MA684 Midterm Project

Dongyuan Zhou

November 13, 2017

Background

In the financial field, making decisions on whether to lend money to borrowers is one of the most important work. A whole cycle of this work usually includes the following two steps.

Firstly, grades the loan according to the borrower credit record as well as the loan amount and loan time, and then make decision whether lend or not.

Secondly, summarize the default rate and improve the former step to avoid default next time.

Objective

In this analysis, we analyze data from Lending club (LC):

Stage1: Initial EDA (Data: 2015-2017)

Analyze how borrower's status inluenced their loan amount? (Linear regression)

Stage2: Before loan was funded (Data: 2015-2017)

Analyze how LC assigned loan grade: How borrower's status as well as the loan amount and loan time influenced loan grade? (Multinomial regreesion)

Stage3: After laon ended (Data: 2007-2011)

Summarize default rate for each state.(Multilevel logistic regression)

Stage4: Summarize and Discussion

Assessment of the result. Discuss about the data limitations and future directions.

Data description

Data source: https://www.lendingclub.com/info/download-data.action

To get reasonable analysis, we only choose the data which had been verified by LC.

In the whole analysis, we transform loan grade to grade number.

grade	A	B	C	D	E	F	G
gradenumber	7	6	5	4	3	2	1

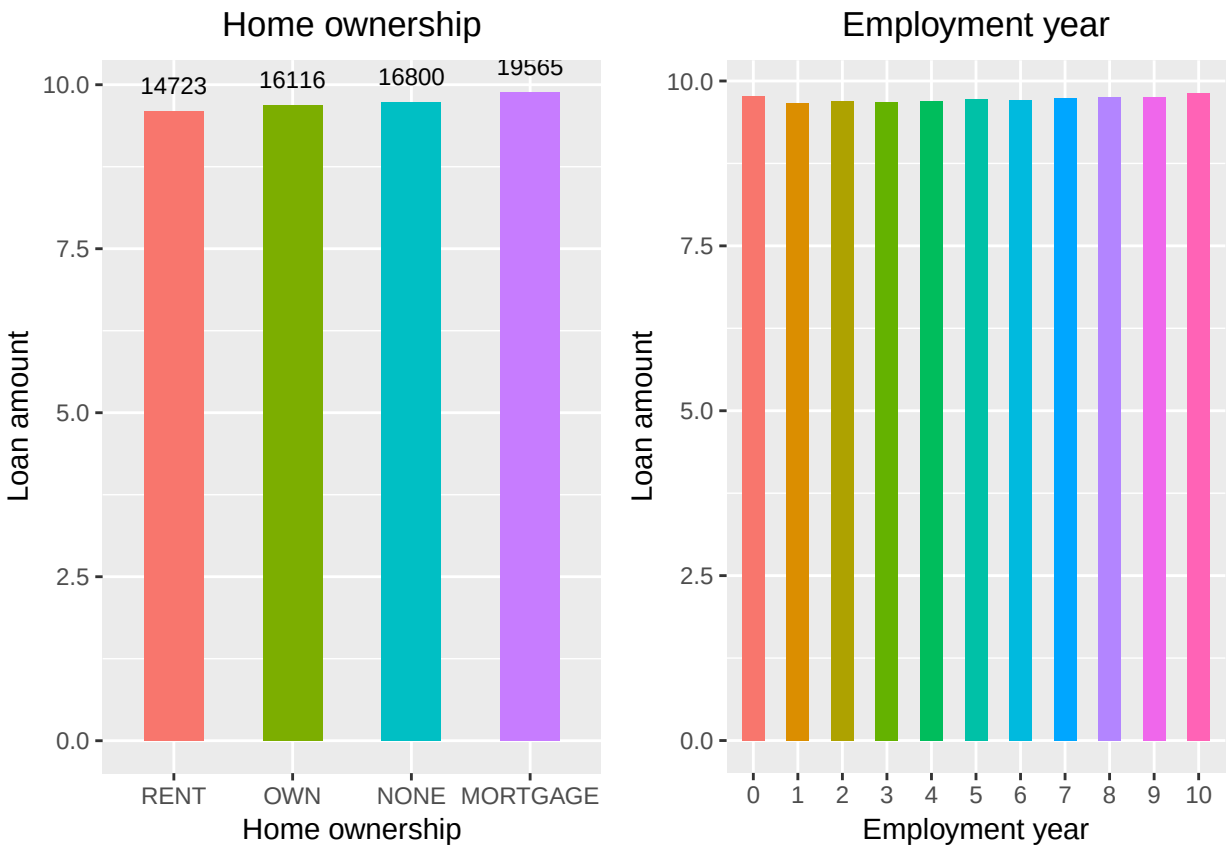
Analysis Part

1. Initial EDA

Analyze how borrower's status inluenced their loan amount? (Linear regression)

