Lending Club Data

Siran Zhao

December 3, 2017

OVERVIEW

As more and more people like to borrow money from lending club. Under this circumstance, I think if we can predict the probability that whether a person is going to charge off his(her) loans, then we can control the risk we are going to face.

So I got the data from the website of Lending Club from 2007 to 2011 to fit a multilevel logistic model and try to predict 2017Q2 results with 2017Q2 filted data.

Variable Description:
loan_amnt: The amount of money who borrowed.
term: 2 terms of loans "36 months" and "60 months", which mean the last period of a loan.
int_rate: Interest rate of a loan.
grade: Grade of a loan.
home_ownership: House ownership of the borrowers.
loan_status: Loan_status means whether a loan had been charged off or fully paid.
pay or not: binary form of loan_status in which 1 means fully paid, 0 means charged off.

Data Cleaning and Models

At first I assigned their format in order to fit the models. Because the scale of interest rate is not appropriate, so I centralized them.

Then I picked up 3 variables to fit the model which I think have the most relation to my topic. I picked loan_status as the Y, because it contain 2 form of output,"Fully Paid" and "Charged off", which can be treat as binomial "1: Fully paid", "0: Charged Off". Then I fitted three models each of them I would add a new kind of variable in it.

The 1st model: $y = \beta_0 + \beta_1 * x_1 + \varepsilon$

2nd model: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \varepsilon$

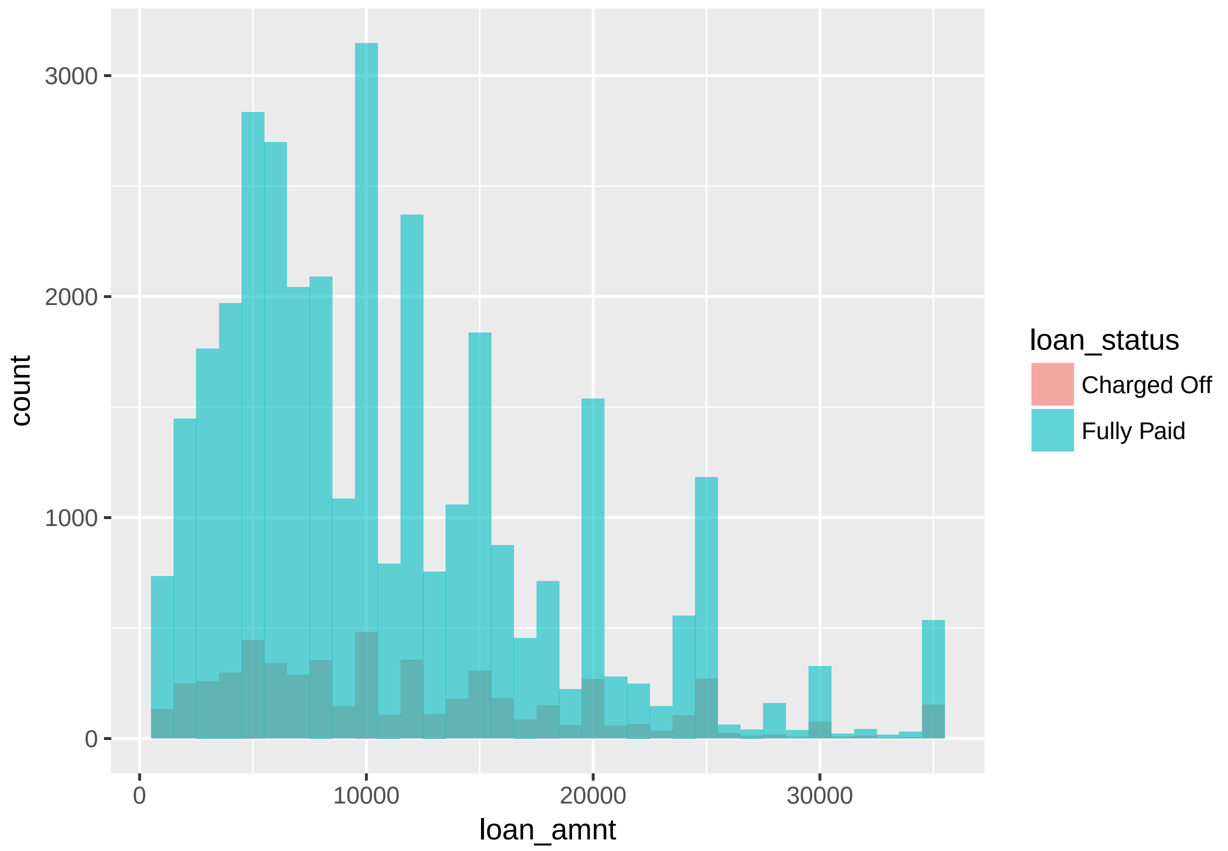
3rd model: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \varepsilon$

