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

```
projectpath <- "/Users/alpha/Documents/Project Data/"
lendingdata<- read.csv(paste(projectpath,'Loan.csv',sep = '/'), header=TRUE) #Origina
l 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"
```

```
dim(lendingdata)
```

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

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

```
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") #identify 'useless' columns
relevent_loan_data <- lendingdata[ , !(names(lendingdata) %in% drops)] #take out 'use
less' columns
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

```
drop_after_data<- c("installment","funded_amnt", "funded_amnt_inv", "loan_status", "t
otal_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "tot_rec_late_fee
","recoveries", "collection_recovery_fee", "last_pymnt_amnt",
                    "out_prncp_inv", "out_prncp", "total_rec_late_fee")
pre_loan_data<- relevent_loan_data [,!(names(relevent_loan_data ) %in% drop_after_dat
a)] #select data before loan given
pre_loan_data<- na.omit(pre_loan_data)
dim(pre_loan_data)
```

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

```
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

```
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
```

```
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
```

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

```
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)
```

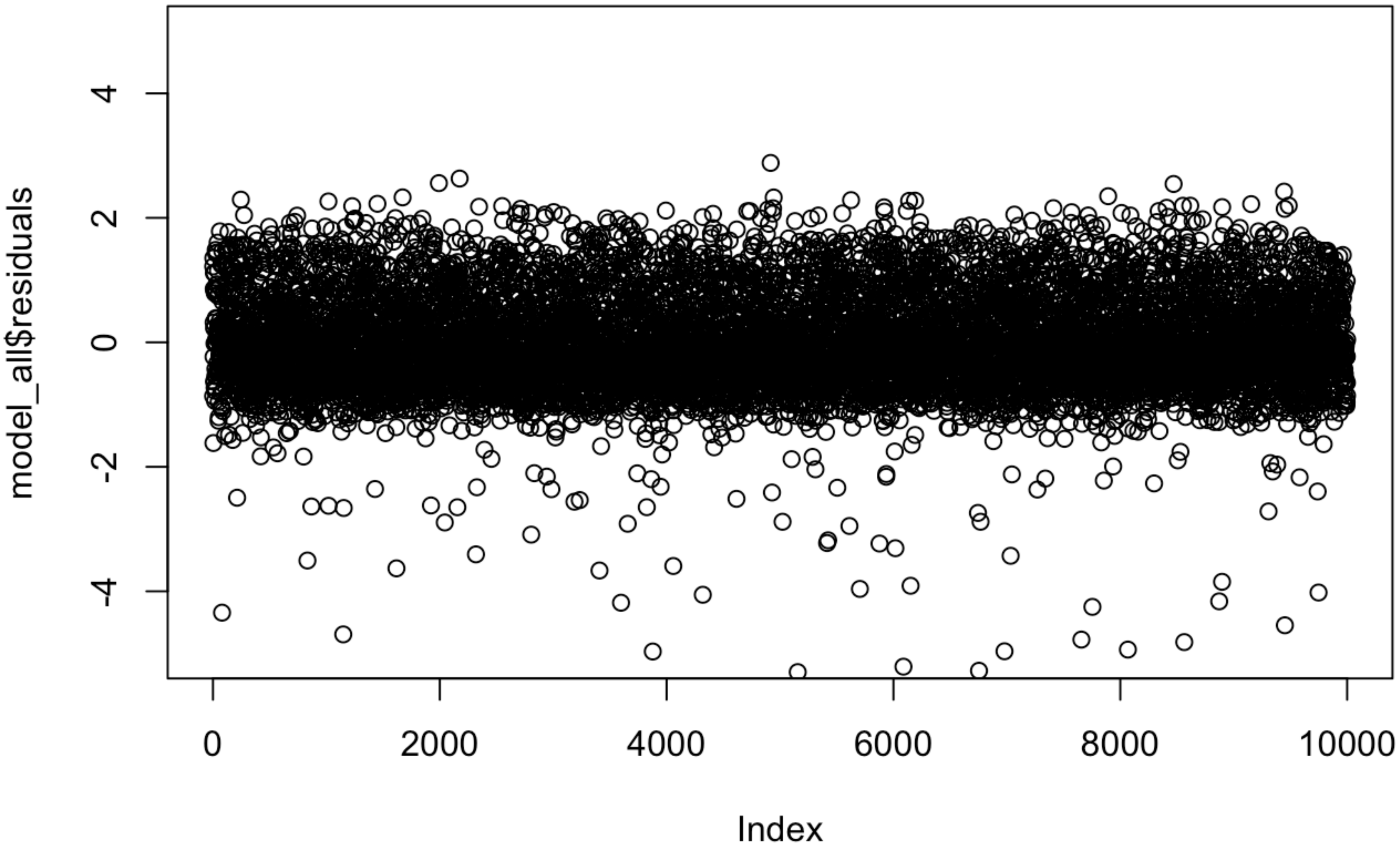
```
## [1] 10000    20
```

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

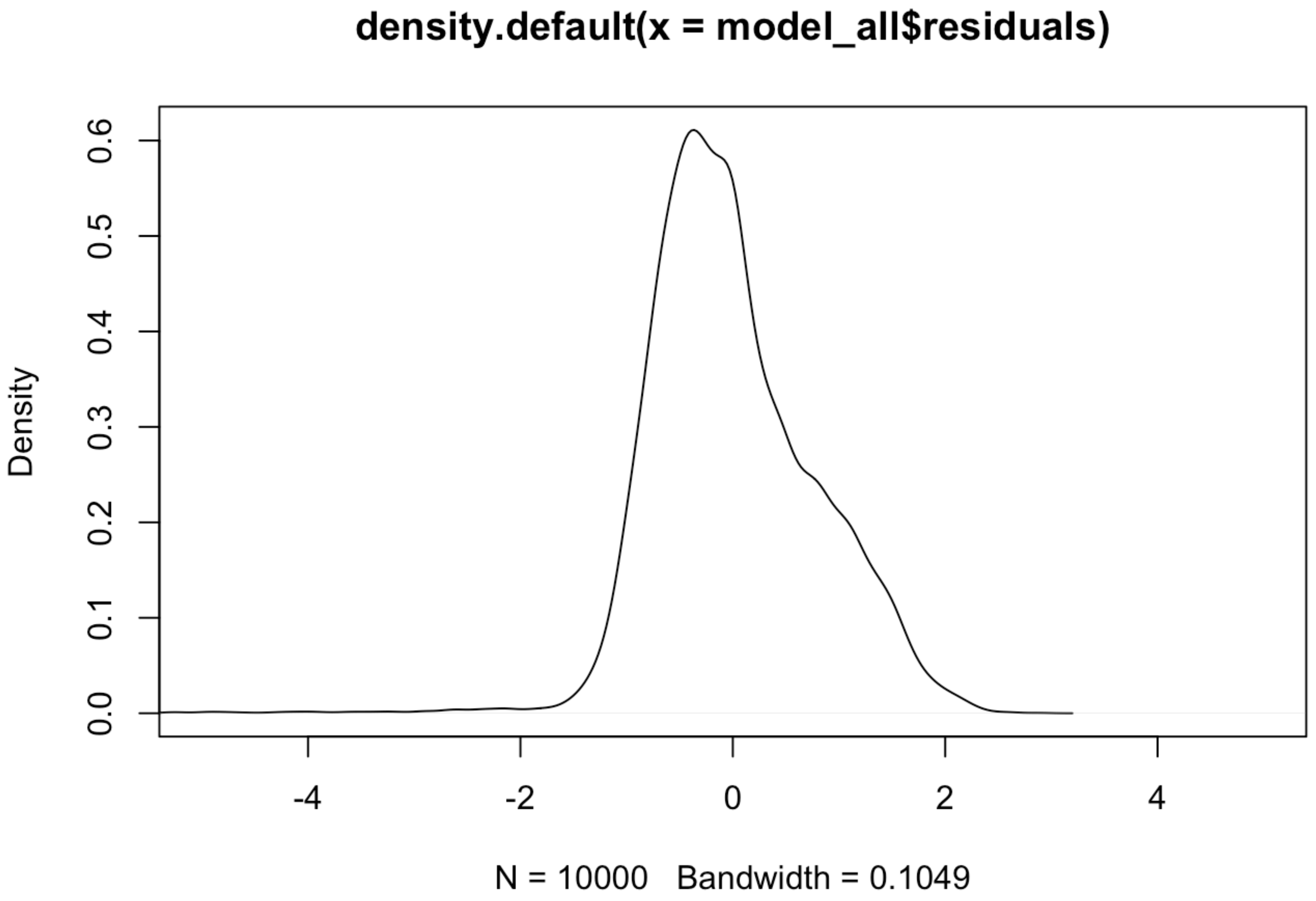
Fitting the Full Model

5.Fit all variables

```
model_all<- lm(int_rate ~ ., data = pre_loan_data_random, na.action = na.exclude) #This model takes a long time to run due to all the variables
plot(model_all$residuals, ylim = c(-5,5))
```



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



```
#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

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

```
## [1] 0.9685161
```

```
anova(model_all)
```

```
## Analysis of Variance Table
##
## Response: int_rate
##                       Df Sum Sq Mean Sq    F value    Pr(>F)    
## loan_amnt              1   4729    4729  6634.7941 < 2.2e-16 ***
## term                   1  32494   32494 45592.8629 < 2.2e-16 ***
## grade                  6 141276   23546 33037.9957 < 2.2e-16 ***
## sub_grade             28   9300     332   466.0503 < 2.2e-16 ***
## emp_length            11     40       4     5.0684 6.011e-08 ***
## home_ownership         3      7       2     3.0616 0.0269856 *  
## annual_inc             1      2       2     2.3132 0.1283154    
## verification_status    2    108      54    75.8203 < 2.2e-16 ***
## purpose               13     48       4     5.1648 2.941e-09 ***
## zip_code             786    596       1     1.0645 0.1130383    
## dti                    1     34      34    47.2848 6.566e-12 ***
## delinq_2yrs            1      6       6     7.8655 0.0050500 ** 
## earliest_cr_line     510    443       1     1.2193 0.0007180 ***
## inq_last_6mths         1      0       0     0.4113 0.5213283    
## open_acc               1      9       9    12.2951 0.0004564 ***
## pub_rec                1     33      33    46.1618 1.161e-11 ***
## revol_bal              1      0       0     0.5505 0.4581479    
## revol_util             1     54      54    75.8018 < 2.2e-16 ***
## total_acc              1      5       5     7.6246 0.0057698 ** 
## Residuals           8629   6150       1                         
## ---
## 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

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

```
## Start:  AIC=-2119.56
## 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
```

