Journal of Internet Banking and Commerce

An open access Internet journal (http://www.icommercecentral.com)

Journal of Internet Banking and Commerce, December 2016, vol. 21, no. 3

PEER TO PEER LENDING, DEFAULT PREDICTION-EVIDENCE FROM LENDING CLUB

SRIHARSHA REDDY

Associate Professor, Institute of Management Technology (IMT), Hyderabad, India, Tel: 9849528676;

Email: srihasrshareddy@imthyderabad.edu.in

KRISHNA GOPALARAMAN

Managing Consultant, enkeyed.com, Hyderabad, India

Abstract

Purpose– The objective of this paper is to outline an approach towards a Classification Problem using R. The focus is on two problem statements as stated below:

1. To combine the data on loans issued and loans declined and build model that replicates Lending Club Algorithm closely
2. Using Lending Club's published data on loans issued and its various attributes, build model that can accurately predict probability of delinquency.

Design/methodology/approach– In order to build a model which replicates lending club algorithm closely various classification techniques such as Logistic Regression, Basic Classification Trees, Generalized Linear Model with Penalization, Ensemble of Decision Trees and Boosted Trees were used using R. Boosted Trees classification method is deployed to build model that can accurately predict probability of delinquency.

Findings– Risk Score variable figures as the top of the variable importance list followed by length of employment as one of the more important variables in determining whether loans where eventually issued. Risk Score (at Origination) figures as the top of the variable importance list. This is followed by Amount Paid as a % of Loan Amount as one of the more important variables in determining whether loans would turn delinquent. The performance (accuracy) on training as well as test set is best given using the xgboost model at 99%.

Practical implication– The paper includes implications for the borrowers to understand the factors influencing the decisions of issuance of loan and for the investors to understand the reasons for delinquency in peer to peer lending.

Originality/value– This paper fulfills an identified need to build a model to predict probability of success in getting loans with identification of reasons for issuance of loans at Lending Club. Similarly, it also attempts to build a model to predict probability of delinquency and reasons contributing to delinquency to benefit investor's community at Lending Club.

Keywords: **Lending Club, Probability of Default, Peer To Peer Lending**

JEL Classification: **G15, G20**

INTRODUCTION

With the emergence of online communities in the past decade a new way of loan origination has entered the credit market: online peer-to-peer (P2P) lending. In this kind of lending model the intermediation of financial institutions is not required [1]. The decision process of loan origination is given to the private lenders and borrowers, and portals such as Prosper.com, Lending Club. Within these platforms borrowers generally describe the purpose of their loan request and provide relevant information about their current financial position. The advantage to the lenders is that the loans generate income in the form of interest, which can often exceed the amount of interest that can be earned by traditional means (such as from saving accounts and CDs). P2P loans give borrowers access to financing that may not have been available from standard financial intermediaries. The platforms often benefit by raising fees for successful realized transactions. Although online P2P lending is a relatively new field of research an increasing amount of scientific contributions has been published in recent years [2-4]. With the emergence of the first online P2P lending platform "Zopa" the new lending model raised attention for the first time in the year 2006 [5]. However it was Prosper.com, who caused a wave of scientific contributions by making the entire platform's data public in 2007. Since then, the topic has attracted researchers from the fields of economics, information technology and social sciences to investigate the relationships between lenders and borrowers in online P2P lending platforms. With the availability of live data from Lending Club, our aim is to predict credit risk in peer to peer lending using appropriate predicting models using ‘R’. Lending club is one of the world’s largest online credit marketplaces, facilitating personal loans, business loans, and financing for elective medical procedures. Borrowers access lower interest rate loans through a fast and easy online or mobile interface. Cumulative amount of loans funded by Lending Club as on 31 March 2016 is $18,732,087,097.

WHAT IS P2P LENDING?

Peer-to-peer finance can be defined as “platforms that facilitate financial services via direct, one-to-one contracts between a single recipient and one or multiple providers” (As per the definition of British Peer-to-Peer Finance Association). Peer-to-peer lending (P2P) is a method of financing debt that enables individuals to borrow and lend money - without the use of an intermediary. Peer-to-peer lending removes the middleman from the process. The advantage to the lenders is that the loans generate income in the form of interest, which can often exceed the amount of interest that can be earned by traditional means (such as from saving accounts and CDs). P2P loans give borrowers access to financing that may not have been available from standard financial intermediaries. Borrowers apply for loans on a P2P portal like Lending Club. P2P platforms evaluate each loan request and lists only those applications that meet credit criteria. Investors have exposure to many different individual loans to diversify their investment. As borrowers make scheduled principal and interest repayments on their loans, investors receive predictable cash flows.

