Capstone Project

Lending Club Loan Data and Default Rate Prediction

Foundations of Data Science Workshop on SpringBoard

Author: Neelesh Lalwani

Mentor: Patrick Grennan

Synopsis:

Lending Club is the world's largest online marketplace connecting borrowers and investors. We're transforming the banking system to make credit more affordable and investing more rewarding. We operate at a lower cost than traditional bank lending programs and pass the savings on to borrowers in the form of lower rates and to investors in the form of solid returns. (information provided by lending club website)

Objective:

The project aims to use lending club data and attempt to predict the risk of loan being default by using loan information from 2010-2011. ***Data can be obtained from LendingClub's website (https://www.lendingclub.com/info/download-data.action)***

Data:

For this project, I used lending club data from 2010-2011 and divided the data into testing and training sets. This data contains all publicly available information about the loans issued from 2010-2011.

Structure of the data:

The data consists of 52 variables with 42536 observations

id	Factor w/ 42536 levels "1000007","1000030",..: 4388 4387 4386 4385 4383 4382 4364 4381 4380 4379 ...
member_id	int 1296599 1314167 1313524 1277178 1311748 1311441 1304742 1288686 1306957 1306721 ...
loan_amnt	int 5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...
funded_amnt	int 5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...
funded_amnt_inv	num 4975 2500 2400 10000 3000 ...
term	Factor w/ 3 levels ""," 36 months",..: 2 3 2 2 3 2 3 2 3 3 ...
int_rate	Factor w/ 395 levels "","10.00%","10.01%",..: 19 162 180 101 76 363 180 263 311 76 ...
installment	num 162.9 59.8 84.3 339.3 67.8 ...
grade	Factor w/ 8 levels "","A","B","C",..: 3 4 4 4 3 2 4 6 7 3 ...

sub_grade	Factor w/ 36 levels "","A1","A2","A3",..: 8 15 16 12 11 5 16 22 28 11 ...
emp_title	Factor w/ 30661 levels ""," old palm inc",..: 1 22922 1 791 28234 28965 24627 17778 1 25138 ...
emp_length	Factor w/ 13 levels "","< 1 year",..: 4 2 4 4 3 6 11 12 7 2 ...
home_ownership	Factor w/ 6 levels "","MORTGAGE",..: 6 6 6 6 6 6 6 6 5 6 ...
annual_inc	num 24000 30000 12252 49200 80000 ...
is_inc_v	Factor w/ 4 levels "","Not Verified",..: 4 3 2 3 3 3 2 3 3 4 ...
issue_d	Factor w/ 56 levels "","Apr-08","Apr-09",..: 15 15 15 15 15 15 15 15 15 15 ...
loan_status	Factor w/ 12 levels "","Charged Off",..: 3 2 9 3 3 3 3 3 2 2 ...
pymnt_plan	Factor w/ 3 levels "","n","y": 2 2 2 2 2 2 2 2 2 2 ...
url	Factor w/ 42536 levels "","https://www.lendingclub.com/browse/loanDetail.action?loan_id=1000007",..: 4389 4388 4387 4386 4384 4383 4365 4382 4381 4380 ...
desc	Factor w/ 28965 levels "","- Pay off Dell Financial: $ 1300.00 - Pay off IRS for 2005: $ 1400.00 - Pay off Mac Comp : $ 1700.00 - Pay off Bill Me Later"| __truncated__,..: 20681 20682 1 20636 20633 1 20512 20417 20632 20418 ...
purpose	Factor w/ 15 levels "","car","credit_card",..: 3 2 13 11 11 15 4 2 13 11 ...
title	Factor w/ 21259 levels "","'08 & '09 Roth IRA Investments",..: 3693 1875 17215 16558 16312 13996 12002 2714 7533 2290 ...
zip_code	Factor w/ 838 levels "","007xx","010xx",..: 728 282 514 765 814 722 253 750 803 653 ...
addr_state	Factor w/ 51 levels "","AK","AL","AR",..: 5 12 16 6 38 5 29 6 6 44 ...
dti	num 27.65 1 8.72 20 17.94 ...
delinq_2yrs	int 0 0 0 0 0 0 0 0 0 0 ...
earliest_cr_line	Factor w/ 531 levels "","Apr-00","Apr-01",..: 203 45 391 172 214 394 223 183 6 489 ...
inq_last_6mths	int 1 5 2 1 0 3 1 2 2 0 ...
mths_since_last_delinq	int NA NA NA 35 38 NA NA NA NA NA ...
mths_since_last_record	int NA NA NA NA NA NA NA NA NA NA ...
open_acc	int 3 3 2 10 15 9 7 4 11 2 ...
pub_rec	int 0 0 0 0 0 0 0 0 0 0 ...
revol_bal	int 13648 1687 2956 5598 27783 7963 17726 8221 5210 9279 ...
revol_util	Factor w/ 1120 levels "","0%","0.01%",..: 944 1013 1106 192 595 278 963 982 337 382 ...
total_acc	int 9 4 10 37 38 12 11 4 13 3 ...
initial_list_status	Factor w/ 2 levels "","f": 2 2 2 2 2 2 2 2 2 2 ...
out_prncp	num 330 0 0 686 1541 ...
out_prncp_inv	num 329 0 0 686 1541 ...
total_pymnt	num 5527 1009 3004 11530 2293 ...
total_pymnt_inv	num 5499 1009 3004 11530 2293 ...
total_rec_prncp	num 4670 456 2400 9314 1459 ...
total_rec_int	num 857 435 604 2198 834 ...
total_rec_late_fee	num 0 0 0 17 0 ...
recoveries	num 0 117 0 0 0 ...
collection_recovery_fee	num 0 1.11 0 0 0 0 0 0 2.09 2.52 ...
last_pymnt_d	Factor w/ 85 levels "","Apr-08","Apr-09",..: 71 7 50 71 71 71 71 71 6 69 ...
last_pymnt_amnt	num 162.9 119.7 649.9 339.3 67.8 ...
next_pymnt_d	Factor w/ 87 levels "","Apr-08","Apr-09",..: 23 1 1 23 23 23 38 23 1 1 ...

