Lending Club Loan Project: Variable Selection For Interest Rate Model

Alpha N.

12/20/2018

This is a complete but preliminary assessment of Lending Club loan data. This exercise is designed to identify the main variables that are used in determining the interest rate to charge borrowers. It could also benefit anyone trying to replicate the Lending Club assessment model.

Layout/Plan:

1.Clean data (~50-60% of work):

```
-Remove NA columns and rows
-Remove all columns obviously not necessary/repetitive in our analysis
-Create separate dataframes of 10,000 rows randomly selected from data
-These samples are to be used for training and prediction
```

2.Fit model for all variables (~25% of work):

```
-Study residuals
-Analyze with ANOVA(preliminary)
```

3.Conduct AIC via step function to choose model (~5-10% of work):

```
-Compare proposed AIC models
-Make any adjustments necessary
```

4.Evaluate final model using a prediction function(~10-15%)

Summary of Findings:

Overall, investors interested in crowdfunding with Lending Club should be advised that in deciding the interest rate to charge borrowers (and therefore determining the investors' return on investment), Lending Club will consider the loan amount, term, credit grade and income verification status (i.e is the declared income actually true) the most. This is probably reassuring because it covers the headline stuff. However, if the usual red flags like revolving account balances, total accounts, debt to income ratio, delinquency in the last two years are important to the investor, then Lending Club may not be the best place to invest because those do not seem to have a large effect on the interest rate decision. This may be in part because people with those were probably not approved in the first place.

Cleaning the Data

1.Load Data and dictionary terms. Have a cursory look

```r
projectpath <- "/Users/alpha/Documents/Loan Data/"
lendingdata<- read.csv(paste(projectpath,'Loan.csv',sep = '/'), header=TRUE) #Original file is 396MB
lendingdict<- read.csv(paste(projectpath, 'LCDataDictionary.csv',sep = '/'), header = TRUE)
names(lendingdata)
```

```
##  [1] "id"                          "member_id"
##  [3] "loan_amnt"                   "funded_amnt"
##  [5] "funded_amnt_inv"             "term"
##  [7] "int_rate"                    "installment"
##  [9] "grade"                       "sub_grade"
## [11] "emp_title"                   "emp_length"
## [13] "home_ownership"              "annual_inc"
## [15] "verification_status"         "issue_d"
## [17] "loan_status"                 "pymnt_plan"
## [19] "url"                         "desc"
## [21] "purpose"                     "title"
## [23] "zip_code"                    "addr_state"
## [25] "dti"                         "delinq_2yrs"
## [27] "earliest_cr_line"            "inq_last_6mths"
## [29] "mths_since_last_delinq"      "mths_since_last_record"
## [31] "open_acc"                    "pub_rec"
## [33] "revol_bal"                   "revol_util"
## [35] "total_acc"                   "initial_list_status"
## [37] "out_prncp"                   "out_prncp_inv"
## [39] "total_pymnt"                 "total_pymnt_inv"
## [41] "total_rec_prncp"             "total_rec_int"
## [43] "total_rec_late_fee"          "recoveries"
## [45] "collection_recovery_fee"     "last_pymnt_d"
## [47] "last_pymnt_amnt"             "next_pymnt_d"
## [49] "last_credit_pull_d"          "collections_12_mths_ex_med"
## [51] "mths_since_last_major_derog" "policy_code"
## [53] "application_type"            "annual_inc_joint"
## [55] "dti_joint"                   "verification_status_joint"
## [57] "acc_now_delinq"              "tot_coll_amt"
## [59] "tot_cur_bal"                 "open_acc_6m"
## [61] "open_il_6m"                  "open_il_12m"
## [63] "open_il_24m"                 "mths_since_rcnt_il"
## [65] "total_bal_il"                "il_util"
## [67] "open_rv_12m"                 "open_rv_24m"
## [69] "max_bal_bc"                  "all_util"
## [71] "total_rev_hi_lim"            "inq_fi"
## [73] "total_cu_tl"                 "inq_last_12m"
```

```r
dim(lendingdata)
```

```
## [1] 887379     74
```

