Prediction of Interest rate - Based on personal credit history

Xinmin Wang [1] ; Yang Liu [2] ; Jonathan Jonker [3]

[1]Department of Statistics, University of Washington;
[2]Department of Chemistry, University of Washington;
[3]Department of Mathematics, University of Washington

Overview

- **Online Credit market has an appealing business model that attracts more users**
- **We can Predict of Interest rate in this market by basic credit & financial information**
- **More than 40,000 cases from Lending Club Company is included in our research**
- **MacroEconomics background are taken into consideration**
- **High leverage points are diagnosed and analyzed**
- **Model is validated by cross validation**

Introduction

- **Background**

Online Credit market has growing business in the past decades. It offers broad range of loan amount and interest rate that allows borrowers to have various choice in the loan market.
Lending Club is one of the largest business in this field that can act as a representative of the market. The study of its historical data can potentially benefit individual borrowers and new business.

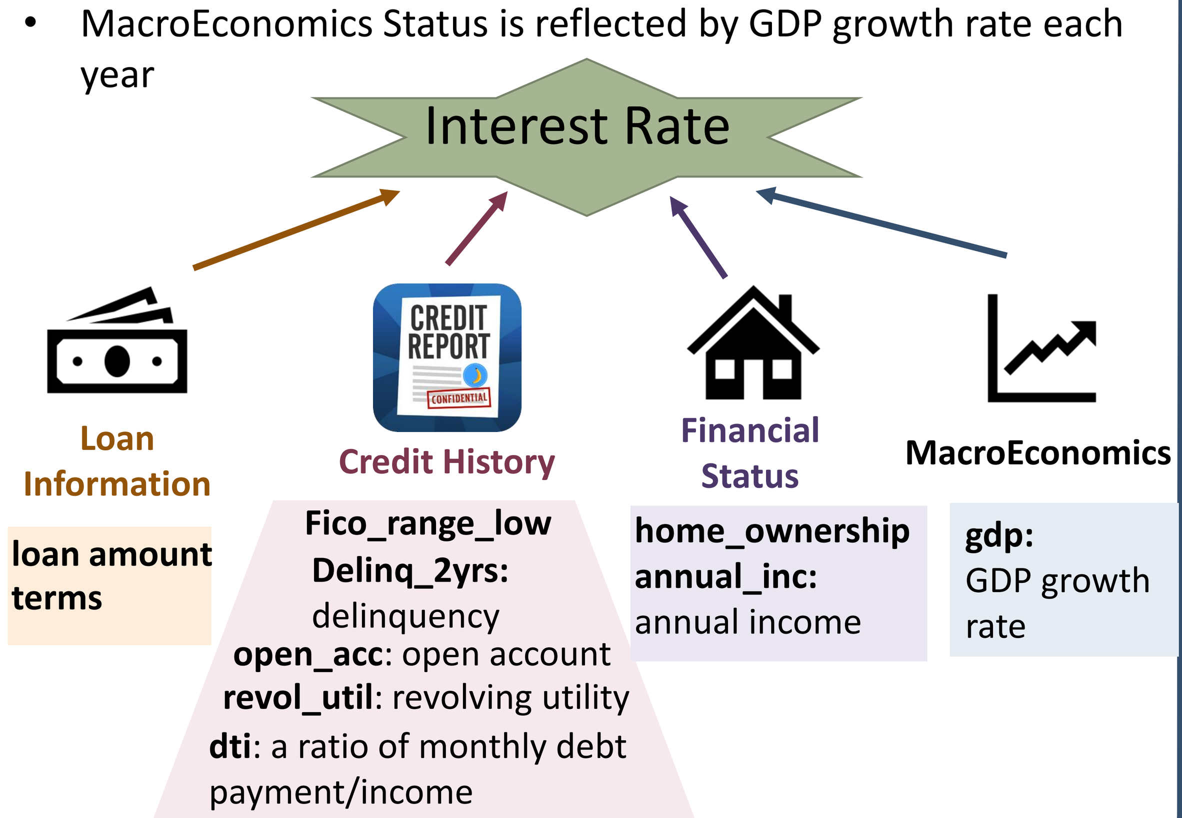
- **Our Research Question**

Given a person's **credit history(such as *Fico Score*), some financial information** and **loan amount**, considering macro econnomics status can we predict the **interest rate** he/ she will get when apply loans from *Lending Club Company*?

Data and Methods

Data Description & Pre-cleaning

- Borrower's Data is collected by *Lending Club* company from 42538 borrowers during the year 2007-2011
- MacroEconomics Status is reflected by GDP growth rate each year



Pre-cleaning included the deletion of highly correlated covariates and non-informative variables

Data and Methods

Methods

Based on our dataset, we are going to apply the following methods to conduct an appropriate Prediction.

