The University of Dayton | ECO 410-01 & 410-02

Predicting Peer-to-Peer (P2P) Loan Defaults through the Regression Analysis of Underlying Loan Application Variables

Stephen Harvey | Ryan Menker

Professor Caporale

15 December 2016

Table of Contents

Introduction

Initial Analysis and Objectives

In order for banks to lend money to individuals or small businesses, they must be able to predict the borrower's risk of default within a degree of certainty. In order to do this, banks have developed complex models to forecast the probability of any particular loan’s default. Bank’s factor in credit history, debt to income ratio, potential collateral, conditions of the loan, and other variables depending on the specific loan. Banks’ consumer lending forecast models have historically been fairly accurate and consistent over the years. Recently, a new form of lending known as peer-to-peer (P2P) lending has captured the interest of lenders and borrowers. According to a PriceWaterhouseCoopers report, “although still in its infancy as a market, US P2P platforms issued approximately $5.5 billion in loans in 2014.” PwC estimates that the market could reach $150 billion or higher by 2025 (1). Unlike consumer lending from banks, peer-to-peer lending leaves the lending up to the discretion of the individual, not an institution. Lenders and borrowers are directly matched through the medium of a website. This new style of lending has created a challenge for people to figure out the most accurate way to determine the default risk of these peer-to-peer loans. The goal of this empirical research paper is to develop a model that predicts the default risk of any peer-to-peer loan application.

Empirical Research Importance and Reasoning

In the mid 2000s, peer-to-peer lending was introduced into the market with the creation of companies like Prosper and Lending Club. The launch of peer-to-peer lending websites allowed not only borrowers an option to take a loan without having to go through a bank, but it also gave individuals seeking passive income an opportunity to lend their personal capital to borrowers and earn interest. At the beginning, there were relatively high borrower default rates. The reason is that on most of the peer-to-peer platforms, there were “few restrictions on borrower eligibility” (2). Additionally, there was little to no regulation or government oversight. People who had a poor credit history and had been denied loans from multiple banks were able to secure loans from individuals using a peer-to-peer platform. The lenders likely did not fully calculate and understand the risks they were taking with these loans. In contrast, banks use trained professionals, special credit software, and fully developed default forecasting models before any loan is given. With peer-to-peer loans, however, any person can arbitrarily lend money to an anonymous borrower for any reason. This can be disastrous for those people who lend money on a peer-to-peer platform that have no knowledge of the credit market and no way to understand and assess a loan’s riskiness. It is important for lenders to understand the potential risk, so that they can accurately determine whether it is in their best interest to invest in a specific loan and what the probability of that loan defaulting is.

Review of Existing Literature

For the purpose of our empirical research project, we used data generated from loans on the Lending Club platform. Lending Club is one of the top peer-to-peer lending companies in terms of the number of loans generated per year and the gross total monetary amount of loans. It is registered with the Securities and Exchange Commision (SEC) with its loan offerings traded as securities on a secondary market called FOLIOfn. The company has been publicly traded on the New York Stock Exchange (NYSE) under the ticker "LC" since it had its initial public offering in December 2014. It was the first peer-to-peer lender to register its offerings as securities with the SEC, and to offer peer-to-peer loan trading on a secondary market. People use Lending Club to fund personal loans for purposes such as paying off credit cards, debt consolidation, or home improvements. Lending Club also allows for auto refinancing and small business loans for up to $300,000. Our research is focused on personal loans generated from 2007-2011 that have either been fully paid off or have defaulted. We chose Lending Club not only because they are one of the most well known and largest, but because they also have one of the most extensive data sets to analyze.

