Lending Club
--- Invest Efficiently and Elegantly

Group 2A

Cheng, Zian
Lian, Xinyi
Pan, Jing
Jiang, Yi
Song, Xianwen

Abstract

As a large amount of peer-to-peer (P2P) lending companies are facing bankruptcy in China, Lending Club, known as the world's largest marketplace for personal loans draws our attention.

Given that the data of the company is posted on their official website, we aim to analyze lending companies based on its loan data from 2007 to 2015. The data is made up of 887 thousand observations and 74 variables, and it consists of lots of information including the amount of loan, the latest payment information, the occupation, gender, credit scores etc. After data cleaning, our data consists of 210 thousand observations and 26 variables.

Our group applied a series of methods, both supervised and unsupervised, to find out the best way to predict the probability of default (LDA, KNN, logistic regression, Clustering, PCA). We separate the data into 50% training set and 25% of validation set and 25% test set, and train based on training set.

In this project, we are standing at the investor side and analysis how an investor of the lending club company can make the most profit of their money. We can also develop operation plan for the company to secure its future growth.

Introduction

P2P Lending. Lending Club uses technology to operate at a lower cost, passing the savings on to borrowers with lower rates and to investors with solid returns. Borrowers who used a personal loan via Lending Club to consolidate debt or pay off high interest credit cards report that the interest rate on their loan was an average of 24% lower.

$100
$150
Pay
06:00
Investors
In exchange for solid returns, investors purchase Notes, which correspond to fractions of loans.
LendingClub
LendingClub screens borrowers, facilitates the transaction, and services the loans.
Borrowers
Borrowers use loans to consolidate debt, improve their homes, finance major purchases, and more.

Loan Amount. From the dataset made up of around nine hundred thousand samples, we can see the average amount of loan outstanding is around $12,000. The median is around $15,000. From the amount issued throughout the years, we find out that the company is growing rapidly and the loan increased per year reaches $0.6 billion. The amount that is issued since the company started has been rising exponentially. Therefore, our group thinks that we need to find out a way to forecast the default probability of a debt. Through building analytical model, we can prevent debts with high default probability and make sure that the company is financially healthy. Our models secure the company's future growth, as well as the funds given by investors of the company.

Figure 1.1 Loan amount distribution

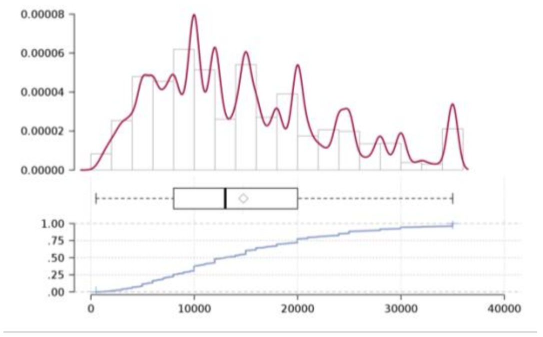
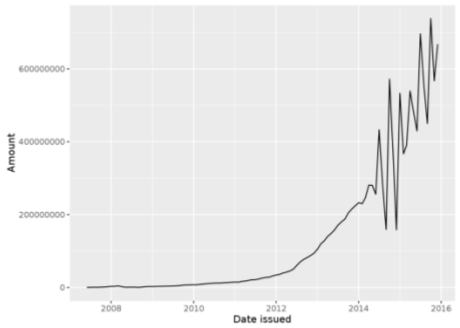


Figure 1.2 Amount issued per years



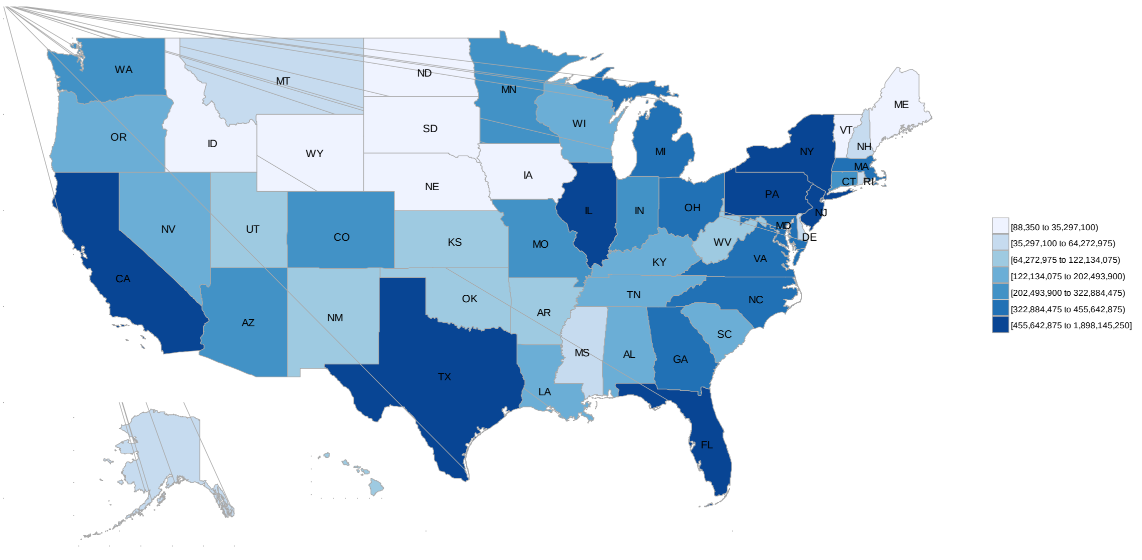
Geometric Distribution. The states with the highest lending amounts are California, Texas, Florida, Illinois, Philadelphia, New Jersey, and New York (Figure 1.3). All of these lending states has loan outstanding at around $0.45 billion to $1.8 billion. The amount is so large that a prediction of the company's revenue is very important for the investors.

Interest Rate and Grade. For the interest rates, we assign 5.31%-7.96% as group A, 9.43%-11.98% as group B, 12.61%-16.01% as group C, 17.47%-21.85% as group D, 22.90%-26.77% as group E, and the rest as group F (Attachment 1).