$(x_1 = Term\ of loans\ x_2 = Home\ ownership\ x_3 = Interest\ Rate)$

The models are fitted upon groups of loans grade, which from A~G as A is the best and G is the worst.

Histgram distribution of loan_amnt and loan_status

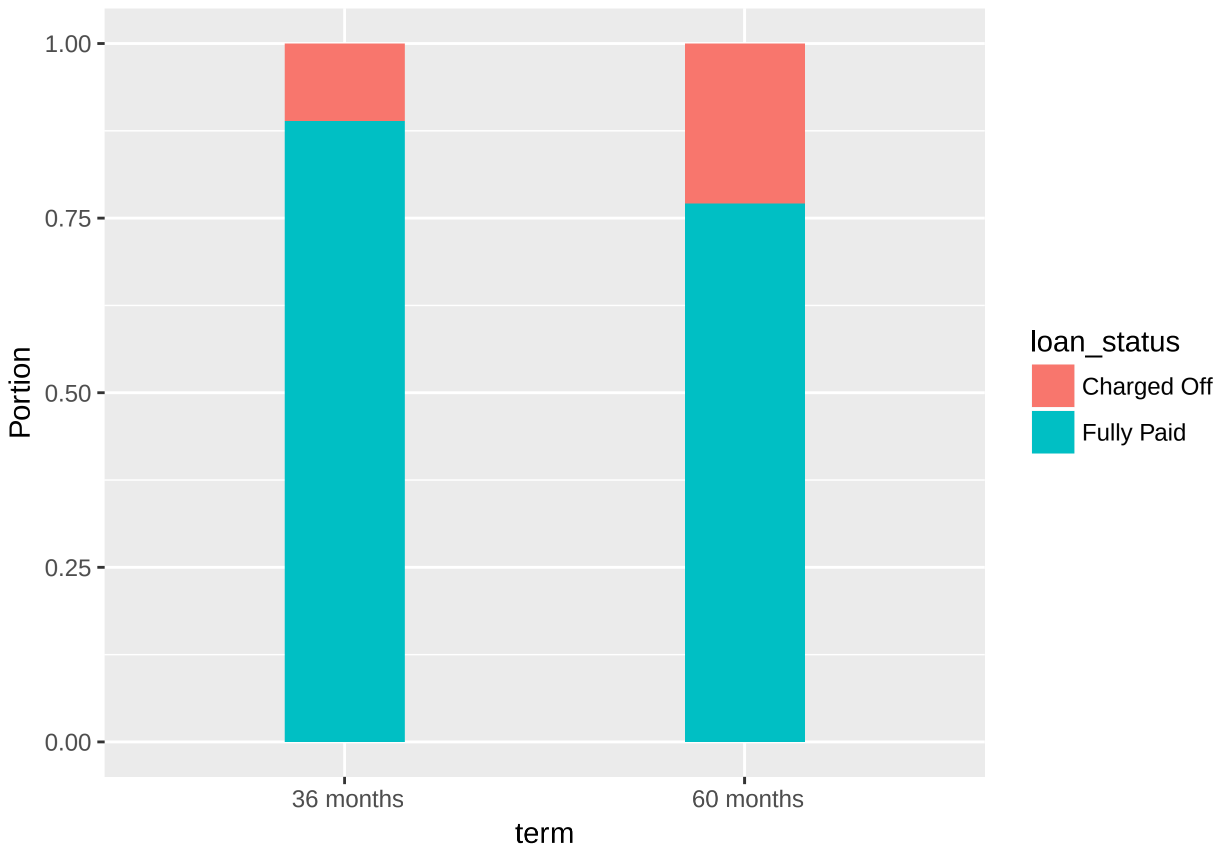
Through this plot we can define the relation between loan amount and loan status, we can see whether their distribution are similar.