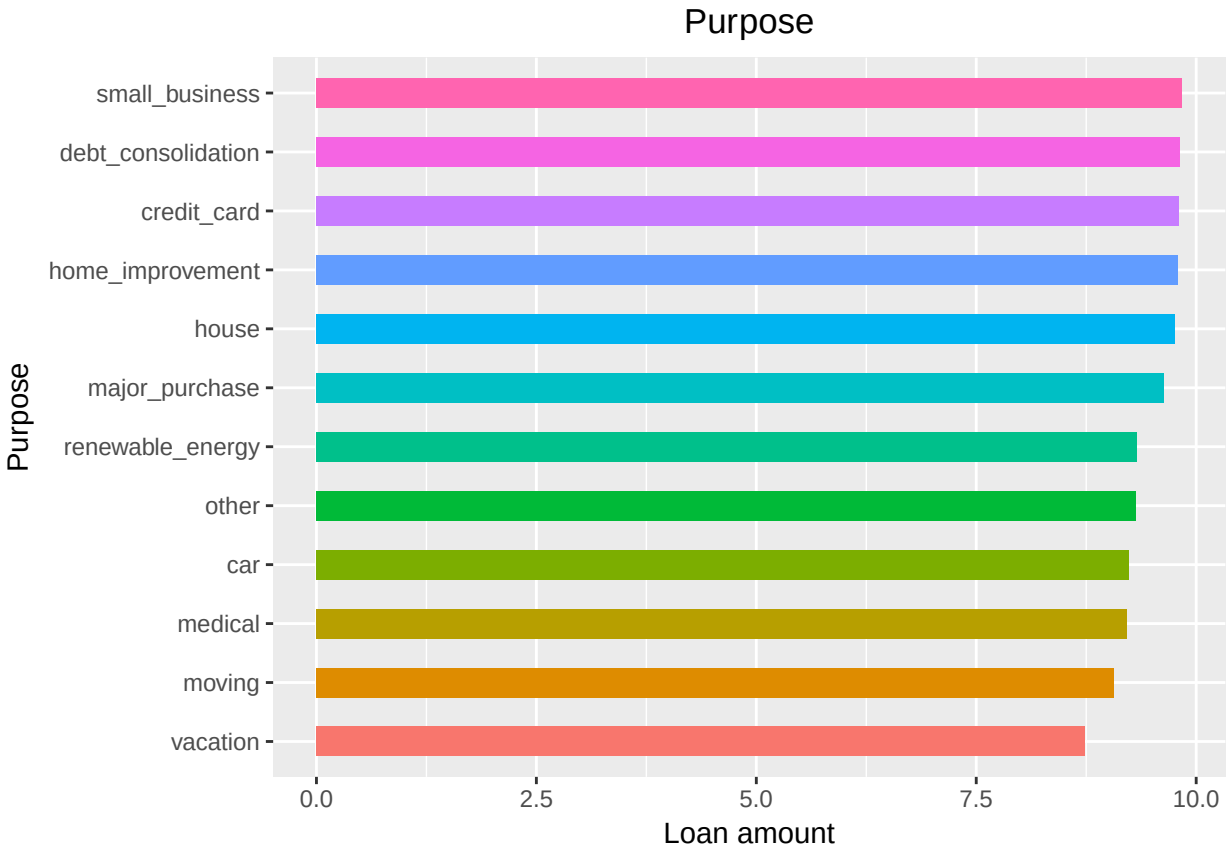
Initial EDA

1. How home ownership influenced loan amount?
2. How employment years inluenced loan amount?
3. How loan purpose influenced loan amount?
4. How state status influenced loan amount?
5. How monthly income influenced loan amount?

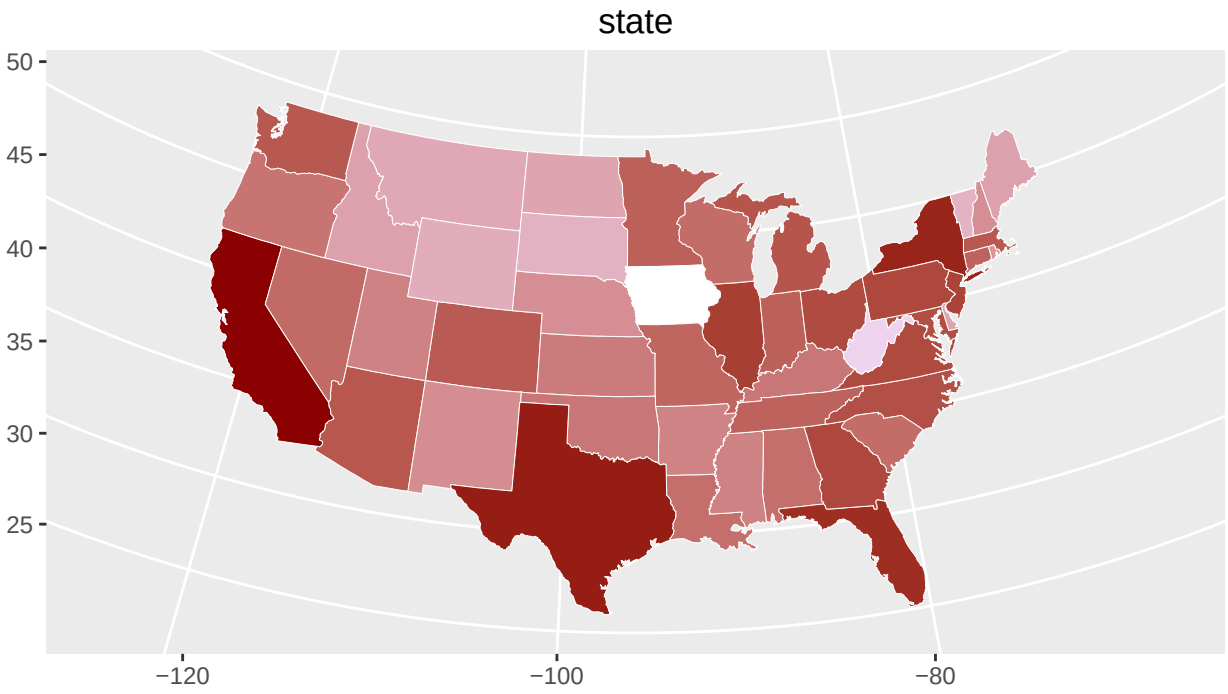


From the chart above, we could see that borrowers whose home is mortgage had greater loam amount than the other three types.

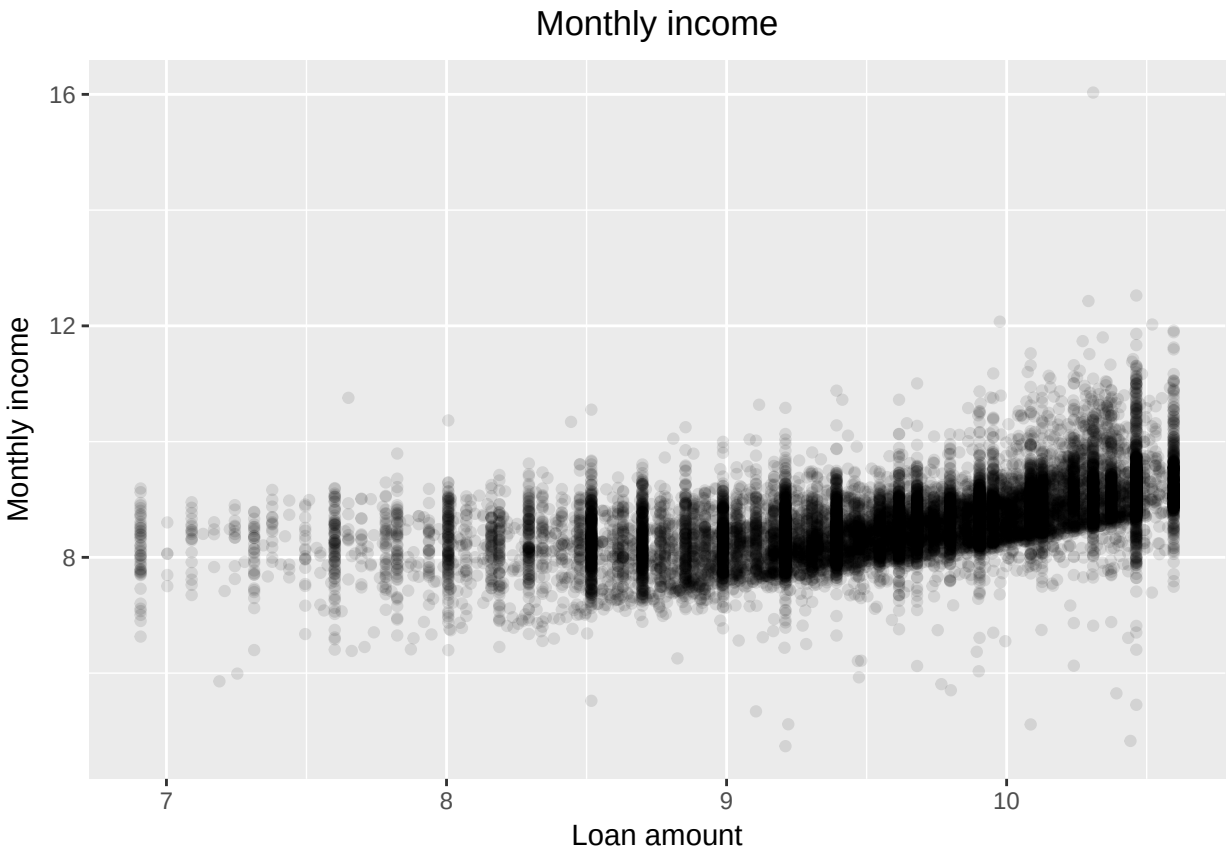
For the employment length, it seems that borrowers who was just get their jobs had more loan amount than those who had already work for years.