Default. While loans that are at the state "Current" or "Issued" are ignored in our project, we are interested in loans that are "Fully Paid" or in other default-like situations. Fully paid is considered as not default (Default = 0). For charged off, default, does not meet the credit policy, in grace period, and late are all considered as default (Default = 1).

Loan Purpose. Most of the lenders embrace the purpose of debt consolidation (with 60% of the lenders). They just want to diversify their portfolio. A slightly above 20% of the lenders lend to pay back their credit card. The other customers lend for home improvement, cars and different kinds of other purchases (Appendix 2).

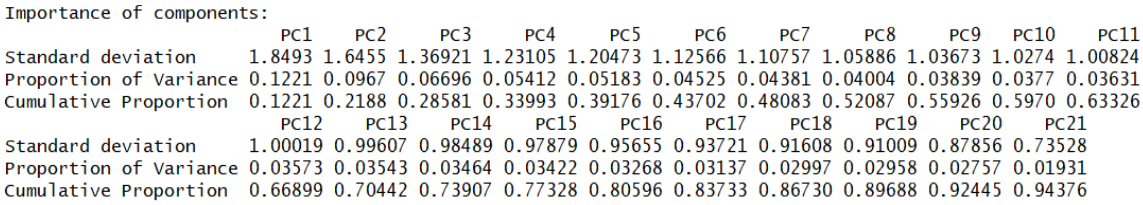
Figure 1.3 Amount Loan Outstanding by State


WA
MT
ND
MN
ME
OR
ID
SD
WI
VT
NH
WY
MI
NY
MA
IA
CT
RI
NE
IL
PA
NJ
NV
UT
CO
IN
OH
MD
DE
KS
MO
WV
VA
CA
KY
OK
AR
TN
NC
AZ
NM
SC
MS
AL
GA
TX
LA
FL
[88,350 to 35,297,100)
[35,297,100 to 64,272,975)
[64,272,975 to 122,134,075)
[122,134,075 to 202,493,900)
[202,493,900 to 322,884,475)
[322,884,475 to 455,642,875)
[455,642,875 to 1,898,145,250]

Method

Data Cleaning. We first plot a correlation matrix, and delete columns that are meaningless in prediction including id, member_id, policy_code, addr_state, title, desc, url etc (Appendix 3). Secondly, we get rid of the columns with more than 50% of its value missing. The next step is to deal with the rows with missing values using na.omit function. Given that the data is clean and tidy, we change classes of purposes into dummies and so does the loan status (default = 1, fully-paid =0). At the same time, we will turn home ownership, verification status, term, employment length and credit grade into categorical variables into categorical variables.

PCA. PCA is a method of fitting an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small, and by omitting that axis and its corresponding principal component from our representation of the dataset, we lose only a commensurately small amount of information. In our PCA analysis, we find that the element which draws the highest importance of data constructs a standard deviation of around 1.8493. All elements unveil a roughly uniformly-decreasing pattern in generating standard deviation (Appendix 3).



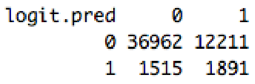
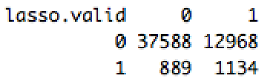
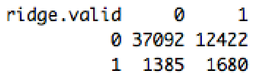
```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9   PC10    PC11
Standard deviation     1.8493 1.6455 1.36921 1.23105 1.20473 1.12566 1.10757 1.05886 1.03673 1.0274 1.00824
Proportion of Variance 0.1221 0.0967 0.06696 0.05412 0.05183 0.04525 0.04381 0.04004 0.03839 0.0377 0.03631
Cumulative Proportion  0.1221 0.2188 0.28581 0.33993 0.39176 0.43702 0.48083 0.52087 0.55926 0.5970 0.63326
                          PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     1.00019 0.99607 0.98489 0.97879 0.95655 0.93721 0.91608 0.91009 0.87856 0.73528
Proportion of Variance 0.03573 0.03543 0.03464 0.03422 0.03268 0.03137 0.02997 0.02958 0.02757 0.01931
Cumulative Proportion  0.66899 0.70442 0.73907 0.77328 0.80596 0.83733 0.86730 0.89688 0.92445 0.94376
```

Clustering. We decided to run k-means clustering on the dataset to understand the data better. We picked k to be 4 because within groups' sum of squares were dropping very quickly before 4 and after 4 it became very flat. Here's the result we obtained. It can be seen that group 3 has the lowest interest rate, highest loan amount, lowest annual income while group 4 is the opposite (Appendix 4).

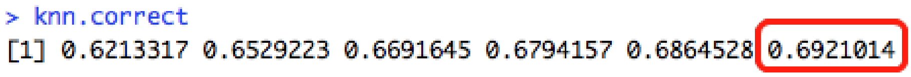
LDA. LDA is looking for linear combinations of variables which best explain the data. In this situation, we find that using the LDA analysis, we can have a forecast correctness of 73.8%.

Logistic Regression, Lasso & Ridge. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Lasso and Ridge are regression analysis methods that perform both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. We use ROC curve to find out the best threshold.

In our logistic regression, we find that interest rate, employment length, self-reported annual income, home ownership and the total number of credit lines are among the most important factors. (Appendix 4)

Type	Logistic	Lasso	Ridge
Prediction	 logit.pred 0 1 0 36962 12211 1 1515 1891	 lasso.valid 0 1 0 37588 12968 1 889 1134	 ridge.valid 0 1 0 37092 12422 1 1385 1680
Correctness	73.89%	73.64%	73.74%

KNN. K-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.



```
> knn.correct
[1] 0.6213317 0.6529223 0.6691645 0.6794157 0.6864528 0.6921014
```

