Lending Club Loan Data Analysis

Rahul Raghatate[1], Saheli Saha[1], Siddharth Thiruvengadam[1], Syam Herle[1]

Abstract
Lending Club is a peer-to-peer lending company, the largest of its kind in the world with $1.1 billion originated loans. It is the first peer-to-peer lender to register its offerings as securities combining with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. In this paper we are trying to understand the reasons people are turning to private lenders instead of banks, effects of variables such as annual income and loan amount influencing the interest rate, group of applicants are more likely to default and/or delay on their payments, whether we can assign the loan interest rates to applications automatically.

Keywords
Loan — Interest Rate — Default Rate Analysis — Exploratory Data Analysis —Linear Regression — Logistic Regression

[1] *Data Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*

Contents

Introduction

Lending Club[5] is headquartered in San Francisco, California. Lending Club has been around since 2007, and operates an online lending platform that matches investors and borrowers in the secondary lending markets.

Peer-to-peer lending networks are typically used to finance small and personal loans for which the bank might not be the ideal source of money. Lending Club competes with other P2P lending platforms, including Prosper and Perform, as well as online direct lenders like Avant. Its original business line is unsecured personal loans for individuals. It also offers unsecured loans to business owners and two niche products - medical loans and auto refinancing loans.

Details of loans offered

- Loan Types: Personal (unsecured), business (unsecured), medical, auto refinancing
- Loan Terms: 36 or 60 months for personal loans; 12, 24, 36, 48, or 60 months for business loans; 24 to 84 months for medical loans; 24 months or longer for auto refinancing loans
- Loan Size: $1,000 to $40,000 for personal loans; $5,000 to $300,000 for business loans; $499 to $50,000 for medical loans; $5,000 to $55,000 for auto refinancing loans
- Rates: 5.89% to 35.89% for personal loans; 9.77% to 35.71% for business loans (subject to change); 3.99% to 26.99% for medical loans; 2.24% to 24.99% for auto refinancing loans
- Origination Fee: 1% to 6%, depending on loan size, term, and borrower profile
- Minimum Investment: $1,000

Lending Club's business model

Lending Club accepts loan applications from potential borrowers. Based on the borrower's risk profile, credit history, the requested loan amount and loan purpose, the club assigns an interest rate and loan repayment term (3 or 5 years). Not all applications are approved, and the ones that are approved are added to the loan listings. 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. Lending Club also makes traditional direct to consumer loans, which are not funded by investors but are assigned to other financial institutions.

Lending Club makes money by charging borrowers an origination fee and investors a service fee.

Investors make money from the interest on the loan. Lending Club also offers them the opportunity to create diversified portfolios that aren't directly tied to bond markets. Its investments offer better yields than CDs, money market accounts, and savings accounts, though it's critical to note that the investments are not FDIC-insured. Borrowers are attracted by the shorter loan approval time and the fact that they have to provide very minimal collateral.

Lending Club focuses on high-credit-worthy borrowers, declining approximately 70% of the loan applications it receives, and assigning higher interest rates to riskier borrowers within its credit criteria. Only borrowers with FICO Scores of 660 or higher can be approved for loans[5].

1. Overview of the analysis

The analysis presented in this report is motivated by following three issues that the Lending Club deals with.

1.1 Existing Issues

- Due to the nature of its business, Lending Club constantly needs to re-evaluate the algorithms it uses to model risk in lending to a borrower. Its algorithm uses borrower and loan attributes such as the length of a loan term, target weighted average interest rate, borrower credit score, employment tenure, and others.
- Lending Club uses pre-defined business rules – based on variables mentioned above such as the loan amount, FICO scores, employment history etc. - and manual inspection to assign interest rates to loan applications that get approved. A model which can approximate these business rules, and also use non-traditional variables (which aren't related to credit history and net worth) to assign interest rates to loans could potentially be very useful.
- Lending Club has been finding it hard to attract investors over the last three years. It might be useful to analyze the reasons behind people choosing peer-to-peer lending networks over traditional sources such as banks. Such analysis can be used in improving marketing efforts and selecting the right incentives to provide to investors and borrowers alike.

1.2 Problem Formulation

To dig deeper into these issues, analysis has been done to answer the following questions.

- What are the reasons people are turning to private lenders instead of banks?
- How do variables such as annual income and loan amount influence the interest rate?
- Which applicants are more likely to default and/or delay on their payments?
- Can the task of assigning loan interest rates to applications be automated through a predictive model?

2. Data Description

The dataset contains records of personal loans that were financed by Lending Club from 2007 to 2015[2].

There are 74 variables in total. Most of the variables fall into one of three categories – loan, borrower's credit profile and borrower's risk profile. Loan data contains variables such as the loan amount, interest rate, payment term and loan purpose. Borrower's credit profile has employment history, income, number of credit lines and so on. Borrower's risk profile has information about missed payments, debt, credit verifications and criminal records.

Here are some summary statistics on the data.

- The dataset contains 887,379 observations of 74 variables. The variables are comprised of 45 numeric and 29 non-numeric variables (categorical, binary and text)
- Totals on the loan amounts reveal that the dataset contains records of 13 billion USD worth of loans from 2007 to 2015
- Many of the variables have missing values, sometimes up for even 80% of the observations
- Many of the variables are stored with incorrect data types and/or inconsistent formats
- After all the data cleaning, transformation and feature selection techniques are applied, the number of variables comes down to 26

3. Data Cleansing

3.1 Missing value treatment

- There are 18 variables for which more than 50% of the observations are missing values (NA). These variables were dropped.
- For the remaining categorical variables, missing values were replaced by the mode.
- For some of the missing numeric variables were replaced by their variable means. For others, an appropriate value had to be filled in based on the description of the variable. An example is 'months since last arrest', which would have a missing value for a person without any arrests, so this missing value is replaced by a very high number.

3.2 Data type and format corrections

- Some numeric variables were stored as strings, and these were converted back.
- The date columns which were also stored as strings had different date formats, so this was fixed.

4. Feature engineering and Feature selection

4.1 Feature Engineering

4 new features were created by applying transformations or formulae to the variables present in the dataset.