From the plot above, we can easily find that whether the loan amounts are high or low, the distribution of people who charged off(the dark part) are not influenced a lot. So I think we can neglect the influence which loan amount gave on the loan status.

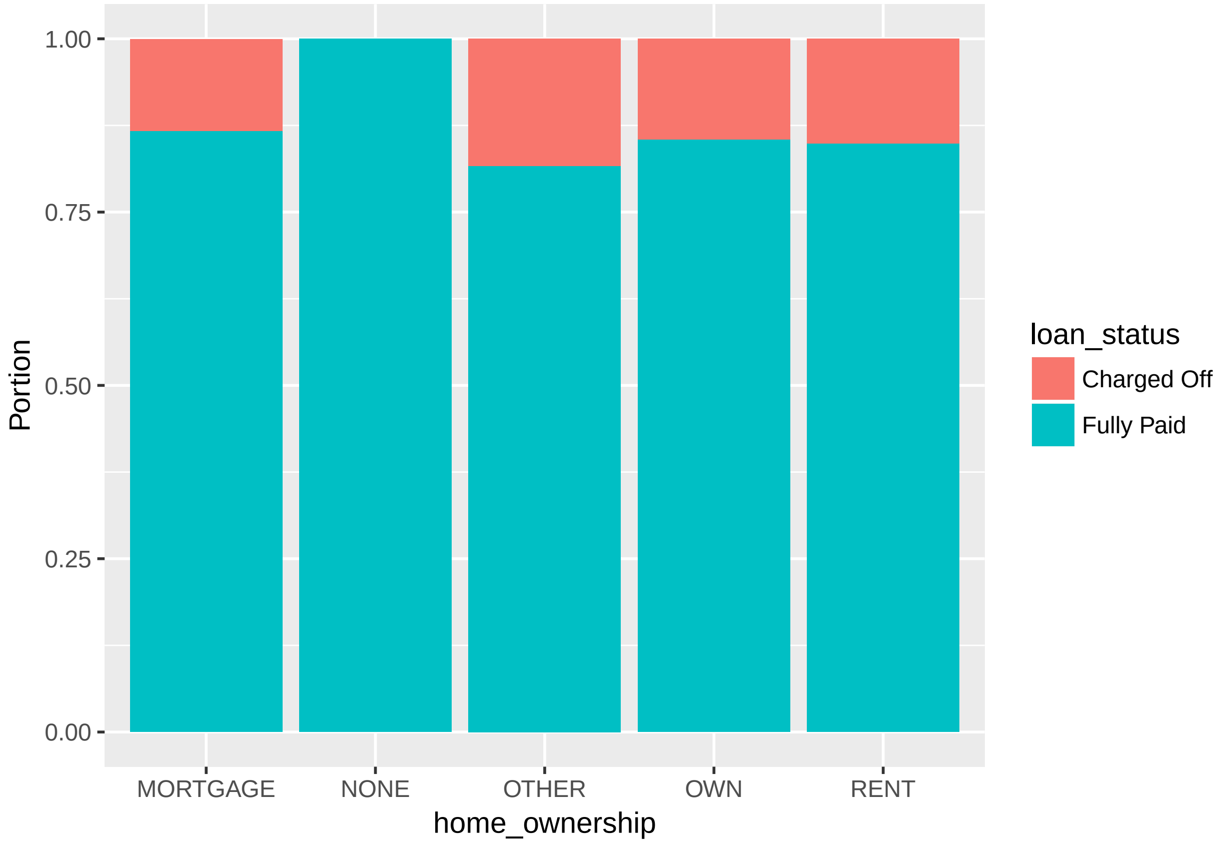
potion Graph between Loan_term and loan_status

At first I plot the graph between loan status and loan terms.



From the plot above 36 months ratio for charged off is lower than 72 months. People may have higher probability to charge off their loans as the loans period get longer. So there must be some connection between terms and results.