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

```
## Start:  AIC=-2119.56
## 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=-2119.56
## 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           784       593  6743 -2766.5
## - earliest_cr_line   510       426  6576 -2469.2
## - home_ownership       3         2  6152 -2121.6
## - loan_amnt            1         0  6150 -2121.2
## <none>                              6150 -2119.6
## - inq_last_6mths       1         4  6154 -2115.6
## - annual_inc           1         4  6154 -2114.7
## - open_acc             1         5  6155 -2113.6
## - revol_bal            1         5  6155 -2113.6
## - total_acc            1         5  6155 -2112.7
## - delinq_2yrs          1         8  6158 -2108.8
## - emp_length          11        32  6182 -2089.9
## - purpose             13        35  6185 -2088.5
## - pub_rec              1        26  6176 -2079.2
## - term                 1        30  6180 -2072.6
## - dti                  1        33  6183 -2068.4
## - revol_util           1        55  6205 -2032.4
## - verification_status  2        91  6241 -1976.3
## - sub_grade           34     87063 93213 24997.0
## 
## Step:  AIC=-2766.5
## 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   512       423   7166 -3182.5
## - home_ownership       3         2   6746 -2768.8
## - loan_amnt            1         0   6743 -2768.1
## <none>                                6743 -2766.5
## - open_acc              1         3   6746 -2764.4
## - inq_last_6mths        1         3   6746 -2763.6
## - total_acc             1         4   6747 -2762.2
## - annual_inc            1         5   6748 -2761.5
## - revol_bal             1         8   6751 -2756.9
## - delinq_2yrs           1        11   6754 -2752.1
## - purpose              13        37   6780 -2738.5
## - emp_length           11        36   6779 -2735.7
## - pub_rec               1        26   6770 -2729.4
## - term                  1        35   6778 -2716.5
## - dti                   1        37   6780 -2713.8
## - revol_util            1        58   6801 -2682.9
## - verification_status   2        96   6839 -2629.7
## - sub_grade            34     96006 102749 24403.1
##
## Step:  AIC=-3182.53
## 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
## - home_ownership        3         2   7168 -3185.2
## - loan_amnt             1         0   7166 -3184.5
## <none>                               7166 -3182.5
## - annual_inc            1         5   7171 -3177.0
## - inq_last_6mths        1         6   7172 -3176.7
## - revol_bal             1         6   7172 -3175.9
## - open_acc              1         8   7174 -3173.6
## - delinq_2yrs           1        10   7175 -3171.2
## - total_acc             1        17   7183 -3161.2
## - pub_rec               1        21   7187 -3155.4
## - purpose              13        41   7207 -3151.2
## - emp_length           11        42   7208 -3146.3
## - term                  1        31   7197 -3141.3
## - dti                   1        44   7210 -3123.6
## - revol_util            1        70   7236 -3087.5
## - verification_status   2       112   7278 -3031.7
## - sub_grade            34    103743 110909 24143.3
##
## Step:  AIC=-3185.21
## int_rate ~ loan_amnt + term + sub_grade + emp_length + 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   7168 -3187.2
## <none>                               7168 -3185.2
## - annual_inc            1         5   7174 -3179.7
## - inq_last_6mths       1         6   7174 -3179.1
## - revol_bal            1         7   7175 -3178.0
## - open_acc             1         8   7176 -3176.5
## - delinq_2yrs          1        10   7178 -3173.6
## - total_acc            1        17   7185 -3163.8
## - pub_rec              1        21   7189 -3157.9
## - purpose             13        42   7210 -3153.3
## - emp_length          11        42   7210 -3148.8
## - term                 1        32   7200 -3143.3
## - dti                  1        44   7212 -3125.9
## - revol_util           1        72   7240 -3087.6
## - verification_status  2       112   7280 -3034.0
## - sub_grade           34    104737 111906 24226.7
## 
## Step:  AIC=-3187.18
## int_rate ~ term + sub_grade + emp_length + 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>                               7168 -3187.2
## - inq_last_6mths       1         6   7174 -3180.9
## - annual_inc           1         6   7174 -3180.8
## - revol_bal            1         7   7175 -3179.2
## - open_acc             1         8   7176 -3178.5
## - delinq_2yrs          1        10   7178 -3175.6
## - total_acc            1        17   7185 -3165.8
## - pub_rec              1        21   7189 -3159.9
## - purpose             13        42   7210 -3155.2
## - emp_length          11        42   7210 -3150.8
## - term                 1        35   7204 -3139.8
## - dti                  1        44   7212 -3127.9
## - revol_util           1        72   7240 -3089.6
## - verification_status  2       113   7282 -3034.4
## - sub_grade           34    105057 112225 24253.2
```

```
## 
## Call:
## lm(formula = int_rate ~ term + sub_grade + emp_length + 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  
##                         5.715e+00                         -1.553e-01  
##                       sub_gradeA2                        sub_gradeA3  
##                         7.120e-01                          1.412e+00  
##                       sub_gradeA4                        sub_gradeA5  
##                         1.793e+00                          2.549e+00  
##                        sub_gradeB1                         sub_gradeB2
##                          3.251e+00                           4.259e+00
##                        sub_gradeB3                         sub_gradeB4
##                          5.092e+00                           5.981e+00
##                        sub_gradeB5                         sub_gradeC1
##                          6.432e+00                           7.139e+00