Borrowers who loaned money for business, they had greater loan amount than those who loaned just for vacation.



From the map we could clear see that loan amount in CA, TX, FL and NY is higher than other states, which reflects the financial demand in each state.

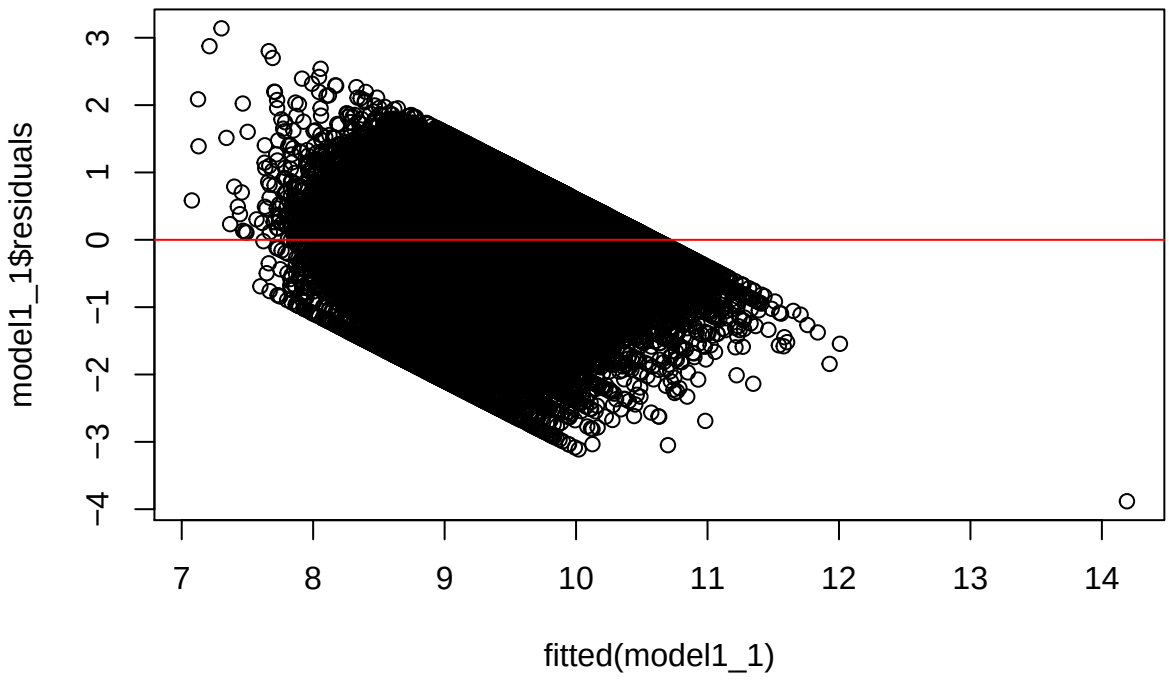


For the monthly income, borrowers who receive higher income per month could have higher loan amount, it is reasonable since they seemd to have higher debt-paying ability, and they seemed to have more source of demand for money.

Model

Then we fit a linear model to show the relationship between loan amount and borrower's status.

model 1 lm(log(loan_amnt) ~ log(monthly_inc) + purpose + home_ownership + emp_year + addr_state)



Check the residual:

Residual standard error: 0.6091.

The residual plot shown above has a trend, which means that this model has problems and maybe other factors should be included into consideration.

Interpret:

For each 1% difference in monthly income, the predicted difference in loan amount is 0.63%.

For each 1% difference in employment year, the predicted difference in loan amount is 0.003%.

The other input, purpose, home ownership, state, is categorical so it does not make sense to take its logarithm.

2. Before loan was funded (Data:2015-2017)

Analyze how LC assigned loan grade: How borrower's status as well as the loan amount and loan time influenced loan grade? (Multinomial regreesion)

Initial EDA: https://dongyuanzhou.shinyapps.io/ma684_shiny/

From the link, we could see that higher loan amount, lower loan grade; lower income, lower loan grade; What's more, longer loan term means higher probability to have lower loan grade and borrowers who rent their home seems to has lower loan grade. There is not much significant diffenrence according the the EDA for loan purose and loan state. Maybe we could use the multilevel regression to show the result.

Model

Then we fit a linear model to show the relationship between loan amount and borrower's status as well as the loan amount and loan time.

model 2 polr(grade ~ log(loan_amnt) + year + log(monthly_inc) + home_ownership + emp_year + purpose + addr_state)

Check residuals: Residual Deviance: 567279.20

AIC: 567425.20

Interpret:

We get emp_year that is positive and insignificant contrary to our expectation. It seems reasonable to remove emp_year variable from our model. Meanwhile, according to the industry regulation, loan purpose and state of loan has little significance compared to monthly income and loan amount.

Therefore, we consider a new model for the grade of loan as a function of loan amount, loan term and borrower's monthly income.

model 3 polr(grade ~ log(loan_amnt) + year + log(monthly_inc))