The idea of private loans is an old business model where private persons borrow money without any mediation [6,7]. Online P2P lending is a recent phenomenon where private persons borrow money using online P2P lending platforms like Lending club.com.

The first lending platform, Zopa was established in Europe (UK) in 2005. Since then various forms of lending platforms followed [8,9] identify 67 platforms existing worldwide, with 17 platforms in Americas, 36 in Europe and 16 in Australasia.

The first lending platform in the United States was launched in February 2006 (prosper.com). Smava (smava.de), the first German P2P lending company, was founded in February 2007. Today most of the existing platforms work on a national level, due to different legal requirements in different countries [10]. The following Table 1 shows list of lending platforms as on September 2016.

Table 1: Peer-to-Peer Lending Platforms (Country, Launch year).

Business Focus	**America**	**Europe**	**Asia/Australia**
Personal Loans	Lending Club (USA, 2007)	Zopa (UK,2005)	CreditEase (China, 2006)
	Prosper (USA, 2006)	Yes Secure (UK, 2012)	PPDai (China, 2007)
	Peerform (USA, 2012)	Smava (Germany, 2007)	Renrendai (China, 2007)
	Wikiloan (USA, 2012)	auxmoney (Germany, 2007)	MoneyAuction (South Korea, 2007)
	Fairplace (Brazil, 2010)	Kokos (Poland, 2008)	Popfunding (South Korea, 2007)
		Finansowo (Poland, 2008)	Donjoy (South Korea, n/a)
		Prestiamoci (Italy, 2007)	maneo (Japan, 2008)
		Boober (Italy, 2007)	Aqush (Japan, 2009)
		Friendsclear (France, 2008)	SBI Social Lending (Japan, 2011)
		Prêt d‘Union (France, 2011)	iGrin (Australia, 2007)
		Cashare (Switzerland, 2008)	Lending Hub (Australia, 2009)
		Fixura (Finland, 2010)	Nexx (New Zealand, n/a)
		isePankur (Estonia, 2009)	Lendit (New Zealand, n/a)
		Comunitae (Spain, 2009)	Indialends (India,2014)
		Lendland (Russia, n/a)	P2PLendingindia (India,n/a)
		Frooble (Netherlands, 2007)	Loanmeet (India,2012)
		FriendCredit (Romania, 2012)	Peerlend (India,2015)
		Noba (Hungary, 2010)	Cashkumar (India,2014)
		RateSetter (UK, 2010)	Ilendclub (India,2014)
		Bondora (Estonia, 2008)	i2ifunding (India,2014)
		Burnley Savings and Loan (UK,2011)	RupaiyaExchange (India,2013)
		Buy2Letcars.com (UK,2012)	Faircent (India,2013)
		SavingStream (UK,2013)	Kickstart (India,n/a)
		Folk2Folk (UK,2013)	
		Lendlinvest (UK,2013)	
		AssetzCapital (UK,2013)	
		Sancus (Jersey,2015)	
		Orchard (Jersey,2016)	
		Crosslend (Germany,2015)	
		Viventor (Latvia,2015)	
		FellowFinance (Sweden,2015)	
		Crowdhouse (UK,2015)	

		Elevate (UK,2015)	
		MoneyThing (UK,2015)	
		Btcpop (UK,2015)	
		Beehive (UAE,2014)	
Payday Loans	YadYap (USA, 2010)	Trustbuddy (Norway, 2010)	
	Kudols (USA, 2012)	Relendex (UK,2013)	
		The Lending Well (UK, 2012)	
Student Loans	People2Capital (USA, 2010)		Qifang (China, 2008)
	Social Finance (USA, 2011)		
Commercial and residential real estate	Money360 (USA, 2010)	Relendex (UK, 2010)	
		RebuildingSociety (UK,2012)	
Business	Rebirth Financial (USA, 2011)	Fundingcircle (UK, 2010)	
	SoMoLend (USA, 2012)	ThinCats (UK, 2010)	
		Platform Black (UK, 2011)	
		Massow’s Angels (UK, 2012)	
		FundingKnight (UK,2012)	
		YouAngel (UK,2011)	
		One Stop Funding (UK,2011)	
		Twino (Latvia,2015)	
		GrowthStreet (UK,2015)	
		GO2BusinessLoans (UK,2015)	
Leasing		Squirrl.com (UK, 2012)	
Factoring		MarketInvoice (UK, 2010)	
Source: Author compilations			

Online P2P lending platforms differ in type and the approach adopted. They can basically be divided into two types: commercial and non-commercial [11]. While commercial platforms in general are limited to national markets, noncommercial platforms often operate globally. The main difference between the two platform types is the lender’s general intention and his expectations concerning returns. A lender who engages in commercial platforms gets a reasonable interest for the risk he is taking. In non-commercial platforms lenders get no or little reward for the risks they are willing to take. Here lenders rather want to “donate” small loans to projects in economically underdeveloped regions in the world.

