Loan-level Credit Risk Assessment for P2P Lending

By Yuqing (Elisa) Yao

I. Goals of Analysis

This report aims to illustrate types of loans that tend to have "bad" performance ("Charged Off", "Default", or "Does not meet the credit policy - Charged Off") and to create models to predict individual loan's performance ("good" or "bad") in order to help investors make investment decisions.

II. Recommendation

Based on descriptive statistics and statistical significance tests, the research suggests that:

- *Higher interest rate reflected risk premium for exposing to higher credit risk which is consistent to the common sense. Individual investors should be aware of the risk-return relationship and choose reasonable return rate that corresponds to their risk capacity.*

- *Although shorter term loans (36 months) may be thought to have higher liquidity risk, it appears to have higher credit risk.*

- *Loans that finance for small businesses, car purchase, credit card payment, debt consolidation, and house purchase have higher risk than other purposes.*

- *Beware of states that tend to have a higher bad loan rates, such as AL, AR, FL, IA, IN, KY, MO, MS, NV, TN.*

- *Choosing borrowers that has higher annual income substantially reduces risk exposure to defaults.*

- *Investors should pay attention to signals that shows the level of eagerness borrow, e.g. inquiries within last 6 months, and number of delinquencies within last 2 years.*

III. Expected Effect of Actions

If the investors make decisions according to the recommendations above, they are expected to:

- *Reduce their unwanted risk exposure*

- *Be able to more accurately construct loan portfolios with desired risk-return profile*

- *More efficiently and scientifically assess a borrower's default probability*

Besides, LendingClub can refer to this analysis to:

- *Adjust interest rates of individual loans to reflect default probability*

- *Assess overall credit risk exposure of the platform*

- *Enhance its current credit rating system*

IV. Conclusions from Data Analysis

The following conclusions can be drawn from data analysis. Please see technical discussion in **Appendix**.

- *Bootstrap Forest has the highest predictive power among models used.*

ROC curve, Lift curve, and Confusion Matrix indicates that the models have satisfactory accuracy. Bootstrap Forest has the highest R-Square among the models.

- Interest rate has the highest indication of defaulting ("Bad" performance).

All of the models used suggest that interest rate has the highest explanatory power among the predictors used. The higher the interest rate, the more likely the loan end up being "Bad".

- Shorter-term (36 months) tend to have a higher rate of bad performance than longer-term (60 months) loans

In all of the prediction models, term ranks very high level among predictors and is significant at 99.9% confidence level.

- Annual income is a significant indicator for loan performance

Annual income usually has large contribution to explanatory power in the models that this analysis used.

- Loan purpose that are small businesses, car purchase, credit card payment, debt consolidation, and house purchase tend to have worse performance

These categories are 99.9% significant in the logistic regression.

- Certain states tend to have higher default rates. This may be correlated to economic status of the states.

By plotting the heat map, we can intuitively find out that AL, AR, FL, IA, IN, KY, MO, MS, NV, TN have more than 20% "bad" rate. However, in the Chi-Square test, the difference of patterns is not statistically significant.

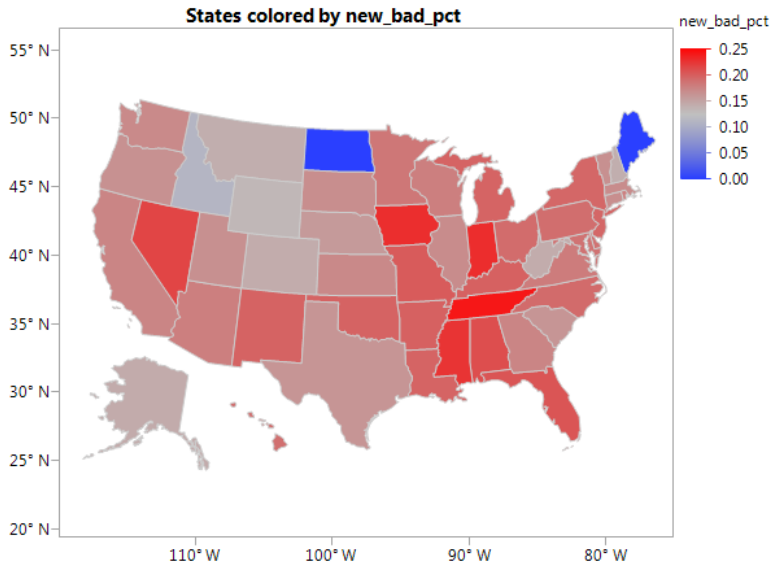

Figure 1. Heatmap of "bad" loan percentage

V. Methodologies

My analysis is based on the fact that good loans and bad loans are qualitative assessment. Popular models that deal with categorical response variable are Logistic Regression and Classification Tree. Therefore, I used both models to conduct analysis. Model set-up and results are listed follows.

new_loan_status = f(loan_amnt, term, int_rate, installment, verification_status, purpose, dti, delinq_2yrs, inq_last_6mths, mths_since_last_delinq, open_acc, pub_rec, revol_util, total_acc, new_home_ownership, new_annual_inc, new_desc_length3, new_cr_sr, new_emp_length)