Even though peer-to-peer lending companies have been around for a relatively short amount of time, there has been extensive research conducted on the topic of default rates, interest rates, and creditworthiness of peer-to-peer loans. One report titled "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending" by Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech & Min Lu, explores credit risk in loans generated from May 2007 to June 2012 on Lending Club. This report concluded that credit score, debt-to-income ratio, FICO score and revolving line utilization had the largest impact in determining a specific loan's default risk. The report focused in large part on FICO score. A panel model was tested in which there were multiple categories of FICO scores. For example, FICO scores with the following ranges of 714-749 and 750-779 were tested separately. They concluded from their models that the first FICO range had a maximum default risk of 0.0950 and a minimum of 0.0787. The FICO range of 750-779 had a maximum default risk of 0.0720 and a minimum of 0.0632. As the FICO score drops into the 600s, loans defaulted at a much higher rate with FICO scores in the range of 640–659 defaulting between 0.5963 and 0.6286 within a 95% confidence interval. The authors suggested Lending Club should only allow users with higher FICO scores to be extended credit because the loans with a lower FICO defaulted at such a high rate. (3)

After reading this piece of literature, we developed a model similar to the one built in the report, except we applied our model to newer loans and analyzed the results. We also left out the factor of FICO. Lending Club's data set no longer has FICO scores available for past loans, so we were forced to work without that variable. More specifically, we created a model forecasting default rates based on Lending Club data from 2007-2011 and tested that model on loans from 2012-2013. Our empirical project adds to the original literature by testing our model on loans outside of the data set and by omitting the factor of FICO in our data set and model.

The Model

Data Collection and Munging

In order to estimate the probability of a loan's default from its initial application, we performed an ordinary least squares (OLS) linear regression. For the initial regression, we utilized the 2007-2011 dataset of loans from Lending club's statistics page (4). The raw data file contained 39,786 observations, each with 115 attributes. Because our dependent variable was a dummy variable indicating whether a loan has defaulted, as defined in the 'loan_status' attribute as 'Charged Off', we removed all loans that were either still active or late on payments but not yet charged off. Additionally, we removed any attributes that were assigned after the loan had been originated, such as interest and loan grade, because regressing these variables would be invalid, as they are variables that had been assigned by Lending Club through their analysis. Further, our purpose is to illustrate whether the underlying loan variables can be regressed in order to estimate for future loan defaults, so it was important to remove any variables that were created after the loan was originated. An example of this would be the attribution of late payments during the loan's lifespan. To continue, we removed any variables that were not either categorical or numeric, such as the description of the loan and the job title of the borrower. Lastly, we converted any alphanumeric variables, for instance converting the loan term from "36 Months" to simply 36 and made dummy variables out of categorical attributes. After the data munging, we had a finalized dataset with 39,318 observations and 28 attributes, including loan identifiers like 'loan id' and 'member_id'.

Regression Analysis and OLS Assumption Violation Correction

Through our data collection and munging, we trimmed the dataset to 28 total variables, of which 19 were capable of being regressed. Our initial regression showed that 6 were insignificant in estimating the success of a loan, leaving 13 remaining variables for our analysis.

Dependent Variable: IS_CHARGED_OFF
Method: Least Squares
Date: 12/14/16 Time: 23:47
Sample: 1 39318
Included observations: 39268

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.125843	0.023673	5.315782	0.0000
ANNUAL_INC	-1.07E-07	2.03E-08	-5.269867	0.0000
DELINQ_2YRS	-0.006659	0.002419	-2.752290	0.0059
DTI	0.000528	0.000200	2.634710	0.0084
HAS_HOME_OWNERSH	0.027004	0.023694	1.139714	0.2544
HAS_MORTGAGE	0.024513	0.023392	1.047913	0.2947
HAS_RENT	0.023642	0.023386	1.010970	0.3120
INQ_LAST_6MTHS	0.006532	0.001110	5.884134	0.0000
INSTALLMENT	0.000879	1.72E-05	51.13154	0.0000
IS_SOURCE_VERIFIED	0.013088	0.002971	4.404585	0.0000
IS_VERIFIED	0.016207	0.003036	5.339096	0.0000
LOAN_AMNT	3.31E-05	5.28E-07	62.69902	0.0000
OPEN_ACC	-0.000848	0.000381	-2.226718	0.0260
PUB_REC	0.020362	0.004962	4.103661	0.0000
REVOL_BAL	-4.83E-08	8.92E-08	-0.541147	0.5884
REVOL_UTIL	-0.000861	0.004873	-0.176785	0.8597
TOTAL_ACC	3.12E-05	0.000152	0.204900	0.8377
TOTAL_PYMNT	1.17E-05	1.87E-06	6.247860	0.0000
TOTAL_REC_INT	1.49E-05	2.08E-06	7.168475	0.0000
TOTAL_REC_PRNCP	-8.60E-05	1.74E-06	-49.33968	0.0000

