

# Midterm\_Project\_Report

Risk analysis of interest rates

*Weiling Li*

*12/6/2019*

## 1. Introduction

When issuing loans, one of the most important thing a issuer needs to consider is the risk of not getting the money back. If a financial entity can calculate the risk at the point of issuing loans. Then it can adjust the interest rate to mitigate the potential loss.

However, there are some doubts that interest rates actually alter the risk of charged off(see Angbazo 1997), resulting a mismatch between the true risk and the “pre-interest-rate” risk. But how big is the interest-rate risk? How far off are we if the risk is ignored?

To answer this question, we proposed a roadmap to examine the pre- and post-interest-rate risk and a way to exaime the association between the risk of charge-off and the interest-rate.

## 2. Method

### 2.1 Dataset

The data used in this project comes from lending club. It included all the issued loans from 2012 to 2015 and detailed information not only the loans themselves also the credit related attributes of the borrowes at the point of issuing.

Originally, there are 42,390 unique issuing record and each has over 150 different variables describing the loan and the borrower. Due to the scope of this project, only 20 variables were picked and eventually, 11 were used to construct the model. The 11 variables are:

- **y**: Binary variables, 1 indicates the loan was determined as a charged off. 0 indicates no charged off
- **fico\_10**: Borrower’s fico score(high bound) at the point of requesting a loan.
  - This variable is measured at:  $\text{fico\_10} = (\text{fico} - \text{median}(\text{fico}))/10$
- **inq\_last\_6\_mths**: The borrower’s number of inquiries during 6 months before requesting a loan.
- **open\_acc**: The number of opening financial accounts of the borrower(bank accounts or fund accounts, etc.) at the time of requesting a loan.
- **dti\_scale**: **dti** is the ratio of borrower’s total amount of debt to total yearly income, after issuing the loan. this variable was scaled such that it is centered at it’s mean and divided by standard deviation estimates.
- **installment\_log**: The log of monthly installment of the loan + interest rate.
- **addr\_state**: The borrower’s state of residence at the point of requesting the loan.
- **sub\_grade**: The borrower’s **sub\_grade** calculated by lending club at the point of issuing the loan. This grade ranges from A to G with each grade being subdivided into 5 grades from 1 to 5.
- **term**: The expected number of months of paying back the issued loan, for our dataset, only 36 months and 90 months are available.
- **purpose**: the purpose of the loan, categorized by lending club. a total of 14 categories: credit card, car, small business, other, wedding, debt\_consolidation, home\_improvement, major\_purchase, medical, moving, vacation, house, renewable\_energy and education.
- **emp\_length**: Categorical variable describing the length of employment at current position. from **less than 1 year** to **over 10 years** and **not applicable** a total of 12 levels.
- **int\_rate**: interest rate calculated at the time of issuing the loan.

## 2.2 Model Selection

The risk of issuing loans are explained as credit risks. It's definition goes: *the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations* (see Labarre, n.d.)

In this project, we access this risk by calculating the probability(not a categorical problem) of a charge off(no repay) based on the borrower's information presented at the time of requesting a loan. Naturally, logistic regression is selected as the model. Multi-level logistic regression model was used because the data itself presented a hierarchical structure(for example, `addr_state`, `sub_grades`, `emp_length` are all group level variables).

In the model, 11 variables can be categorized as, outcome, individual level predictors(fixed effect) and group level predictors(random effect). The detailed explanation is listed below:

- outcome:  $P(y = 1)$
- Fixed effect:
  - `fico_10`, `inq_last_6mths`, `open_acc`, `dti_scale`, `installment_log`
  - `int_rate`
- Random effect:
  - Random intercept: `addr_state`, `sub_grade`, `term`, `purpose`, `emp_length`
  - Random slope: `int_rate:sub_grade`

The first model will use all the variable except `int_rate` to evaluate the risk, then `int_rate` will be added into the second model and then compare the model fit and it's coefficients.

## 2.3 Experimental Design

The whole dataset was divided into training set and validation set. The model was trained on training set data and it's performance was validated in the testing set to avoid over training. The acquisition of the training set and the validation set is done in the following manner:

- The ratio of training vs testing was set to 7:3.
- Separate the original dataset according to `y` into `y = 0` and `y = 1`
- Within the two table, randomly draw 30% of the rows and combine them as validation set.
- Combine the remaining rows of the two tables as training set

The resulting training dataset has 29,637 records and the validation set has 12,717 records

## 2.4 Result Validation and Inference

There are different ways to access a logistic regression model's fit. Since the purpose of this project is to study the probability of charge off conditioned on the borrower's credit history. It is unnecessary to make categorical predictions, meaning setting threshold to predict 0 response or 1 response. In fact during the model validation process, it is impossible to make such predictions because the overall probability estimated of charge off is quite low.

