DATA MINING TECHNIQUES TO PREDICT DEFAULT IN LENDING CLUB

Haolun Xu

Shanghai Foreign Language School

*Corresponding Author's Emails: 982301244@qq.com

This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 12 March 2018

Accepted 12 April 2018

Available online 15 May 2018

ABSTRACT

This study aims to build a predictive model for default in Lending Club using artificial neural network and compare its performance to logistic regression model using the data downloaded from <https://www.kaggle.com/wendykan/lending-club-loan-data/data>. All the records who were eligible were randomly assigned into 2 groups: training sample and testing sample. Two models were built using training sample: artificial neural network and logistic regression. We used these two models to predict the risk of default in Lending Club in the testing sample. Receiver operating characteristic (ROC) were calculated and compared for these two models for their discrimination capability and a curve using predicted probability versus observed probability were plotted to demonstrate the calibration measure for these two models.

KEYWORDS

artificial neural network, logistic regression and ROC curve.

1. INTRODUCTION

Based on a study, peer-to-peer lending, also abbreviated as P2P lending, is the practice of lending money to individuals or businesses through online services that match lenders with borrowers. Lending Club is the world's largest peer-to-peer lending platform [1]. The company claims that \$15.98 billion in loans had been originated through its platform up to December 31, 2015. ⁱ Lending Club enables borrowers to create unsecured personal loans between \$1,000 and \$40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee [2].

To reduce default risk, Lending Club focuses on high-credit-worthy borrowers, declining approximately 90% of the loan applications it received as of 2012 and assigning higher interest rates to riskier borrowers within its credit criteria.ⁱⁱ Only borrowers with FICO score of 660 or higher can be approved for loans. Since inception, Lending Club's default rate ranges from 1.4% for top-rated three-year loans to 9.8% for the riskiest loans in year 2012. ⁱⁱⁱ Peer-to-peer lending's future is depending on the successful management of default rate in company like in Lending Club [3].

This study aims to 1) examine the predictors of Default in Lending Club 2) build a predictive model for Default in Lending Club using artificial neural network and compare its performance to logistic regression model.

2. RESULTS

Table 1: Logistic Regression for Default in Lending Club

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-6.015	0.447	-13.454	< 2e-16	***
loan_amnt	0.000	0.000	7.578	0.000	***
int_rate	0.038	0.015	2.495	0.013	*
installment	-0.002	0.000	-6.134	0.000	***
annual_inc	0.000	0.000	-2.011	0.044	*
dti	0.046	0.004	10.656	< 2e-16	***
delinq_2yrs	0.148	0.026	5.615	0.000	***
inq_last_6mths	-0.060	0.030	-1.998	0.046	*
open_acc	0.048	0.008	5.976	0.000	***
pub_rec	0.253	0.050	5.060	0.000	***
revol_bal	0.000	0.000	-1.059	0.290	
revol_util	0.000	0.001	0.296	0.768	
total_acc	-0.021	0.004	-5.606	0.000	***
Grade_AB	-1.254	0.318	-3.947	0.000	***
Grade_CD	-0.398	0.265	-1.502	0.133	
Grade_EF	-0.073	0.238	-0.308	0.758	
emp7y	-0.032	0.080	-0.399	0.690	
emp10y	0.098	0.071	1.369	0.171	

owned	-0.404	0.064	-6.339	0.000	***
verified	0.133	0.076	1.738	0.082	.
debt_conso	0.121	0.086	1.397	0.162	
credit_c	0.140	0.106	1.320	0.187	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 1: Artificial Neural Network in training sample

In above plot, line thickness represents weight magnitude and line color weight sign (black = positive, grey = negative). The net is essentially a black box so we cannot say that much about the fitting, the weights and

the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used. The loan grade and ownership of the home were also significant predictors.

Figure 2: Variable Importance in Artificial Neural Network

According to this neural network, the top 5 most important predictors were annual income, total credit revolving balance, number of derogatory public records, revolving line utilization rate as well as loan grade. All five together accounted for more than 60% the weights in the model.

For training sample, the ROC was 0.73 for the Logistic regression and 0.65 for the artificial neural network. Logistic regression performed better clearly. In testing sample, the ROC was 0.75 for the Logistic regression and 0.66 for the artificial neural network.

Figure 3: ROC in training sample for Logistic Regression (Red) vs Neural Network (Blue)

REFERENCES

- [1] Lending Club Statistics. Lending Club. <https://www.lendingclub.com/info/statistics.action>
- [2] Lending Club Statistics. Lending Club. 2012. <https://www.lendingclub.com/info/demand-and-credit-profile.action>
- [3] Barth, C. 2012. Looking for 10% yields? Go online for peer to peer lending. Forbes. <https://www.forbes.com/sites/chrisbarth/2012/06/06/looking-for-10-yields-go-online-for-peer-to-peer-lending/#544de9493a8f>