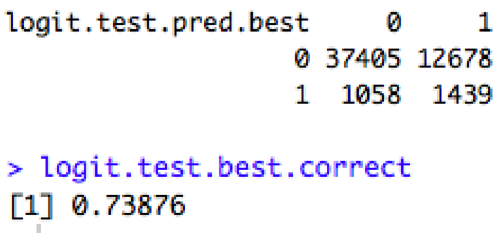
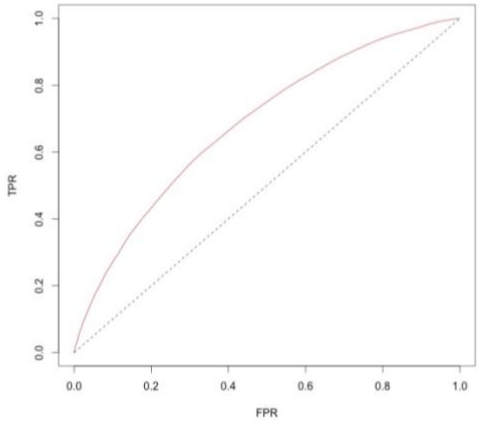
Revenue Estimation. We calculate the revenue from using the net present value. If the debt defaults, the lost is the amount of debt. If the debt is fully paid, the net present value of the debt is the present value of the principal and all interest payments discounted to the day when the debt is issued, and minus the amount paid by the investor at the date of debt issuance. Finally, we sum up the net present value according to their perspective gradings.

$$\text{Net Present Value of a loan} = \begin{cases} -\text{principal} & \text{if the loan defaults} \\ -\text{principal} + \text{PV(principal)} + \text{PV(interest)} & \text{if the loan is fully paid} \end{cases}$$

Results

Default Probability. After fitted all the models using training data, we find out the best prediction model by applying those models again on the validation set. Logistic regression performs extremely well throughout the process. Therefore, logistic regression is selected and fitted again in the data combing the validation set and the training set. We have 73.88% of probability that our default prediction is right.

3.1 Test Results	3.2 ROC Curve

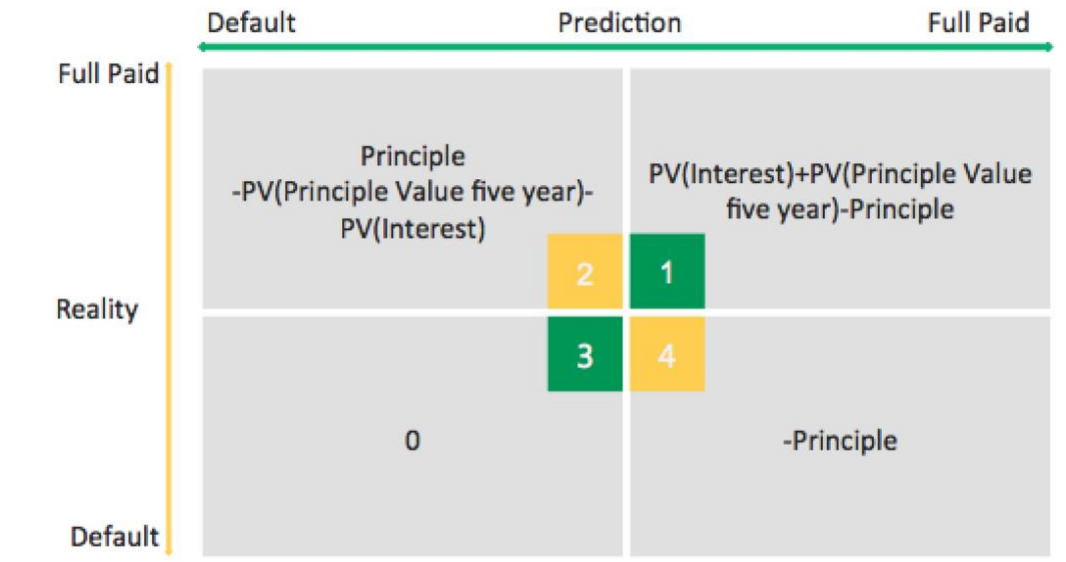

`logit.test.pred.best      0     1`
`                    0 37405 12678`
`                    1  1058  1439`
 	
`> logit.test.best.correct`
`[1] 0.73876`
	

Revenue. We calculate the revenue generated by different grades of loaners using the Revenue Estimation method. At the same time, we also generate the revenue forecasted by our default probability model, and the result is as follow:

Most of the investors invested in grade 2, grade 3 and grade 4 loans, and these loans are regarded as the safest. However, according to our earning estimation, we find out that grade 6 have a higher return. Our group also compared the earning per dollar using our estimation and the real default data, as well as number of transaction and the earning per transaction.

Company Valuation. In this case, we applied the four situations: fully paid debt versus estimated as default; fully paid versus estimated as not default; default versus estimated as default; default versus estimated as not default. Our group will calculate the perspective income according to different states.

Figure 2.1 Profit Estimation



Earnings Estimated versus Real Data. We argue that our estimation might work perfectly and we believe that we are able to capture the earnings changes based on grade. When we applied different seed during the data separation procedure, our model works perfectly and stably when estimating the level 1 bonds, level 3 bonds, level 6 bonds and level 7 bonds. Therefore, we might suspect that the grading of level 2, 4 and 5 are around the boundary of 1, 3 and 6, therefore might be hard to find out its default value.(Figure 3.1)

Insights

Our default probability model can help the investor better estimate the default probability of a loan, neglecting its official grading. Third-party estimation helps investors to eliminate risk. At the same time, our revenue generation method will help the investors and fund managers to better-select which loans to invest.

Investment Suggestions. We suggest that during the investment procedure, we can invest in level 6 bond first. As there are only limited level 6 bonds on the P2P platform (figure 3.3), we would suggest you to invest the rest of your portfolio on level 3 bonds. If your total investment is more than the sum of numbers of total transactions of level 3 bonds and level 6 bonds, we further suggest that we can invest a basket of bonds consisting level 1, 3 and 6 bonds. Given that a mutual fund can invest a basket of bonds, we would believe that our analysis works perfectly for them.

Advices for Management. It seems extremely clear from our estimate and the reality that the interest rates for the Grade 7 bonds are too low. Our group suggests that the interest rate of Grade 7 needs to rise to the highest level under regulation of the

State and Federal Law. The company should also keep working on its grading system to provide a comfortable investing environment for the P2P funds.

Shareholders Perspective. Our group finds out that the grading system of the company is working well regarding control default risk and preparing the value at risk. However, there is still room to improvement. The risk evaluation can improve by changing the existing 7 grades system into fewer grades or other kinds of more clear-cut expressions.

