Josh Eysenbach

Sabrina Purvis

Paritosh Rai

STAT 6372-402

Spring 2020

# LendingClub Loan Analysis

## Introduction

Anyone who has made a major purchase has likely had to consider borrowing funds from a financial institution. Whether purchasing a vehicle, starting a small business or going back to school, those decisions are at least partially made based on finances. Not only does the borrower need to simply receive approval for the loan, they also are assigned an interest rate against the amount requested. The borrowed amount, term of the loan and interest rate together drive payment terms that materially impact borrower decisions. Lending institutions factor in many variables when deciding with whom they do business. We are seeking to understand what factor or factors determine interest rates.

## Data Description

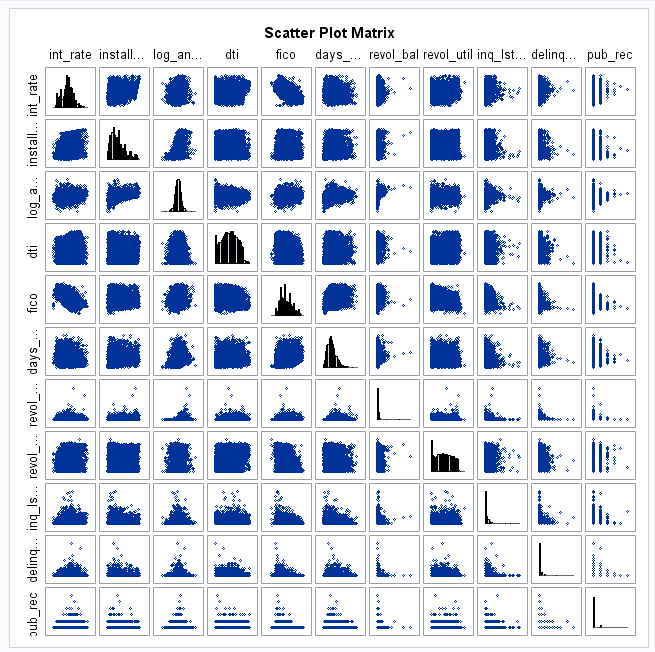
Our dataset is based on data provided by LendingClub. “LendingClub is the world's largest peer-to-peer lending platform. The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015.” [(1)](#_Background_on_Lending) The dataset we analyzed was requisitioned from Kaggle [(2)](#_Data_Set) and specifically focuses on loan data between 2007-2010. We were provided with continuous and categorical variables that we could include in our analysis.

|  |  |  |
| --- | --- | --- |
| **Original Variable** | **Edited Variable** | **Description** |
| credit.policy | credit\_policy | Identifies whether loan applicant met a prescribed set of criteria for lending set by Lending Club |
| purpose | purpose | Purpose/type of Loan acquired |
| **int.rate** | **int\_rate** | **Interest Rate** |
| installment | installment | Monthly installment amount $ |
| log.annual.inc | log\_annual\_inc | log(Annual Income $) |
| dti | dti | Debt-to-Income Ratio |
| fico | fico | FICO credit score |
| days.with.cr.line | days\_with\_cr\_line | Days with a credit line |
| revol.bal | revol\_bal | Revolving Credit Balance $ |
| revol.util | revol\_util | Revolving Credit Utility (% of total credit line used) |
| inq.last.6mths | inq\_lst\_sixmths | Credit Inquiries in last 6 months |
| delinq.2yrs | delinq\_twoyrs | Delinquencies on payments in last 2 years |
| pub.rec | pub\_rec | # of derogatory public records |