R-squared	0.563191	Mean dependent var	0.143603
Adjusted R-squared	0.562980	S.D. dependent var	0.350691
S.E. of regression	0.231833	Akaike info criterion	-0.085088
Sum squared resid	2109.447	Schwarz criterion	-0.080719
Log likelihood	1690.622	Hannan-Quinn criter.	-0.083704
F-statistic	2663.350	Durbin-Watson stat	1.977735
Prob(F-statistic)	0.000000		

Initial OLS regression with all 19 variables.

With the model in place, we were able to build the hypothesis test as follows:

Ho: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=\beta_9=\beta_{10}=\beta_{11}=\beta_{12}=\beta_{13}=\beta_{14}=\beta_{15}=\beta_{16}=\beta_{17}=\beta_{18}=\beta_{19} = 0$

Ha: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=\beta_9=\beta_{10}=\beta_{11}=\beta_{12}=\beta_{13}=\beta_{14}=\beta_{15}=\beta_{16}=\beta_{17}=\beta_{18}=\beta_{19} \neq 0$

Through the Ordinary Least Squares regression, the model obtained an F-Statistic of 2663.35, which is much greater than the critical region of 1.5705 necessary to reject the null-hypothesis at a 0.05 level of significance, thus the overall model is significant.

Dependent Variable: IS_CHARGED_OFF
Method: Least Squares
Date: 12/14/16 Time: 23:55
Sample: 1 39318
Included observations: 39318

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.150453	0.003921	38.36950	0.0000
ANNUAL_INC	-1.09E-07	1.95E-08	-5.595698	0.0000
DELINQ_2YRS	-0.006483	0.002398	-2.703019	0.0069
DTI	0.000504	0.000188	2.679474	0.0074
INQ_LAST_6MTHS	0.006640	0.001102	6.027265	0.0000
INSTALLMENT	0.000878	1.68E-05	52.39019	0.0000
IS_SOURCE_VERIFI	0.013033	0.002965	4.395948	0.0000
IS_VERIFIED	0.016329	0.003035	5.379781	0.0000
LOAN_AMNT	3.32E-05	5.16E-07	64.20693	0.0000
OPEN_ACC	-0.000839	0.000287	-2.923287	0.0035
PUB_REC	0.020222	0.004936	4.096880	0.0000
TOTAL_PYMNT	1.18E-05	1.87E-06	6.291742	0.0000
TOTAL_REC_INT	1.48E-05	2.07E-06	7.135399	0.0000
TOTAL_REC_PRNCP	-8.61E-05	1.75E-06	-49.35859	0.0000

R-squared	0.562813	Mean dependent var	0.143827
Adjusted R-squared	0.562668	S.D. dependent var	0.350919
S.E. of regression	0.232066	Akaike info criterion	-0.083231
Sum squared resid	2116.709	Schwarz criterion	-0.080176
Log likelihood	1650.232	Hannan-Quinn criter.	-0.082263
F-statistic	3892.158	Durbin-Watson stat	1.978169
Prob(F-statistic)	0.000000		

OLS regression with the remaining 13 significant variables.

Although the overall model was significant with regards to the F-Test, and as earlier stated, we reduced the number of variables to only those which were significant on the basis of their individual probability. After removing the insignificant variables, we formed a new hypothesis test:

Ho: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=\beta_9=\beta_{10}=\beta_{11}=\beta_{12}=\beta_{13} = 0$

Ha: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=\beta_9=\beta_{10}=\beta_{11}=\beta_{12}=\beta_{13} \neq 0$

After which, we performed a Breusch-Pagan-Godfrey test for heteroskedasticity, in order to understand the underlying error variance of the model.