##                        sub_gradeC2                         sub_gradeC3
##                          7.632e+00                           8.248e+00
##                        sub_gradeC4                         sub_gradeC5
##                          8.888e+00                           9.666e+00
##                        sub_gradeD1                         sub_gradeD2
##                          1.035e+01                           1.118e+01
##                        sub_gradeD3                         sub_gradeD4
##                          1.157e+01                           1.214e+01
##                        sub_gradeD5                         sub_gradeE1
##                          1.281e+01                           1.320e+01
##                        sub_gradeE2                         sub_gradeE3
##                          1.362e+01                           1.432e+01
##                        sub_gradeE4                         sub_gradeE5
##                          1.508e+01                           1.590e+01
##                        sub_gradeF1                         sub_gradeF2
##                          1.687e+01                           1.766e+01
##                        sub_gradeF3                         sub_gradeF4
##                          1.819e+01                           1.901e+01
##                        sub_gradeF5                         sub_gradeG1
##                          1.952e+01                           1.931e+01
##                        sub_gradeG2                         sub_gradeG3
##                          1.972e+01                           2.030e+01
##                        sub_gradeG4                         sub_gradeG5
##                          1.924e+01                           2.191e+01
##                   emp_length1 year                 emp_length10+ years
##                          1.743e-02                           1.147e-01
##                  emp_length2 years                   emp_length3 years
##                          2.919e-02                           2.418e-02
##                  emp_length4 years                   emp_length5 years
##                          1.850e-02                           1.449e-01
##                  emp_length6 years                   emp_length7 years
##                          2.492e-01                           1.681e-01
##                  emp_length8 years                   emp_length9 years
##                          5.951e-02                           7.934e-02
##                      emp_lengthn/a                          annual_inc
##                         -8.310e-03                          -5.412e-07
## verification_statusSource Verified         verification_statusVerified
##                         -8.448e-02                           1.770e-01
##                 purposecredit_card           purposedebt_consolidation
##                         -5.625e-02                          -5.756e-02
##                 purposeeducational             purposehome_improvement
##                         -3.467e+00                          -1.367e-01
##                       purposehouse               purposemajor_purchase
##                         -1.571e-01                          -9.693e-02
##                      purposemedical                        purposemoving
##                          -8.704e-02                           -2.610e-01
##                        purposeother              purposerenewable_energy
##                          -1.564e-01                           -3.984e-01
##               purposesmall_business                      purposevacation
##                          -2.158e-01                           -1.228e-01
##                      purposewedding                                  dti
##                           4.937e-01                           -9.351e-03
##                         delinq_2yrs                       inq_last_6mths
##                          -3.619e-02                            2.565e-02
##                            open_acc                              pub_rec
##                          -7.689e-03                           -8.559e-02
##                           revol_bal                           revol_util
##                          -1.629e-06                            4.202e-03
##                           total_acc
##                           5.093e-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

```
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_6
mths +
                         open_acc + pub_rec + revol_bal + revol_util + total_acc, dat
a = pre_loan_data_random)

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, da
ta = pre_loan_data_random) #takes out "verification status"

summary(aic_model_forward)$r.squared
```

```
## [1] 0.9685161
```

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

```
## [1] 0.9680489
```

9.Compare Models AIC models

```
anova(aic_model_forward,aic_model_backward) #So it appears we will go with the backwa
rd 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 6149.9
## 2   8631 6241.1 -2   -91.256 64.021 < 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

```
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 6149.9
## 2   8631 6241.1 -2   -91.256 64.021 < 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.Final model validation: now let us use our model to predict interest rates for the second sample(i.e out version of out-of-sample)

```
prediction <- predict(aic_model_backward, data = pre_loan_data_random2, na.action = n
a.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  9.9900 12.99000 13.16947 15.8800 28.99000
## predicted_rate 4.794042 10.0382 12.98937 13.24159 16.0767 27.86852
```

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

```
## [1] 4.357458
```

```
sd(prediction)
```

```
## [1] 4.348692
```

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 cross-validation than just ANOVA and summary statistics. 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.