- Loan issue year: The year of the loan was extracted from its issue date, to see if there were any yearly trends in loan approval.
- Credit age: The number of days of credit history in the borrower's profile was calculated using the date on which they opened their first credit line[6].
- Credit inquiries: The number of credit inquiries in a borrower's profile was bucketed. This resulted in a categorical variable with 5 levels[7].
- Default: A binary variable indicating whether a member defaulted on the loan repayment was created using the loan status and payment history records.

4.2 Feature Selection

Several categorical variables had too many classes to be meaningful in the analysis being done in this report, so they were dropped. For example, the members' 'job title' had close to 30000 unique entries, and thus proved to be too granular to be useful. Some variables were not required after they were used to engineer features (as described in the previous section), and they were dropped as well.

To further reduce the number of variables, statistical methods were applied.

4.2.1 Correlation Analysis

Based on the data descriptions it appeared that many of the variables, typically connected to loan amounts and payments, would be correlated. The correlation matrix for all the numeric variables was computed, and indeed there were 8 variables that had correlations of 70The first 4 rows and columns of the correlation matrix are shown in Figure 1. It is seen that the installment is very highly linked to the loan amount.

```
> correlation[1:4,1:4]
            loan_amnt int_rate installment annual_inc
loan_amnt        1.00     0.13        0.94       0.33
int_rate         0.13     1.00        0.11      -0.08
installment      0.94     0.11        1.00       0.33
annual_inc       0.33    -0.08        0.33       1.00
```

Figure 1. Correlation Analysis

4.2.2 Variable importance from regression and classification trees

Regression trees were built to predict interest rate. The trees and their variable importance scores were studied to choose important variables. Classification trees were similarly used, with the loan grade as the target variable. Various combinations of the variables and parameters (tree depth, node size etc.) were used to construct both the regression and classification tree models. At this point, after all the data cleaning, feature engineering and feature selection steps are applied, the dataset has 26 variables, which is much lower than the original number of variables (74). One of the regression trees output gave variable importance scores as shown in Figure 2.

```
round(regression_tree$variable.importance/10000,2)
          term          total_rec_int        total_rec_prncp              loan_amnt
        288.71                 278.70                 155.50                 135.66
       issue_d          last_pymnt_amnt          issue_d_year             annual_inc
         76.11                  73.80                  52.70                  37.12
    grade_fnq       verification_status            revol_bal            tot_cur_bal
         22.12                  16.54                  14.87                   6.88
home_ownership                recovries                    dti                purpose
          3.03                   2.57                   0.87                   0.50
     open_acc mths_since_last_major_derog             revol_util mths_since_last_record
          0.21                   0.09                   0.07                   0.06
application_type                pub_rec             delinq_2yrs mths_since_last_delinq
          0.04                   0.01                   0.00                   0.00
```

Figure 2. Variable Importance

5. Text features Analysis

5.1 Exploratory Data Analysis

The dataset contains 74 variables, 45 numeric and 29 non numeric. We cleaned the data and the from the final dataset we tried to understand which variables are important. At first we tried to understand the loan amount distribution throughout the country and the relation with interest rate distribution.

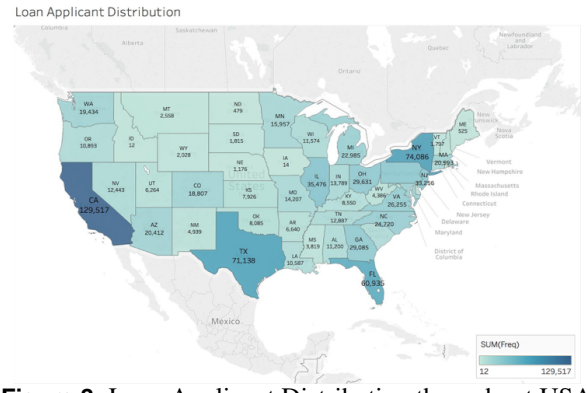

Figure 3. Loan Applicant Distribution throughout USA

As we can see above California has highest loan applicants where as states like Texas New York has medium rate of applicants and states like Iowa, Idaho has low rates of applicants. Then we plotted the interest rate distribution below.

Here we can see same pattern of distribution as loan applicants. California has highest interest rate where as Idaho, Iowa has the lowest.

For further analysis we divided grouped the data year wise to understand the flow of loan applicants year wise.

In Figure 3 and 4 we are comparing the loan amount distribution year wise, we can see every year the amount is increasing, people are preferring applying for greater loan amount every year from Lending club instead of bank.

We grouped the interest rate and loan amount in 3 parts- Low, Middle and High to understand the interaction between these 2 variables.

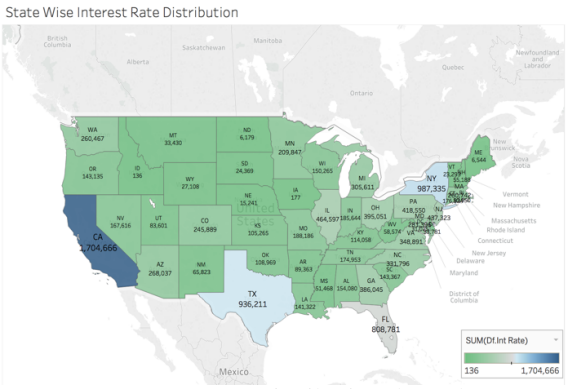

Figure 4. Interest Rate Distribution throughout USA

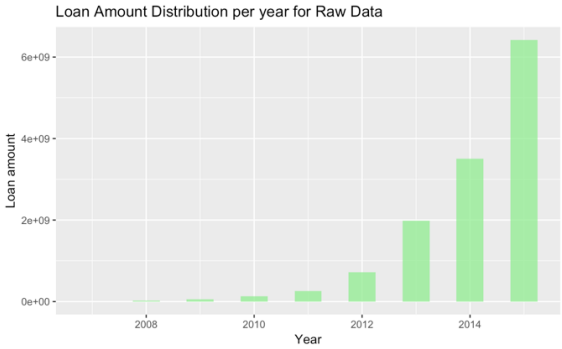

Figure 5. Loan Amount Distribution per year

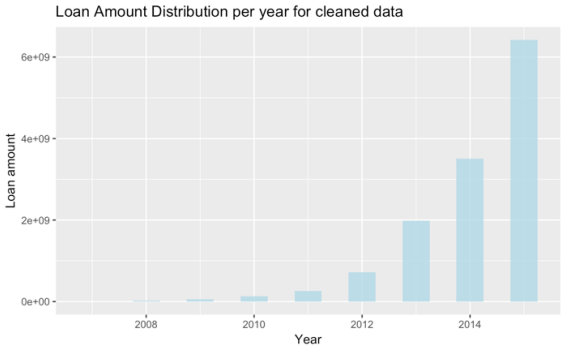

Figure 6. Loan Amount Distribution per year

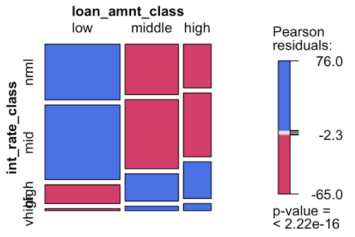

Figure 7. Interaction between Loan Amount and Interest Rate

We noticed from the mosaic plot that there are non-linear relationship between the two variables, there are other factors as well which is effecting the interest rate.