- **Multiple Linear Regression**

To study the relationship between interest rate and one's credit history, we applied multivariate linear regression model.

$$Y_{(\text{interest rate})} = X\beta + e$$

- **Stepwise Variable Selection**

After fitting initial model, we use stepwise algorithm to select variables for our model based on AIC (Akaike Information Criterion). This method will make sure our model fitting observations sufficiently well in the least complex way.

- **Multivariate Box-Cox Transformation**

To transform our predictors so that all bivariate scatterplots have a linear mean function, we used Box-Cox method. We transformed our covariates and response based on the power transformation hypothesis results. The estimation process for λ is based on maximizing a likelihood of the data when errors are assumed to be Normally distributed

- **Model Diagnostics**

-Assumption Check
-Influential and leverage points detection
-Outliers Detection

- **Model Validation**

10-fold cross-validation is used by us as a model validation technique for assessing how the results of our statistical analysis will generalize to an independent data set. We partitioned our data in 10 equal sized subsamples. A single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data.

Data Analysis Results

Initial Analysis

On the partial correlation heatmap, we can see positive correlation between interest rate and loan amount, dti, delinqency, revolving utility and negative correlation with fico score.

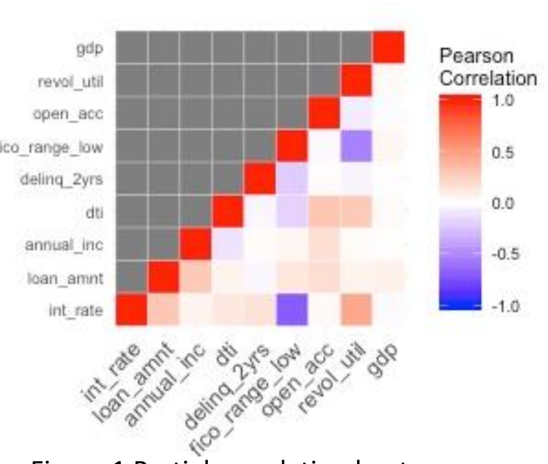


Figure 1 Partial correlation heatmap

Variable selection

-Initial linear model on all predictors showed variables that are not significant.
-Further select variables by stepwise methods to reduce complexity.
-Update full data (- annual income)

AIC=50772.9

loan amount
terms
fico_range_low
delinq_2yrs
open_acc
revol_util
dti
home_ownership
gdp
~~annual_inc~~

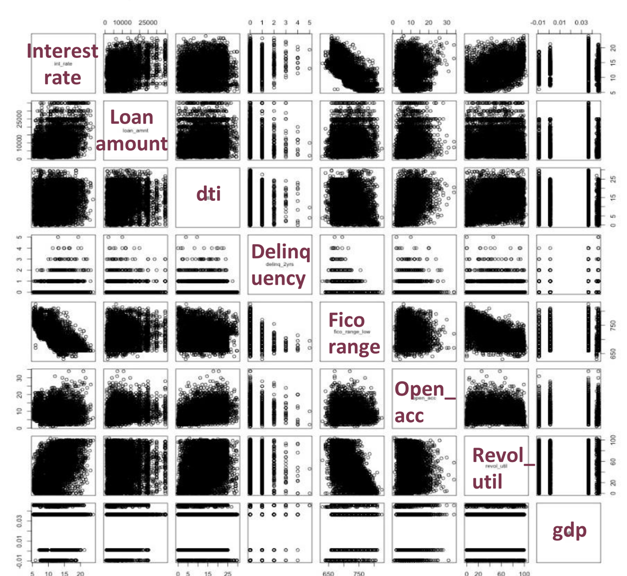


Figure 2 scatterplot matrix on numerical variables

Data Analysis Results

Box-Cox Transformation

Multivariate Power Transform

We need to transform the predictors so that all bivariate scatterplots have a linear mean function (approximately). The combination of **λ** in the table was tested and within 95% confidence interval.

	loan_amnt	dti+1	delinq_2yrs+0.1	fico_range_low	open_acc	revol_util+1	gdp+0.0093
λ	0	1	-3	-2	0	1	1

Table 1 Transformation Summary

Transformation of Response Variables

Noticing non homogeneous variance, we did transformation to response varaible. We choose the best value of λ = 0.6. In this way, the errors looks more normal.