last_credit_pull_d	Factor w/ 90 levels "","Apr-09","Apr-10",..: 74 89 74 74 74 74 74 74 13 58 ...
collections_12_mths_ex_med	int 0 0 0 0 0 0 0 0 0 0 ...
mths_since_last_major_derog	logi NA NA NA NA NA NA ...
policy_code	int 1 1 1 1 1 1 1 1 1 1 ...

Exploratory Data Analysis :

There are 12 loan statuses: Charged Off, Current, Default, Fully Paid, In Grace Period, Late (16-30 days), Late (31-120 days).

```
my.data %>% group_by(loan_status) %>% summarise(count = n())
```

Loan Status	Count
Charged Off	5200
Current	5379
Default	3
Does not meet the credit policy. Status:Charged Off	752
Does not meet the credit policy. Status:Current	85
Does not meet the credit policy. Status:Fully Paid	1906
Does not meet the credit policy. Status:Late (31-120 days)	6
Fully Paid	28890
In Grace Period	118
Late (16-30 days)	28
Late (31-120 days)	168

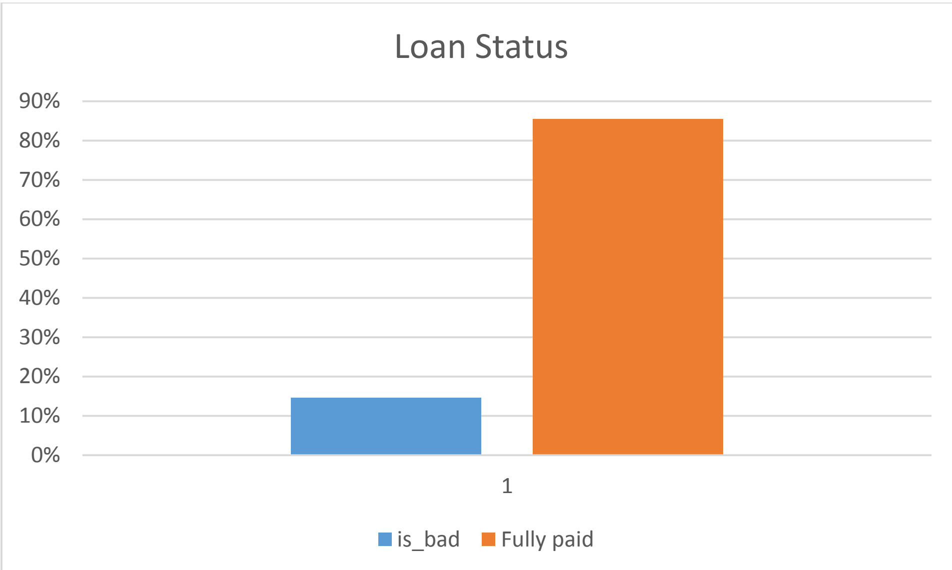
Bad definition:

For loans having a higher risk of default, the following loan status will be used as bad:

1. Charged Off
2. Late (31-120 days)
3. Late (16-30 days)
4. Default

A new variable is_bad is created based on the following above bad definition and others are classified as Fully Paid

See below for the distribution:



Looking at home ownership

table(my.data$home_ownership)

MORTGAGE	NONE	OTHER	OWN	RENT
18959	8	136	3251	20181

Data Cleaning and Preparation:

Next step is to identify variables with more 0s and NAs. Have to remove missing data

1) Find columns with missing data:

tmp = sort(sapply(my.data, function(x) sum(length(which(is.na(x)))))/nrow(my.data),decreasing = TRUE)

2) Next we remove columns with NAs and 0s

Based on the observed data values description, url, employee title, issue date, policy code, last credit pull , id, zip code, earliest credit line information and months last paid are removed since they do not add additional value to the bad definition and the objective of this project.

my.data$desc = NULL

my.data$url = NULL

my.data$emp_title = NULL

my.data$issue_d = NULL

my.data$zip_code = NULL

my.data$policy_code = NULL

my.data$mths_since_last_major_derog = NULL

my.data$last_credit_pull_d <- NULL

my.data$next_pymnt_d <- NULL

my.data$last_pymnt_d = NULL

my.data$earliest_cr_line = NULL

my.data$initial_list_status = NULL

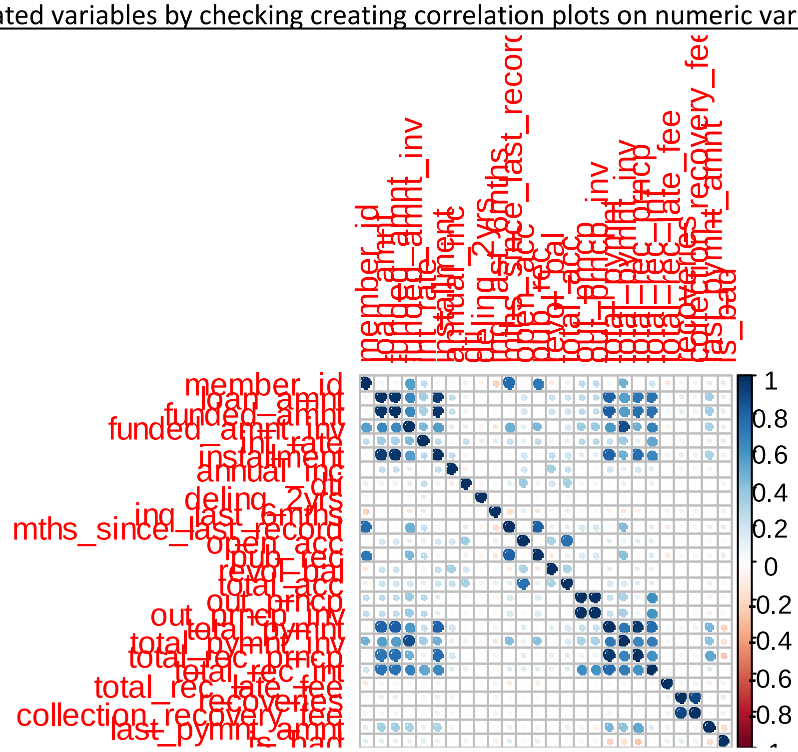
my.data$collections_12_mths_ex_med = NULL

my.data$mths_since_last_major_derog = NULL

my.data$mths_since_last_delinq = NULL

my.data$id = NULL

3) <u>Parse Interest rate to numeric and remove the % value</u>
4) <u>Annual Inc is set to numeric</u>
5) <u>Find Corelated variables by checking creating correlation plots on numeric variables.</u>



Removing variables with correlation greater than 0.75 since they wont add variance to the model objective.