As we noticed in Figure 3 and 4 that loan amount is increasing every year we tried to understand the purpose behind this. We plotted word cloud of purpose variable.

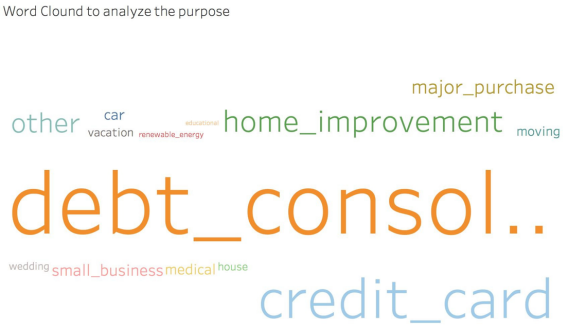

Figure 8. Word Cloud for analysis of purpose of taking loan

As we can see from the word cloud people have taken loan mostly for debt consolidation and credit card. For further verification we made word cloud of description variable where customers has described in sentence the reason behind taking the loan.

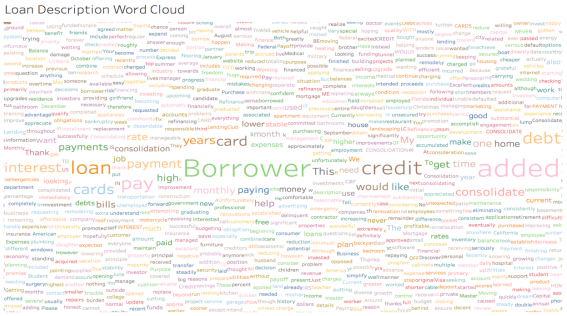

Figure 9. Word Cloud of loan description for purpose analysis

We extended our project and applied NLP on the 'DESC' variable to build corpus by removing stop words, tokenizing and other techniques to get a clean corpus. As we can see in the word cloud same words like debt consolidation and credit cards have greater frequency.

We tried to analyze the impact of grade in loan amount distribution.

AS we can see it has clean distribution grade wose from Grade A to Grade G. For further investigation we plotted the sub grade wise loan amount distribution.

We can clearly see a contradiction from our previous conclusion here loan amount is higher for Sub grades E1-G5.

After noticing these kind of fluctuation we decided to understand the effect of other variables on loan amount using grades.

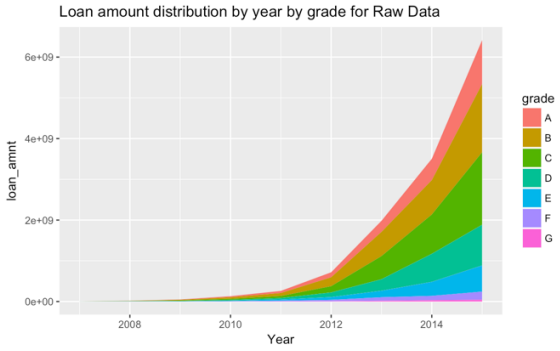

Figure 10. Loan Amount Distribution by year and grade before Data Engineering

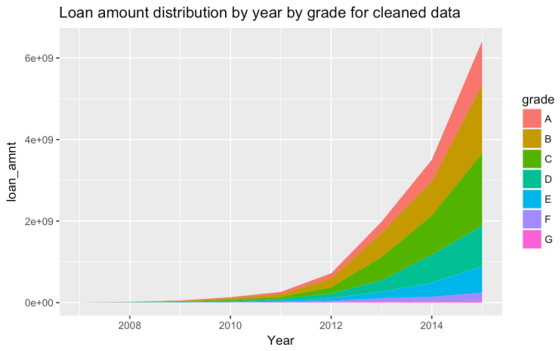

Figure 11. Loan Amount Distribution by year and grade after Data Engineering

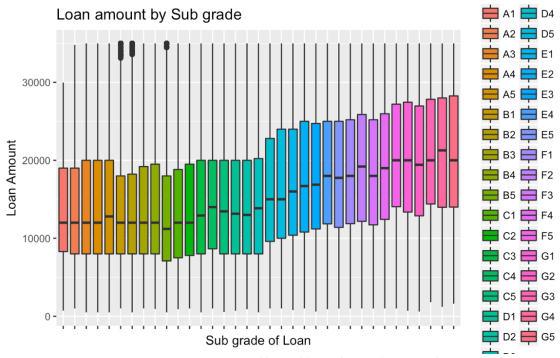

Figure 12. Loan Amount distribution based on Sub-Grade

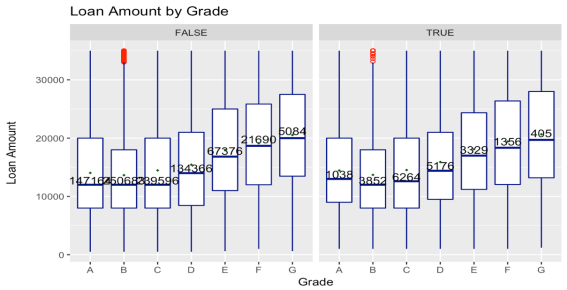

Figure 13. Loan Amount distribution based on Grade and Grouped by Default-Rate

We choose to see the impact of default rate on loan amount. For True and False we got very similar pattern. On an average loan amount is higher for G grades but loan applicants are really less in that group and most of the customers are coming under False group of default rate. To analyze this unexpected result we applied logistic regression which is discussed below in section 6.1.