Figure 3.1 Earnings Per Dollar (by grade)

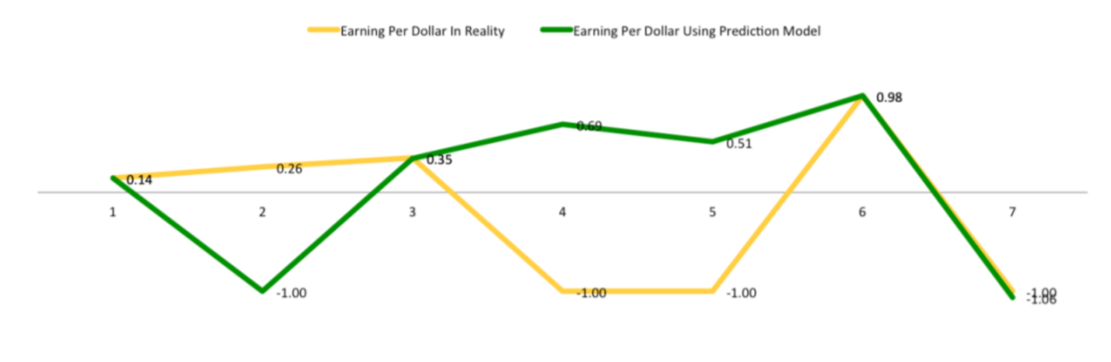


Figure 3.2 Number of Transaction (by grade)

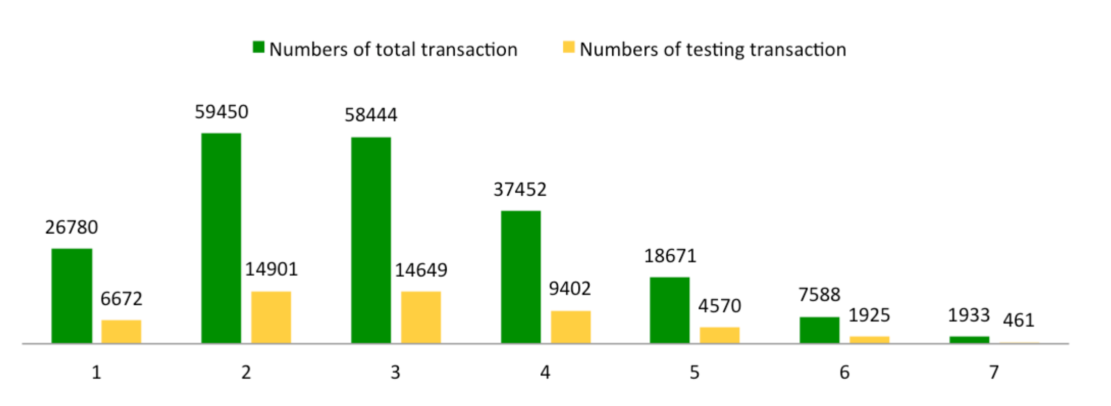
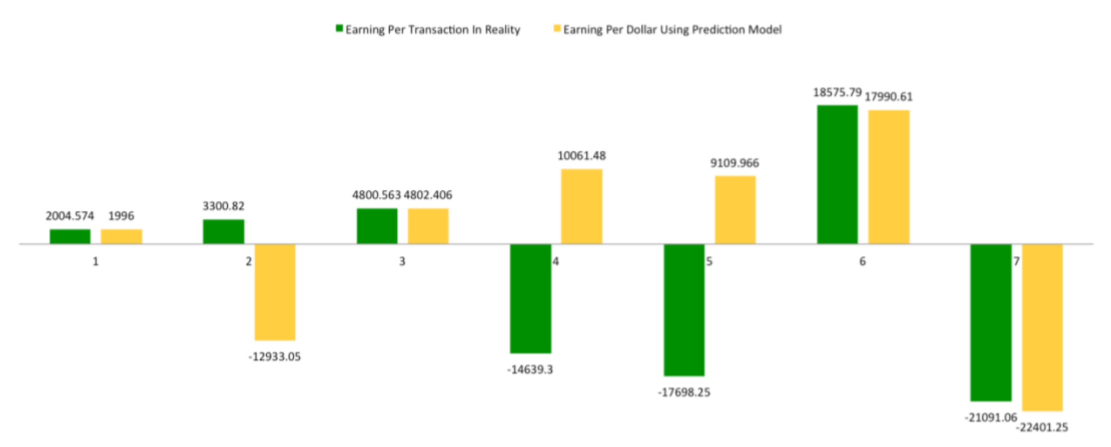
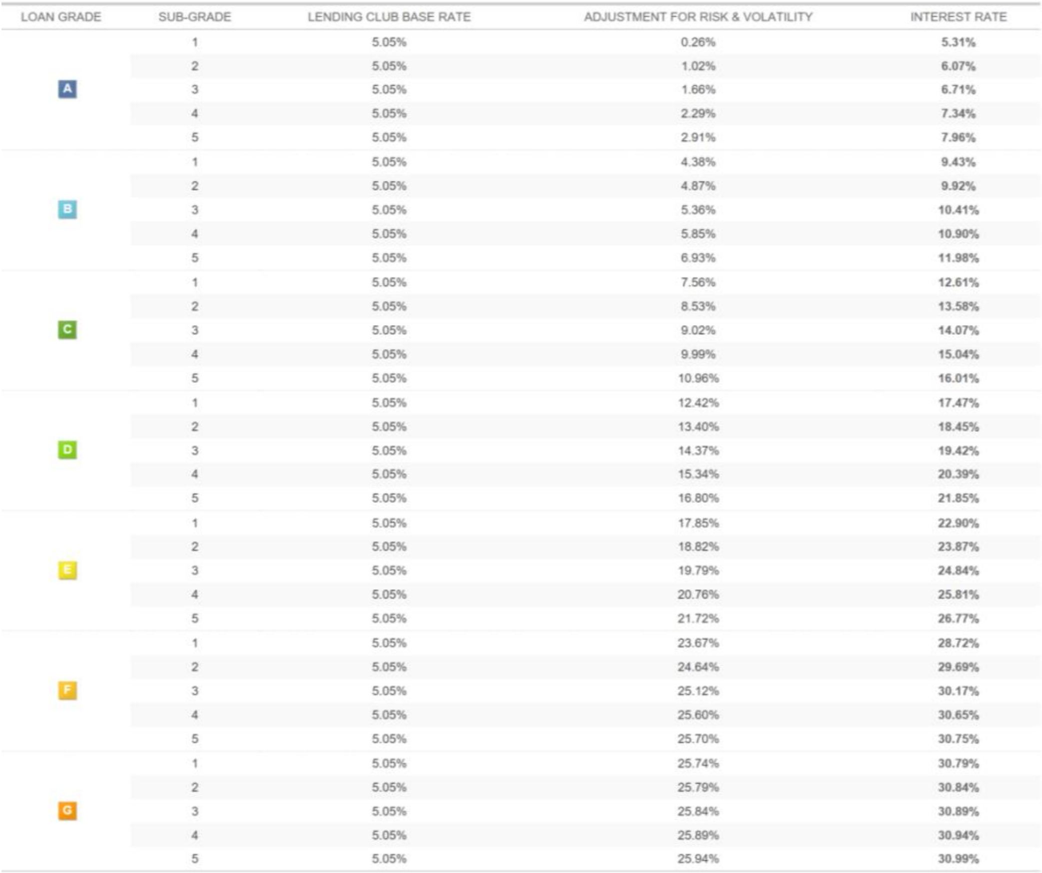