LITERATURE REVIEW: P2P LENDING, SUCCESS RATE AND DELINQUENCY

Studies on Success Rate

There are very few studies in existing literature on factors contributing to borrowers' creditworthiness. Lin et al. [12] found that borrowing requests with lower credit ratings are less likely to be funded and more likely to default and end with higher interest rates using data collected from Prosper.com. Lin et al. [12] found that bank card utilization has a curve linear effect on lending outcomes: while bank card utilization at low and medium levels signals the creditworthiness of borrowers, very high utilization of bank cards leads to decreased funding probability and increased interest rates due to the risk of high leverages and vulnerability to shocks. Iyer et al. [2], found that borrower's default rate, debt-income ratio, and the number of loan requests in the last six months has had a salient negative effect on a lender's decision.

Although there are no conclusive findings concerning the impact of credit rating on lending outcome for online P2P lending websites in China, Chen [13] reported that credit rating in Ppdai.com in China is influential in determining funding probability, but less of a determinant for interest rates. However, default rate is much lower for borrowers with higher credit levels. Success rate of a loan is negatively correlated with the interest rate. Moreover, the size of loan is associated with lower success rate and higher interest rate; therefore, it is possible for borrowers to increase the success rate of a loan by paying higher interest rates and/or reducing the loan size [14].

Studies have also revealed that lenders would use some subjective, non-standardized information to derive the borrower's credit ratings. For instance, the highest interest rate that the borrowers are willing to pay is a valuable, positive signal for potential lenders [2]. For the lending websites in China, information asymmetry is found to moderate the impact of social capital on trust, which is critical to willingness to lend [15].

Studies on Delinquency

Lin et al. [16] estimated that friends in a borrower's social network with verified identities as lenders decreased the odds of default by 9% on average. In addition, by analyzing 6-month secondary data on lenders, borrowers and loan repayments collected from Prosper.com, Kumar [17] showed that credit grade and account verification were associated with lower probability of loan default while loan size was positively associated with default rate. Interestingly, certain factors which affect interest rates and risk premiums, such as debt to income ratio, home ownership and group leader endorsement, demonstrated no significant effects on default rates.

A comprehensive analysis of Lending Club loan data by Emekter et al. [18] reveals that there exists a selection bias in the sense that high-income borrowers with the highest FICO credit scores (A FICO score is a type of credit score created by the Fair Isaac

Corporation. Lenders use borrowers' FICO scores along with other details on borrowers' credit reports to assess credit risk and determine whether to extend credit) do not borrow from Lending Club. In particular, top one third of the consumers with respect to FICO scores do not create any loan listings on Lending Club. They also observed that higher interest rates charged on the higher risk borrowers are not worth the risk. Specifically, higher rates charged for the borrowers with low credit grade of lending club are not high enough to overcome the greater default risk that the lenders take.

The above findings are important for investors participating in social lending to identify those who will pay back their loan in full within due time. Profitability of investors is a critical component in overall sustainability of the social lending market. In this regard, Emekter et al. [18] suggest that ‘‘the lenders would be better off to lend only to the safest borrowers with the highest Lending Club grades’’.

In order to improve identification of good borrowers within the context of social lending, this study proposes and presents comparisons of different machine learning methods including Classification Tree (rpart), Logistic Regression (glm), Generalized Regression Models (glmnet), Random Forests (randomForest) and Gradient Boosted Trees (xgboost). Our computational results on Lending Club data between January 2015 and March 2016 indicate that Gradient Boosted Trees (xgboost) outperform the other classification methods and stand as a scalable and powerful approach for predicting borrower status.

Lopez [19] used Gaussian mixture models on the Prosper data set containing loan transactions between November 2005 and December 2008. Lopez [19] found that if an individual with a high-risk FICO score belongs to a trusted social community, then this individual’s social membership can still help secure a loan. Thus, even though a high-risk credit score usually means lack of access to traditional bank-mediated financial markets, a positive social feature can outweigh a highly negative financial feature in socially mediated markets. Complex behavioral dynamics further complicate the social lending process. For example, the simple auction mechanism used in some social lending platforms can lead to unpredictable payments for the borrower. An incentive compatible mechanism might be more suitable to eliminate this inefficiency where lenders report their true interest rate and do not change their rate dynamically [20]. Otherwise, such inefficiencies enable users with adversarial interests to use the lending platform as an arbitrage opportunity: borrow at 10% and then loan at 20% [21].