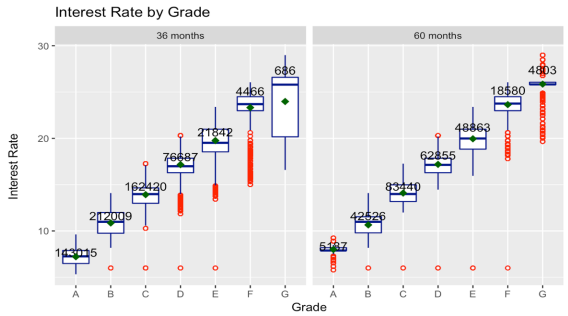

Figure 14. Effect of grade on the Interest Rate grouped by term

We also tried to understand the effect on interest rate for the grade. We can see a clear pattern here. People has taken loan mostly for long term that is 60 months, although the grade G has less people but their interest rate is higher than Grade A people and distribution is also high. This clearly indicates that there are other factors apart from number of loan applicants that are deciding the grade and interest rate for any customer.

We have 'Home Ownership' variable to understand the effect of it on interest rate.

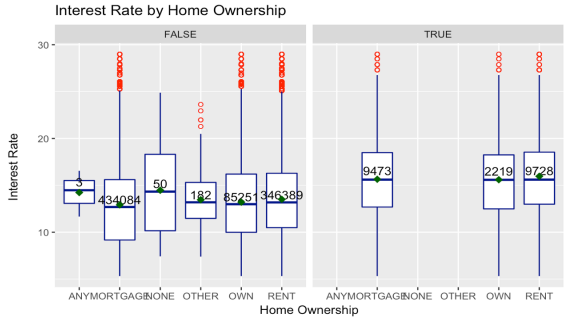

Figure 15. Interest Rate distribution based on Home Ownership and Grouped by Default-Rate

Here interest rate shows no pattern so we can say that it does not depend on home ownership. People has taken loan mostly for Mortgage loan and Rent.

6. Models Methodology

6.1 Classification Model for Predicting Default Loan

In the previous section we have done exploratory data analysis to analyze the default rate in more details we built our own classification model. Our main goal is the prediction of loan defaults from a given set of observations by selecting independent variables or features that result in an acceptable model performance as quantified by a pre-defined measure.

Defining Default For loans data the usual variable of interest is a delay or a default on required payments. So we will try to define a default variable making use of the potential variables which indicates loan will be default or not. From our exploratory data analysis we got a good intuition that the loan status variable seems to be an indicator of the current state of a particular customer.

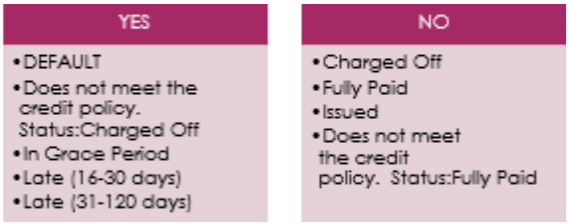

Figure 16. Defining Default

Based on the values of the loan status attribute we have defined the target variable default as in the Figure 16.

Class Imbalance After defining the target variable we analyzed the dataset and found it imbalanced. The distribution of the negative class of default was very high compared to the positive class of default as shown in the Figure 17.

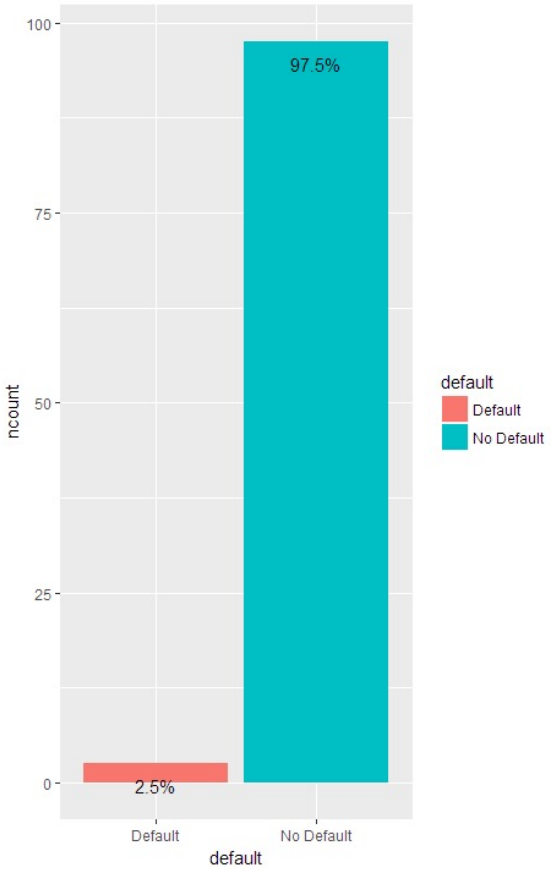

Figure 17. Defining Default

The possible issues of having imbalance data for classification problems are getting a poor classification rates on the minor class and classifier being biased towards to the majority class and classifying all the records as majority class. To address the imbalance data we have to do an under-sample of the majority class with cross validation on each fold independently to get an honest estimate of model performance. The under-sample of the majority class is done using Multivariate Adaptive Regression Splines.

6.2 Regression Model for Interest Rate Prediction

Our main aim is to automate the process of int_rate assignment for loan application by Lending Club using machine learning techniques which was earlier done manually and tediously on case by case basis. To exhibit the relationship between the interest rate and the explanatory variables, we employed series of multivariate linear regression models considering the trade-off between number of variables vs regression statistics[metrics] to reach final model. Below is the flow diagram for the models:

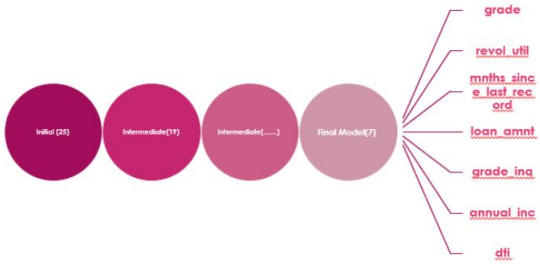

Figure 18. Series of Regression Models

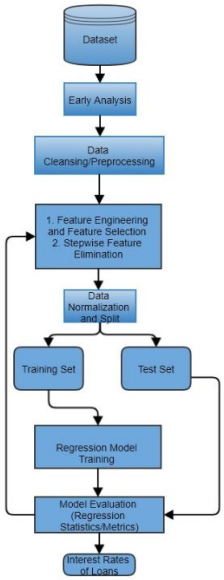

Figure 19. Methodology for Stepwise Regression

The explanatory variables and the model were selected based on the outcome of our exploratory analysis and recursive feature elimination[RFE]/stepwise regression.