Linear Model after Transformation

tr. variables	Estimate	Std. Error	t value	Pr(>|t|)
Log(loan_amnt)	0.2801	2.95E-03	95.10	0.000
Dti+1	-0.0038	3.26E-04	-11.65	0.000
$(\text{delinq\_2yrs}+0.1)^{-3}$	0.0000	1.96E-05	-1.84	0.065
$(\text{fico\_range})^{-2}$	-6267576	2.58E+04	-242.93	0.000
Log(open_acc)	-0.1609	4.42E-03	-36.44	0.000
revol_util+1	0.0010	8.79E-05	11.12	0.000
gdp+0.0093	-5.1140	1.21E-01	-42.18	0.000
terms 60 months	0.7047	4.85E-03	145.15	0.000
home_ownershipOWN	0.0450	7.69E-03	5.86	0.000
home_ownershipRENT	-0.0016	4.28E-03	-0.37	0.714

Observations	Residual St.Err	R^2	Adjusted R^2
42307	1.817	0.7649	0.7649

Table 2 Regression Summary

Delinquency is not significantly related to interest rate after controlling for other variables. We then carry out regression analysis without delinquency, the coefficient of other variables does not change much in the new model, and still remain highly significant at 0.001 level.

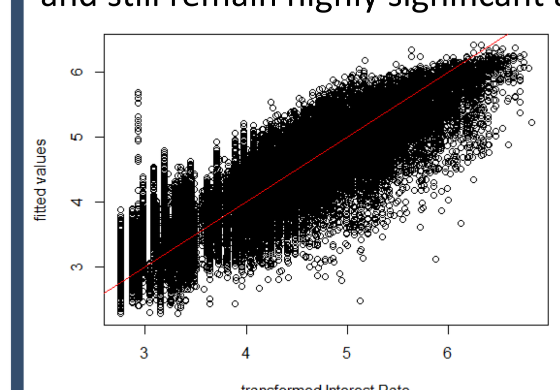


Figure 3 fitted value vs. transformed interest rate

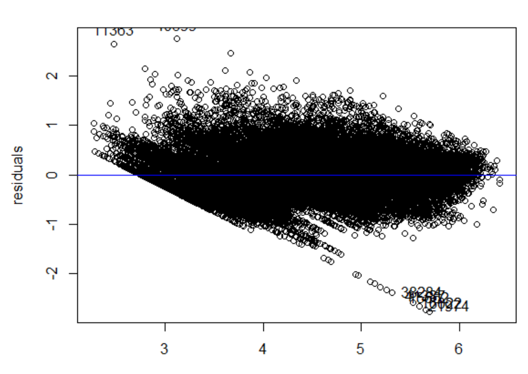


Figure 4 Residual Plot vs. fitted value (transformed interest rate)

Variance of transformed model look more normal than nontransformed model (not shown here). In order to check whether our model fits the data well, we carry out regression diagnostics.

Regression Diagnostics

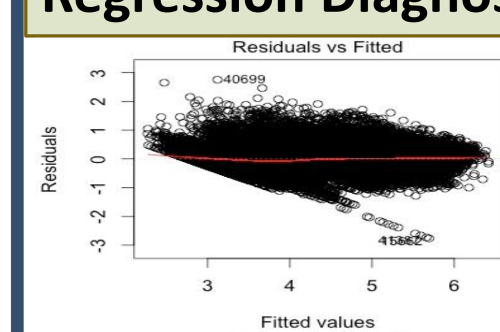


Figure 5 residual plot

From the residual plot on the left, we can see that the residuals are even distributed around 0, hence the **homoscedasticity assumption** is well satisfied.

From the scale-location plot on the right, we can get the same conclusion since there is not any pattern shown on the plot between standardized residuals and fitted values. The residuals appear randomly spread.

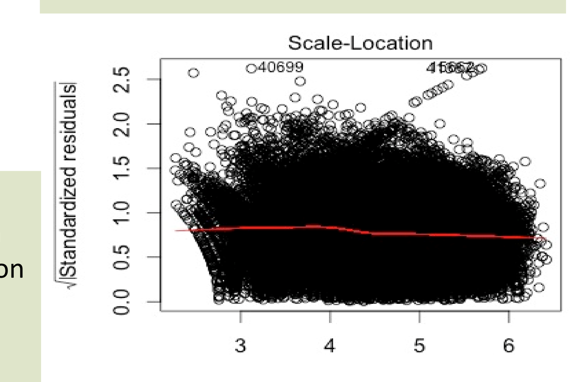


Figure 6 scale location plot

Data Analysis Results

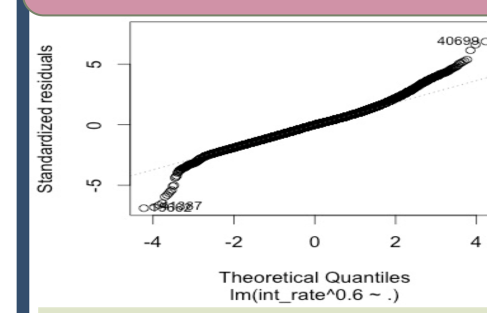


Figure 7(left) QQ-plot

From the normal Q-Q plot, we can see there is only a slight deviation from the normality assumption, which does not matter since our main dataset basically satisfied **normality assumption** very well

Then we analyzed influential point using cook's distance. We can see the largest value is 0.0025, which is << 0.5. Hence there is no influential point in our model.