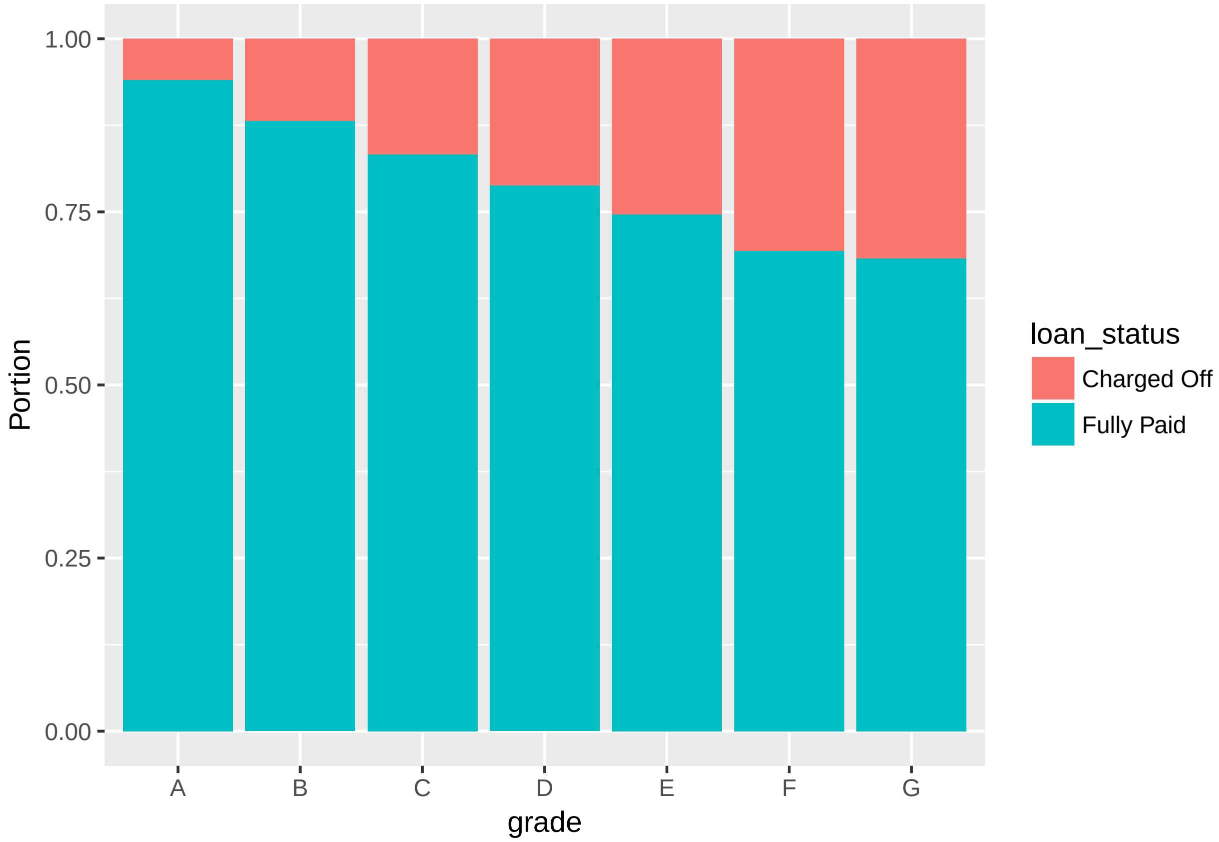
potion graph between home_ownership and loan_status



From the graph above, we can find slightly difference in charge off rate among different kinds of ownership. People who have mortgage kind of house with the smallest rate to charge off their loans, and the other kind and people who rent their house will have higher rate to charge off their loans as people who own their houses are between these three kinds above.

According to the reality, we think people who have their own house or have mortgages would have less possible to charge off their loans compairing with people who rent their house.(Number of people who in NONE kind is just one, so I didn't made any conclusion.)

Ratio of Charged off and fully paid loan among Loan Grade.



From this graph, we can find that as the Grade of the loans went down, the ratio of charged off became higher. So I think we can fit the models with the elements above.

Fitting Models

At first we tried to fit a model only with term of loans in different grade levels. Then we added two more into the model to see whether it can make my models more accurate.