Empirical studies show that when a group leader in a lending platform mediates the group actively, the risk factor drops considerably. In addition, if a group leader recommends a loan listing put together by one of the group members, this endorsement increases the chance of the loan being issued and also decreases the final interest rate [10].

There exist several studies proposing a set of guidelines in order to make purely rational investment decisions in social lending. In one such study on Prosper loan data that includes loan transactions between November 2005 and March 2007, irrespective of the financial credit rating categories, three simple rules help decrease the risk of a default [22]. These investment rules are as follows:

1. Invest only in borrowers without any delinquent accounts.
2. Invest only in borrowers that satisfy Rule 1 and that have a debt-to-income (DTI) ratio less than 20%.
3. Invest in borrowers that satisfy Rule 2 and that have no credit inquiry reports during the last 6 months.

In studies conducted on social communities, herding (denser clustering following a power law regime) effects usually prevail [23-25]. Empirical studies show that the tendency of an individual to join a given community is effected by the number of friends in this community and the inter-connectedness of this individual's friends within the community. Such behavioral bias also exists in investment decisions of lenders at Prosper. The loan data between 2006 and 2008 show that previous lender decisions effected subsequent lender decisions and lender decisions were not made purely rationally [26]. For the interested reader, there exist other real-world networks (such as airports and power grid transmission lines) and other social networks (such as DBLP and LiveJournal) that also exhibit a herding behaviour [27,28].

The closest study to ours is the work of Emekter et al. [18] where the authors analyze Lending Club data between May 2007 and June 2012 and present a logistic regression (LR) model for predicting default probability of a borrower. Their model includes FICO scores as well as Lending Club grades in default prediction [29].

We were motivated to undertake present study due to availability of fewer studies on predicting delinquency using Lending Club data. Present study has following two objectives:

1. To combine the data on loans issued and loans declined and build model that replicates Lending Club Algorithm closely
2. Using Lending Club's published data on loans issued and its various attributes, build model that can accurately predict delinquency.

METHODOLOGY: OVERVIEW OF CLASSIFICATION TECHNIQUES USED

When outcome or response variables are categorical or qualitative we can adopt various methods to predict such classification. The methods used in our study are explained below.

Logistic Regression (glm package in R): Logistic regression belongs to the class of generalized linear model and it measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities.

Basic Classification Trees (rpart package in R): These are decision trees that partition data into smaller homogenous groups with nested if-then statements.

Generalized Linear Model with Penalization (glmnet package in R): These models fits a generalized linear model via penalized maximum likelihood viz. shrinking the coefficients as well as number of predictors used.

Ensemble of Decision Trees (randomForests package in R): It extends the concept of decision trees to build an ensemble of trees, with each tree built using a sample of predictors to mitigate over-fitting.

Boosted Trees (xgboost package in R): It extends the concept of ensemble of decision trees, except that each tree built is based on the previous tree and seeks to minimize the residuals.

Data Preparation and Sources

The data is available for the period from 2007 until 2016Q1. There are over 8 Million records of which about 12% constitute loans issued and rest 88% for which loans were declined. There are a total of 115 variables associated with each record of issued loans and 9 variables associated with each record of rejected loans.

For purposes of data preparation, a number of steps were undertaken. Duplicate rows, if any were removed from the data. Also, wherever case IDs were missing, such rows were dropped. These formed a very insignificant portion of the total data set.

There were a number of records for which certain variables has no data. This could be because they are not applicable for the specific record and/or the variable was introduced at a later date and hence earlier records have missing data and/or data was simply not available or not recorded. There are a number of options available. One of them would be to eliminate rows with missing data – this would reduce the dataset to almost 3% of its size and valuable information would be lost. Second option would be to identify specific columns that contribute to missing data and eliminate them. However, we have chosen to do neither and have retained all of the data, since some of the machine learning algorithms can handle missing values and larger the data set available for training, better is the model performance.

Since some of the machine learning packages do not work with missing data, we decided to impute some values so that we train a uniform data set across all algorithms. There are a couple of ways to do this. One would be to impute some sort of mean, median value to missing cells of a column or another option would be use packages such as MICE (R Package) to impute values based on nearest neighbor or other such logic. In our case we decided to preserve the information that value is missing by assigning a value furthest from the values present. E.g. if the values of a column range between 1 to 20, we impute a value of -9999 for representing the missing values of numeric variables and “NODATA” for categorical variables.