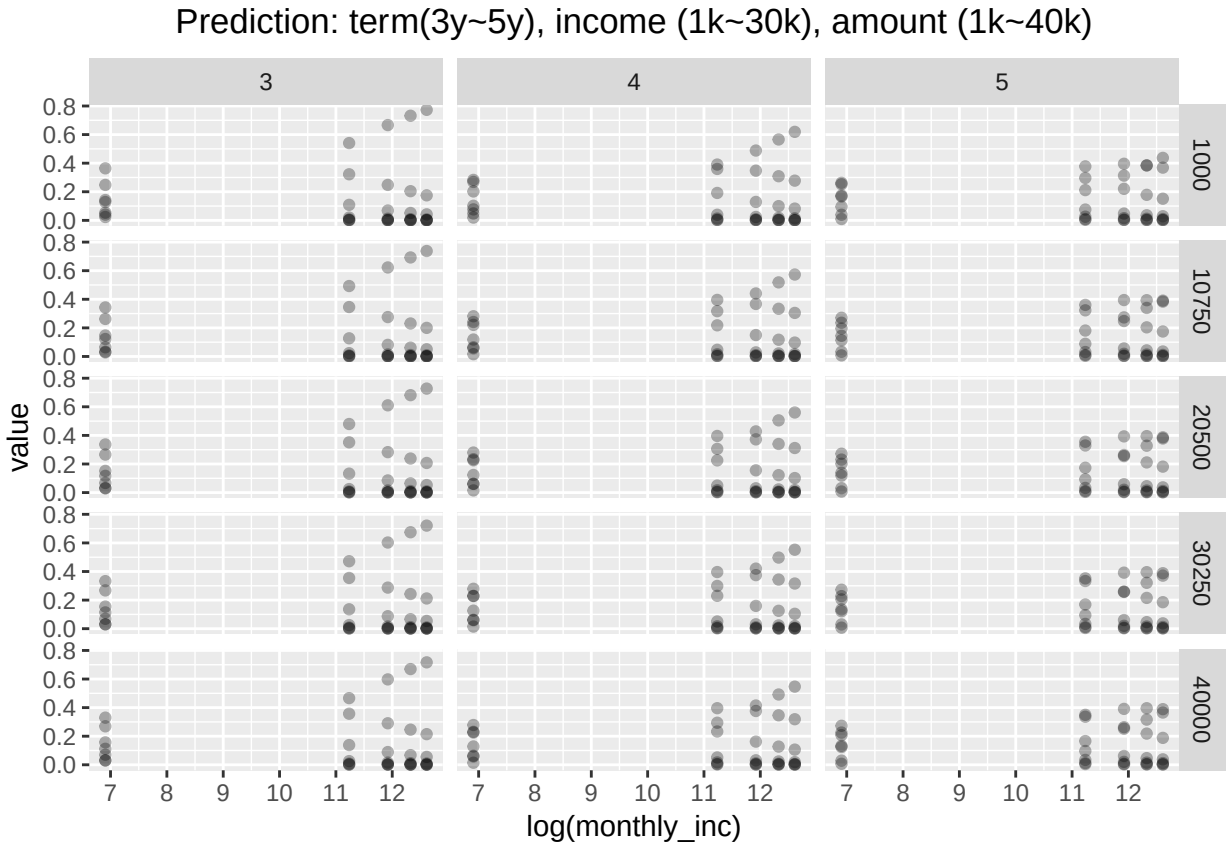
```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = grade ~ log(loan_amnt) + year + log(monthly_inc),
##     data = Loan1517)
##
## Coefficients:
##                      Value Std. Error t value
## log(loan_amnt)    0.08061   0.006914   11.66
## year              0.73920   0.005199  142.18
## log(monthly_inc) -0.77201   0.008807  -87.66
##
## Intercepts:
##     Value    Std. Error t value
## A|B  -5.7369   0.0705   -81.3740
## B|C  -4.0606   0.0696   -58.3036
## C|D  -2.3739   0.0693   -34.2518
## D|E  -1.2098   0.0693   -17.4471
## E|F  -0.0846   0.0698    -1.2120
## F|G   1.1674   0.0714    16.3443
##
## Residual Deviance: 575377.17
## AIC: 575395.17
```

Check rediduals: Residual Deviance: 575377.17

AIC: 575395.17

Then we predict the loan grade for loan term between 3 years to 5 years, monthly income equals 1000 to 40000 and loan amount equals 1000 to 300000.

