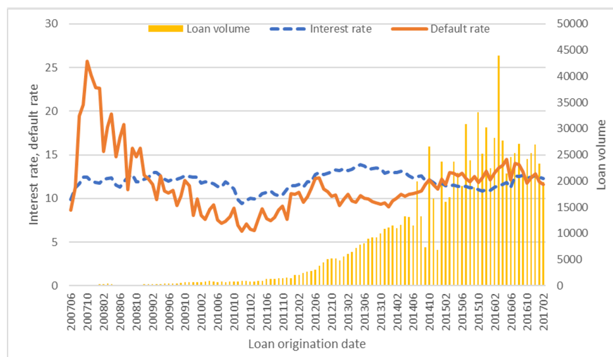

Figure 1. Loan volume, pricing, and performance over time

Figure 1 plots the loan volume, pricing, and performance of the LC loans along the loan issuance date. The loan volume was quite small in the beginning with only 23 loans issued in the LC debut month of June 2007. The loan volume gradually picked up and reached 1,000 loan issuances in June 2011. LC business grew rapidly in the next 4-5 years. Since 2015 the monthly origination volume has been around 25,000. At the peak, 44,000 loans were originated in the single month of March 2016. The interest rate has been relatively stable, fluctuating in the range of 10%-14%. Conversely, the default rate was quite volatile. It was high up to 2009, which could be due to the nascency of the business and the financial crisis of 2007-2009. Afterwards the default rate decreased and stayed below the interest rate line. In the extreme month, the interest rate was as much as 4% higher than the default rate. But starting from 2015 when the loan volume stabilized at the high level, the default rate became higher than the interest rate.

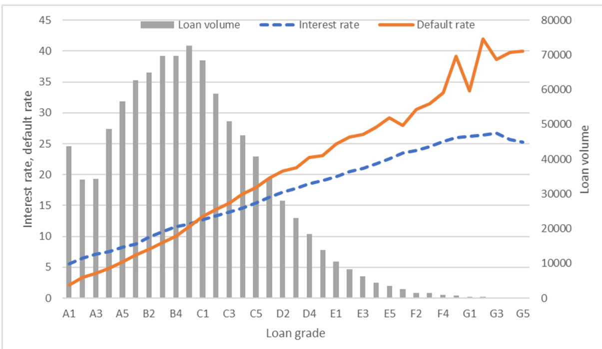

Figure 2. Loan volume, pricing, and performance by loan grade

Figure 2 presents the loan volume, pricing, and performance along the dimension of LC loan grade. LC assigns grades to inform investors of its assessed risks. The grade system contains 6 letter groups from A to G with increasing risks. Within each grade, there are 5 numeric subgrades from 1 to 5 with increasing risks. For simplicity this paper refers to the letter and number combined grade as grade unless otherwise noted. Figure 2 shows that the loan distribution is right skewed with grades A through D taking almost 96% of the loans and grade B being the most populated with 35% of the loans. The interest rate is the same for loans with the same grade at the same issuance time and is adjusted periodically. Figure 2 shows a very clear monotonic increasing trend of the interest rate with grade, confirming the mapping between grade and interest rate. The best grade A1 has an average interest rate of approximately 2%, and the worst grade G5 has an average interest rate as high as 40%. The default rate also monotonically increases with grade, suggesting that LC grade rank orders risks. However, the interest rate line and the default line have different slopes. For loans with good grades, the interest rate is higher than default rate on average, so on aggregate investing highly rated loans returns positive cash flow. But for riskier loans, the interest rate is not high enough to cover the losses. This echoes the finding by Emekter et al. (2015) that the high risks of low-grade loans are not covered by the interest rate charged.