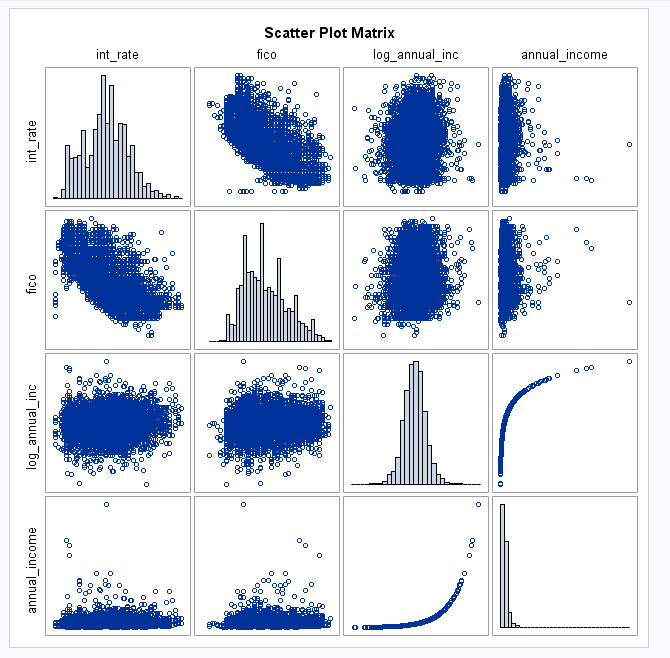
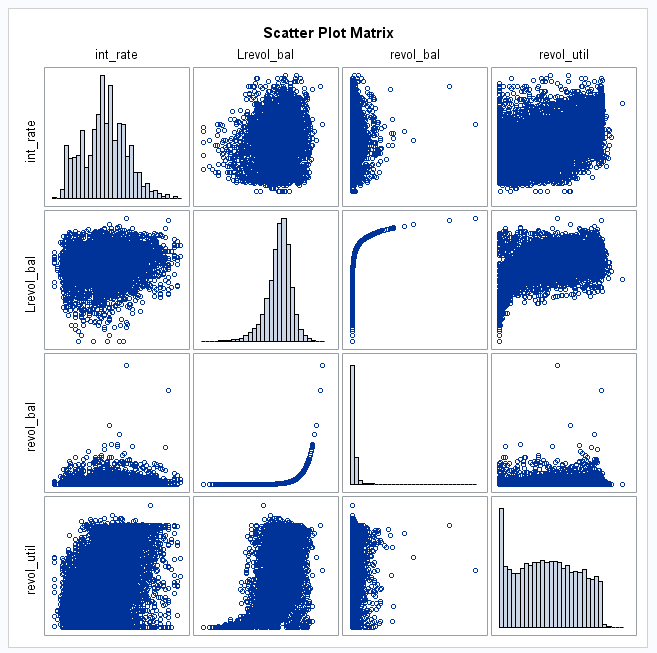
\*Note – variable names edited to convert “.” to “\_” for ease of use in SAS

## Exploratory Analysis

Initial data exploration showed that FICO score was the only clearly correlated variable with interest rate. Due to sheer volume of observations and outliers, many of the scatterplots are indecipherable without further manipulation.



Annual income was provided in log transformation. Because of the size of the dataset, this is not a surprising adjustment. We did validate that this is the correct treatment by taking exp^(Log\_Annual\_Income) and confirmed that the data is in fact very right skewed when plotted. Similar to annual income, the spread of revolving balance on credit/loan accounts appears very right skewed, so we should check on a log transformation there as well. The distribution of the log transformed revolving balance is more normally distributed; again, however it does not appear to have much correlation with interest rates. The variable revol\_util (the % of total credit line being used) looks more correlated with interest rate and is likely more useful as a determinant of interest rate than the total balance by itself.

Overall, the data requires no transformation and can be used as provided.

Addressing Objective 1:

### Restatement of Problem

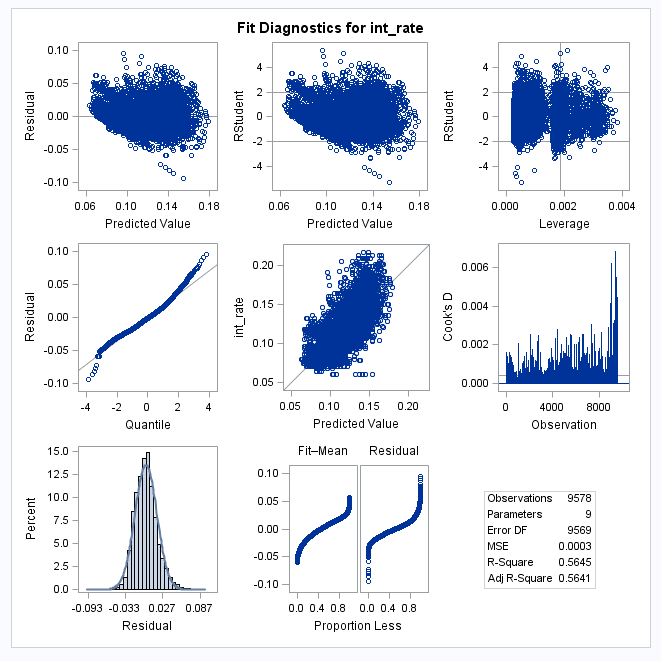
The first task is to determine which variable or variables influence and can predict interest rate on various types of personal and small business loans.