7. Experiments and Results

7.1 Classification Model

7.1.1 Variable Importance

Exploratory data analysis helped to analyze the important features and based on that we preformed the following preliminary data transformation steps,

- Changing the character variables to factors
- Dropping variable which seems to be unfit for modelling based on high NA ratios and unique ratios
- Dropping one of the highly correlated variable by applying correlation on numeric attributes
- Handling the imbalance data by undersample of the majority class
- Train/Test Split [75 % / 25 %].

After going through the above mentioned steps we downsized the data attributes to 26 from the initial attributes set.

For the classification problem of loan default, we need to select features which are seems to have high importance for the prediction of the dependent variable. Accordingly, we applied 'Multivariate Adaptive Regression Splines' model on the dataset to select potential candidates for the problem. We applied MARS function on the dataset and we can see the result as follows in Figure 20.

```
Selected 10 of 11 terms, and 5 of 191 predictors
Termination condition: RSq changed by less than 0.001 at 11 terms
Importance: total_rec_late_fee, int_rate, last_pymnt_amnt, recoveries, issue_d_year, ...
Number of terms at each degree of interaction: 1 9 (additive model)
GCV 0.02423634  RSS 18757.23  GRSq 0.03777786  RSq 0.03782261  CVRSq 0.03766651

                 nsubsets   gcv    rss
total_rec_late_fee        9 100.0  100.0
int_rate                  8  64.5   64.6
last_pymnt_amnt           7  48.7   48.8
recoveries                6  34.1   34.2
issue_d_year              5  22.6   22.7
```

Figure 20. Results from MARS

The independent important variables for the classification problem from the Multivariate Adaptive Regression Splines models are,

- total_rec_late_fee
- int_rate
- last_pymnt_amnt
- recoveries
- issue_d_year

After selecting the important features, we tried to fit a logistic regression model with five predictors for classifying the loan record is default or not.

7.1.2 LR Model

While linear regression often used for (continuous) quantitative variables, logistic regression is its counterpart for (discrete) qualitative responses also referred to as categorical. Rather than modeling the outcome directly as default or not, logistic regression models the probability that the response belongs to a particular category or not. The logistic function will ensure that probabilities are within the range [0, 1]. The model coefficients are estimated via the maximum likelihood method. For our problem we have used complex logistic regression model supported by 'caret' package of R, where a train function is defined to streamline the model building and evaluation process. The train function can be used for the followings,

- Evaluate, using re-sampling, the effect of model tuning parameters on performance.
- Choose the 'optimal' model across these parameters.
- Estimate model performance from a training set.

The controlling parameters are defined in a separate function and passed to the training function [3]. A 10-fold cross-validation was repeated five times and the class probabilities of two classes are returned.

7.1.3 Results

For evaluating the logistic regression model, we have used the fitted model to predict the target variable 'default' on unseen test data. We have used the following model evaluation metrics,

- Confusion Matrix : It is a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. In our work we have used the function 'caret::ConfusionMatrix()', which gives confusion matrix along with confidence interval and kappa value. The confusion matrix along with the accuracy for our model is Figure 21

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  96295  1081
       yes 54496  2921

               Accuracy : 0.641
                 95% CI : (0.6386, 0.6433)
    No Information Rate : 0.9741
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0492
```

Figure 21. Confusion Matrix

- ROC Curve : Receiver Operating Characteristic (ROC) [4] summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1 - specificity). For plotting ROC, it is advisable to assume $p > 0.5$ since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of $p > 0.5$. The area under curve (AUC), referred to as index of accuracy(A), is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model. The ROC of our problem is the Figure 22. In our case the area under curve is of accuracy is approximately 70%.

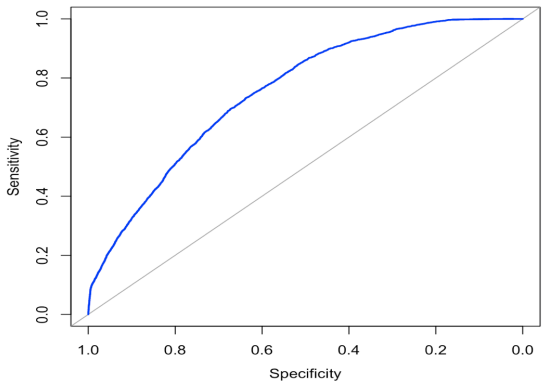

Figure 22. ROC curve

7.2 Regression Model

Linear Regression:

Linear regression is used to predict the value of an outcome variable Y based on one or more input predictor variables X [1].

In simple terms, establish relationship between the predictor variable(s) and the response variable as a function/formula. Then estimate the value of the response Y using predictor values.

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_{n-1} X_n$$

Data Preparation: Preliminary Analysis and data summary and structure allowed us to do some basic data transformations and follow below standard data preparation steps for modeling.

- Convert all categorical 'chr' features to 'factor' for modeling.
- Omit records with 'NA' values as we have lot of data.
- Train/Test Split:
  - train : 75% of data
  - test : 25% of data
- Use *Preprocess[caret package]*[8]: Data Normalization using mean differencing and dividing by standard deviation.

Regression Model Metrics: For comparing models and feature selection, we used RMSE, Min-Max accuracy, and MAPE in addition to Model Summary metrics. These are the standard metrics generally used for regression models [1].

1. For Model:
   (a) Multiple R-squared
   (b) Adjusted R-squared
   (c) Correlation Accuracy
   Correlation between predicted values and actual values
   (d) Root Mean Squared Error[RMSE]
   (e) Min-Max Accuracy

$$= mean\left(\frac{min(actuals, predicteds)}{max(actuals, predicteds)}\right)$$

(f) Mean Accuracy Prediction Error[MAPE]

$$= mean\left(\frac{abs(predicteds - actuals)}{actuals}\right)$$

2. For Feature Elimination:
   (a) p-values for statistical significance
   (b) Variance Inflation Factors(vifs) It is used to explain amount of multicollinearity (correlation between predictors) exists in a regression analysis. Need to remove features based on vifs¿5.
   (c) Coefficients and Standard Errors