Another important step is identifying variables that do not qualify to be predictors e.g. Customer or Loan ID. Many of the simpler models do not perform well when you have highly correlated variables as predictors. We removed highly correlated predictors and/or those that with only one unique value. Cleaning numeric data e.g. Removal of “%” sign from “40%” was done. Predictors which represent dates were converted to date format and extracted year, month as separate columns. Ordering levels where applicable in case of categorical (or factor) data, e.g. Years of experience were ordered from “<1” to over “>10”. This facilitates ease in visualization as well processing.

There are a number of categorical variables with many levels or unique values. This data has to be converted into numeric. The options available to process such variables are either feature hashing or hot-encoding or binning or a combination thereof. Where you cannot reduce the number of levels and yet want to retain size of data set within reasonable limits we would undertake feature hashing. In this case we reduced number of levels in case of categorical (or factor) variables and created dummy variables from categorical (or factor) data. E.g. Loan Term has only two options viz. 18 months and 36 months and therefore amenable to creating dummy variables. In other cases, dummy variables are created with a cut-off for cumulative frequencies, beyond which all values default to “other”.

Couple of variables included text data. We applied Natural Language Processing methods to derive terms (words) and their frequencies in that text and add these to the list of predictors. In this we followed a simple process of converting text of each cell by removing punctuation and numbers, convert to lower case, convert to base words, remove common words like “to”, etc. and then constructing a document term matrix by removing sparse terms.

Data centering and scaling is recommended when amongst other factors, variable importance is derived from value of coefficients while using some of the simpler models like logistic regression. Centering and scaling is not necessary when using some of the more advanced or tree-based models. Better option is to leave the data as it is, scaling option can be invoked (as needed) for specific algorithms while training them.

The success of any machine learning exercise is the ability to do feature engineering. Basically, we need to make it easy for algorithms to find basis to bifurcate data. In this case we added predictors that are derived from a combination of one or more columns and/or grouping data. E.g. for each row we introduce a column to represent number of months since beginning of term, percentage principal paid till date, etc..

And lastly, we formulated the problem statement and evaluation criteria, based on which an outcome variable was defined (modified from original data set).

DISCUSSION AND ANALYSIS

Problem Statement 1

Using Lending Club's published data on loans issued and its various attributes, build model that can accurately classify loans issued and loans declined.

The objective of this exercise was to seek to replicate as closely as possible the underlying model of LendersClub.com. Towards this end we needed to set-up data for training and validation/test. The key points thrown up by the analysis of data that determined this were the following

1. It is observed that the number of loan applications increased exponentially over the years. A high percentage of applicants were declined and this percentage reduces over the years. Over the last two years the loans declined reduced from over 80% to around 55%.
2. In addition, risk score prior to November, 2013 was FICO score and post-Nov 2013 it was vantage score.

Therefore, in order to build the model, data of 2015 was used to train the models and data of 2016 was used to test the model. This resulted in split of 75:25 for train: test, which is reasonable.

Data on loans issued and loans declined were combined. While building the model it is assumed that denied loans include those were either not offered for investors by Lending Club and/or for which investments were not forthcoming, the decision to deny loan was based on these 9 predictors only and loan issued month is considered as nearest proxy for loan applied month.

Interface used to build models was the Caret package in R. The training function in caret currently supports 192 different modelling techniques and has several functions that attempt to streamline the model building and evaluation process.

5 Models were developed by training 5 different algorithms on 2015 data consisting of over 900,000 cases (rows) and over 40 Columns (or Predictors). The data on which model was tested was from 2016Q1 and consisted of over 300,000 cases (Table 2).

The algorithms (R Package) used included Classification Tree (rpart), Logistic Regression (glm), Generalized Regression Models (glmnet), Random Forests (randomForest) and Gradient Boosted Trees (xgboost).

Cross Validation was done to derive true estimate of model performance. For all models 5-fold validation was used and for xgboost a 10-fold validation was used. By turns the model is trained on all but one fold and the held out fold are predicted by the model to estimate performance measures likely on unseen test.

Table 2: Comparison Performance Measures (Loan Declined vs. Issued).