2.Clean up the data AND create separate dataframes of cleaned data. Preserve data forms as you go

```r
lendingdata <- lendingdata[,-c(48:74)]
drops <- c("last_pymnt_d","initial_list_status", "mths_since_last_record","mths_since_last_delinq",
           "desc", "url","title", "pymnt_plan", "emp_title","id", "member_id", "issue_d", "addr_state")
relevent_loan_data <- lendingdata[ , !(names(lendingdata) %in% drops)] #take out columns not needed
dim(relevent_loan_data)
```

```
## [1] 887379     34
```

3.Create pre-disbursement dataframe. This is all the information available before loan is disbursed to borrower. Omit NA and save CSV file for future use

```r
drop_after_data<- c("installment","funded_amnt", "funded_amnt_inv", "loan_status", "total_pymnt", "tota
                     "out_prncp_inv", "out_prncp", "total_rec_late_fee")
```

```r
pre_loan_data<- relevent_loan_data [,!(names(relevent_loan_data ) %in% drop_after_data)] #select data b
pre_loan_data<- na.omit(pre_loan_data)
dim(pre_loan_data)
```

```
## [1] 886877     20
```

```r
write.csv(pre_loan_data, 'Pre_Loan_Data_Large.csv') # File is now 133MB
```

4.Randomly select samples of 10,000 of 887,379 rows to reduce computation time. Sample size is ideal for local machine. Repeat NA omission and explicitly use data.frame for good measure. Save CSV file for future use. Please note that due to the large volume of data, we do not need to impute any data with missing variables. This is not necessary for the scope of this project. Instead, we will select use observations(rows) that are complete.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
pre_loan_data_random<- sample_n(pre_loan_data, 10000) # load dplyr for sample_n to work
pre_loan_data_random<- na.omit(pre_loan_data_random) #remove NA cells
pre_loan_data_random<- data.frame(pre_loan_data_random)
dim(pre_loan_data_random)
```

```
## [1] 10000    20
```

```r
write.csv(pre_loan_data_random, 'Pre_Loan_Data_Sample1.csv')#Final file is 1.5MB!!
```

```r
pre_loan_data_random2<- sample_n(pre_loan_data, 10000)
pre_loan_data_random2<- na.omit(pre_loan_data_random2) #remove NA cells
pre_loan_data_random2<- data.frame(pre_loan_data_random2)
dim(pre_loan_data_random2)
```

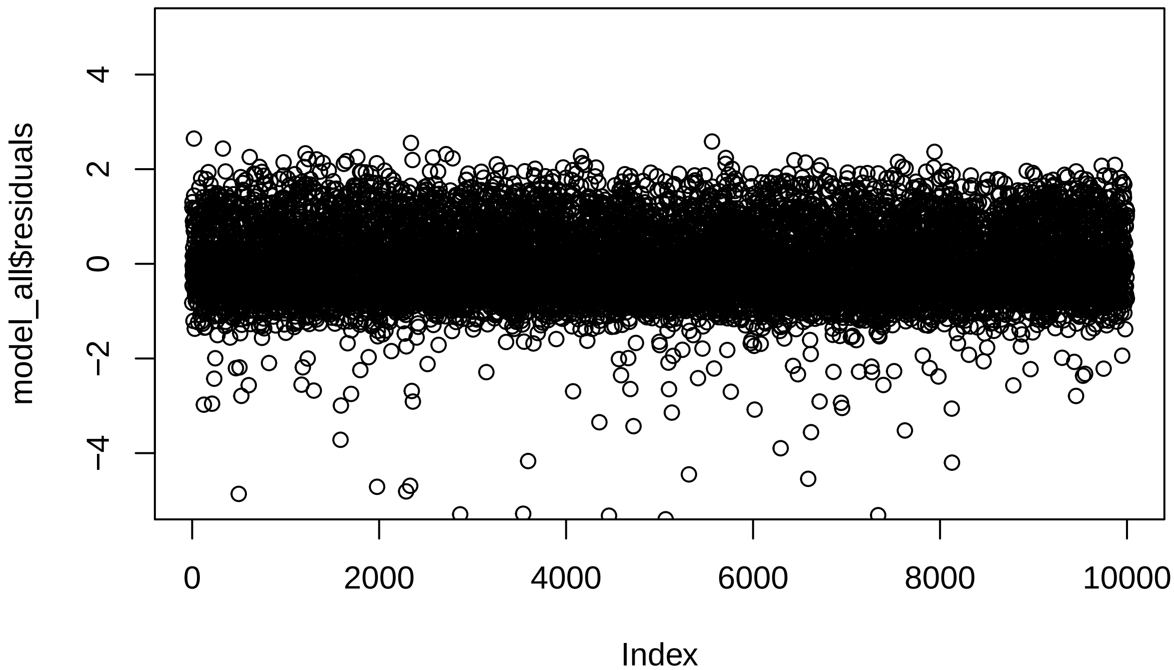
```
## [1] 10000    20
```