Feature Engineering:

Create new factors on Grade, Sub Grade, Home ownership, payment plan. For home ownership which is "Is Rent", create a new factor to check if rented ownership adds significant value to the prediction of bad loans.

Model Building:

Backward elimination is used in this model building, where we use all variables in the initial phase and based on significance testing eliminate variables which do not add value to the model.

In order to do that, creating testing and training data based on randomization and check for the prorpotion of is_bad in training and testing sets.

```
prop.table(table(my.data.trng$is_bad))

        0         1 
0.8720957 0.1279043 
```

```
prop.table(table(my.data.test$is_bad))

        0         1 
0.8753989 0.1246011 
```

We can see that the is_bad has been equally distributed based on randomization.

Created logistic regression model 1 to check for all numeric variables

```
Call:
glm(formula = is_bad ~ loan_amnt + annual_inc + grade + int_rate + 
    dti + addr_state, family = "binomial", data = my.data.trng)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.0960  -0.5724  -0.4659  -0.3404   3.7268  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -3.703e+00  3.663e-01 -10.108  < 2e-16 ***
```

loan_amnt	1.104e-05	2.699e-06	4.088	4.35e-05	***
annual_inc	-6.686e-06	5.660e-07	-11.812	< 2e-16	***
gradeB	-9.482e-02	8.293e-02	-1.143	0.252878	
gradeC	-2.499e-01	1.141e-01	-2.191	0.028450	*
gradeD	-4.456e-01	1.446e-01	-3.081	0.002063	**
gradeE	-6.348e-01	1.735e-01	-3.660	0.000253	***
gradeF	-8.143e-01	2.095e-01	-3.886	0.000102	***
gradeG	-1.121e+00	2.499e-01	-4.485	7.29e-06	***
int_rate	1.973e-01	1.585e-02	12.444	< 2e-16	***
dti	5.204e-04	2.711e-03	0.192	0.847809	
addr_stateAL	-3.152e-01	3.810e-01	-0.827	0.408112	
addr_stateAR	-4.133e-01	4.193e-01	-0.986	0.324179	
addr_stateAZ	-2.602e-01	3.609e-01	-0.721	0.471006	
addr_stateCA	2.897e-03	3.418e-01	0.008	0.993239	
addr_stateCO	-3.334e-01	3.638e-01	-0.916	0.359512	
addr_stateCT	-3.231e-01	3.653e-01	-0.884	0.376519	
addr_stateDC	-1.111e+00	4.990e-01	-2.226	0.026011	*
addr_stateDE	-7.508e-01	5.237e-01	-1.434	0.151691	
addr_stateFL	2.834e-02	3.449e-01	0.082	0.934506	
addr_stateGA	-1.280e-01	3.517e-01	-0.364	0.715913	
addr_stateHI	-2.664e-02	4.172e-01	-0.064	0.949087	
addr_stateIA	-1.164e+01	1.764e+02	-0.066	0.947361	
addr_stateID	2.985e-01	1.154e+00	0.259	0.795892	
addr_stateIL	-3.065e-01	3.518e-01	-0.871	0.383735	
addr_stateIN	-1.177e+01	1.539e+02	-0.077	0.939017	
addr_stateKS	-7.610e-01	4.295e-01	-1.772	0.076449	.
addr_stateKY	-2.122e-01	3.890e-01	-0.545	0.585439	
addr_stateLA	-4.510e-01	3.886e-01	-1.161	0.245820	
addr_stateMA	-3.412e-01	3.550e-01	-0.961	0.336514	
addr_stateMD	-1.626e-01	3.564e-01	-0.456	0.648099	
addr_stateME	-1.162e+01	3.085e+02	-0.038	0.969950	
addr_stateMI	-2.971e-01	3.654e-01	-0.813	0.416297	
addr_stateMN	-1.777e-01	3.672e-01	-0.484	0.628349	
addr_stateMO	-1.830e-01	3.651e-01	-0.501	0.616115	
addr_stateMS	-2.231e-01	8.292e-01	-0.269	0.787857	
addr_stateMT	-6.523e-01	5.505e-01	-1.185	0.236062	
addr_stateNC	-1.365e-01	3.617e-01	-0.377	0.705875	
addr_stateNE	5.394e-02	1.135e+00	0.048	0.962089	
addr_stateNH	3.694e-02	4.200e-01	0.088	0.929919	
addr_stateNJ	-4.931e-02	3.485e-01	-0.142	0.887474	
addr_stateNM	-2.008e-02	4.130e-01	-0.049	0.961212	
addr_stateNV	1.933e-01	3.661e-01	0.528	0.597568	
addr_stateNY	-2.926e-01	3.448e-01	-0.849	0.396103	
addr_stateOH	-3.372e-01	3.543e-01	-0.952	0.341209	
addr_stateOK	-1.743e-01	3.921e-01	-0.445	0.656660	
addr_stateOR	-4.985e-02	3.746e-01	-0.133	0.894120	
addr_statePA	-4.335e-01	3.530e-01	-1.228	0.219386	
addr_stateRI	-1.454e-01	4.223e-01	-0.344	0.730568	
addr_stateSC	-1.305e-02	3.736e-01	-0.035	0.972137	
addr_stateSD	-4.851e-02	5.368e-01	-0.090	0.927992	
addr_stateTN	-5.290e-01	8.146e-01	-0.649	0.516079	
addr_stateTX	-3.745e-01	3.469e-01	-1.080	0.280275	
addr_stateUT	-2.731e-01	4.026e-01	-0.678	0.497532	
addr_stateVA	-3.401e-01	3.539e-01	-0.961	0.336512	
addr_stateVT	-1.297e-02	5.618e-01	-0.023	0.981578	
addr_stateWA	-2.001e-01	3.610e-01	-0.554	0.579339	
addr_stateWI	-9.888e-02	3.718e-01	-0.266	0.790272	

```
addr_stateWV -1.078e-01  4.270e-01  -0.252 0.800748
addr_stateWY -1.736e+00  7.979e-01  -2.175 0.029593 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22941  on 29994  degrees of freedom
Residual deviance: 21799  on 29935  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 21919

Number of Fisher Scoring iterations: 12
```