### Model Selection

The team anticipated that we would likely find only a few powerful predictors related to interest rate based on correlation seen in the scatterplots. Because of this, we opted to utilize forward selection as our method for selection.

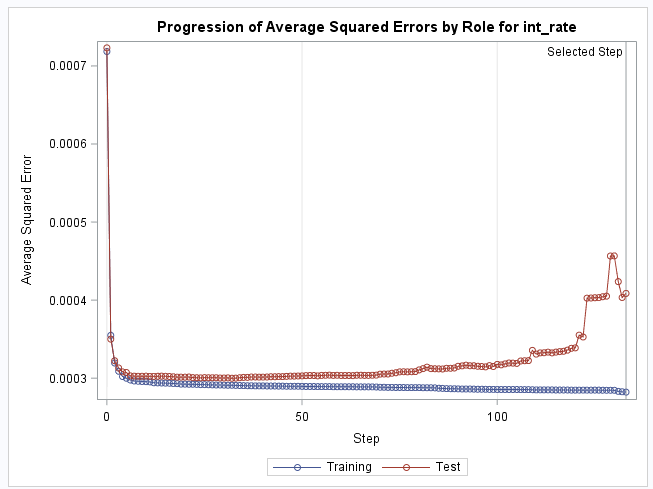
### Checking Assumptions

We evaluated the residual plots of Interest rate. We found a notable void of residuals in the lower left section of the residual scatter plots. This makes sense as rates have a minimum threshold with which they are lent. We have a sample size large enough to assume normal distribution. Additionally, the histogram and QQ plot shows that we have normal standard deviation. Finally, we are assuming independence in observations based on the fact that each observation is a unique loan. While a single person could possibly have requisitioned multiple loans across the span of this dataset, there isn’t sufficient evidence to find violation of independence. Evaluation of the leverage plot does show that there are points that have meaningful leverage but they are counterbalanced in both directions. We also didn’t find concern in the Cook’s D Diagnostic.

****

### Compare Competing Models

A look at the progressive Average Square Error plot of all variables and all possible interactions (using forward selection) shows that most of the prediction power will be in the first one or two steps, and then many other variables can be added without much penalty (to an extent) but are likely not aiding the prediction. Since the first two terms (FICO score and purpose) are the most important terms for prediction and everything after has little effect, we can likely omit interaction terms for anything other than purpose in our selection runs.



### Parameter Interpretation

We found that when predicting interest rate, the following should be used:

Interest rate = .4510 + FICO Score\*(-0.0004669) + DTI\*(0.000151) + Revol\_Util\*(0.00009859) + DebtConsolidation\*(-0.003329) + CreditCard\*(-0.00735) + AllOther\*(-0.005823) + Educational\*(-0.0059) + SmallBusiness\*(-0.01766) + MajorPurchase\*(-0.00422).

We tested with 95% confidence limits. The starting interest rate with no other factors is 45.1% [with confidence interval of (44.29%, 45.95%)]. Each 100 points increase in FICO Credit score reduces the rate by 4.6% [with confidence interval of (-4.55%, -4.78%)]. As Debt to income ratio increase by 1%, so does the interest rate by .015% [confidence interval of (0.0095%, 0.020%)]. The purpose of the loan also impacts interest rate, with small business loans increasing interest rate by 1.7% [confidence interval of (1.57%, 1.96%)]. Lending related to credit cards brings in the biggest impact to interest rate reduction at a reduction of 7.35% [confidence interval of (-5.6%, -9.06%)].