Summaries for 3 models.

```
##                 Estimate Std. Error  z value     Pr(>|z|)
## term 36 months 1.724866  0.1997917 8.633322 5.959545e-18
## term 60 months 1.209807  0.1994507 6.065696 1.313835e-09
```

```
##                         Estimate Std. Error    z value     Pr(>|z|)
## term 36 months        1.83570935  0.2000168  9.1777760 4.400864e-20
## term 60 months        1.28204782  0.1976409  6.4867530 8.770595e-11
## home_ownershipOTHER -1.32088178  0.4477341 -2.9501477 3.176221e-03
## home_ownershipOWN   -0.06900173  0.1179271 -0.5851219 5.584658e-01
## home_ownershipRENT  -0.17097681  0.0620626 -2.7549090 5.870845e-03
```

```
##                         Estimate Std. Error    z value      Pr(>|z|)
## int_rate            -0.40590568 0.04518259 -8.9836741  2.618726e-19
## term 36 months       2.10764214 0.06449299 32.6801737 2.987807e-234
## term 60 months       1.61754573 0.07089995 22.8144822 3.293293e-115
## home_ownershipOTHER -1.32691452 0.44394191 -2.9889372  2.799496e-03
## home_ownershipOWN   -0.06031475 0.11800102 -0.5111375  6.092548e-01
## home_ownershipRENT  -0.16439867 0.06213464 -2.6458457  8.148699e-03
```

From above summaries, we can find that:

1. Interest rate have negative influence on Fully paid probability. Which means that if the interest of loans goes higher, borrower are more likely to charge off their loans. (1 unit of interest rate's improvement will decline 0.4 units of log odds)