From this we can see, address variable does not contribute significant value to the model and is removed.

2) We create logistic regression 2 and check for summary

```
Call:
glm(formula = is_bad ~ loan_amnt + annual_inc + grade + int_rate,
    family = "binomial", data = my.data.trng)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0373  -0.5731  -0.4729  -0.3477   3.6910

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.895e+00  1.283e-01 -30.362  < 2e-16 ***
loan_amnt    1.079e-05  2.669e-06   4.041 5.32e-05 ***
annual_inc  -6.627e-06  5.577e-07 -11.884  < 2e-16 ***
gradeB      -9.126e-02  8.254e-02  -1.106  0.26890
gradeC      -2.560e-01  1.134e-01  -2.256  0.02404 *
gradeD      -4.493e-01  1.437e-01  -3.127  0.00177 **
gradeE      -6.390e-01  1.724e-01  -3.707  0.00021 ***
gradeF      -8.261e-01  2.080e-01  -3.973 7.11e-05 ***
gradeG      -1.129e+00  2.485e-01  -4.542 5.56e-06 ***
int_rate     1.985e-01  1.574e-02  12.605  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22941  on 29994  degrees of freedom
Residual deviance: 21913  on 29985  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 21933

Number of Fisher Scoring iterations: 5
```

We see that most variables contribute significantly and pass the significance test (p value < 0.05)

Checking on the confusion matrix:

Logistic model 2 cross table :

```
   |                   my.data.test$prediction
my.data.test$is_bad |      Good | Row Total |
--------------------|-----------|-----------|
                  0 |     10874 |     10974 |
                    |     0.82  |           |
--------------------|-----------|-----------|
                  1 |      1662 |      1562 |
                    |     0.17  |           |
--------------------|-----------|-----------|
       Column Total |     12536 |     12536 |
--------------------|-----------|-----------
```

3) We create another logistic regression model based on other variables and check for significance testing

```
my.data.logModel.3 = glm(is_bad~

                           funded_amnt +
                           funded_amnt_inv +
                           int_rate +
                           installment +
                           grade +
                           sub_grade +
                           emp_length +
                           pymnt_plan +
                           delinq_2yrs +
                           inq_last_6mths +
                           revol_bal +
                               is_rent

                         ,data=my.data.trng,family=binomial())
```

- Summary(my.data.logModel.3)

```
Call:
glm(formula = is_bad ~ funded_amnt + funded_amnt_inv + int_rate +
    installment + grade + sub_grade + emp_length + pymnt_plan +
    delinq_2yrs + inq_last_6mths + revol_bal + is_rent, family = binomial(),
    data = my.data.trng)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6336  -0.5642  -0.4669  -0.3459   2.8907
```

```
Coefficients: (6 not defined because of singularities)
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -5.103e+00  2.556e-01 -19.968  < 2e-16 ***
funded_amnt          1.347e-04  1.153e-05  11.689  < 2e-16 ***
funded_amnt_inv     -8.219e-05  7.410e-06 -11.092  < 2e-16 ***
int_rate             2.595e-01  2.045e-02  12.691  < 2e-16 ***
installment         -2.246e-03  2.935e-04  -7.651 1.99e-14 ***
gradeB               1.551e-01  2.635e-01   0.588  0.55623    
gradeC              -1.793e-01  3.009e-01  -0.596  0.55139    
gradeD              -4.648e-01  3.276e-01  -1.419  0.15592    
gradeE              -7.493e-01  3.661e-01  -2.047  0.04071 *  
gradeF              -6.951e-01  4.217e-01  -1.648  0.09929 .  
gradeG              -1.705e+00  5.455e-01  -3.125  0.00178 ** 
sub_gradeA2          5.463e-01  2.638e-01   2.071  0.03840 *  
sub_gradeA3          5.969e-01  2.536e-01   2.353  0.01860 *  
sub_gradeA4          5.224e-01  2.433e-01   2.147  0.03177 *  
sub_gradeA5          6.574e-01  2.430e-01   2.706  0.00681 ** 
sub_gradeB1          1.169e-01  1.265e-01   0.924  0.35524    
sub_gradeB2          2.019e-01  1.136e-01   1.777  0.07551 .  
sub_gradeB3          1.558e-01  9.966e-02   1.564  0.11792    
sub_gradeB4          1.133e-01  1.013e-01   1.118  0.26342    
sub_gradeB5                 NA         NA      NA       NA    
sub_gradeC1          2.009e-01  1.232e-01   1.630  0.10304    
sub_gradeC2          1.637e-01  1.206e-01   1.357  0.17491    
sub_gradeC3          2.647e-01  1.224e-01   2.163  0.03057 *  
sub_gradeC4          8.604e-02  1.292e-01   0.666  0.50560    
sub_gradeC5                 NA         NA      NA       NA    
sub_gradeD1          2.484e-01  1.455e-01   1.707  0.08777 .  
sub_gradeD2          2.435e-01  1.283e-01   1.897  0.05781 .  
sub_gradeD3          9.141e-02  1.310e-01   0.698  0.48519    
sub_gradeD4          1.316e-01  1.331e-01   0.989  0.32272    
sub_gradeD5                 NA         NA      NA       NA    
sub_gradeE1          4.201e-01  1.647e-01   2.550  0.01076 *  
sub_gradeE2          1.747e-01  1.699e-01   1.029  0.30364    
sub_gradeE3         -1.660e-01  1.809e-01  -0.917  0.35889    
sub_gradeE4         -1.035e-01  1.849e-01  -0.560  0.57560    
sub_gradeE5                 NA         NA      NA       NA    
sub_gradeF1         -2.234e-01  2.483e-01  -0.900  0.36832    
sub_gradeF2         -1.679e-01  2.569e-01  -0.654  0.51331    
sub_gradeF3         -4.273e-01  2.751e-01  -1.553  0.12036    
sub_gradeF4         -4.405e-01  2.820e-01  -1.562  0.11822    
sub_gradeF5                 NA         NA      NA       NA    
sub_gradeG1          5.167e-01  4.574e-01   1.130  0.25861    
sub_gradeG2          8.465e-01  4.703e-01   1.800  0.07189 .  
sub_gradeG3          6.125e-01  4.929e-01   1.243  0.21398    
sub_gradeG4         -2.650e-01  5.201e-01  -0.509  0.61041    
sub_gradeG5                 NA         NA      NA       NA    
emp_length1 year    -1.686e-02  7.947e-02  -0.212  0.83195    
emp_length10+ years  8.338e-02  6.554e-02   1.272  0.20331    
emp_length2 years   -8.811e-02  7.502e-02  -1.174  0.24021    
emp_length3 years   -2.873e-02  7.598e-02  -0.378  0.70535    
emp_length4 years   -2.007e-02  7.983e-02  -0.251  0.80145    
emp_length5 years    1.800e-02  8.067e-02   0.223  0.82348    
emp_length6 years   -2.136e-02  9.266e-02  -0.231  0.81770    
emp_length7 years    2.547e-02  9.971e-02   0.255  0.79840    
emp_length8 years    4.011e-02  1.055e-01   0.380  0.70384    
```

```
emp_length9 years   -1.175e-03  1.151e-01  -0.010  0.99186    
emp_lengthn/a        7.864e-01  1.060e-01   7.417 1.20e-13 ***
pymnt_plany          1.666e+00  7.305e-01   2.281  0.02255 *  
delinq_2yrs         -5.692e-02  3.467e-02  -1.642  0.10068    
inq_last_6mths      -5.524e-02  1.283e-02  -4.304 1.67e-05 ***
revol_bal           -4.984e-06  1.127e-06  -4.424 9.70e-06 ***
is_rentTRUE          9.388e-02  3.776e-02   2.487  0.01290 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22935  on 29973  degrees of freedom
Residual deviance: 21759  on 29919  degrees of freedom
  (26 observations deleted due to missingness)
AIC: 21869

Number of Fisher Scoring iterations: 6
```