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	605.4034	Prob. F(13,39304)	0.0000
Obs*R-squared	6559.560	Prob. Chi-Square(13)	0.0000
Scaled explained SS	21255.34	Prob. Chi-Square(13)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 12/15/16 Time: 00:12
Sample: 1 39318
Included observations: 39318

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.072130	0.002115	34.10606	0.0000
ANNUAL_INC	-9.70E-09	1.05E-08	-0.921142	0.3570
DELINQ_2YRS	0.001709	0.001294	1.320968	0.1865
DTI	0.000192	0.000101	1.894612	0.0582
INQ_LAST_6MTHS	0.002811	0.000594	4.731468	0.0000
INSTALLMENT	3.82E-05	9.04E-06	4.231341	0.0000
IS_SOURCE_VERIFI	0.000740	0.001599	0.462985	0.6434
IS_VERIFIED	0.001906	0.001637	1.164185	0.2444
LOAN_AMNT	8.82E-06	2.79E-07	31.65678	0.0000
OPEN_ACC	-0.000773	0.000155	-4.990701	0.0000
PUB_REC	0.012765	0.002662	4.795065	0.0000
TOTAL_PYMNT	1.19E-05	1.01E-06	11.74681	0.0000
TOTAL_REC_INT	-7.40E-06	1.12E-06	-6.635297	0.0000
TOTAL_REC_PRNCP	-2.61E-05	9.41E-07	-27.72089	0.0000

R-squared	0.166834	Mean dependent var	0.053836
Adjusted R-squared	0.166558	S.D. dependent var	0.137101
S.E. of regression	0.125164	Akaike info criterion	-1.318028
Sum squared resid	615.7372	Schwarz criterion	-1.314973
Log likelihood	25925.12	Hannan-Quinn criter.	-1.317061
F-statistic	605.4034	Durbin-Watson stat	1.979267
Prob(F-statistic)	0.000000		

Breusch-Pagan-Godfrey Heteroskedasticity Test

The Breusch-Pagan-Godfrey Test for Heteroskedasticity showed that the model suffered from an inconsistent error variance. The cause for this inconsistency most likely arises from the variance in datasets origination periods from 2007 to 2011, wherein Lending Club's loan acceptance criteria would have changed. In order to correct for this, we adjusted our model's estimation method by running the regression with a Huber-White covariance method, often referred to as a white correction.

Dependent Variable: IS_CHARGED_OFF
Method: Least Squares
Date: 12/14/16 Time: 23:55
Sample: 1 39318
Included observations: 39318
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.150453	0.004236	35.51568	0.0000
ANNUAL_INC	-1.09E-07	2.78E-08	-3.927636	0.0001
DELINQ_2YRS	-0.006483	0.002615	-2.479265	0.0132
DTI	0.000504	0.000195	2.589195	0.0096
INQ_LAST_6MTHS	0.006640	0.001157	5.739465	0.0000
INSTALLMENT	0.000878	2.08E-05	42.22494	0.0000
IS_SOURCE_VERIFI	0.013033	0.002978	4.376681	0.0000
IS_VERIFIED	0.016329	0.002929	5.574978	0.0000
LOAN_AMNT	3.32E-05	7.57E-07	43.82131	0.0000
OPEN_ACC	-0.000839	0.000300	-2.801724	0.0051
PUB_REC	0.020222	0.005855	3.453708	0.0006
TOTAL_PYMNT	1.18E-05	4.95E-06	2.377889	0.0174
TOTAL_REC_INT	1.48E-05	5.11E-06	2.888233	0.0039
TOTAL_REC_PRNCP	-8.61E-05	4.68E-06	-18.42076	0.0000

R-squared	0.562813	Mean dependent var	0.143827
Adjusted R-squared	0.562668	S.D. dependent var	0.350919
S.E. of regression	0.232066	Akaike info criterion	-0.083231
Sum squared resid	2116.709	Schwarz criterion	-0.080176
Log likelihood	1650.232	Hannan-Quinn criter.	-0.082263
F-statistic	3892.158	Durbin-Watson stat	1.978169
Prob(F-statistic)	0.000000	Wald F-statistic	1132.023
Prob(Wald F-statistic)	0.000000		