Instead of categorical predictive power, the more adequate validation method is to use the model to simulate the original dataset, then examine the simulated distribution of charge off rate compared to the original data.

When the model is deemed to be valid, statistical inference will be drawn from the coef estimates. especially, the two models mentioned above will be compared to explore the association of `int_rate` and the risks as well as it's potential impact.

## 2.5 Model Estimation Package

For this project, `glmer` function with `bimomial` family and `logit` link function in `lmer` package was used. Alternatively, `winBUGS` and `rstanarm` packages can also be used to evaluate the model coefficients.

## 2.6 Limitations

There are 3 major limitations associated with the project.

1. One of the biggest assumptions associated with this project comes with the limitation of the dataset. In the real-world situation, one is able to apply for more than one loans. However, because the dataset masked all the member identification info, the assumption is that: when conditioned at all the credit information at the point of applying a loan, the risk is independent of the member id variable.
2. The second biggest limitation of this model is predictive power into the future. because the estimation only took into account the fiscal year, then it is in theory not adequate to access the change of risk due to time. In other words, the model assumes the risk is independent of time.
3. Because of the nature of the `lmer` package, coefficients and standard errors estimated is not the most accurate. To achieve more accurate results, one can use `stan_glmer` from `rstanarm` instead of `glmer` and set the `intercept_prior = NULL`. Because of the computing cost of this approach, this project will not include this approach into the analysis.

However, this does not affect our research goal if we are only examine the inference as a year average from 2012 to 2015. The deviation of this projects findings to the future true effects depends on the risk's rate of change with time which can be estimated in future works.

## 3. EDA and Model Building

### 3.1 Data Wrangling and EDA

During the data wrangling and EDA process, all 150+ variables were examined and filtered based on the completeness, informativeness and relevance. Generally, variables contain over 70% NA values are dropped. It is not advised to do so in a more serious settings, but for the scope of this project, unless the model fit is too poor, these variables will be ignored. Variables providing the same information in different coding or context will be discarded while keeping only one (for example, variables like `purpose` and `purpose_detail`, unless estimating the difference within each purpose is desired, the detailed info will be discarded unless doing so resulting a poor model fit).

After selecting the relevant variables, EDA was performed to get a feel of the overall data. Mosaic plot and histogram was used to examine the distribution of each variable by its own or conditioned by the outcome.

The source code can be found in the following files:

- Data wrangling: `./Data_Wrangling_&_EDA/Lendingclubreaddata.R`
- EDA: `./Data_Wrangling_&_EDA/EDA.R`

### 3.2 Model Building

After EDA, 18 predictors was selected and cleaned (includes scaling, taking log, etc.). the full list of variables is shown below:

variables
term
grade
sub_grade
emp_length
verification_status
purpose
addr_state
delinq_2yrs
inq_last_6mths
open_acc
chargeoff_within_12_mths
delinq_amnt
loan_amnt_log
cr_length_log
fico_10
installment_log
dti_scale
int_rate_scale

Using the above variables, 2 models were constructed.

The 1st model building is constructed of the following steps:

1. Further drop un-informative variables using no-pooling logistic regression's AIC.
2. Use the model selected by AIC and turn it into a multi-level model as described in previous sections.
3. Examine the model fit via binned residual plot and the conditional distribution of the simulated "fake original dataset"

The 2nd model building is constructed of the following steps:

1. Use the formula of model 1, add `int_rate` as fixed effect and it's interaction with `sub_grade` as random effect.
2. Access model fit using the same method and model 1.

After model validation, two models were compared and it's inference were drawn. Then, the effect of `int_rate` was also examined.