```r
write.csv(pre_loan_data_random2, 'Pre_Loan_Data_Sample2.csv') #Also 1.5MB
```

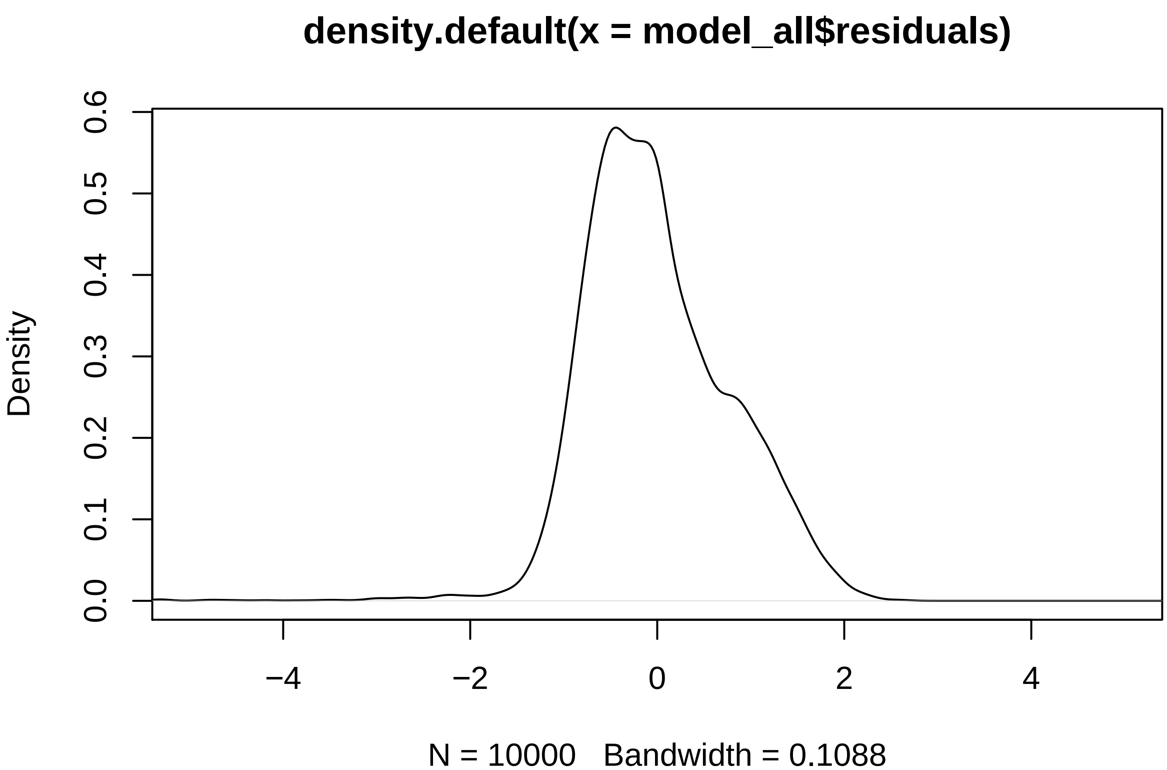
Fitting the Full Model

5.Fit all variables available

```r
model_all<- lm(int_rate ~ ., data = pre_loan_data_random, na.action = na.exclude) #This model takes a l
plot(model_all$residuals, ylim = c(-5,5))
```