Description	**Classification Tree**	**Logistic Regression**	**Generalized Regression Models**	**Random Forest**	**Boosted Trees**
R Package	rpart	glm	glmnet	randomForest	xgboost
Using Caret	Yes	Yes	Yes	Yes	Yes
No of Predictors	42	42	42	42	42
Train Data – Year	2015	2015	2015	2015	2015
Train Data - No of Observations	929,883	929,883	929,883	464,941	929,883
Test Data – Year	2016Q2	2016Q3	2016Q4	2016Q5	2016Q6
Test Data - No of Observations	324,425	324,425	324,,425	324,425	324,425
Cross Validation – Folds	5	5	5	5	10
Cross Validation – Repeats	5	5	5	1	1
Tunelength	10	NA	10	5	10
Performance on Training Data					
Accuracy	0.980	0.950	0.945	0.980	0.992
Performance on Test Data					
Accuracy	0.965	0.891	0.892	0.962	0.985
Sensitivity	0.973	0.930	0.925	0.975	0.988
Specificity	0.954	0.835	0.845	0.943	0.981
Positive Class	Declined	Declined	Declined	Declined	Declined

Performance tuning was done to a limited extent to extract best model performance. The measure used to evaluate model performance was accuracy. In the caret package, for each algorithm there are a certain number of parameters than can be tuned manually or auto-search from a grid of values. In this exercise we used the latter option.

The classification by the best model in comparison to reference (LendersClub.com) is given below (Table 3).

Table 3: The best model in comparison to reference (LendersClub.com).

	Reference	
Prediction	**Declined**	**Issued**
Declined	1,88,276 (TP)	2,520 (FP)
Issued	2,262(FN)	1,31,367 (TN)

Accuracy is match of classification over the total number of observations. viz. 98.53%. Sensitivity or True Positive Rate is 98.81%. Specificity or True Negative Rate is 98.12% The criteria or measure for evaluation of performance of models depends on the objective of the exercise. In this case we wanted to be able to classify loans as declined or issued as closely as was done by LendersClub.com. Therefore, we used Accuracy as the measure. However, if we wanted to ensure that we needed to closely align our model to match classification of declined cases of LendingClub.com, then we would choose Sensitivity as our evaluation criteria.

Problem Statement 2

Using Lending Club's published data on loans issued and its various attributes, build model that can accurately predict delinquency.

The objective of this exercise was to seek to replicate as closely as possible the underlying model of LendersClub.com. Towards this end we needed to set-up data for training and validation/test. The key points thrown up by the analysis of data that determined this were the following

- Risk score prior to November, 2013 was FICO score and post-Nov 2013 it was vantage score.
- The term of loan is either 18 or 36 months. We needed to have data that covers loan terms at various stages of completion.

Therefore, in order to build the model, data of 2014 and 2015 was used to train the models and test the model. The split between train and test was 70:30, which is the general convention followed.

Data on loans issued was utilized. To ensure that the model does not have advantage of after-the-fact predictors, these were eliminated. E.g. While Risk Score at Origination was retained, but latest Risk Score were not included. Close to 88 variables were dropped.

Interface used to build models was the Caret package in R. The training function in caret currently supports 192 different modelling techniques and has several functions that attempt to streamline the model building and evaluation process.

Boosted Trees (Xgboost package in R) was trained on data consisting of over 450,000 cases (rows) and over 343 Columns (or Predictors). The data on which model was tested consisted of about 200,000 cases (Table 4).

Table 4: Final Model Performance Measures (Predict Delinquency).

Description	**Boosted Trees**
R Package	xgboost
Using Caret	Yes
No of Predictors	451
Train Data - Year	2014-15
Train Data - No of Observations	4,59,706
Test Data - Year	2014-15
Test Data - No of Observations	1,97,018
Cross Validation - Folds	10
Cross Validation - Repeats	1
Tunelength	5
Tuning Parameters	nrounds=814, max_depth=4, eta=0.3284755, gamma=3.901233, colsample_bytree=0.5827233, min_child_weight=2
Performance on Training Data	
Accuracy	0.994
Kappa	0.938
Performance on Test Data	
Accuracy	0.989
Kappa	0.900
Sensitivity	0.989
Specificity	0.982
Pos Prediction Value	0.999
Neg Prediction Value	0.841
Prevalance	0.944
Detection Rate	0.933
Detection Prevalance	0.934
Balanced Accuracy	0.986
Positive Class	Declined

Cross Validation was done to derive true estimate of model performance. For xgboost a 10-fold validation was used. By turns the model is trained on all but one fold and the held out fold are predicted by the model to estimate performance measures likely on unseen test.

Performance tuning was done to a limited extent to extract best model performance. The measure used to evaluate model performance was accuracy. In the caret package, for each algorithm there are a certain number of parameters than can be tuned manually or auto-search from a grid of values. In this exercise we used the latter option.

Probability of Default is always with respect to a period. To be able to derive these we need different time period snapshots of each loan. This is not available. This analysis does not seek to derive this probability over different time periods.

A loan is said to be delinquent or in Default if it is “overdue” or “charged off” and it is said to be Standard if it is “Fully Paid” or “Current” or “Issued” or “In Grace Period”.

The classification by the best model in comparison to reference (LendersClub.com) is given below (Table 5).