White Correction (white heteroskedasticity-consistent standard errors & covariance)

Through our analysis we successfully tested for, and corrected, the model's heteroskedasticity. Of course, this is not the only possible violation of OLS assumptions, thus we tested for the possibility of autocorrelation (serial correlation), through a Durbin-Watson test. As shown above, the Durbin-Watson statistic for the model was 1.978169. For this test, the null-hypothesis (Ho) was that the statistic is equal 2, indicating no autocorrelation, and that the alternative-hypothesis (Ha) was that the statistic is not equal to 2, thus indicating either a positive correlation (closer to 0) or a negative correlation (closer to 4). Utilizing a Durbin-Watson table (5), we were able to test whether the model's statistic illustrates a positive, negative, or non-correlated Durbin Watson statistic. The lower (dL) and upper (dU) bounds for a model with 13 regressors and over 200 observations, at the .05 significance level, were 1.632 and 1.908 respectively. Thus, our model's Durbin-Watson statistic of 1.978169 was outside the bounds, closer to 2, and was found to be possess no autocorrelation.

Model Interpretation and Prediction

Model Results Interpretation

As previously stated, the overall model was significant, with an F-Statistic of 3892.158 (after reducing the number of variables), which indicates the accuracy of the model with regards to its ability to predict a loan's probability of defaulting. To begin, the constant was .150453, which describes that the initial probability of any given loan defaulting over its lifespan as 15.0453%. This is consistent with the mean of the dataset, which describes a default rate of 14.3827%. The following 13 independent variables, shown below in the form of confidence intervals, each add or subtract from the probability of a loan defaulting, thus allowing us to accurately depict the outlook of any given loan based upon its underlying attributes, or loan application.

Coefficient Confidence Intervals
Date: 12/15/16 Time: 01:31
Sample: 1 39318
Included observations: 39318

Variable	Coefficient	95% CI Low	95% CI High
C	0.150453	0.142150	0.158756
ANNUAL_INC	-1.09E-07	-1.64E-07	-5.47E-08
DELINQ_2YRS	-0.006483	-0.011608	-0.001358
DTI	0.000504	0.000122	0.000885
INQ_LAST_6MTHS	0.006640	0.004373	0.008908
INSTALLMENT	0.000878	0.000837	0.000918
IS_SOURCE_VERIFI	0.013033	0.007196	0.018870
IS_VERIFIED	0.016329	0.010588	0.022071
LOAN_AMNT	3.32E-05	3.17E-05	3.46E-05
OPEN_ACC	-0.000839	-0.001427	-0.000252
PUB_REC	0.020222	0.008746	0.031698
TOTAL_PYMNT	1.18E-05	2.07E-06	2.15E-05
TOTAL_REC_INT	1.48E-05	4.74E-06	2.48E-05
TOTAL_REC_PRNCP	-8.61E-05	-9.53E-05	-7.70E-05

95% Confidence Intervals of Final Regression

Prediction Model Setup

After performing the regression analysis, we needed to develop a means of predicting the probability of future loans defaulting, as well as a method for testing the accuracy of these predictions. To begin, we loaded and munged the Lending Club dataset for loans from 2012-2013, in order to make the data match our 2007-2011 dataset in terms of variables and format. After processing through the data to remove any variables that were not used in the regression, as well as removing any loans that were still active, there were 145,675 observations to apply our estimations to. We utilized the 2012-2013 dataset because it was important to apply our estimations to an out of sample dataset, as simply applying the estimations to the same dataset as the regression would yield extremely biased results. Therefore, we applied the 13 estimations to the underlying parameters of each loan in the 2012-2013 dataset, thus giving us the sum of the estimated probability of each loan defaulting.