2. As loans period are longer, it have weaker positive influence on Fully paid probability means that a person who take longer loans will have more probability to charge off their loans than people who had shorter time loans. (When period from 36 months become 60 months, it will have 0.5 reduce effect on log odds.)

3. Basis of home_ownership is MORTGAGE, and we can find who have a Other or Rent contracts of the houses are more likely to Charge off. (Other kinds have -1.0 negative influence on log odds comparing with mortgage kind, kinds of own have -0.06 influence than mortgage kind on log odds and rent kind have -0.16 influence on log odds than mortgage kind.)

4. From the coefficients of above 3 models, at first I fitted the term of loans into the model_1, then I added home_ownership into the model, and we can find that the P-value of former coefficient became smaller base on the home ownership. At last, I fitted centralized interest_rate into the model, which reduce the P-value of former coefficients again.

AIC Compairision

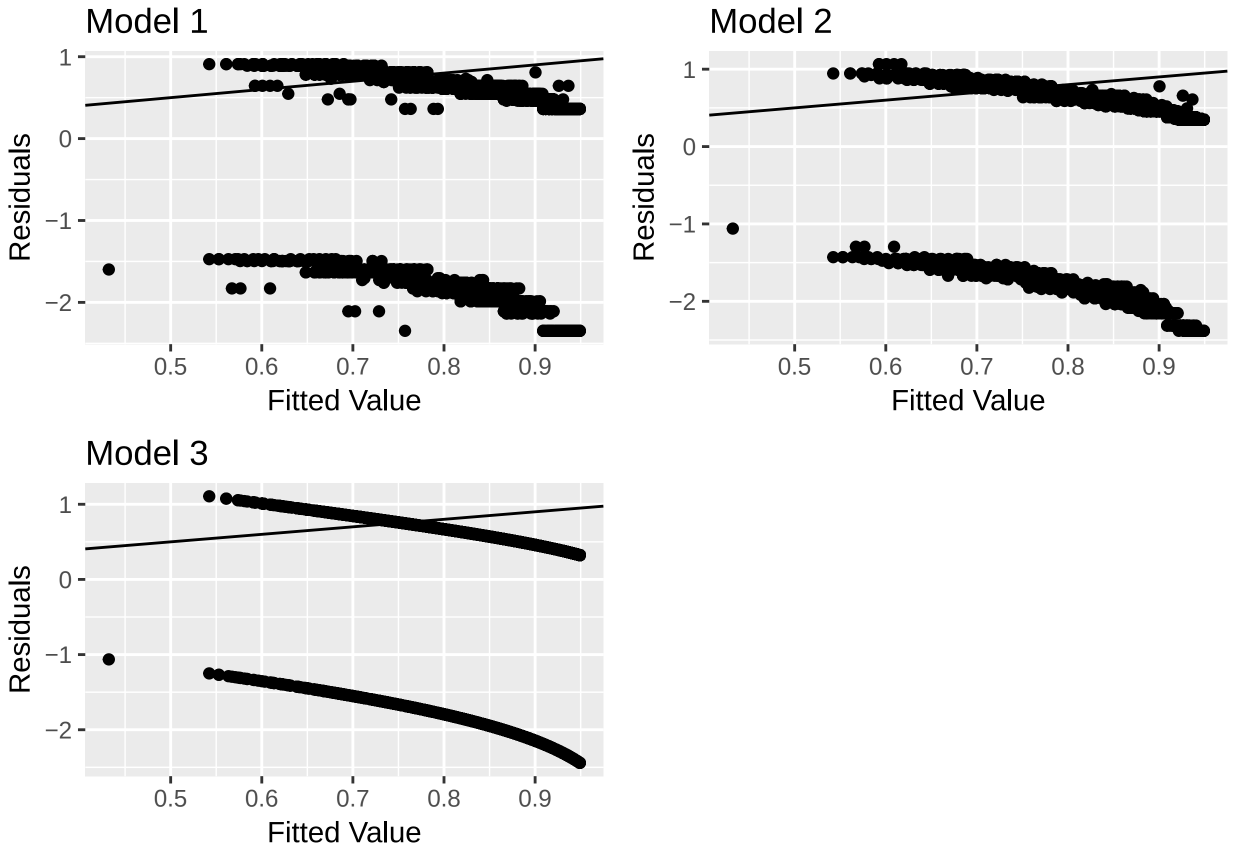
```
##       AIC       BIC    logLik  deviance  df.resid
##  7881.668  7903.299 -3937.834  7875.668  9997.000
```

```
##       AIC       BIC    logLik  deviance  df.resid
##  7873.619  7916.881 -3930.810  7861.619  9994.000
```

```
##       AIC       BIC    logLik  deviance  df.resid
##  7853.875  7904.347 -3919.937  7839.875  9993.000
```