```
Call:
lm(formula = int_rate ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-4.2915 -0.1836  0.0058  0.1823  1.1723

Coefficients:
                                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)                          -1.0803817  0.0027949 -386.559  < 2e-16 ***
pub_rec                               0.0022324  0.0003919    5.696 1.23e-08 ***
total_rec_late_fee                    0.0014662  0.0003376    4.343 1.41e-05 ***
issue_d_year2013                     -0.0115788  0.0018550   -6.242 4.32e-10 ***
issue_d_year2014                     -0.2052529  0.0018637 -110.130  < 2e-16 ***
issue_d_year2015                     -0.3483174  0.0020058 -173.659  < 2e-16 ***
home_ownershipOTHER                   0.0430226  0.0298436    1.442    0.149
home_ownershipOWN                     0.0136992  0.0010726   12.772  < 2e-16 ***
home_ownershipRENT                    0.0132649  0.0007552   17.565  < 2e-16 ***
defaultNo Default                    -0.0225468  0.0019110  -11.798  < 2e-16 ***
gradeB                                0.7409741  0.0009798  756.226  < 2e-16 ***
gradeC                                1.4416909  0.0010674 1350.671  < 2e-16 ***
gradeD                                2.1455088  0.0012618 1700.382  < 2e-16 ***
gradeE                                2.7591887  0.0015756 1751.157  < 2e-16 ***
gradeF                                3.5712352  0.0023020 1551.369  < 2e-16 ***
gradeG                                4.0762090  0.0041693  977.665  < 2e-16 ***
total_rec_int                         0.0622073  0.0005556  111.974  < 2e-16 ***
total_rec_prncp                      -0.0236186  0.0008029  -29.415  < 2e-16 ***
purposepurchase                       0.0281970  0.0008741   32.257  < 2e-16 ***
purposesmall_business                 0.0707030  0.0031852   22.197  < 2e-16 ***
last_pymnt_amnt                       0.0292008  0.0005975   48.870  < 2e-16 ***
loan_amnt                            -0.0224079  0.0004890  -45.828  < 2e-16 ***
annual_inc                           -0.0060761  0.0003442  -17.652  < 2e-16 ***
revol_util                            0.0289731  0.0003455   83.853  < 2e-16 ***
term60 months                         0.0368975  0.0008885   41.528  < 2e-16 ***
grade_inqB                            0.0359294  0.0006536   54.976  < 2e-16 ***
grade_inqC                            0.0673200  0.0013757   48.936  < 2e-16 ***
grade_inqD                            0.1219610  0.1354679    0.900    0.368
initial_list_statusw                 -0.0095988  0.0006440  -14.905  < 2e-16 ***
revol_bal                            -0.0040281  0.0003579  -11.256  < 2e-16 ***
credit_age                           -0.0086807  0.0003174  -27.348  < 2e-16 ***
verification_statusSource Verified    0.0227832  0.0007692   29.617  < 2e-16 ***
verification_statusVerified           0.0462293  0.0008216   56.265  < 2e-16 ***
tot_cur_bal                          -0.0089418  0.0004140  -21.600  < 2e-16 ***
open_acc                              0.0025884  0.0003265    7.928 2.23e-15 ***
recoveries                            0.0084354  0.0003444   24.494  < 2e-16 ***
dti                                   0.0021223  0.0002647    8.017 1.09e-15 ***
mths_since_last_record                0.0050868  0.0003964   12.834  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2346 on 612503 degrees of freedom
Multiple R-squared:  0.9456,    Adjusted R-squared:  0.9456
F-statistic: 2.88e+05 on 37 and 612503 DF,  p-value: < 2.2e-16
```

Figure 23. Initial Model Summary

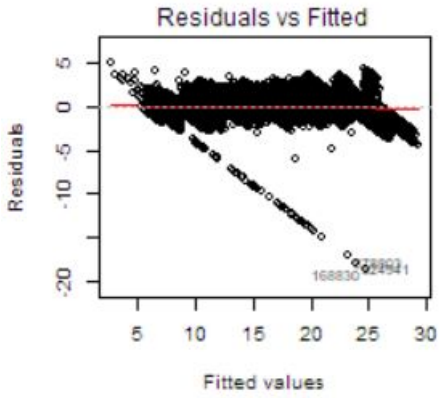

Figure 24. Initial Model-Residual vs. Fitted Values

7.2.1 Initial Model

We started our regression modeling from the final subset of 26 features obtained after feature selection process. Initially removed "issue_d" as newly created feature "issue_d_year" is also present. And undoubtedly, correlation will be very high for two variables.

Dependent Variable:

- continuous: int_rate

Independent Variables:

- continuous[15]: ”pub_rec”, ”open_acc”, ”recoveries”, ”total_rec_prncp”, ”last_pymnt_amnt”, ”loan_amnt”, ”total_rec_int”, ”total_rec_late_fee”, ”annual_inc”, ”revol_util”, ”revol_bal”, ”credit_age”, ”tot_cur_bal”, ”dti”, ”mths_since_last_record”
- categorical[9]:”issue_d”, ”issue_d_year”, ”home_ownership”, ”default”, ”grade”, ”purpose”, ”term”, ”grade_inq”, ”initial_list_status”, ”verification_status”

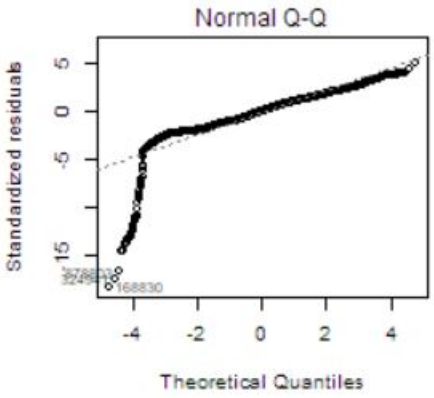

Figure 25. Initial Model-Residual Normality Check

7.2.2 Intermediate Models

After analyzing the results of Initial Model, we keep on removing following variables based on model summary and our metrics defined in previous section for feature elimination.
Removed Variables[5]: ”pub_rec”, ”issue_d_year”, ”default”, ”home_ownership”, ”total_rec_late_fee”,.
Here are results for one of the intermediate model with 19 independent variables:

We can see that, though the metrics for model comparison shows overall decrease/loss, it is acceptable at the cost of eliminated 6 features. Model efficiency has very minimally affected. Also, there are insignificant and correlated features still present in model.

After running manually a series of regression models and continuing the elimination of features, we stopped at below final model as we removed most of the insignificant features and also took care of multicollinearity among the predictors.

7.2.3 Final Model

Dependent Variable:

- **continuous**: int_rate

Independent Variables:

- **continuous[5]**:”annual_inc”, ”revol_util”, ”dti”, ”mths_since_last_record”