```r
plot(density(model_all$residuals), xlim=c(-5,5))
```




```r
#plots residuals on the Y axis and fitted values on the X axis.
```

Residuals appear concentrated at the center. The good news here is that the model works as expected - the density plot suggests normality as well. It is a little concerning that that there is a slight negative skew on the residuals. That may need a little more analysis. Still, most of our variance is where we want it to be for our purposes - around zero.

6.Test of relative significance with ANOVA to help trim the model to only efficient variables

```r
summary(model_all)$r.squared
```

```
## [1] 0.967658
```

```r
anova(model_all)
```

```
## Analysis of Variance Table
##
## Response: int_rate
##                      Df Sum Sq Mean Sq    F value    Pr(>F)    
## loan_amnt             1   3447  3447.4  4819.7807 < 2.2e-16 ***
## term                  1  31574 31574.2 44143.3501 < 2.2e-16 ***
## grade                 6 138977 23162.8 32383.4433 < 2.2e-16 ***
## sub_grade            28   9248   330.3   461.7450 < 2.2e-16 ***
## emp_length           11     25     2.3     3.2283 0.0002083 ***
## home_ownership        3     53    17.7    24.6838 6.748e-16 ***
## annual_inc            1      7     7.3    10.2164 0.0013970 ** 
## verification_status   2    134    67.2    94.0209 < 2.2e-16 ***
## purpose              13     27     2.1     2.8837 0.0003546 ***
## zip_code            788    634     0.8     1.1252 0.0109135 *  
## dti                   1     26    25.9    36.2765 1.782e-09 ***
## delinq_2yrs           1     11    10.8    15.0424 0.0001059 ***
## earliest_cr_line    508    414     0.8     1.1406 0.0180267 *  
## inq_last_6mths        1     18    17.6    24.5923 7.219e-07 ***
## open_acc              1      8     8.4    11.7572 0.0006089 ***
## pub_rec               1     36    35.7    49.8530 1.785e-12 ***
## revol_bal             1      2     2.3     3.1537 0.0757917 .  
## revol_util            1     22    22.4    31.3799 2.187e-08 ***
## total_acc             1      0     0.0     0.0004 0.9849248    
## Residuals          8629   6172     0.7                         
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA comparison above, it appears loan amount, term, grade, and subgrade have the biggest bearing on our data. We need to trim our model accordingly.

Notice that income verification status has a greater variance on interest rate than actual income. The usual suspects - revolving account balances, total accounts, debt to income ratio, delinquency in the last two years - all seem not to have that large of an effect, comparatively speaking. It is possible that there is a survival bias i.e what we have here are people who were approved already, so the data above may already be favorable.