From the AIC Score above, we can find that as we added home_ownership and annual interest of the loan into the fitting models, the AIC Score went down. This meant the model fit better with this 3 indicators than just one of them.

Residual Plot

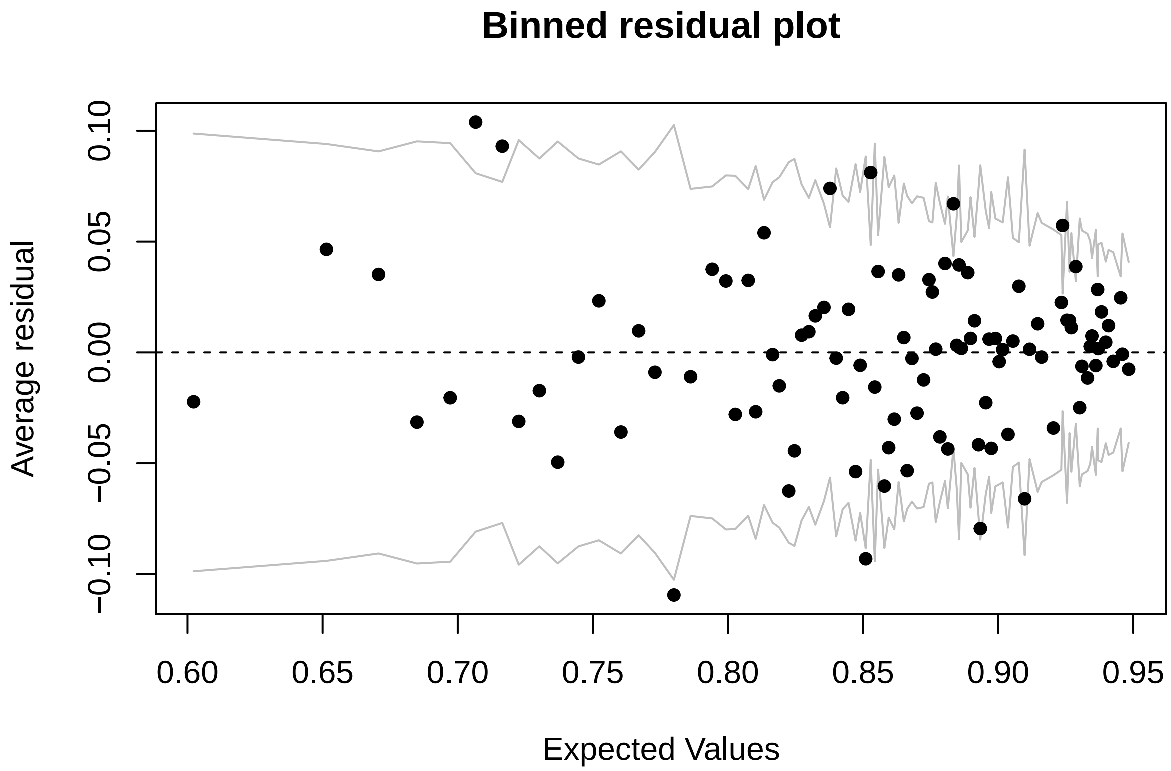


To compare the models we had in another way, I drew the residual plots for these 3 models.

From three residuals plots, we can find the residuals in the first plot which present model 1 which just contain terms of loans are much scatter than the second and the third plots, while the residuals of model 2 are more scatter than model 3. I think they showed that as I added interest rate of loans and the home ownership into models, it help optimize my models.

Binned residual plot

Use binned plot we can also justify the model fitting.



In the binned residuals plot, most of the points lying between two lines and concentrate around the x = 0 line. This means that our model is reliable.

Predict the probability of Fully Paid and Charged off among customers' loan whose status under Fully paid and Charged off in 2017 Q2.

```
## [1] 0.8801498
```

The 0.88 is the ratio of the predicted results which are "Fully Paid" the same as the reality. Though it is lower than 90%, but in my opinion this maybe because that the sample size of real data is too small to make a more accurately predict.

Predict the probability of Fully Paid and Charged off among 9227 customers' loan whose status under Current.

After finishing the models, I tried to predict loan status within 10000 data from 2017 Quarter 2 in which customers' status are still pending.

```
##    loan_amnt      funded_amnt            term        int_rate   
##  Min.   : 1200   Min.   : 1200            :  0   Min.   :3.419  
##  1st Qu.:12000   1st Qu.:12000    36 months:124   1st Qu.:3.807  
##  Median :17913   Median :17913    60 months:488   Median :4.971  
##  Mean   :19048   Mean   :19048                    Mean   :4.789  
##  3rd Qu.:25000   3rd Qu.:25000                    3rd Qu.:5.261  
##  Max.   :40000   Max.   :40000                    Max.   :6.200  
```

```
##
##      grade       sub_grade    home_ownership   annual_inc
##  D      :280   D3     :115   RENT    :281    Min.   :   2
##  E      :195   E5     :100   MORTGAGE:265    1st Qu.:2300
##  F      : 94   D5     : 60   OWN     : 66    Median :4832
##  G      : 43   D4     : 56           :  0    Mean   :4552
##         :  0   E4     : 51   35000   :  0    3rd Qu.:6951
##  A      :  0   E3     : 35   40000   :  0    Max.   :9066
##  (Other):  0   (Other):195   (Other) :  0
##           loan_status    addr_state       prob         pre.status
##  Current        :612   CA     : 76   Min.   :0.2571   Length:612
##                 :  0   NY     : 63   1st Qu.:0.3849   Class :character
##  Charged Off    :  0   FL     : 51   Median :0.4350   Mode  :character
##  Default        :  0   TX     : 50   Mean   :0.4190
##  Fully Paid     :  0   IL     : 28   3rd Qu.:0.4758
##  In Grace Period:  0   NJ     : 28   Max.   :0.5000
##  (Other)        :  0   (Other):316
```

From the summary above, we can find several conclusions:
1. For those who have more probability to charge off their loans, 60 months' loan take larger portion.
2. The least probability for a borrower to paid his loan is 0.257,and the mean is 0.41.

I filtered the people who are likely to charge off in the future with Probability larger than 50%. And count the number are 612. We should be aware of these peoples.

What can we do with this prediction?

From the prediction we can get probability for each borrower. Then we can assigh levels for these borrowers base on their probability for paying their loans or charging off their loans.