```
Call:
lm(formula = int_rate ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-18.1001  -0.8350   0.0166   0.7968   5.2692

Coefficients:
                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      7.507464   0.004718 1591.265  < 2e-16 ***
gradeB                           3.256713   0.004568  712.902  < 2e-16 ***
gradeC                           6.276542   0.004966 1263.839  < 2e-16 ***
gradeD                           9.337608   0.005863 1592.690  < 2e-16 ***
gradeE                          11.970420   0.007314 1636.682  < 2e-16 ***
gradeF                          15.512247   0.010701 1449.635  < 2e-16 ***
gradeG                          17.639310   0.019414  908.604  < 2e-16 ***
total_rec_int                    0.478374   0.002446  195.543  < 2e-16 ***
total_rec_prncp                  0.339637   0.003345  101.538  < 2e-16 ***
purposepurchase                  0.127379   0.004070   31.296  < 2e-16 ***
purposesmall_business            0.326454   0.014860   21.969  < 2e-16 ***
last_pymnt_amnt                 -0.049722   0.002707  -18.371  < 2e-16 ***
loan_amnt                       -0.361948   0.002038 -177.597  < 2e-16 ***
annual_inc                      -0.027978   0.001598  -17.504  < 2e-16 ***
revol_util                       0.145226   0.001608   90.330  < 2e-16 ***
term60 months                    0.262724   0.004113   63.880  < 2e-16 ***
grade_inqB                       0.186274   0.003046   61.162  < 2e-16 ***
grade_inqC                       0.367153   0.006408   57.300  < 2e-16 ***
grade_inqD                       1.030773   0.632115    1.631   0.1030
initial_list_statusw            -0.157506   0.002966  -53.110  < 2e-16 ***
revol_bal                       -0.015057   0.001661   -9.063  < 2e-16 ***
credit_age                      -0.027457   0.001464  -18.751  < 2e-16 ***
verification_statusSource Verified  0.018152  0.003570   5.084 3.69e-07 ***
verification_statusVerified      0.203085   0.003818   53.191  < 2e-16 ***
tot_cur_bal                     -0.055611   0.001713  -32.458  < 2e-16 ***
open_acc                        -0.001014   0.001522   -0.666   0.5052
recoveries                       0.129079   0.001565   82.470  < 2e-16 ***
dti                              0.002323   0.001235    1.881   0.0599 .
mths_since_last_record           0.015723   0.001412   11.137  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.095 on 612512 degrees of freedom
Multiple R-squared:  0.9383,    Adjusted R-squared:  0.9383
F-statistic: 3.329e+05 on 28 and 612512 DF,  p-value: < 2.2e-16
```

Figure 26. Intermediate Model Summary

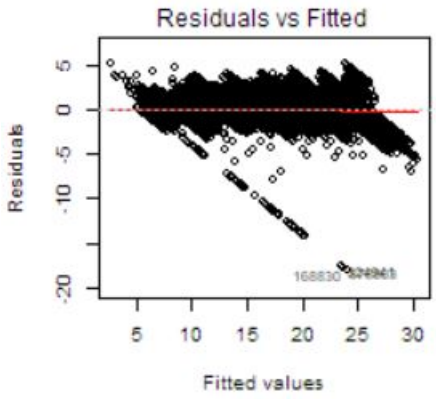

Figure 27. Intermediate Model-Residual vs. Fitted Values

- **categorical[2]**: ”grade”, ”grade_inq”

At the cost of minimal decrease in prediction efficiency based on regression metrics shown in Figure 33, we have reduced the model predictors significantly from 25 to 7. This allows us to conclude with the most important features required for interest rate prediction along with model validity.

From exploration of relationship between interest rate and few important variables such as ’grade’, ’annual_income’, ’loan_amount’, ’sub_grade’ and series of regression models based on feature elimination shows that the variables included in final model are relevant both from statistical and business perspective.

From Summary plot for Final Regression Model, we can see the standard errors are of $10^{-}3$ scale compared to the coefficient estimates as expected. Otherwise, model would have been invalid as the errors in estimates will be high.

It makes sense, that int_rate will be highly dependent on

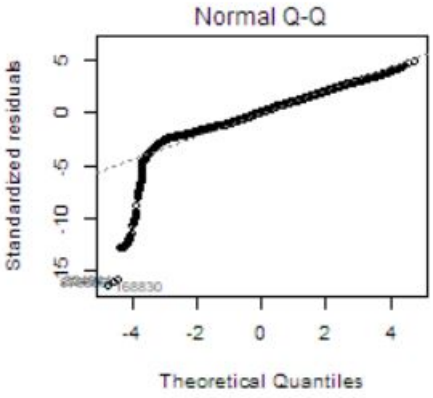

Figure 28. Intermediate Model-Residual Normality Check

```
Call:
lm(formula = int_rate ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7495  -0.9223   0.0459   0.8099  11.0963