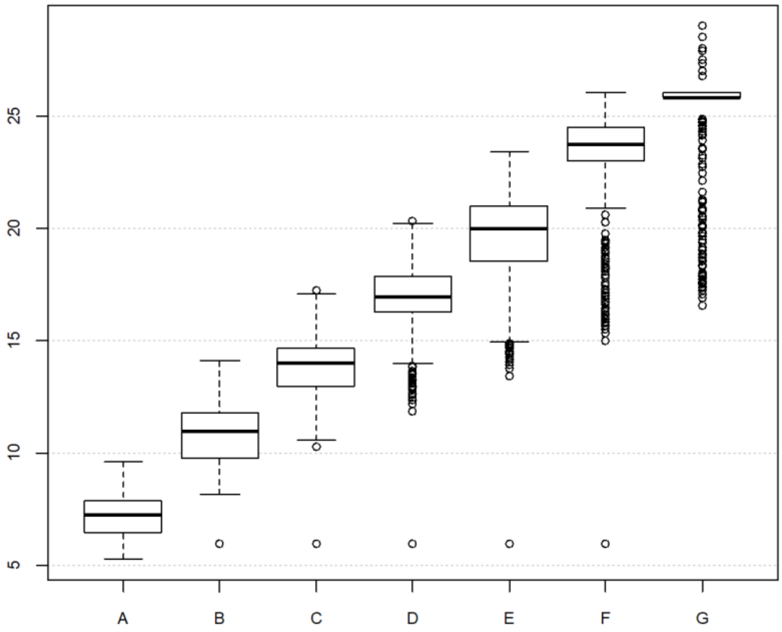
Figure 3.3 Earnings Per Transaction



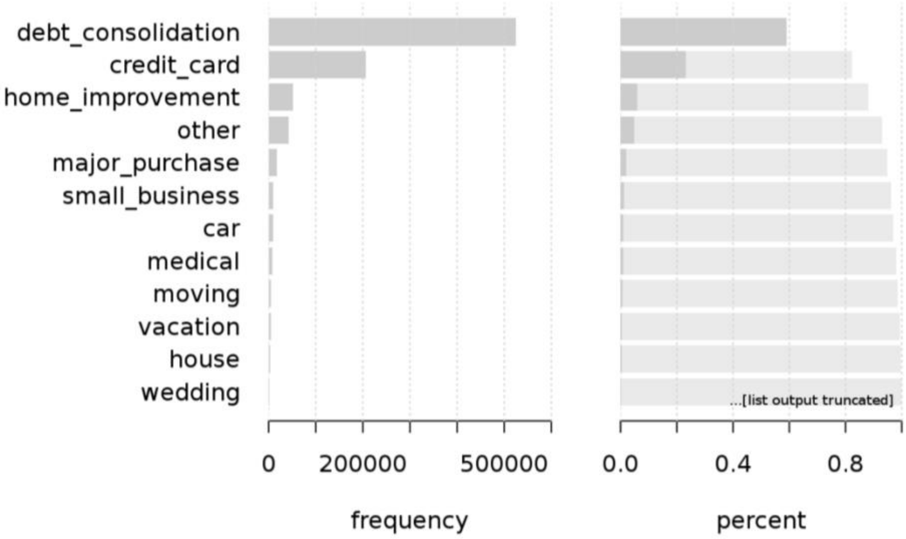
Appendix 1
Loan Grade



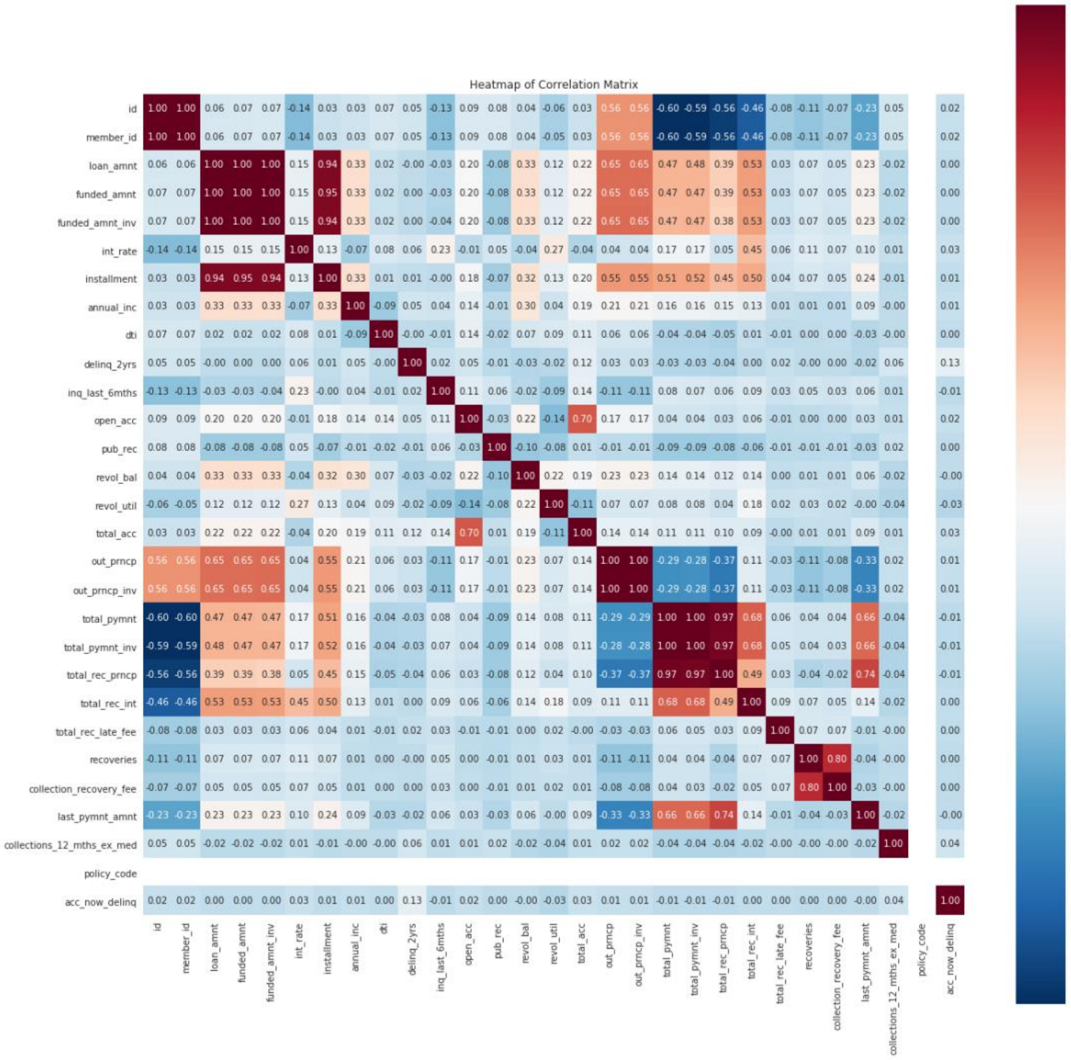
LOAN GRADE	SUB-GRADE	LENDING CLUB BASE RATE	ADJUSTMENT FOR RISK & VOLATILITY	INTEREST RATE
A	1	5.05%	0.26%	5.31%
	2	5.05%	1.02%	6.07%
	3	5.05%	1.66%	6.71%
	4	5.05%	2.29%	7.34%
	5	5.05%	2.91%	7.96%
B	1	5.05%	4.38%	9.43%
	2	5.05%	4.87%	9.92%
	3	5.05%	5.36%	10.41%
	4	5.05%	5.85%	10.90%
	5	5.05%	6.93%	11.98%
C	1	5.05%	7.56%	12.61%
	2	5.05%	8.53%	13.58%
	3	5.05%	9.02%	14.07%
	4	5.05%	9.99%	15.04%
	5	5.05%	10.96%	16.01%
D	1	5.05%	12.42%	17.47%
	2	5.05%	13.40%	18.45%
	3	5.05%	14.37%	19.42%
	4	5.05%	15.34%	20.39%
	5	5.05%	16.80%	21.85%
E	1	5.05%	17.85%	22.90%
	2	5.05%	18.82%	23.87%
	3	5.05%	19.79%	24.84%
	4	5.05%	20.76%	25.81%
	5	5.05%	21.72%	26.77%
F	1	5.05%	23.67%	28.72%
	2	5.05%	24.64%	29.69%
	3	5.05%	25.12%	30.17%
	4	5.05%	25.60%	30.65%
	5	5.05%	25.70%	30.75%
G	1	5.05%	25.74%	30.79%
	2	5.05%	25.79%	30.84%
	3	5.05%	25.84%	30.89%
	4	5.05%	25.89%	30.94%
	5	5.05%	25.94%	30.99%