Variable Selection for Model

7.Use forward and backward AIC

```r
step(model_all, direction = "forward")$AIC #output narrowed down to prevent massive multi-page printout
```

```
## Start:  AIC=-2083.58
## int_rate ~ loan_amnt + term + grade + sub_grade + emp_length + 
##     home_ownership + annual_inc + verification_status + purpose + 
##     zip_code + dti + delinq_2yrs + earliest_cr_line + inq_last_6mths + 
##     open_acc + pub_rec + revol_bal + revol_util + total_acc
```

```
## NULL
```

```r
step(model_all, direction = "backward")
```

```
## Start:  AIC=-2083.58
## int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
##     home_ownership + annual_inc + verification_status + purpose +
##     zip_code + dti + delinq_2yrs + earliest_cr_line + inq_last_6mths +
##     open_acc + pub_rec + revol_bal + revol_util + total_acc
##
##
## Step:  AIC=-2083.58
## int_rate ~ loan_amnt + term + sub_grade + emp_length + home_ownership +
##     annual_inc + verification_status + purpose + zip_code + dti +
##     delinq_2yrs + earliest_cr_line + inq_last_6mths + open_acc +
##     pub_rec + revol_bal + revol_util + total_acc
##
##                        Df Sum of Sq   RSS     AIC
## - zip_code            787       601  6773 -2728.9
## - earliest_cr_line    508       406  6578 -2462.8
## - total_acc             1         0  6172 -2085.6
## - loan_amnt             1         0  6172 -2085.1
## - open_acc              1         1  6173 -2083.7
## <none>                               6172 -2083.6
## - revol_bal             1         6  6178 -2075.7
## - annual_inc            1         6  6178 -2075.3
## - purpose              13        27  6199 -2066.7
## - delinq_2yrs           1        14  6186 -2063.4
## - term                  1        19  6191 -2054.4
## - dti                   1        20  6192 -2053.8
## - emp_length           11        34  6206 -2051.4
## - revol_util            1        22  6194 -2049.3
## - inq_last_6mths        1        27  6199 -2041.6
## - pub_rec               1        33  6205 -2032.2
## - home_ownership        3        45  6217 -2017.1
## - verification_status   2       109  6281 -1912.0
## - sub_grade            34     87301 93473 25024.9
##
## Step:  AIC=-2728.89
## int_rate ~ loan_amnt + term + sub_grade + emp_length + home_ownership +
##     annual_inc + verification_status + purpose + dti + delinq_2yrs +
##     earliest_cr_line + inq_last_6mths + open_acc + pub_rec +
##     revol_bal + revol_util + total_acc
##
##                        Df Sum of Sq    RSS     AIC
## - earliest_cr_line    509       424   7197 -3139.4
## - total_acc             1         0   6773 -2730.9
## - loan_amnt             1         0   6773 -2730.3
## - open_acc              1         1   6774 -2728.9
## <none>                                6773 -2728.9
## - revol_bal             1         7   6779 -2721.3
## - annual_inc            1         7   6779 -2721.2
## - purpose              13        28   6801 -2713.0
## - delinq_2yrs           1        16   6788 -2707.6
## - emp_length           11        31   6804 -2705.4
## - term                  1        19   6792 -2702.9
```

```
## - dti                  1        19   6792 -2702.4
## - revol_util           1        25   6798 -2693.5
## - inq_last_6mths       1        34   6807 -2680.3
## - pub_rec              1        35   6808 -2679.3
## - home_ownership       3        48   6820 -2664.8
## - verification_status  2       115   6887 -2565.0
## - sub_grade           34     95537 102310 24354.2
##
## Step:  AIC=-3139.43
## int_rate ~ loan_amnt + term + sub_grade + emp_length + home_ownership +
##     annual_inc + verification_status + purpose + dti + delinq_2yrs +
##     inq_last_6mths + open_acc + pub_rec + revol_bal + revol_util +
##     total_acc
##
##                       Df Sum of Sq    RSS     AIC
## - loan_amnt            1         0   7197 -3141.4
## <none>                               7197 -3139.4
## - revol_bal            1         4   7201 -3135.9
## - total_acc            1         5   7202 -3134.8
## - open_acc             1         5   7202 -3134.8
## - purpose             13        27   7224 -3128.2
## - annual_inc           1        10   7206 -3128.1
## - delinq_2yrs          1        13   7210 -3123.3
## - term                 1        18   7214 -3117.1
## - emp_length          11        32   7229 -3116.8
## - dti                  1        25   7222 -3106.8
## - pub_rec              1        27   7224 -3103.4
## - inq_last_6mths       1        34   7230 -3094.8
## - revol_util           1        39   7236 -3087.7
## - home_ownership       3        49   7246 -3076.9
## - verification_status  2       129   7326 -2966.0
## - sub_grade           34    102860 110057 24066.1
##
## Step:  AIC=-3141.39
## int_rate ~ term + sub_grade + emp_length + home_ownership + annual_inc +
##     verification_status + purpose + dti + delinq_2yrs + inq_last_6mths +
##     open_acc + pub_rec + revol_bal + revol_util + total_acc
##
##                       Df Sum of Sq    RSS     AIC
## <none>                               7197 -3141.4
## - revol_bal            1         4   7201 -3137.9
## - open_acc             1         5   7202 -3136.8
## - total_acc            1         5   7202 -3136.7
## - purpose             13        27   7224 -3129.7
## - annual_inc           1        10   7207 -3128.9
## - delinq_2yrs          1        13   7210 -3125.2
## - emp_length          11        32   7229 -3118.6
## - term                 1        19   7216 -3117.2
## - dti                  1        25   7222 -3108.8
## - pub_rec              1        28   7225 -3105.0
## - inq_last_6mths       1        34   7230 -3096.8
## - revol_util           1        39   7236 -3089.7
## - home_ownership       3        49   7246 -3078.9
## - verification_status  2       131   7328 -2965.1
```