## 4. Results

### 4.1 Model 1

#### 4.1.1 Model Description and Validation

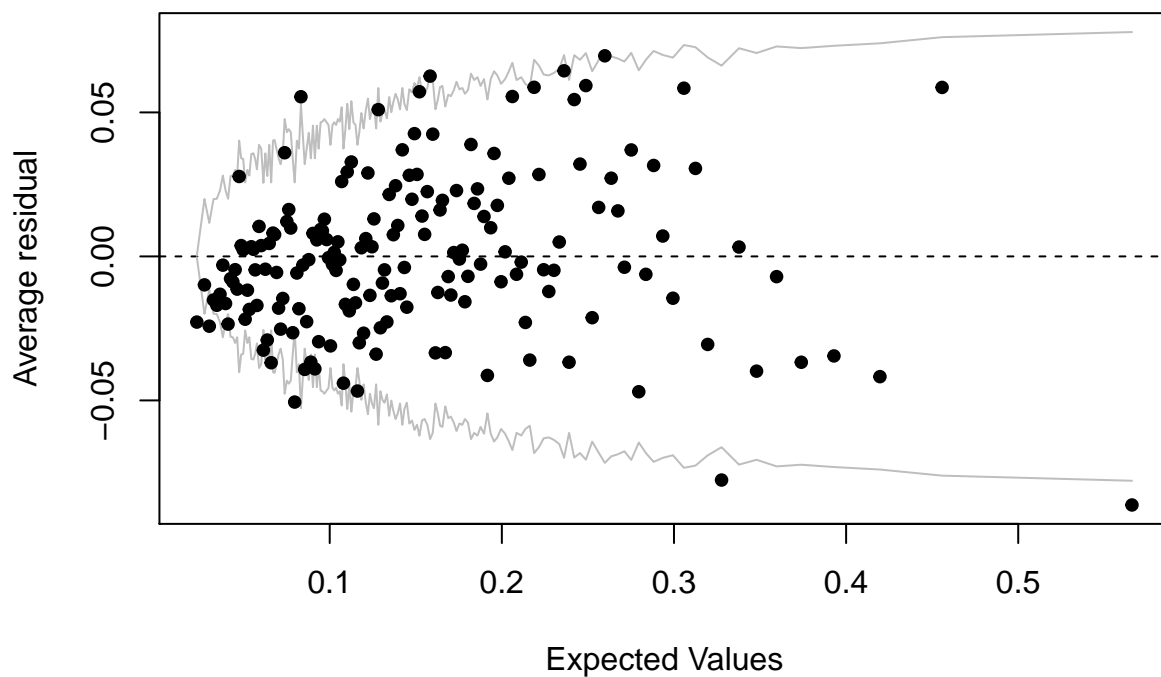
Following the described steps the first model was built. The results are shown below:

```
## glmer(formula = y ~ fico_10 + inq_last_6mths + open_acc + dti_scale +
##      installment_log + delinq_2yrs + (1 | addr_state) + (1 | sub_grade) +
##      (1 | term) + (1 | purpose) + (1 | emp_length), data = lendingclub_model_train,
##      family = binomial)
##               coef.est coef.se
## (Intercept)    -1.38    0.33
## fico_10         -0.07    0.01
## inq_last_6mths  0.14    0.01
## open_acc        -0.02    0.00
## dti_scale        0.08    0.02
## installment_log -0.02    0.03
## delinq_2yrs     -0.06    0.03
```

```
##
## Error terms:
## Groups      Name      Std.Dev.
## addr_state  (Intercept) 0.15
## sub_grade   (Intercept) 0.31
## purpose     (Intercept) 0.32
## emp_length  (Intercept) 0.18
## term        (Intercept) 0.36
## Residual    1.00
## ---
## number of obs: 29673, groups: addr_state, 50; sub_grade, 35; purpose, 14; emp_length, 12; term, 2
## AIC = 23442.2, DIC = 22924.7
## deviance = 23171.5
```

The binned residual plot is shown below

**Binned residual plot**



There are some issue with the residual plot, we can observe a trend rising and then declining pattern in the residuals

Under current model, we have the following performance:

```
## expected occurence of charged of under model 1 is 1846
## true observed occurence of charged off in validation set is 1921
```

The predicted occurence is really close to the original dataset. After examine the distribution using mosaic plot, the model is able to capture the overall trend of the data. One example is shown below:



#### 4.1.2 Inference

#### 4.2 Model 2

##### 4.2.1 Model Description and Validation

##### 4.2.2 Inference

#### 4.3 Model Comparison

### 5. Conclusions

#### 5.1 Risk Estimation

#### 5.2 int\_rate Association Estimation

#### 5.3 Future Directions

### Reference

Angbazo, Lazarus. 1997. "Commercial Bank Net Interest Margins, Default Risk, Interest-Rate Risk, and Off-Balance Sheet Banking." *Journal of Banking & Finance* Volume 21, Issue 1, January 1997, Pages 55-87

Labarre, Olivia. n.d. "Credit Risk." <https://www.investopedia.com/Terms/c/Creditrisk.asp>.