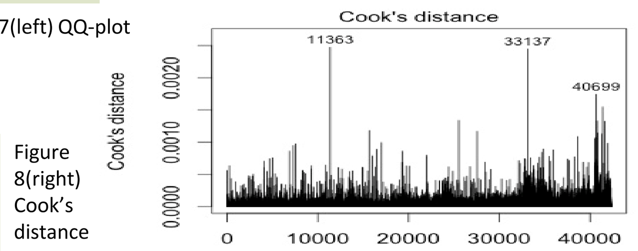


Figure 8(right) Cook's distance

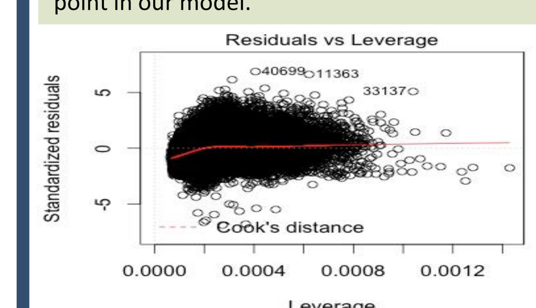


Figure 9 Residuals vs. Leverage

Finally, we look at residuals VS leverage plot.
About 2315 out of 42307 datapoints have leverages larger than 2p/n=0.0005; 282 points has leverages larger than 3p/n =0.0007. We examined those high leverage cases and compared to the whole sample.
Removing the high leverage points(>3p/n) did not cause the model to change much.

	median loan amnt	home ownership(OWN%)	dti median	fico_low median	open_acc median	revol_util median
leverage > 0.0007	3000	46	8.3	680	3	10.1
Whole dataset	9800	7	13.49	710	9	49.7

Table 2 high leverage points summary

10-fold Cross Validation

To check if our prediction model has overfitting problems, we used 10-fold cross validation method to evaluate predictive performance. We split the whole data into 10 folds and examined the overall RMSE (root mean squared error). We also listed the RMSE of other models.
Hence our model will perform accurately in practice.

Model	mod0	mod1	mod2	mod3
RMSE	0.40114	0.40115	0.40131	4.9277

Mod 1 is our chosen model!!

model	Transformation on response	Transformation on covariates	# of covariates	Deleted variables (compare to full data)
Mod 0	Y	Y	9	-
Mod 1	Y	Y	8	delinquency
Mod 2	Y	Y	7	delinquency, home_ownership
Mod 3	N	N	9	-

Table 3 RMSE of different model

Without transformation (mod 3), the RMSE is much larger; After transformation, with delinquency does not lower the RMSE significantly (compare mod 0 vs. 1); However, mod2 deleting home_ownership (with largest p-value) causes the RMSE to increase, lowers the prediction accuracy.

Discussions

- **Outliers**

Outliers are identified in residual plot that does not fit in our model:
High revolving utility but resulted in low interest rate;
Low revolving utility but resulted win high interest rate.

- **Sharp cut in residual plot**

We've noticed a sharp cut at the lower end of fitted values. Those are caused by the minimum interest rate 5.42% offered by the company. Also, interest rates are not strictly continuous so that have a step-wised pattern.

- **How to get a lower rate?**

According to our model, controlling for the Macroenvironment, credit history information and loan information, personal financial status such as annual income and home ownership is not significant.
Having higher dti, fico score, more open account and lower revolving utility can possibly end up with lower interest rate.

Conclusions

- $(\text{Interest Rate})^{0.6} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\log(\text{loan\_amnt}) + \boldsymbol{\beta_2}(\text{dti}+1) + \boldsymbol{\beta_3}(\text{fico\_range})^{-2} + \boldsymbol{\beta_4}\log(\text{open\_acc}) + \boldsymbol{\beta_5}(\text{revol\_util}+1) + \boldsymbol{\beta_6}(\text{gdp}+0.0093) + \boldsymbol{\beta_7}\,\text{terms\_60\_months} + \boldsymbol{\beta_8}\,\text{OWN} + \boldsymbol{\beta_9}\,\text{RENT} + \mathbf{e}$

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
3126405	0.2800	-0.0038	-6252829	-0.1608	0.0010	-5.118	0.7046	0.0452	-0.0011

Observations	Residual St.Err	R^2	Adjusted R^2
42307	0.401	0.765	0.765

Reference

https://www.lendingclub.com/info/download-data.action
http://www.multpl.com/us-gdp-growth-rate/table/by-year