```
## - sub_grade           34    102999 110196 24076.7
##
## Call:
## lm(formula = int_rate ~ term + sub_grade + emp_length + home_ownership +
##     annual_inc + verification_status + purpose + dti + delinq_2yrs +
##     inq_last_6mths + open_acc + pub_rec + revol_bal + revol_util +
##     total_acc, data = pre_loan_data_random, na.action = na.exclude)
##
## Coefficients:
##                 (Intercept)                    term 60 months
##                   5.695e+00                        -1.132e-01
##                 sub_gradeA2                       sub_gradeA3
##                   7.098e-01                         1.429e+00
##                 sub_gradeA4                       sub_gradeA5
##                   1.754e+00                         2.546e+00
##                 sub_gradeB1                       sub_gradeB2
##                   3.188e+00                         4.216e+00
##                 sub_gradeB3                       sub_gradeB4
##                   5.154e+00                         6.040e+00
##                 sub_gradeB5                       sub_gradeC1
##                   6.514e+00                         7.168e+00
##                 sub_gradeC2                       sub_gradeC3
##                   7.595e+00                         8.229e+00
##                 sub_gradeC4                       sub_gradeC5
##                   8.815e+00                         9.556e+00
##                 sub_gradeD1                       sub_gradeD2
##                   1.035e+01                         1.115e+01
##                 sub_gradeD3                       sub_gradeD4
##                   1.159e+01                         1.219e+01
##                 sub_gradeD5                       sub_gradeE1
##                   1.272e+01                         1.303e+01
##                 sub_gradeE2                       sub_gradeE3
##                   1.361e+01                         1.401e+01
##                 sub_gradeE4                       sub_gradeE5
##                   1.509e+01                         1.580e+01
##                 sub_gradeF1                       sub_gradeF2
##                   1.689e+01                         1.741e+01
##                 sub_gradeF3                       sub_gradeF4
##                   1.828e+01                         1.883e+01
##                 sub_gradeF5                       sub_gradeG1
##                   1.887e+01                         1.947e+01
##                 sub_gradeG2                       sub_gradeG3
##                   2.049e+01                         2.017e+01
##                 sub_gradeG4                       sub_gradeG5
##                   1.985e+01                         2.079e+01
##            emp_length1 year                emp_length10+ years
##                   2.853e-02                         7.677e-02
##           emp_length2 years                  emp_length3 years
##                  -1.456e-02                         8.912e-03
##           emp_length4 years                  emp_length5 years
##                   7.114e-02                         1.144e-01
##           emp_length6 years                  emp_length7 years
##                   1.805e-01                         1.321e-01
```

```
##                  emp_length8 years                        emp_length9 years
##                          9.815e-02                                1.907e-02
##                      emp_lengthn/a                      home_ownershipOTHER
##                         -7.556e-02                               -5.297e+00
##                  home_ownershipOWN                       home_ownershipRENT
##                         -5.948e-02                                6.446e-03
##                         annual_inc  verification_statusSource Verified
##                         -7.851e-07                               -1.025e-01
##        verification_statusVerified                       purposecredit_card
##                          1.766e-01                                2.755e-02
##          purposedebt_consolidation                       purposeeducational
##                          6.630e-02                               -1.156e+00
##           purposehome_improvement                             purposehouse
##                          3.060e-02                                3.163e-01
##             purposemajor_purchase                            purposemedical
##                          1.135e-01                               -1.024e-01
##                     purposemoving                             purposeother
##                          1.248e-01                               -4.112e-02
##           purposerenewable_energy                    purposesmall_business
##                         -4.938e-02                                3.544e-03
##                   purposevacation                           purposewedding
##                         -4.790e-02                                4.237e-01
##                                dti                              delinq_2yrs
##                         -7.015e-03                               -4.213e-02
##                     inq_last_6mths                                 open_acc
##                          6.203e-02                               -6.054e-03
##                            pub_rec                                revol_bal
##                         -9.169e-02                               -1.143e-06
##                         revol_util                                total_acc
##                          3.082e-03                                2.711e-03
```