The dataset is then split into training and testing by issue year. In earlier years of LC business, the number of loans were few, the screening models were still at their learning stage, and the default rates were high. We decide to drop this small volume of loans and use loans issued during January 2011- December 2015 for training the screening models, and loans issued afterwards (January 2016 - February 2017) for testing. Our testing sample is not only out of sample but also out of time, enabling a more rigorous and meaningful testing. The training and testing datasets have 603,497 and 361,319 observations, respectively. To facilitate the tuning speed, we randomly sample 30% of the training sample to construct the hyperparameter tuning sample of the XGB models. The tuning dataset contains 181,049 loans, which is sufficiently large for the tuning purpose.

We develop screening models using two types of algorithms: the LR model and the XGB model. In predicting credit risk of retail loans with binary outcomes, Logistic regression is a commonly used conventional statistical model, while machine learning algorithms like XGB are gaining popularity. Variable contribution to screening is shown by variable importance. The model performance is evaluated by model rank ordering of loan defaults.

3. Research Methodology

3.1. Logistic Regression

The Logistic regression model is a parametric statistical model used to predict the probability of a binary outcome. It takes inputs (or independent variables, explanatory variables) and estimates a linear function of them, then the weighted sum of the inputs will be transformed through a Logistic function to output a continuous number between 0 and 1 as the predicted probability of the outcome (or dependent variable). The functional form of the Logistic regression can be written as

$$logit(Y) = \log log\left(\frac{Y}{1-Y}\right) = \beta^{'} X$$

where Y is the binary dependent variable which equals 1 for an event and 0 for a non-event, and X is the vector of the independent variables. The β are the model coefficients which can be estimated by maximizing the log-likelihood function. With the estimated $\hat{\beta}$ the response probability can be predicted as

$$P(Y = 1) = \frac{e^{\hat{\beta}^{'} X}}{1+e^{\hat{\beta}^{'} X}}$$

The modeling analysis of this paper is implemented in *R* 3.6.1. We fit the Logistic regression model using the *glm* function available in base *R*.

3.2. XGB

XGB (Chen and Guestrin, 2016) is a type of ensemble model. Arguing that one single learner might be unstable, ensemble models train multiple learners with a common objective to generate a more robust learner. XGB leverages the boosting technique under which learners are built sequentially until no further improvement can be made. Boosting creates multiple new training datasets via random sampling with replacement over weighted data and the previous learner informs the weight. XGB is shown to be effective in reducing variations and improving stability and is gaining wide popularity. This paper will calibrate and estimate XGB models using *R caret* package with the *xgboost* package running in the background.

In machine learning models, the dependent variable is often called the label and the independent variables are called the features. Corresponding to variable coefficients in parametric models, machine learning models are specified by hyperparameters. The XGB hyperparameters include nrounds (number of iterations), max_depth (maximum tree depth), eta (learning rate), gamma (minimum split loss), min_child_weight (minimum child weight), lambda (L2 regulation on leaf weights, equivalent to Ridge regression if equals 1), alpha (L1 regulation on leaf weights, equivalent to LASSO regression if equals 1), colsample_bytree (subsample ratio of columns), and subsample (subsample percentage). XGB models with larger number of iterations, deeper trees, smaller learning rate, smaller split loss, and smaller child weight can achieve better fitting of the development sample; but might overfit with less comparable performance on the testing sample. We calibrate the key hyperparameters to reduce variations of model performance across samples. A grid search is performed for combinations of the hyperparameter values to identify a set of hyperparameters leading to the best performance. In addition, this paper uses the 5-fold cross-validation to address potential overfitting issues. The training dataset is divided equally into five random samples and each time four of them are used for model training and the remaining one sample is used for validation. For a combination of hyperparameters, the overall performance is the average of the five performances on the validation sample as it takes one of the five samples. In the end, the combination of hyperparameters resulting in the best performance is selected as the optimal hyperparameters. The final model is then estimated using the optimized hyperparameters on the entire training sample.

Table 2: Hyperparameter tuning

Hyper parameters	max tree depth	eta (learning rate)	gamma (min split loss)
search range	5, 6, 7, 8	0.01, 0.05, 0.1	0.01, 0.05, 0.1
XGB-X0	5	0.05	0.05
XGB-X1	5	0.05	0.01

Notes: The XGB hyperparameters are tuned using the tuning sample of a 30% random sample of the training dataset with total observations of 181,049 observations.