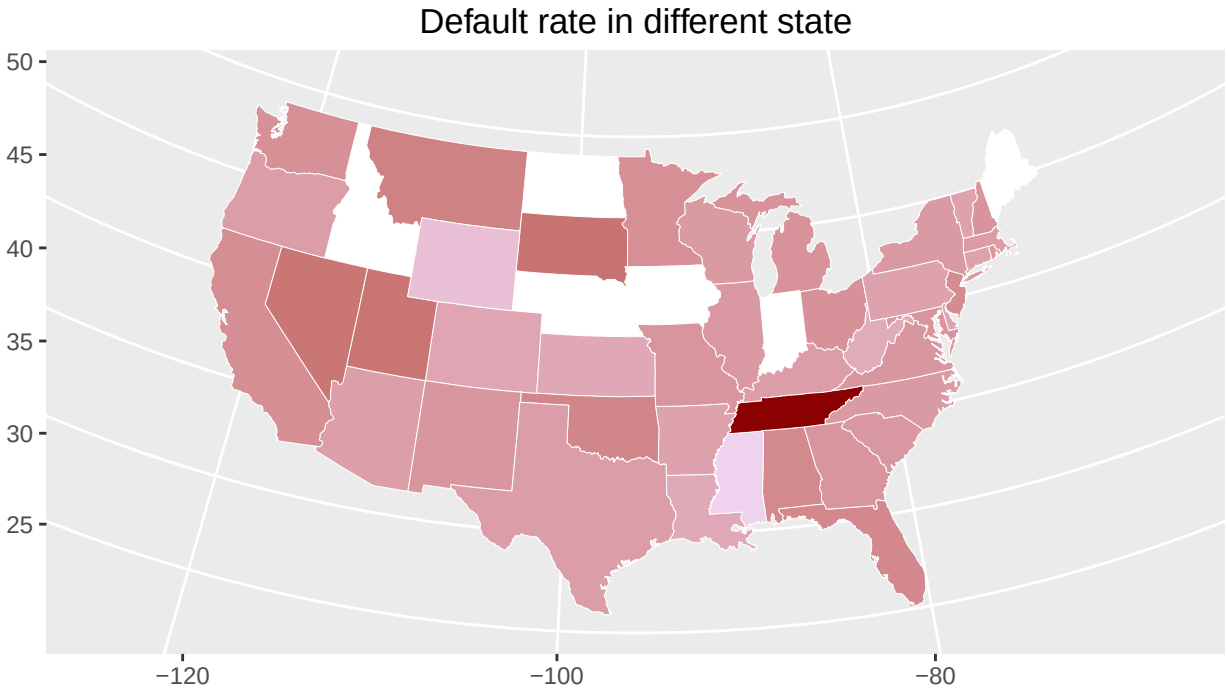


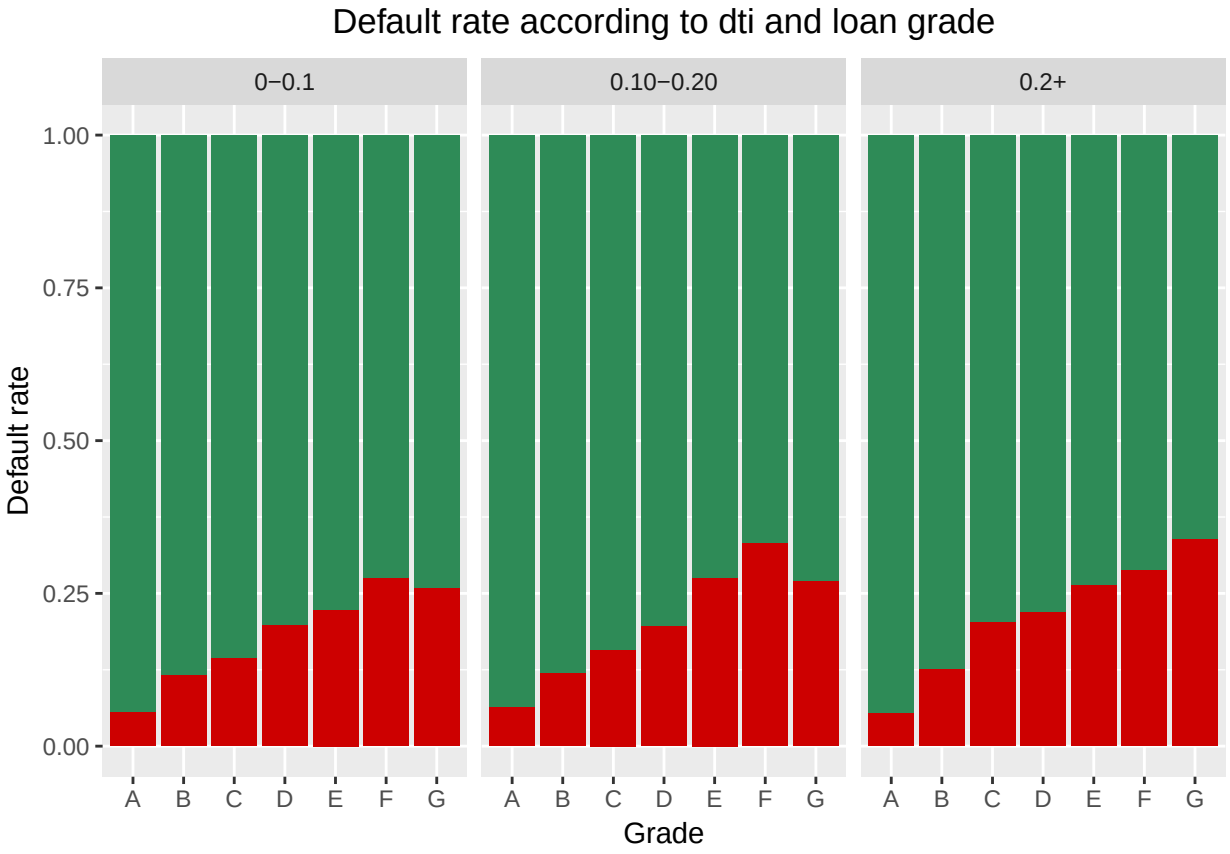
3. After laon ended (Data: 2007-2011)

Summarize default rate for each state.(Multilevel logistic regression)