1. Logistic Regression

Table 1. Parameters estimation and corresponding significance

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	3.74843563	0.2492423	226.18	<.0001*
loan_amnt	7.92115e-6	7.013e-6	1.28	0.2587
term[36 months]	-0.229923	0.0221662	107.59	<.0001*
int_rate	0.08035571	0.0029019	766.80	<.0001*
installment	0.00047474	0.0002163	4.82	0.0282*
verification_status[Not Verified]	-0.0437696	0.0131188	11.13	0.0008*
verification_status[Source Verified]	0.04074365	0.0119552	11.61	0.0007*
purpose[car]	-0.2763171	0.0823808	11.25	0.0008*
purpose[credit_card]	-0.223115	0.0345416	41.72	<.0001*
purpose[debt_consolidation]	-0.1062622	0.0300939	12.47	0.0004*

2. Decision Tree

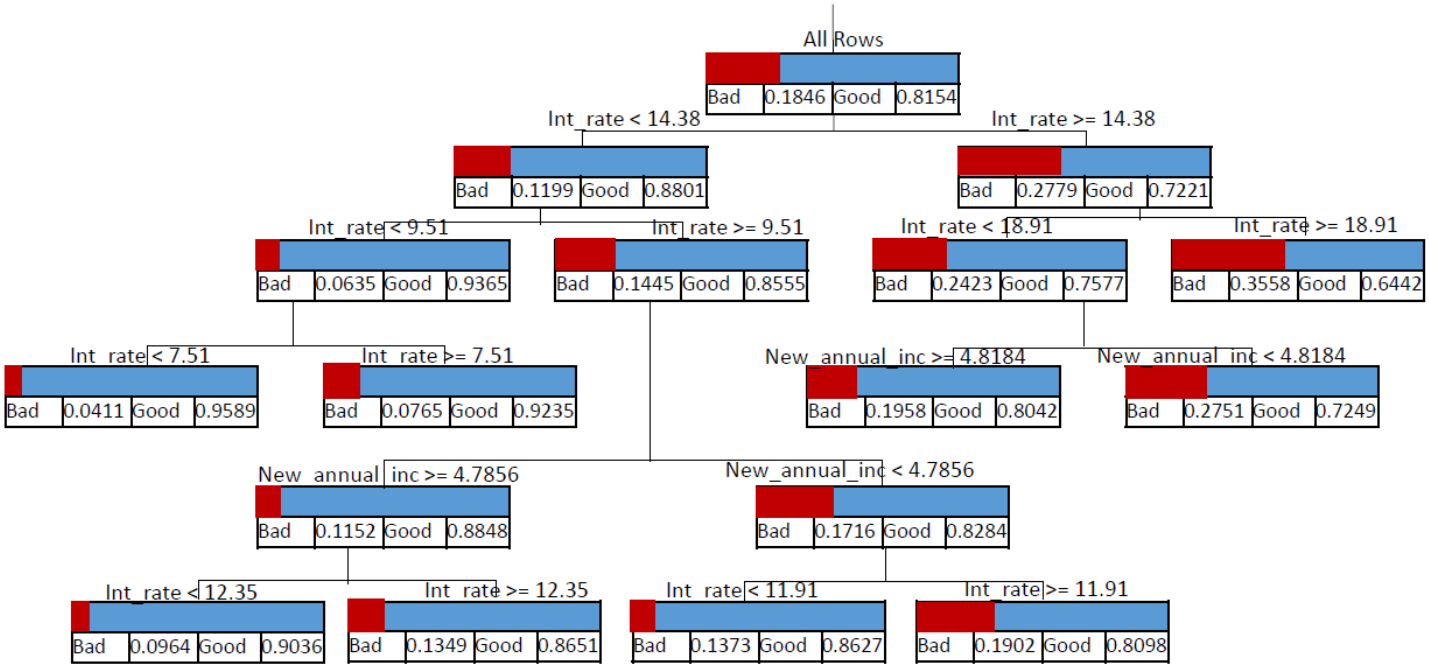

Figure 2. Top portion of Decision Tree

VI. Research Outlook

Further efforts could be made on continuing the research, such as considering structural change, time dependence, and applying ensemble learning. Extended topic using the same dataset could be constructing P2P loan portfolios, Copula of individual loans, and effects of economic factors on loan performance.

Reference

Tsai, Ramiah, Singh (2014). Peer Lending Risk Predictor. Stanford University. Retrieved from: http://cs229.stanford.edu/proj2014/Kevin%20Tsai,Sivagami%20Ramiah,Sudhanshu%20Singh,Peer%20Lending%20Risk%20Predictor.pdf

Davenport, K. (2013). Gradient Boosting: Analysis of LendingClub's Data. Retrieved from: https://www.r-bloggers.com/gradient-boosting-analysis-of-lendingclubs-data/

Walczak, E. (2016). Initial loan book analysis. Kaggle. Retrieved from: https://www.kaggle.com/erykwalczak/d/wendykan/lending-club-loan-data/initial-loan-book-analysis/comments

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLoS One*, 10(10) doi:http://dx.doi.org/10.1371/journal.pone.0139427

Milad Malekipirbazari, Vural Aksakalli. Risk assessment in social lending via random forests, *Expert Systems with Applications*, Volume 42, Issue 10, 15 June 2015, Pages 4621-4631, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2015.02.001.

Guo, Y., Zhou, W. et al. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*. 249: 417–426