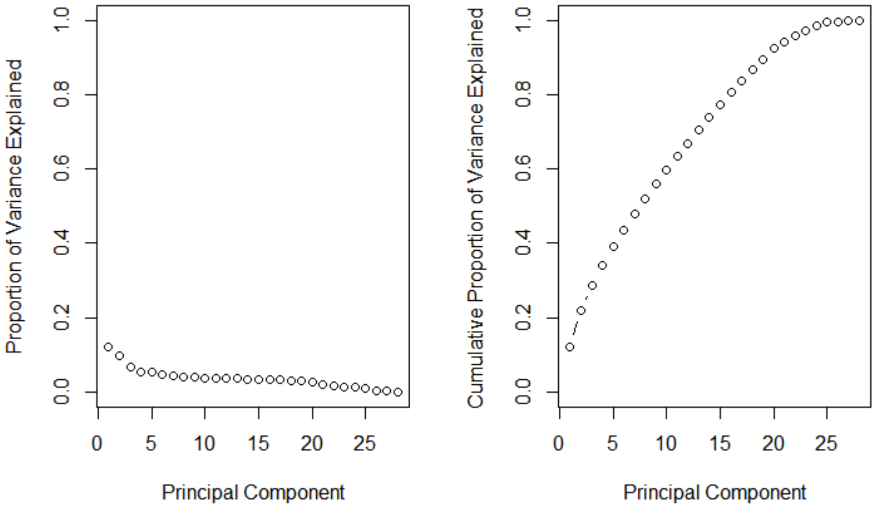
Appendix 2
Loan Purpose



Appendix 3
Correlation Matrix

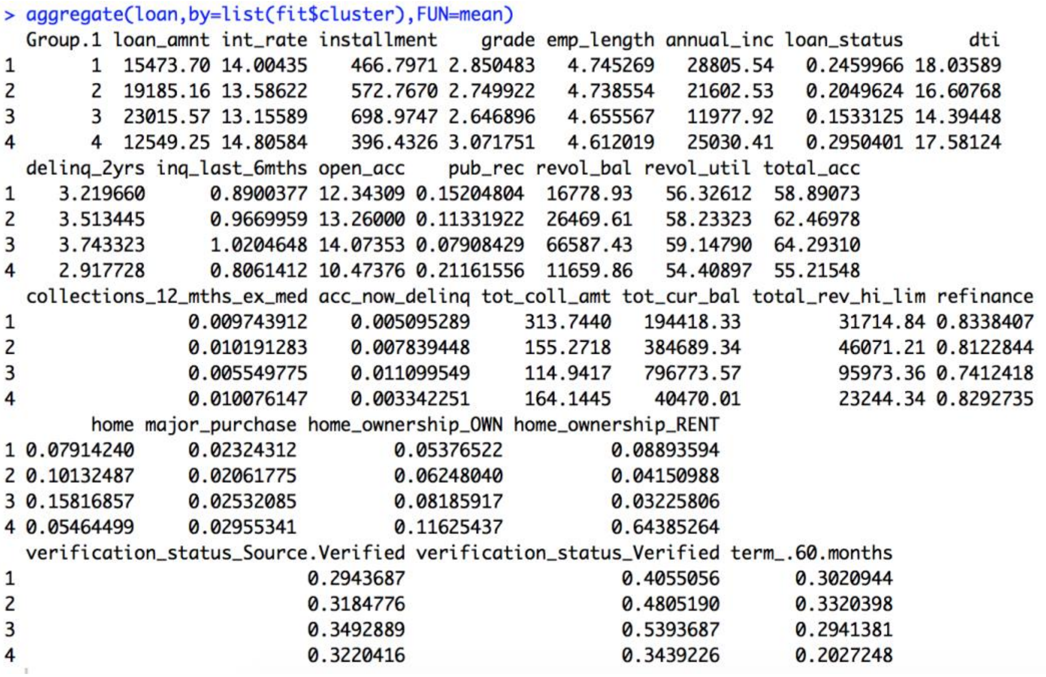


PCA Analysis



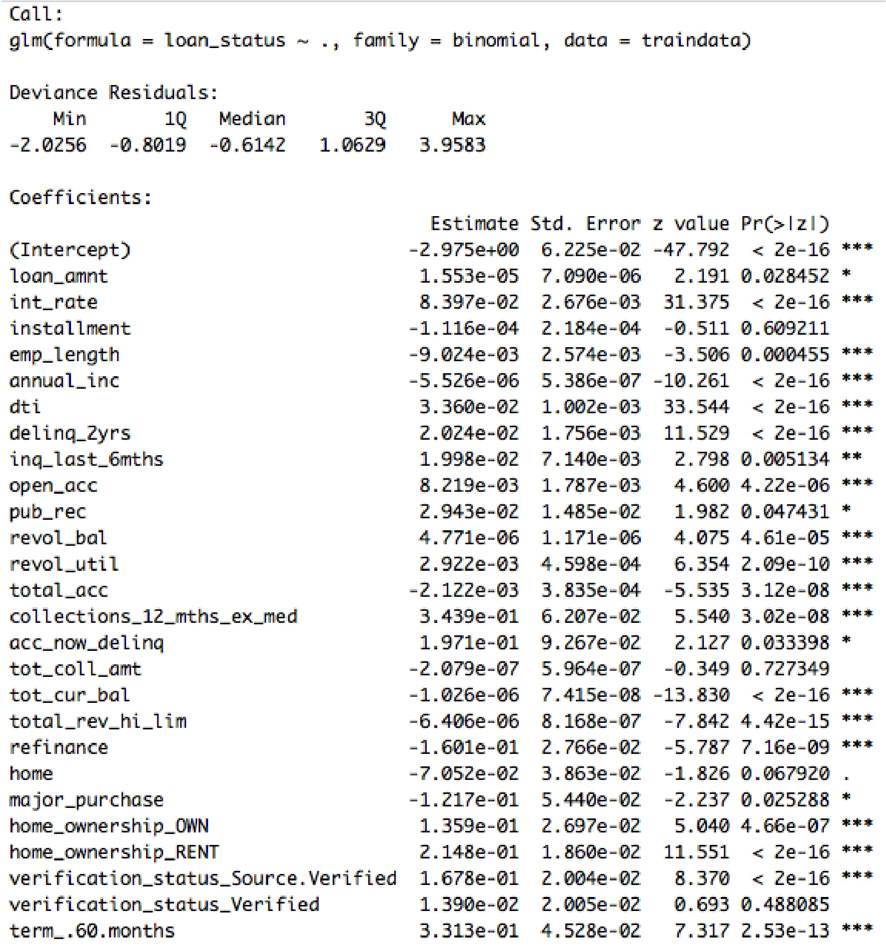
Appendix 4

Clustering Result



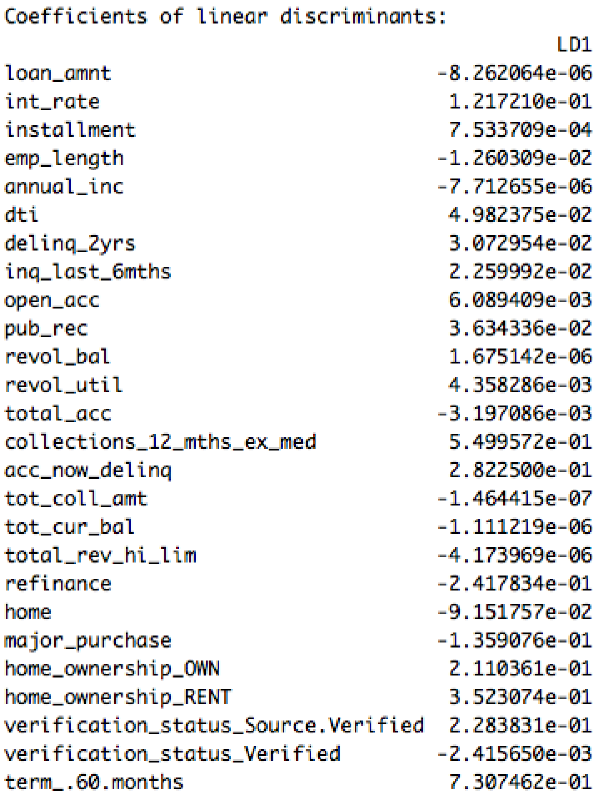
```
> aggregate(loan,by=list(fit$cluster),FUN=mean)
  Group.1 loan_amnt int_rate installment    grade emp_length annual_inc loan_status      dti
1       1  15473.70 14.00435    466.7971 2.850483   4.745269   28805.54   0.2459966 18.03589
2       2  19185.16 13.58622    572.7670 2.749922   4.738554   21602.53   0.2049624 16.60768
3       3  23015.57 13.15589    698.9747 2.646896   4.655567   11977.92   0.1533125 14.39448
4       4  12549.25 14.80584    396.4326 3.071751   4.612019   25030.41   0.2950401 17.58124
  delinq_2yrs inq_last_6mths open_acc    pub_rec revol_bal revol_util total_acc
1    3.219660      0.8900377 12.34309 0.15204804  16778.93   56.32612  58.89073
2    3.513445      0.9669959 13.26000 0.11331922  26469.61   58.23323  62.46978
3    3.743323      1.0204648 14.07353 0.07908429  66587.43   59.14790  64.29310
4    2.917728      0.8061412 10.47376 0.21161556  11659.86   54.40897  55.21548
  collections_12_mths_ex_med acc_now_delinq tot_coll_amt tot_cur_bal total_rev_hi_lim refinance
1                0.009743912    0.005095289     313.7440   194418.33         31714.84 0.8338407
2                0.010191283    0.007839448     155.2718   384689.34         46071.21 0.8122844
3                0.005549775    0.011099549     114.9417   796773.57         95973.36 0.7412418
4                0.010076147    0.003342251     164.1445    40470.01         23244.34 0.8292735
        home major_purchase home_ownership_OWN home_ownership_RENT
1 0.07914240     0.02324312         0.05376522          0.08893594
2 0.10132487     0.02061775         0.06248040          0.04150988
3 0.15816857     0.02532085         0.08185917          0.03225806
4 0.05464499     0.02955341         0.11625437          0.64385264
  verification_status_Source.Verified verification_status_Verified term_.60.months
1                           0.2943687                    0.4055056       0.3020944
2                           0.3184776                    0.4805190       0.3320398
3                           0.3492889                    0.5393687       0.2941381
4                           0.3220416                    0.3439226       0.2027248
```

Logistic Regression Result

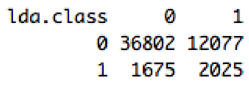


```
Call:
glm(formula = loan_status ~ ., family = binomial, data = traindata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0256  -0.8019  -0.6142   1.0629   3.9583  

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         -2.975e+00  6.225e-02 -47.792  < 2e-16 ***
loan_amnt                            1.553e-05  7.090e-06   2.191 0.028452 *  
int_rate                             8.397e-02  2.676e-03  31.375  < 2e-16 ***
installment                         -1.116e-04  2.184e-04  -0.511 0.609211    
emp_length                          -9.024e-03  2.574e-03  -3.506 0.000455 ***
annual_inc                          -5.526e-06  5.386e-07 -10.261  < 2e-16 ***