Prediction Model Results

As previously described through our Excel based model, we were able to calculate the estimated probability each of loan defaulting. In order to make predictions, we compared the estimated probability of each loan defaulting to the mean probability of default from the 2007-2011, in order to use the probability as a binary indicator with the following logic (1 indicates the loan is estimated to default, 0 indicates that the loan is estimated to be fully paid):

IF Estimated_Default_Probability > Mean_Default_Probability
THEN Estimate_Will_Default = 1
ELSE
Estimate_Will_Default = 0

After making the prediction as either that the loan will default (1) or that it will be paid off (0), we then compared the results to the actual historical value of the attribute, thus testing the accuracy of our prediction. Utilizing this methodology, we were able to make predictions on a per loan basis of whether or not we expect it to default, as shown below.

Variable	Loan_Application_Data	Regression_Coefficient	Estimation	**Sum_of_Estimations**
id	10224583			**0.075879984**
c	1	0.150453	0.150453	estimate_is_charged_off
loan_amnt	11100	0.0000332	0.36852	0
installment	384.68	0.000878	0.33774904	actual_is_charged_off
annual_inc	90000	-0.000000109	-0.00981	0
is_source_verified	0	0.013033	0	estimate_accuracy
is_verified	0	0.016329	0	1
dti	3.73	0.000504	0.00187992	
delinq_2yrs	1	-0.006483	-0.006483	
inq_last_6mths	0	0.00664	0	
open_acc	9	-0.000839	-0.007551	
pub_rec	0	0.020222	0	
total_pymnt	13575.64001	0.0000118	0.160192552	
total_rec_prncp	11100	-0.0000861	-0.95571	
total_rec_int	2475.64	0.0000148	0.036639472	

In the case above, loan 10224583 has an estimated default probability of 0.075879984, which is not greater than the mean of the regressed 2007-2011 dataset of 0.143827, thus our model predicts that the loan will be fully paid off (estimate_is_charged_off = 0). Comparing our prediction to the actual loan's performance (actual_is_charged_off = 0), we can conclude that our model is correct with regards to this single loan.

Finally, in order to test the overall accuracy of the model, we ran the above prediction and accuracy test on all 145,675. From this, we were able to calculate the mean accuracy of the model, as well as an overall estimate for the populations default rate.

Model Statistic	Calculated Value
Estimated_Mean_Accuracy	0.919251759
2007_to_2011_Actual_Mean_Charged_Off	0.143827
2012_Estimated_Mean_Charged_Off	0.25444311
2012_Actual_Mean_Charged_Off	0.175424747

As shown above, our model was able to accurately predict whether a loan will default with a success rate of 91.9252% and a variance from the true mean of 7.9018% (0.25444311-0.175424747). Through this analysis, we can conclude that the probability of a loan defaulting can be accurately estimated to certain degrees of certainty. However, more prediction would be necessary to have an implementable model that could be applied to today's current loan listings.

Summary

Through multiple regression analysis we attempted to estimate the probability of a peer-to-peer loan's default from its application. Initially, we performed an ordinary least squares (OLS) linear regression, which we found to suffer from heteroskedasticity, leading us to adjust our model through a White Correction. This method, coupled with an Excel based prediction model, allowed us to regress loan applications from Lending Club's 2007 to 2011 dataset, and then use the resulting estimations to predict the probability of future loans defaulting. To test the accuracy of our predictions, we compared the actual values of the of the 2012 to 2013 loans with what we estimated them to be (either paid off or defaulted). Applying this methodology, we were able to achieve a success rate of 91.9252%, thus allowing us to identify and avoid loans that show a high probability of defaulting in the future.

Works Cited

1. PWC article. “Peer pressure: How peer-to-peer lending platforms are transforming the consumer lending industry”
2. Bradley, Christine; Burhouse, Susan; Gratton, Heather; Miller, Rae-Ann (2009). "Alternative Financial Services: A Primer". *FDIC Quarterly*. **3** (Q1). Federal Deposit Insurance Corporation. Retrieved July 30, 2012
3. Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech & Min Lu. *Applied Economics*. “Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending”
4. Lending Club Data, lendingclub.com/info/download-data.action
5. Durbin Watson Tables, http://www.dm.unibo.it/~simoncin/Durbin_Watson_tables.pdf