We tune three key hyperparameters of XGB: maximum tree depth, learning rate, and minimum loss split. Table 2 Row 1 shows the search ranges of the hyperparameter tuning. We test maximum tree depth from 5 through 8, and learning rate and minimum loss split for the values of 0.01, 0.05, and 0.1, respectively. In total, 36 combinations of hyperparameters are evaluated for each model. We set the number of iterations to be 500, which is reasonably large. The minimum child weight is set fixed at the value of 5. The hyperparameter lambda is set at 1, alpha is set at 0, subsample ratio of columns is set at 0.8, and subsample is set at 1. Rows 2-3 of Table 2 report the optimized hyperparameters of the XGB model with feature set X0 and X1. Details on the feature sets are provided in the next sub-section.

3.3. Model Specification

We use the LC assigned interest rate as the baseline for default prediction. As the interest rate is assigned by the loan grade, the predictive power of interest rate reflects the quality of grade in signaling loan quality. To calibrate the difference between the 2-year default probability and the interest rate of 3-year maturity, we estimate a Logistic regression of default with interest rate as the only explanatory variable. We record the benchmark model as LR-X0. Similarly, we develop a XGB benchmark model, XGB-X0, with interest rate as the only feature. We expect interest rate is better calibrated under XGB-X0 than LR-X0 since ML can capture nonlinear relationships better than LR, which is a generalized linear regression.

The feature set used to develop the screening models contains interest rate, loan and application information (funded loan amount, amount difference between the applied and funded loan amount, PTI, employment length, loan purpose, home ownership, listing status, and verification status), key bureau variables (DTI, the lowest FICO), and the 15 selected bureau variables. We also control for loan issue year and residence state. The corresponding screening models are labeled as LR-X1 and XGB-X1.

3.4. Performance Measure and Model Interpretation

Model performance is evaluated for its ranking ordering. The performance metric used for model rank ordering is AUC. With a range from 0 to 1, the AUC reflects the area under the receiver operating characteristic (ROC) curve. The larger the AUC, the better the model classifies events. The ROC curve plots the true positive rate (TPR) on the y axis and false positive rate (FPR) on the x axis, so ideally the ROC curve should be pushed to the edge of the upper left corner with 100% TPR and 0% FPR and area under the curve is maximized to be 1.

Unlike parametric statistical models whose model specification can reveal the relationship between model inputs and model outputs, dubbed as a "black box", ML is more complicated and less transparent. We report variable importance of the models to help understand the contributions of variables to loan screening. The variable importance of LR is ranked by the magnitude of the standardized coefficients, and the variable importance of XGB is ranked by the impurity importance which is defined as the improvement in the performance measure (the area under the curve, AUC, is used in this paper) attributed to the splitting variable at each split in each tree, weighted by the number of observations the node is responsible for. The feature importance is then accumulated over all the trees in the model for each variable. The variable importance is scaled to the variable with the top importance with a value of 0 to 100.

4. Results

Table 3 reports the model performance in rank ordering defaults. If calibrated by LR, the interest rate has a training AUC of 66.0 and a testing AUC of 67.7. The performance is similar under XGB with a training AUC is 67.0 and a testing AUC of 66.9. If investors develop customized screening models supplementing interest rate with publicly available information on loan and borrower, their screening ability can be improved even more. The LR-X1 has a training AUC of 69.7 and a testing AUC of 69.9, and the XGB-X1 has a training AUC of 72.8 and a testing AUC of 70.7. Comparing XGB-X1 with LR-X0 shows that by leveraging both the full set of granular information and loan grade, and the more sophisticated ML algorithm, the performance of screening is improved by 10.3% (AUC 72.8 vs. 66.0) in sample and 4.4% (AUC 70.7 vs. 67.7) out of sample. In utilizing the full information set to develop the screening model, the XGB model performs better than the traditional statistical model, with an AUC gain of 4.4% (72.8 vs. 69.7) in sample and 1.1% (70.7 vs. 69.9) out of sample.