Table 5: The best model in comparison to reference (LendersClub.com).

	Reference	
Prediction	**Default**	**Standard**
Default	10,132 (TP)	242 (FP)
Standard	1,002 (FN)	1,85,642 (TN)

Accuracy is match of classification over the total number of observations. viz. 99.37%. Sensitivity or True Positive Rate is 91.00%. Specificity or True Negative Rate is 99.87%. The criteria or measure for evaluation of performance of models depends on the objective of the exercise. In this case we wanted to be able to classify whether loans would default or remain standard during course of the loan term. Therefore, we used Accuracy as the measure. Another alternative is AUC or Area under The Curve.

FINDINGS AND CONCLUSION

Problem Statement 1

The performance (accuracy) on training as well as test set is best given using the xgboost model at 99.2% and 98.5% respectively. Not surprisingly Risk Score figures as the top of the variable importance list. This is followed by Length of Employment as one of the more important variables in determining whether loans where eventually issued. Surprisingly the “Debt to Income” ratio does not seem to figure in the list of top 20 variables of importance. It is possible for us to use the text included by applicant in the

“Loan Title” column to replicate the lending club outcomes with near perfect accuracy. The terms that figure high of the list of important variables include – consolidation, debt, card, credit, refinance, home, improve (Table 6).

Table 6: Final Model Variable Importance (Loan Declined vs. Issued).

Variable	Imp.	Description
Risk_Score	100.0	Borrower’s FICO at loan origination (Avg.)
consolid	94.7	The loan title provided by the borrower
debt	94.7	The loan title provided by the borrower
Emp.Len_F.10. years	77.7	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Card	71.5	The loan title provided by the borrower
Credit	71.5	The loan title provided by the borrower
Refinanc	71.5	The loan title provided by the borrower
Amount.Requested	66.9	The total amount requested by the borrower
Month_F.Jul	61.5	The date which the borrower applied
Emp.Len_F.2.years	61.4	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Month_F.Oct	60.9	The date which the borrower applied
Emp.Len_F.3.years	60.8	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Month_F.Apr	60.2	The date which the borrower applied
Month_F.Aug	60.2	The date which the borrower applied
Month_F.May	60.1	The date which the borrower applied
Emp.Len_F.1.year	59.9	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Home	59.7	The loan title provided by the borrower
Improv	59.5	The loan title provided by the borrower
Emp.Len_F.4.years	59.4	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Month_F.Sep	59.4	The date which the borrower applied

Under the current model, applying for loan with a view to consolidate loan obligations, having a particular risk score cut-off and employment length would result in favorable outcome viz loan issue.

Problem Statement 2

The performance (accuracy) on training as well as test set is best given using the xgboost model at 99.4% and 98.9% respectively. Risk Score (at Origination) figures as the top of the variable importance list. This is followed by Amount Paid as a % of Loan Amount as one of the more important variables in building model to determine whether loans would turn delinquent. Surprisingly the "Debt to Income" ratio does not seem to figure in the list of top 20 variables of importance. Risk Score (Latest) gives the best indication of possibility of default and was not included in the building of the model, so that we are able to proactively predict upfront the possibility of loan turning delinquent during the term of the loan. It is possible to understand the variable importance and how they influence delinquency across time periods to determine upfront possibility of default. The median return for investors was around 9% based on diversification of loan portfolio, while the interest rates on loan grades A to G were ranging from 7% to 23%. As an investor, use of this model can provide significantly superior returns (Table 7).

Table 7: Final Model Variable Importance (Predict Delinquency).

Variable	Imp.	Description
Risk_Score	100	Borrower's FICO at loan origination (Avg.)
Paid_P_Loan	62.4	Amount Paid as a % of Loan Amount (Predictor Added)
last_pymnt.2016	24.4	Last month payment was received (Year)
last_pymnt.2015	19.8	Last month payment was received (Year)
out_prncp	14.5	Remaining outstanding principal for total amount funded
total_rec_prncp	13.9	Principal received to date
TermLeft	13.9	Term of loan left (Predictor Added)
last_pymnt_amnt	11.9	Last total payment amount received
AMTD	11.3	Months of Term completed till date
int_rate	5.7	Interest Rate on the loan
term..60.months	5.46	The number of payments on the loan. Values are in months and can be either 36 or 60.
month_next_pymnt.Jun	5.25	Next scheduled payment date
fico_range_high	3.85	The upper boundary range the borrower's FICO at loan origination belongs to.
total_rec_int	2.44	Interest received to date
installment	2.4	The monthly payment owed by the borrower if the loan originates.
last_pymnt.2014	1.45	Last month payment was received (Year)
funded_amnt	1.32	The total amount committed to that loan at that point in time.

month_iss.Oct	0.9	The month which the loan was funded
year_iss.2015	0.64	The month which the loan was funded (Year)
bc_open_to_buy	0.47	Total open to buy on revolving bankcards.