Model validation:

Checking for anova testing using chi-square testing to check significance testing and lift provided by the new variables

Analysis of Deviance Table

Model: binomial, link: logit

Response: is_bad

Terms added sequentially (first to last)

```
                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                            29973      22935              
funded_amnt      1    45.26     29972      22890 1.727e-11 ***
funded_amnt_inv  1    26.48     29971      22864 2.665e-07 ***
int_rate         1   769.25     29970      22094 < 2.2e-16 ***
installment      1    85.55     29969      22009 < 2.2e-16 ***
grade            6    68.03     29963      21941 1.037e-12 ***
sub_grade       28    57.34     29935      21884 0.0008812 ***
emp_length      11    64.04     29924      21819 1.630e-09 ***
pymnt_plan       1     4.26     29923      21815 0.0390971 *  
inq_last_6mths   1    21.22     29921      21792 4.102e-06 ***
revol_bal        1    26.86     29920      21765 2.185e-07 ***
is_rent          1     6.19     29919      21759 0.0128737 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Checking the confusion matrix on the testing set:

	predicted		
actual	Good	Bad	Row Total
0	10960 0.875	10 0.001	10970
1	1552 0.124	10 0.001	1562
Column Total	12512	20	12532

Comparing model 2 and model 3 based on confusion matrix test to check for the lift provided, we can see model 3 is a better predictor for loan default dataset:

- Sensitivity = 10/1562 = 0.001
- Specificity = 10960/12512 = 0.87

Conclusion:

It is essential to build a more robust system so peer- to peer lending can offer healthy loans to investors. The solution presented above can check with a good accuracy for loan applications but a stronger risk profile will need better validation of credentials before the loan is approved. Model deployed here can predict with more than 80% accuracy of separating a good loan from a bad one. Lending club must utilize strengthened risk controls and models to enhance their approval/decline process.