The better performance of machine learning algorithm XGB over LR could be attributed to its ability to fit complicated data patterns such as nonlinearities and interactions. For example, the current specification of LR assumes a linear relationship between default and FICO, but the relationship could potentially be nonlinear. Sourced from the same borrower credit profile, FICO could be related to other bureau variables, which is not accounted for explicitly in the LR model used. We may specify more complicated parametric models to account for possible nonlinearities or interactions, but ML models can learn the data patterns by themselves, resulting in more flexible and better model performance.

Table 3: Model performance comparison

	LR-X0	XGB-X0	LR-X1	XGB-X1
Training	66.0	67.0	69.7	72.8
Testing	67.7	66.9	69.9	70.7

Table 4: Model variable importance comparison

LR-X1		XGB-X1	
int_rate	100	int_rate	100
pti	49	pti	15
acc_open_past_ 24mths	43	acc_open_past_ 24mths	14
dti	42	dti	10
fico_range_low	28	issue_y2015	9
mths_since_ recent_bc	24	fico_range_low	8
percent_bc_gt_75	20	tot_hi_cred_lim	6
inq_last_6mths	18	emp_length	5
mo_sin_old_ rev_tl_op	18	total_bc_limit	5
emp_length	18	avg_cur_bal	5

Table 4 lists the 10 variables with top importance. Both LR-X1 and XGB-X1 list interest rate as the most important variable. Both models list PTI, number of accounts opened in the past 24 months, and DTI as the next three most important variables. The FICO score is not ranked as high with consideration of more granular bureau variables. Overall, this paper finds that LC loan grade is highly effective in risk screening, however, application and borrower information can supplement loan grade to further improve loan screening. Such variables are PTI, number of accounts opened in the past 24 months, DTI, and FICO.

5. Conclusion

In P2P lending, the platforms do not serve as the intermediator and therefore do not bear the credit risks of the listed loans. Investors take on the credit risks directly and therefore the objectives of platforms and investors could potentially be misaligned. This creates a motivation for P2P investors to actively screen loans rather than depend on the suggestive grade provided by platforms. Platforms are required to share with investors the same information they used in approving loans and assigning grades, making it feasible for investors to screen loans beyond loan grade. This paper shows the possibility and value of active screening. Leveraging both the full set of granular information and loan grade, and under the more advanced ML algorithm, the performance of screening is improved by 10.3% in sample and 4.4% out of sample in rank ordering. XGBoost also outperforms the conventional parametric Logistic regression model with a performance gain of 4.4% in sample and 1.1% out of sample.

References

Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing Journal*, 73: 914–920. https://doi.org/10.1016/j.asoc.2018.09.029

Chang, S., Kim, S. D., & Kondo, G. (2015). Predicting default risk of lending club loans. *Machine Learning*, 1–5.

Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785–794. ACM. https://doi.org/10.1145/2939672.2939785

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1): 54–70. https://doi.org/10.1080/00036846.2014.962222

IBISWorld (2020, December 17). *Peer-to-Peer Lending Platforms Industry in the US – Market Research Report*. https://www.ibisworld.com/united-states/market-research-reports/peer-to-lending-platforms-industry/

Iyer, R., A. Khwaja, E. Luttmer, and K. Shue. (2015). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, 62:1554–77. https://doi.org/10.1287/mnsc.2015.2181

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10): 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

Prime Meridian (2015, December 28). *Marketplace lending in 2015 – a year of performance and growth.* https://www.pmifunds.com/marketplace-lending-in-2015-a-year-of-performance-and-growth/

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. PLoS ONE, 10(10), 1–22. https://doi.org/10.1371/journal.pone.0139427

Vallee, B., Zeng, Y., (2019). Marketplace lending: a new banking paradigm? *Review of Financial Studies*, 32(5): 1939-1982. https://doi.org/10.1093/rfs/hhy100