REFERENCES

1. Herzenstein M, Andrews RL, Dholakia UM, Lyandres E (2008) The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. Boston University School of Management Research Paper 14.
2. Iyer, R, Khwaja AI, Luttmer EF, Shue K (2009) Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? In AFA 2011 Denver meetings paper.
3. Pope DG, Sydnor JR (2011) What's in a Picture? Evidence of Discrimination from Prosper. com. Journal of Human Resources 46: 53-92.
4. Ravina E (2007) Beauty, personal characteristics, and trust in credit markets. Personal Characteristics, and Trust in Credit Markets.
5. Hulme MK, Wright C (2006) Internet based social lending: Past, present and future. Social Futures Observatory 115.
6. Everett CR (2010) Group membership, relationship banking and loan default risk : The case of online social lending. *Group*. West Lafayette, IN. Retrieved from Available at SSRN: http://ssrn.com/abstract=1114428
7. Herrero-Lopez S (2009) Social Interactions in P2P Lending. Proceedings of the 3rd Workshop on Social Network Mining and Analysis, Paris: pp: 1-8. ACM. Retrieved from http://portal.acm.org/citation.cfm?id=1731011.1731014
8. Frerichs A, Schumann M (2008) Peer to peer banking–state of the art. University of Göttingen. Institute of Business Informatics. Work report 2.
9. Moenninghoff SC, Wieandt A (2013) The future of peer-to-peer finance. Journal of Business Research 466-487.
10. Berger SC, Gleisner F (2009) Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. BuR - Business Research, Official Open Access Journal of VHB 2: 39-65.
11. Ashta A, Assadi D (2009) An Analysis of European Online micro-lending Websites. EMN 6th Annual Conference. Milan: Nantik Lum Foundation, 33: 4-28. Retrieved from http://www.european-microfinance.org/data/file/microlendingwebsites
12. Lin MF, Prabhala NR, Viswanathan S (2012) Judging borrowers by the company they keep: Social networks and adverse selection in online peer-to peer lending, Management Science (forthcoming),
13. Chen DY (2012) Is online peer-to-peer lending market effective? A study on herding behaviour in China, Working Paper (School of Management, Fuzhou University),

14. Collier B, Hampshire R (2010) Sending mixed signals: Multilevel reputation effects in peer-to-peer lending markets, Proceedings of the CSCW, Savannah, Georhia, USA, 197-206.
15. Chen DY, Lai FJ, Nie FQ (2012) Social capital, transaction trust, and information asymetry--an experimental study on peer-to-peer lending, Journal of Beihang University (Social Science Edition) (forthcoming)
16. Lin M (2009) Peer-to-Peer Lending: An Empirical Study. 15th Americas Conference on Information Systems. San Francisco: Association for Information Systems.
17. Kumar S (2007) Bank of one: Empirical analysis of peer-to-peer financial marketplaces, Proceedings of the Americas Conference on Information Systems 1-8.
18. Emekter R, Tu Y, Jirasakuldech B, Lud M (2015) Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. Applied Economics 47: 54-70.
19. Lopez SH (2009) Social interactions in p2p lending. In 3rd Workshop on social network mining and analysis 1-8.
20. Chen N, Ghosh A, Lambert N (2009) Social lending. In 10th ACM conference on electronic commerce 335-344.
21. Steelman A (2006) Bypassing banks. In Region focus 37-40.
22. Klafft M (2008) Online peer-to-peer lending: A lender's perspective. In International conference on e-learning 371-375.
23. Gao R, Feng J (2014) An overview study on P2P lending. International Business and Management 8: 14-18.
24. Lee E, Lee B (2012) Herding behaviour in online P2P lending: An empirical investigation. Electronic Commerce Research and Applications 11: 495-503.
25. Yum H, Lee B, Chae M (2012) From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. Electronic Commerce Research and Applications 11: 469-483.
26. Shen D, Krumme C, Lippman A (2010) Follow the profit or the herd? Exploring social effects in peer-to-peer lending. In International conference on social computing.
27. Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of small world networks. PNAS 97: 11149-11152.
28. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: Membership, growth, and evolution. In 12th ACM International conference on knowledge discovery and data mining, 44-54.
29. Catuneanu O, Abreu V, Bhattacharya JP, Blum MD, Dalrymple RW, et al. (2009) Towards the standardization of sequence stratigraphy. Earth-Science Reviews 92: 1-33.