dti                                  3.360e-02  1.002e-03  33.544  < 2e-16 ***
delinq_2yrs                          2.024e-02  1.756e-03  11.529  < 2e-16 ***
inq_last_6mths                       1.998e-02  7.140e-03   2.798 0.005134 ** 
open_acc                             8.219e-03  1.787e-03   4.600 4.22e-06 ***
pub_rec                              2.943e-02  1.485e-02   1.982 0.047431 *  
revol_bal                            4.771e-06  1.171e-06   4.075 4.61e-05 ***
revol_util                           2.922e-03  4.598e-04   6.354 2.09e-10 ***
total_acc                           -2.122e-03  3.835e-04  -5.535 3.12e-08 ***
collections_12_mths_ex_med           3.439e-01  6.207e-02   5.540 3.02e-08 ***
acc_now_delinq                       1.971e-01  9.267e-02   2.127 0.033398 *  
tot_coll_amt                        -2.079e-07  5.964e-07  -0.349 0.727349    
tot_cur_bal                         -1.026e-06  7.415e-08 -13.830  < 2e-16 ***
total_rev_hi_lim                    -6.406e-06  8.168e-07  -7.842 4.42e-15 ***
refinance                           -1.601e-01  2.766e-02  -5.787 7.16e-09 ***
home                                -7.052e-02  3.863e-02  -1.826 0.067920 .  
major_purchase                      -1.217e-01  5.440e-02  -2.237 0.025288 *  
home_ownership_OWN                   1.359e-01  2.697e-02   5.040 4.66e-07 ***
home_ownership_RENT                  2.148e-01  1.860e-02  11.551  < 2e-16 ***
verification_status_Source.Verified  1.678e-01  2.004e-02   8.370  < 2e-16 ***
verification_status_Verified         1.390e-02  2.005e-02   0.693 0.488085    
term_.60.months                      3.313e-01  4.528e-02   7.317 2.53e-13 ***
```

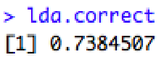
LDA Result



```
Coefficients of linear discriminants:
                                              LD1
loan_amnt                           -8.262064e-06
int_rate                             1.217210e-01
installment                          7.533709e-04
emp_length                          -1.260309e-02
annual_inc                          -7.712655e-06
dti                                  4.982375e-02
delinq_2yrs                          3.072954e-02
inq_last_6mths                       2.259992e-02
open_acc                             6.089409e-03
pub_rec                              3.634336e-02
revol_bal                            1.675142e-06
revol_util                           4.358286e-03
total_acc                           -3.197086e-03
collections_12_mths_ex_med           5.499572e-01
acc_now_delinq                       2.822500e-01
tot_coll_amt                        -1.464415e-07
tot_cur_bal                         -1.111219e-06
total_rev_hi_lim                    -4.173969e-06
refinance                           -2.417834e-01
home                                -9.151757e-02
major_purchase                      -1.359076e-01
home_ownership_OWN                   2.110361e-01
home_ownership_RENT                  3.523074e-01
verification_status_Source.Verified  2.283831e-01
verification_status_Verified        -2.415650e-03
term_.60.months                      7.307462e-01
```



```
lda.class     0     1
        0 36802 12077
        1  1675  2025
```



```
> lda.correct
[1] 0.7384507
```