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              7.269208   0.004065 1788.355  < 2e-16 ***
gradeB                   3.470421   0.004927  704.432  < 2e-16 ***
gradeC                   6.553117   0.005095 1286.185  < 2e-16 ***
gradeD                   9.704262   0.005819 1667.811  < 2e-16 ***
gradeE                  12.395435   0.007091 1747.963  < 2e-16 ***
gradeF                  16.053848   0.010776 1489.716  < 2e-16 ***
gradeG                  18.105077   0.020581  879.697  < 2e-16 ***
grade_inqB               0.255549   0.003372   75.793  < 2e-16 ***
grade_inqC               0.468281   0.006884   68.021  < 2e-16 ***
grade_inqD              -2.516122   0.074380  -33.828  < 2e-16 ***
grade_inqE              -4.496682   0.204621  -21.976  < 2e-16 ***
revol_util               0.166480   0.001666   99.939  < 2e-16 ***
dti                     -0.014192   0.001413  -10.041  < 2e-16 ***
loan_amnt                0.037001   0.001714   21.582  < 2e-16 ***
mths_since_last_record   0.007671   0.001591    4.822 1.42e-06 ***
annual_inc              -0.070672   0.001701  -41.557  < 2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.278 on 665141 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.9149
F-statistic: 4.769e+05 on 15 and 665141 DF,  p-value: < 2.2e-16
```

Figure 29. Final Model Summary

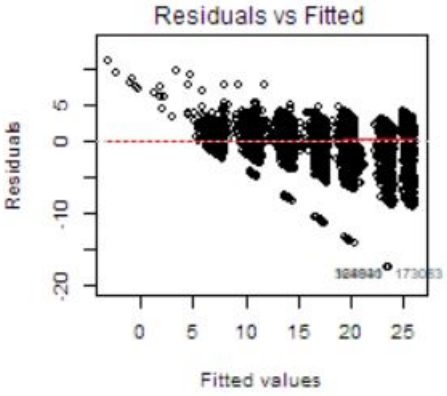

Figure 30. Final Model-Residual vs. Fitted Values

grade of loan, number of credit inquiries made, revolving utilization, debt to income ratio, and annual income. For example, if person's annual income is low, and debt to income ratio is high, such loan case might get low grade and his loan will be very likely to be defaulted in future and hence high interest_rate will be allocated as there is huge risk factor.

7.3 Build Specifications

- Processor: Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz
- Memory: 16.0 GB

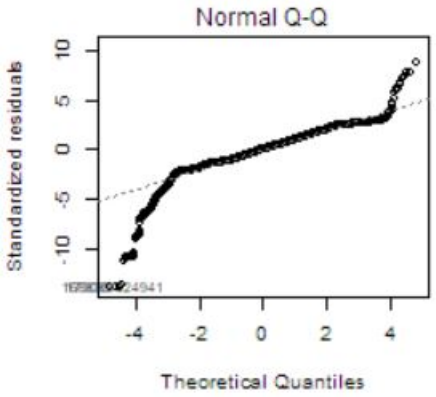

Figure 31. Final Model-Residual Normality

	vif.reg_model2.
gradeB	2.0254
gradeC	2.1168
gradeD	1.8282
gradeE	1.5030
gradeF	1.1901
gradeG	1.0509
grade_inqB	1.0893
grade_inqC	1.0923
grade_inqD	1.0028
grade_inqE	1.0003
revol_util	1.1317
dti	1.0173
loan_amnt	1.1979
mths_since_last_record	1.0308
annual_inc	1.1614

Figure 32. Final Model-Variance Inflation Factor Scores

- Software Platform: R Studio Version 1.0.153 with R version 3.4.2 (2017-09-28) [x86_64-w64]
- OS: Windows 10 (64-bit)

8. Conclusion

This report presented an approach to automate the process of interest rate allocation for loan application. Our state-of-art, 'Final Regression Model' can be used to predict the interest_rate approximately 92% accurately. Also, the model consist of only 7 variables for achieving such high prediction efficiency allowing us to reduce the data load on large scale.

For predicting future default loans, we established classification model using logistic regression based on 5 features providing accuracy approx. 64% which was badly affected due to class imbalance and absence of significant features like 'FICO_Score'.

9. Limitations and Future Work

Based on regression modeling, there is huge scope of improvement in the model and achieve higher accuracy. Currently, the data is very messy with high proportion of missing values for relevant features. Also, important feature like FICO_Score is missing which might prove in different model results and

Model	Variables [Ind]	Model Statistics				
		Adj_R^2	RMSE	Corr-Accuracy	Min-Max Accuracy	MAPE
Initial_Model	25	0.9456	0.0020	97.22%	93.19%	0.073
Intermediate	19	0.9383	0.0018	96.83%	92.93%	0.076
Final	7	0.9149	0.0022	95.64%	92.08%	0.087

Figure 33. Regression Model Comparison

significance of other features. So, missing data and class imbalance renders accuracy/prediction metrics as an insufficient for evaluation. Though, the models seems valid under given circumstances, one should seek domain experts input.

Below are few ideas for future work on interest rate prediction modeling and as well as default loan prediction modeling.

1. Feature Engineering and feature integration from Lending club site for FICO Score which is very significant.
2. Reduce categories of few variables like purpose, title to consider them and find significance
3. Advanced Automated Feature Selection Using Forward & Backward Selection, Caret Package [VarImp, Recursive Feature Elimination(RFE), Genetic Algorithms], and Boruta Package
4. Year, Grade based bucket Regression Model for more efficient predictive modeling as there is high grade and year based data imbalance
5. Use Robust Regression, XGBoost,Ensemble Model, Random Forest Regression, and Neural Network Modeling for regression and classification
6. Try dimensionality reduction using PCA, Factor analysis techniques
7. Try different classification model like SVM, Random Forest after dimensionality reduction

Acknowledgments

Many thanks to Prof. Yen-Ning Huang at Indiana University Bloomington for her academic as well as professional guidance. We would also like thank Miguel Raul Pebes Trujillo for guiding us along our project milestones.

References

[1] Prabhakaran, S. (n.d.). Linear Regression. Retrieved December 13, 2017, from http://r-statistics.co/Linear-Regression.html

[2] https://www.kaggle.com/wendykan/lending-club-loan-data

[3] Alexander Wagner, https://triamus.github.io/project/lending-club-loan-data-in-r/

[4] Analytic Vidhya, Tavish Shrivastava https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/

[5] Lending Club Website https://www.lendingclub.com/

[6] http://www.magnifymoney.com/blog/personal-loans/lendingclub-review-borrowers-insiders-reveal578

[7] http://cs229.stanford.edu/proj2014/Kevin%20Tsai,Sivagami%20Ramiah,Sudhanshu%20Singh,Peer%20Lending%20Risk%20Predictor.pdf

[8] Kuhn, M.(2008). Caret package. Journal of Statistical Software, 28(5)