dti: Debt to income ratio

Initial EDA

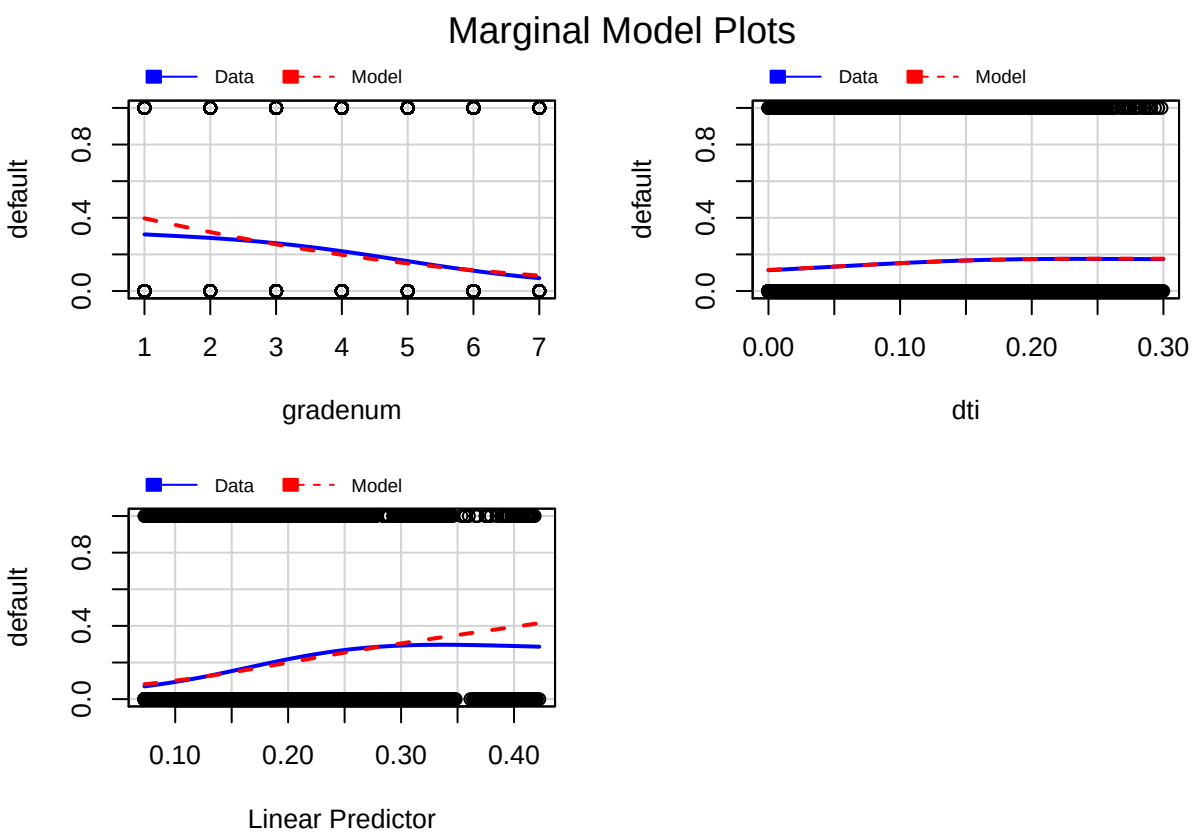

Default rate in different state
50
45
40
35
30
25
−120
−100
−80

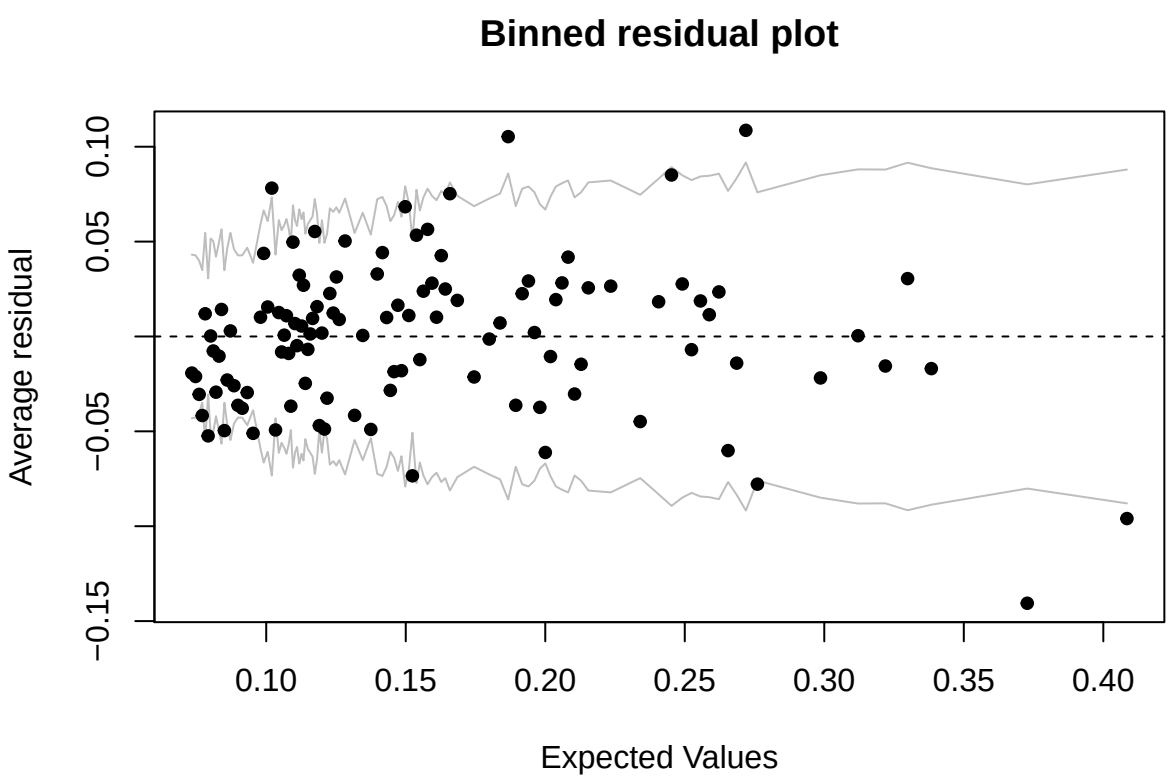


Model

First, we fit a logistic model to show the relationship between the default rate and loan grade as well as the Debt to income ratio.

model 4 glm(default ~ gradenum + dti, family=binomial)


Marginal Model Plots
Data
Model
default
0.0
0.4
0.8
1
2
3
4
5
6
7
gradenum
Data
Model
default
0.0
0.4
0.8
0.00
0.10
0.20
0.30
dti
Data
Model
default
0.0
0.4
0.8
0.10
0.20
0.30
0.40
Linear Predictor



Check model:

From the binned plot, we could see a disturbing pattern, with an extreme negative residual in the last bins: people in the lowest bin are about 10% less likely to default than is predicted by the model.

The error rate for the model is 16%.

interpret:

1 sd increase in debt-to-income ratio has a multiple effect of exp(0.01)=1.01 on odds od default,controling grade in same level.

The odds ratio of default for grade A vs grade G is exp(1.8)= 6.05

The loan in grade B is 0.75/4=19% more likely to default than loan in grade A.

The loan in grade C is 1.13/4=28% more likely to default than loan in grade A.

The loan in grade D is 1.38/4=35% more likely to default than loan in grade A.

The loan in grade E is 1.70/4=43% more likely to default than loan in grade A.

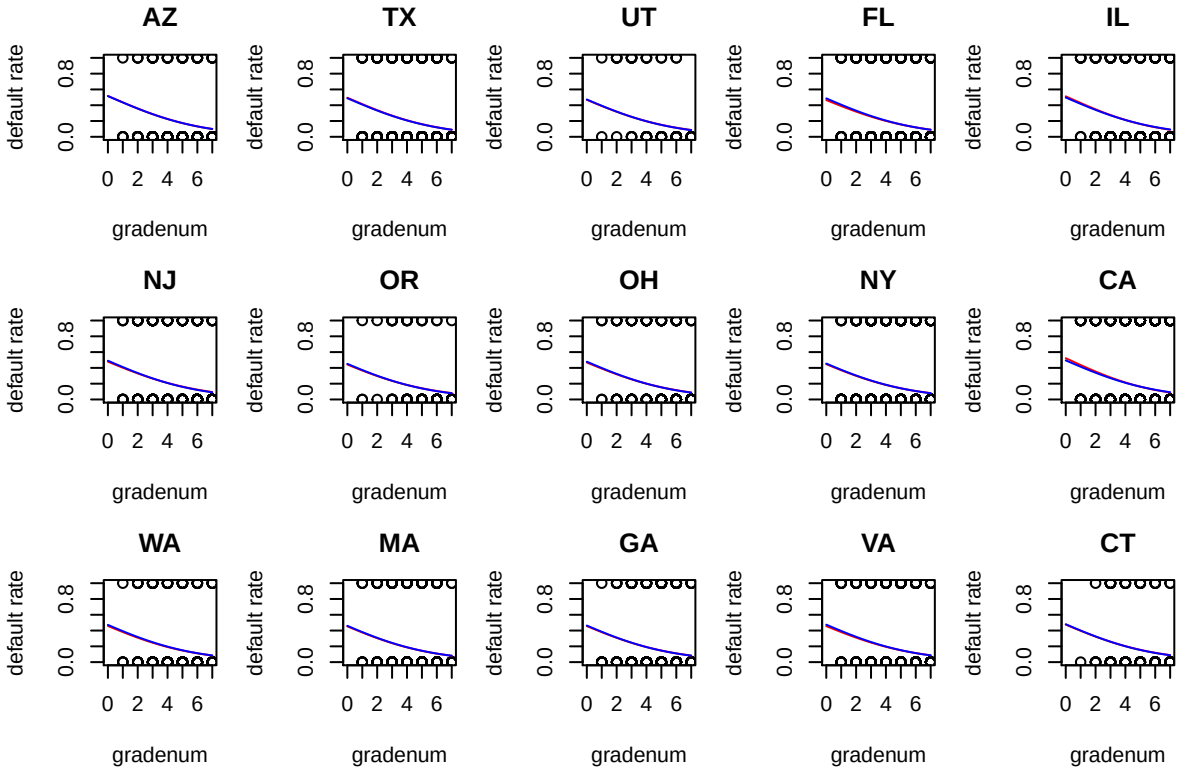
The loan in grade F is 1.94/4=49% more likely to default than loan in grade A.