And we can assign 4 levels as below:
1. $0.00_{0.20}$ 1st level: Need to be checked each month.
2. $0.20_{0.50}$ 2nd level: Need to be checked each quarter.
3. $0.50_{0.80}$ 3rd level: Checking each year.
4. $0.80_{1.00}$ 4th level: Randomly checked.

Which from level 1 to level 4 represent the risk we are facing become lower.

term	int_rate	grade	home_ownership	prob	pre.status
60 months	3.010820	C	RENT	3rd level: Checking each year	Will Fully Paid
36 months	1.986541	B	OWN	3rd level: Checking each year	Will Fully Paid
36 months	1.888514	B	RENT	3rd level: Checking each year	Will Fully Paid
36 months	1.442393	A	OWN	4th level: Randomly checked	Will Fully Paid
36 months	1.594434	A	MORTGAGE	4th level: Randomly checked	Will Fully Paid
36 months	1.414385	A	RENT	4th level: Randomly checked	Will Fully Paid

Shortcoming and Future development

1st: Is that to make more accurate predictions, we need to have more data in the future.

2nd: I should lift the 50% rate to a higher place if I can fit a better model, because in real world, a higher than 50% of probability who will pay their loans is not reliable enough.

Appendix

1 Summaries of 3 models I fitted.

1st model with just terms of loans.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: pay.or.not ~ term + (1 | grade) - 1
##    Data: sample.loan.data
##       AIC       BIC    logLik  deviance  df.resid
##  7881.668  7903.299 -3937.834  7875.668      9997
## Random effects:
##  Groups Name        Std.Dev.
##  grade  (Intercept) 0.5067
## Number of obs: 10000, groups:  grade, 7
## Fixed Effects:
## term 36 months  term 60 months
##          1.725           1.210
```

2nd model with terms and home ownership

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: pay.or.not ~ term + home_ownership + (1 | grade) - 1
##    Data: sample.loan.data
##       AIC       BIC    logLik  deviance  df.resid
##  7873.619  7916.881 -3930.810  7861.619      9994
## Random effects:
##  Groups Name        Std.Dev.
##  grade  (Intercept) 0.4967
## Number of obs: 10000, groups:  grade, 7
## Fixed Effects:
##      term 36 months       term 60 months  home_ownershipOTHER
##               1.836                1.282               -1.321
##   home_ownershipOWN   home_ownershipRENT
##              -0.069               -0.171
```

3rd model with term, home ownership and interest rate of loans

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: pay.or.not ~ int_rate + term + home_ownership + (1 | grade) -
##     1
##    Data: sample.loan.data
##       AIC       BIC    logLik  deviance  df.resid
##  7853.875  7904.347 -3919.937  7839.875      9993
## Random effects:
##  Groups Name        Std.Dev.
##  grade  (Intercept) 0.07612
## Number of obs: 10000, groups:  grade, 7
## Fixed Effects:
##            int_rate       term 36 months       term 60 months
##           -0.40591              2.10764              1.61755
## home_ownershipOTHER    home_ownershipOWN   home_ownershipRENT
##           -1.32691             -0.06031             -0.16440
```

2 Prediction of 2017 Q2 data which are under fully paid and charged offf status:

term	int_rate	grade	home_ownership	loan_status	prob	same
36 months	1.470400	A	MORTGAGE	Fully Paid	0.8306310	1
36 months	1.470400	A	OWN	Fully Paid	0.8219758	1
36 months	4.971354	E	RENT	Fully Paid	0.4936472	0
36 months	1.888514	B	MORTGAGE	Fully Paid	0.7891816	1
36 months	2.398653	B	MORTGAGE	Fully Paid	0.7526751	1
36 months	2.398653	B	MORTGAGE	Fully Paid	0.7526751	1

3 Prediction of 2017 Q2 data which are under current status:

term	int_rate	grade	home_ownership	loan_status	prob	pre.status
60 months	4.291169	D	MORTGAGE	Current	0.4465464	Will Charge off
60 months	6.131670	F	RENT	Current	0.2622023	Will Charge off
36 months	5.261433	E	RENT	Current	0.4642716	Will Charge off
36 months	6.131670	F	MORTGAGE	Current	0.4061128	Will Charge off
36 months	5.939617	F	RENT	Current	0.3854445	Will Charge off
60 months	5.165407	E	MORTGAGE	Current	0.3941556	Will Charge off