It appears the best model is the one that is suggested by backward selection, but without 'verification status.' The forward and backward methods do not yield very different results though

8.Use the AIC model suggested by the step function

```r
aic_model_forward<- lm(int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
                         home_ownership + annual_inc + verification_status + purpose +
                         zip_code + dti + delinq_2yrs + earliest_cr_line + inq_last_6mths +
                         open_acc + pub_rec + revol_bal + revol_util + total_acc, data = pre_loan_data_

aic_model_backward<- lm(int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
                          home_ownership + annual_inc + purpose +
                          zip_code + dti + delinq_2yrs + earliest_cr_line + inq_last_6mths +
                          open_acc + pub_rec + revol_bal + revol_util + total_acc, data = pre_loan_data_

summary(aic_model_forward)$r.squared
```

```
## [1] 0.967658
```

```r
summary(aic_model_backward)$r.squared
```

```
## [1] 0.9670852
```

9.Compare Models AIC models

```r
anova(aic_model_forward,aic_model_backward) #So it appears we will go with the backward model
```

```
## Analysis of Variance Table
##
## Model 1: int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
##     home_ownership + annual_inc + verification_status + purpose +
##     zip_code + dti + delinq_2yrs + earliest_cr_line + inq_last_6mths +
##     open_acc + pub_rec + revol_bal + revol_util + total_acc
## Model 2: int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
##     home_ownership + annual_inc + purpose + zip_code + dti +
##     delinq_2yrs + earliest_cr_line + inq_last_6mths + open_acc +
##     pub_rec + revol_bal + revol_util + total_acc
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   8629 6172.0
## 2   8631 6281.3 -2   -109.32 76.419 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confirming our interpretation of the AIC numbers, an ANOVA comparison of both models confirms the backward model to be marginally better.

10.Compare AIC model with original model

```r
anova(model_all,aic_model_backward)
```

```
## Analysis of Variance Table
##
## Model 1: int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
##     home_ownership + annual_inc + verification_status + purpose +
##     zip_code + dti + delinq_2yrs + earliest_cr_line + inq_last_6mths +
##     open_acc + pub_rec + revol_bal + revol_util + total_acc
## Model 2: int_rate ~ loan_amnt + term + grade + sub_grade + emp_length +
##     home_ownership + annual_inc + purpose + zip_code + dti +
##     delinq_2yrs + earliest_cr_line + inq_last_6mths + open_acc +
##     pub_rec + revol_bal + revol_util + total_acc
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   8629 6172.0
## 2   8631 6281.3 -2   -109.32 76.419 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the same vein, variable selection not only results in a more efficient model, but the predictive value is also increased relative to the original model, 'model_all.'

Using our model for prediction

11.Preliminary model validation: now let us use our model to predict interest rates for the second sample(i.e out-of-sample). This is a preliminary look at the model's quality. Please note that model quality testing is a far more extensive task beyond the scope of this project.

```r
prediction <- predict(aic_model_backward, data = pre_loan_data_random2, na.action = na.action)
raw_data_rate<- c(summary(pre_loan_data_random2$int_rate))
predicted_rate<- c(summary(prediction))
rbind(raw_data_rate, predicted_rate) #summary comparison
```

```
##                     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

```
## raw_data_rate  5.320000  9.99000 12.99000 13.23177 15.88000 28.99000
## predicted_rate 4.890174 10.10638 13.06015 13.31237 16.18514 26.95149
```

```r
sd(pre_loan_data_random2$int_rate)
```

```
## [1] 4.34621
```

```r
sd(prediction)
```

```
## [1] 4.296203
```

Final Thoughts and Future Work

As you can see, the SD and the summary all appear to be very close to the raw data. This is a good sign for anyone that might attempt to replicate the Lending Club system. Still, this model may need more robust validation than just ANOVA and summary statistics, so it is not a . A confirmation of what we see via PCA analysis might be a good plan as well.

We also need to study the effect of zip code on the interest rate. While evaluating the summary models, there appeared to be an odd relationship between interest rate and some zip codes in counties in CT and NJ. This quirk is definitely worth exploring.