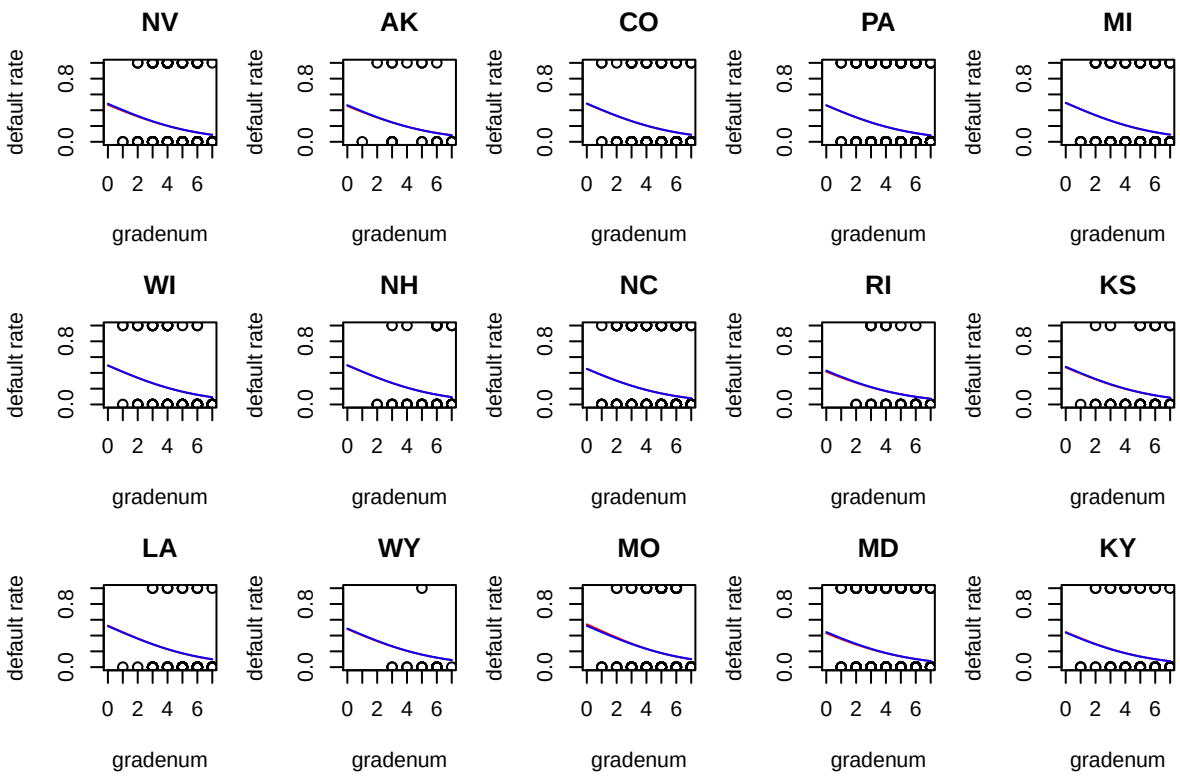
The loan in grade G is 1.82/4=46% more likely to default than loan in grade A.

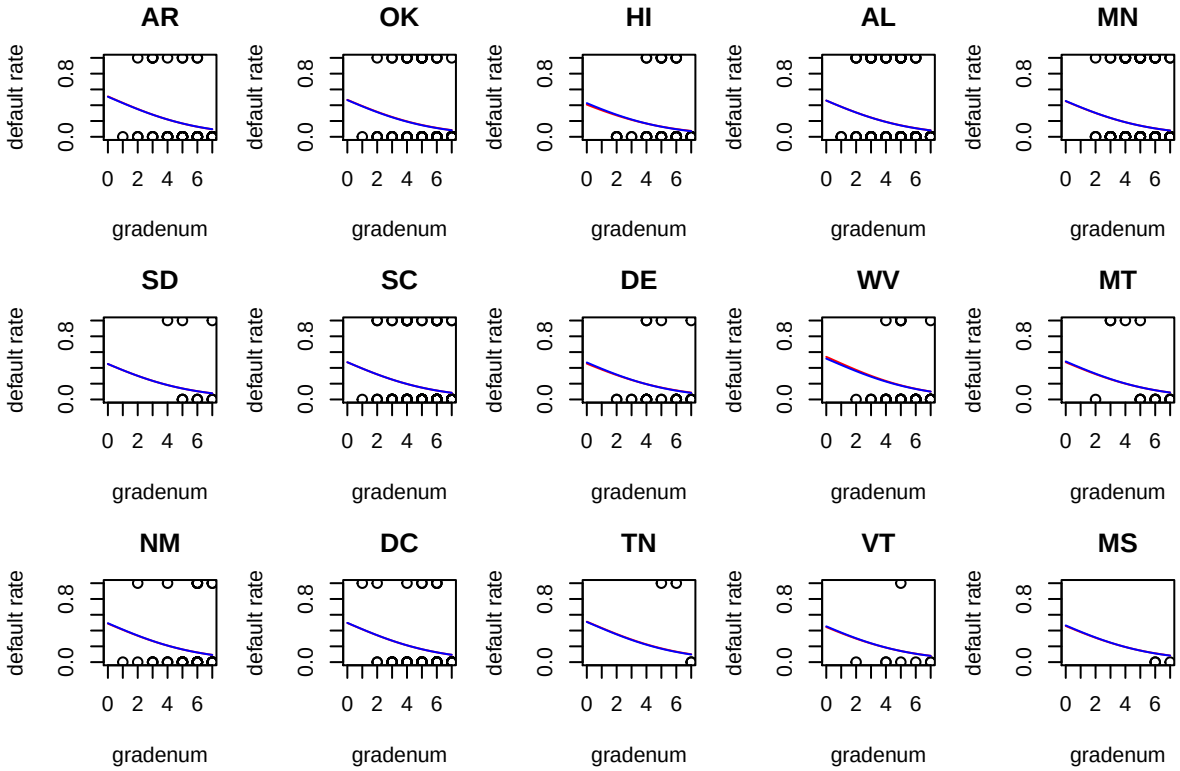
Since state and purpose may also has influence on the default rate, we fit multilevel model to see if there is some difference between state as well as purpose.

model 5 glmer(default ~ gradenum + dti + (1|addr_state), family=binomial(link="logit"))

model 6 glmer(default ~ gradenum + dti + (1 + gradenum|addr_state), family=binomial(link="logit"))