### 

### Final conclusions

A few key factors drive interest rate determination by LendingClub. The primary factor is the FICO Credit Score, followed by the stated purpose of the loan, Debt ratios and revolving utilization of credit. As our data was observational data and was not randomly selected, we can only apply the findings to the dataset itself. We feel that, if given the opportunity to control the dataset collected, we could have incorporated a few other interesting factors to understand further drivers in interest rates. Age demographics, location and marital status would all be interesting factors to next evaluate for significance, allowing us to enhance this model.

## Addressing Objective 2

### Type of Analysis

State what route you are going to take 2way ANOVA or Time series and summarize the goal.

### Main Analysis Content

This will depend on the route you take. I’m leaving it open here to see what you do.

### Conclusion/Discussion

The conclusion should reprise the questions and conclusions of objective 2.

# Appendix

## Objective 1 SAS Code

/\*-------------Import data--------------------\*/

**proc** **import** out=work.loans

datafile="\\smu.edu\Files\users$\jeysenbach\Apps.SMU\Desktop\SAS\loan\_data2.csv"

DBMS= csv replace;

getnames=yes;

datarow=**2**;

**run**;

**proc** **print** data=loans; **run**;

/\*---------Scatter plot matrices-------------\*/

title 'Scatter Plot Matrix';

**proc** **sgscatter** data=loans;

matrix int\_rate installment log\_annual\_inc dti fico days\_with\_cr\_line revol\_bal revol\_util inq\_lst\_sixmths delinq\_twoyrs pub\_rec

/ diagonal = (histogram);

**run**;

**quit**;

/\*fewer variables\*/

**proc** **sgscatter** data=loans;

matrix int\_rate dti fico days\_with\_cr\_line revol\_util inq\_lst\_sixmths delinq\_twoyrs pub\_rec

/ diagonal = (histogram);

**run**;

/\*check unlogged annual income\*/

**data** loans;

set loans;

annual\_income = exp(log\_annual\_inc);

**run**;

**proc** **print** data=loans; **run**;

**proc** **sgscatter** data=loans;

matrix int\_rate fico log\_annual\_inc annual\_income / diagonal = (histogram);

**run**;

/\*looking at log of revolving balance\*/

**data** loans;

set loans;

Lrevol\_bal = log(revol\_bal);

**run**;

**proc** **sgscatter** data=loans;

matrix int\_rate Lrevol\_bal revol\_bal revol\_util

/ diagonal = (histogram);

**run**;

/\*Check some negative determinants of credit score\*/

**proc** **sgscatter** data=loans;

matrix int\_rate inq\_lst\_sixmths delinq\_twoyrs pub\_rec

/ diagonal = (histogram);

**run**;

/\*--------------Exploring ASE with all variables--------------\*/

**proc** **glmselect** data=loans plots(stepaxis = number) = (criterionpanel ASE) seed = **1**;

partition fraction(test = **.5**);

class purpose credit\_policy;

model int\_rate = fico | dti | revol\_util | days\_with\_cr\_line | revol\_bal | delinq\_twoyrs | pub\_rec inq\_lst\_sixmths | credit\_policy | purpose / selection=forward (stop=none) showpvalues stats=all STB;

**run**;

/\*---------------MLR Runs ------------------------------\*/

/\*Forward Selection - no interactions\*/

**proc** **glmselect** data=loans plots = (criterionpanel ASE) seed = **1**;

partition fraction(test = **.5**);

class purpose credit\_policy;

model int\_rate = credit\_policy fico dti revol\_util days\_with\_cr\_line delinq\_twoyrs pub\_rec inq\_lst\_sixmths purpose

/ selection=forward showpvalues stats=all STB;

**run**;

/\*checking diagnostics on using only 3 variables\*/

**proc** **glm** data=loans plots(maxpoints=**10000**) =diagnostics;

class purpose;

model int\_rate = purpose fico revol\_util;

**run**;

## Graphs from EDA

## 

# References

## Background on Lending Club

<https://en.wikipedia.org/wiki/LendingClub>

## Data Set

<https://www.kaggle.com/sarahvch/predicting-who-pays-back-loans>

## Data Source:

<https://www.lendingclub.com/>

## Statistics Theory

Ramsey, F. L., and Schafer, D. W. (2013), The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed.), Boston, MA: Brooks/Cole, with associated website www.statisticalsleuth.com