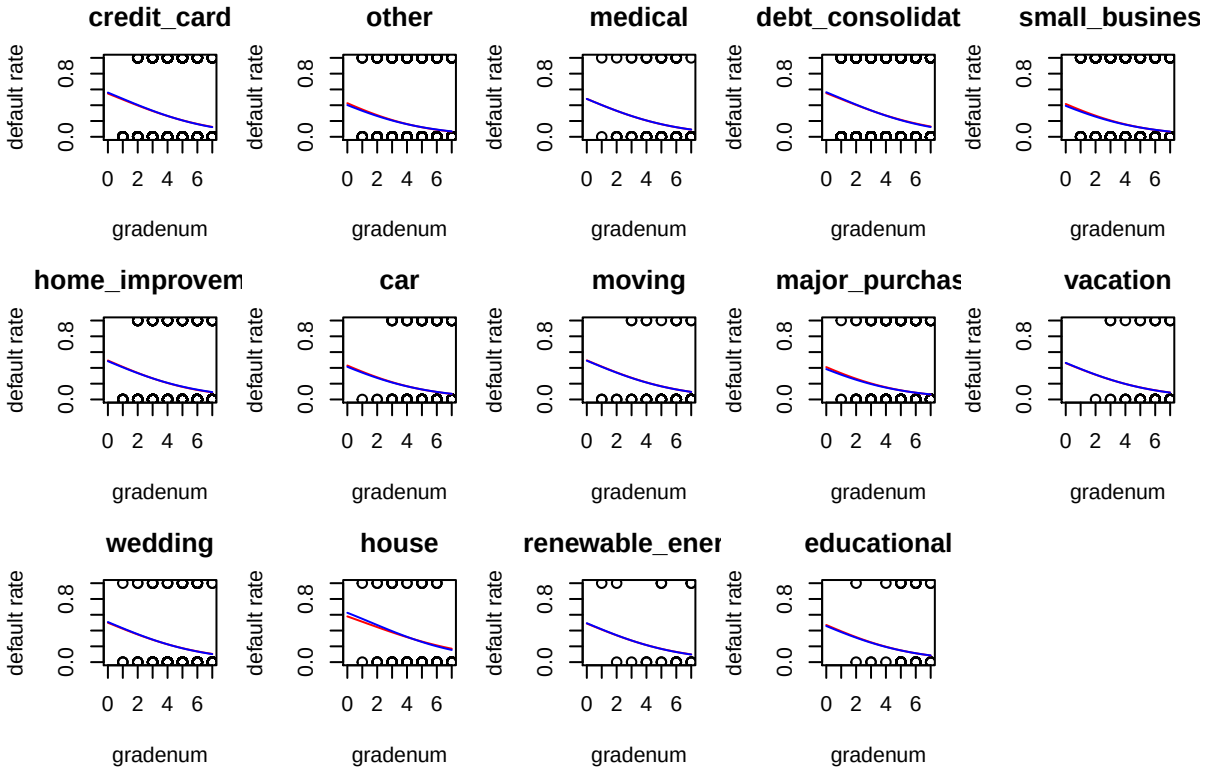

AZ
TX
UT
FL
IL
NJ
OR
OH
NY
CA
WA
MA
GA
VA
CT
default rate
0.0
0.8
0 2 4 6
gradenum


NV
AK
CO
PA
MI
WI
NH
NC
RI
KS
LA
WY
MO
MD
KY
default rate
0.0
0.8
0 2 4 6
gradenum



model 7 glmer(default ~ gradenum + dti + (1|purpose), family=binomial(link="logit"))

model 8 glmer(default ~ gradenum + dti + (1 + gradenum|purpose), family=binomial(link="logit"))

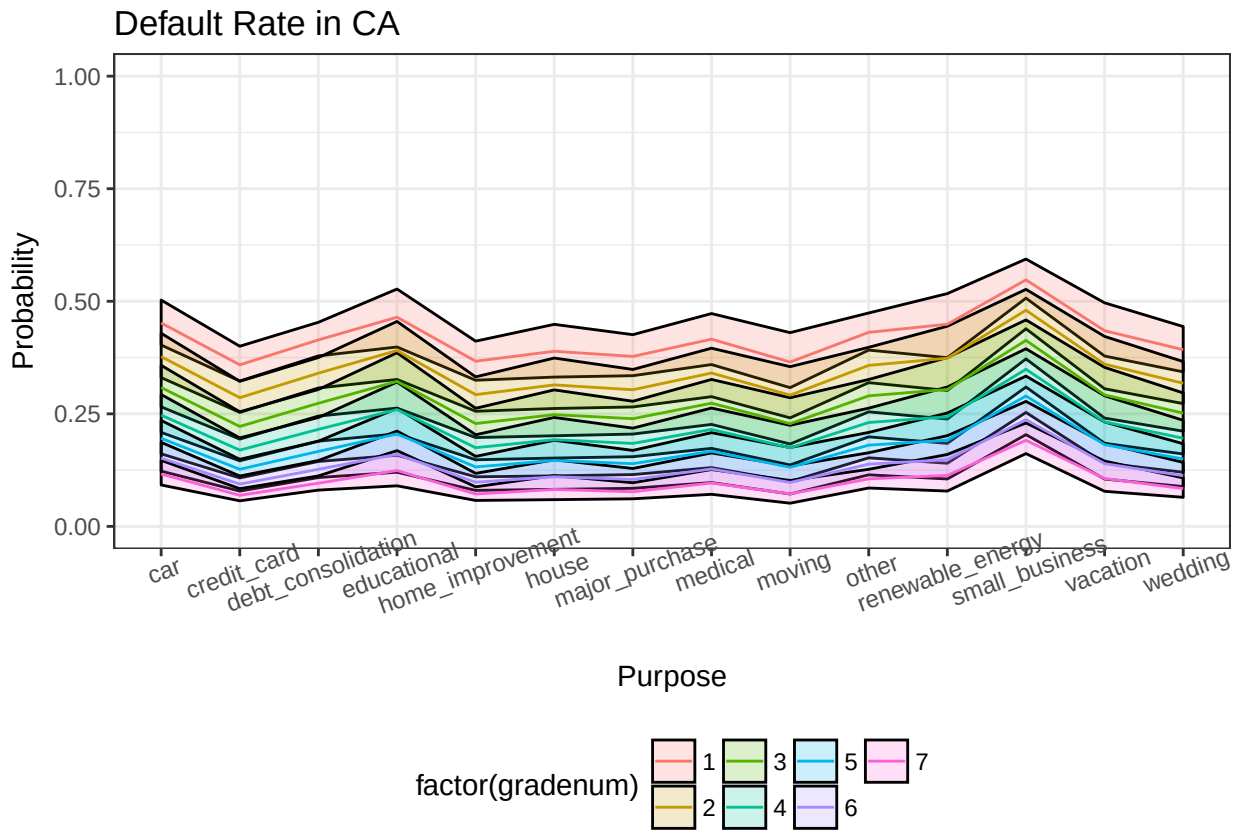


From the plot above, we could see that there is some difference if we include state as random effect, especially in CA.

Also, there is difference for the house purpose when we include it as random effect.

Then we fit a mixed effect model and show the result in CA, we could see that the difference is not very precise.

```
## Number of observations < 20, random effect quantiles may not be well-defined.
```



4. Discussion

1. Result

From this analysis report, we could know that the loan grade is significantly related to loan amount and loan term. What's more, if borrower had another mortgage, the grade would be higher as expected. This is reasonable, since their credit quality had checked by other institutions. However, the employment year is not such related as common sense, and this is a point that could be analyze deeper.

For default rate, the higher grade reflect the lower default rate. What's more, the initial debt to income ratio is also important. From the analysis, we could give suggestion that grade F and grade G has really high default rate, and reject their loan request may be a better desicion.

2. Data limitation

Since the credit information is private, we cannot track the whole process for each loan.

What's more, the FICO Score should be an important predictor in this analysis,however, since it is personal privacy, we could not get the data. Therefore, just consider about the factor as home ownership and income has its limitation.

Finally, since the financial market and the interest rate policy has changed during years, credit quality could not be the only factor of default. Other investment chances and bank loan should be considered and the time span is important.

3. Future directions

Combine the bank interest rate in each years analysis is what we could consider deeper, and basic financial knowledge is required. Moreover, since the loan is always span 3-5 years or more, maybe we could track the

result years later.

Appendix and reference

Data source: https://www.lendingclub.com/info/download-data.action

Code for data cleaning: “data clean.R”

Code for model: “MA684 Project.Rmd”

Initial EDA: https://dongyuanzhou.shinyapps.io/ma684_shiny/