Credit Score

Haoran Su

12/8/2020

Abstract

Normally the credit scoring algorithm takes the linear generalized model as the calculating method of probability of credit bankruptcy. From records of credit history of normal information, by taking age and income as groups, we explore the correlation between the likelihood of having delinquent between groups. The age groups seems to have random effects in predicting the possibility of credit bankruptcy.

Introduction

Individuals and companies require access to credit for investments or consumption.Financial institutions have to decide how many credits should be given to them. The banks commonly use Credit scoring algorithm to determine whether or not a loan should be granted. The Credit scoring algorithm makes a guess at the probability of default. This project is aimed to improve credit scoring by predicting the probability that somebody will experience financial distress in the next two years. It considers the factors of personal background, the credit return history.

Methods

Data Source

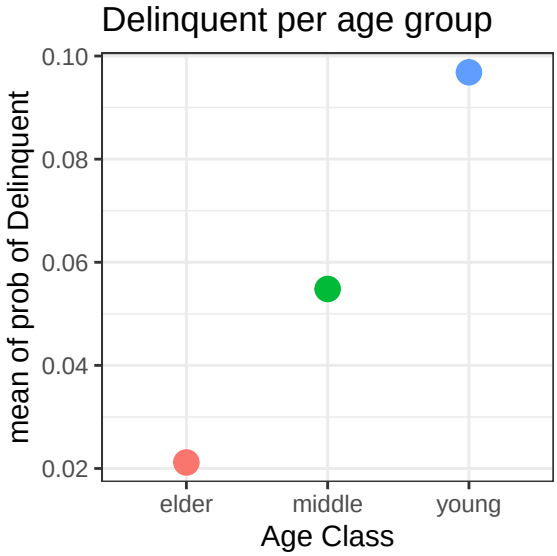
As the project I choose is from a kaggle competition“ Give me some credit”, so the data source is from the competition organizers.

By visualizing the data set and taking a look at the summary of the each columns. I removed the outliers having age smaller than 16. Also, for the number recording the times people past the due dates, I removed the number larger than the maximum possible times.

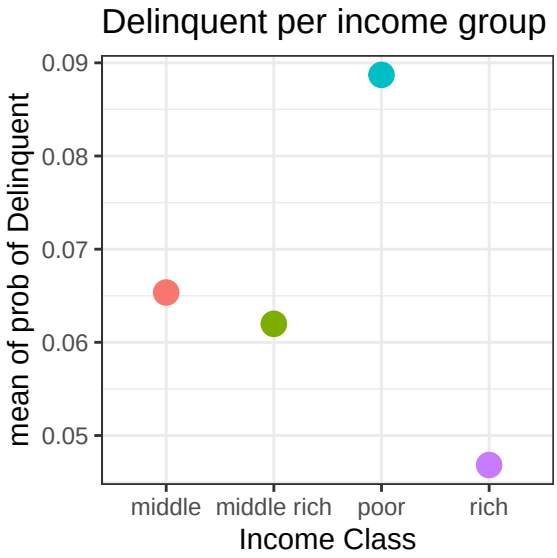
```
## # A tibble: 2 x 2
## # Groups:   SeriousDlqin2yrs [2]
##   SeriousDlqin2yrs      n
##              <int>  <int>
## 1                0 135005
## 2                1   9559
```

In 144564 lines of people's credit, 6.6% of them have record of serious delinquent in 2 years time. Among the records, the distribution of age is similar to the normal distribution. Besides age, the distribution of other factors(income, number of times exceeding the due dates 30 days more, number of dependents, balance divided by credit limit) are extreme left skewed. 75% of the people have income lower than 7500, 25% of people have income between 7500 and 3000000. With the maximum of 54 of the balance on credit cards divided by the credit limits, 75% of people is under 2.

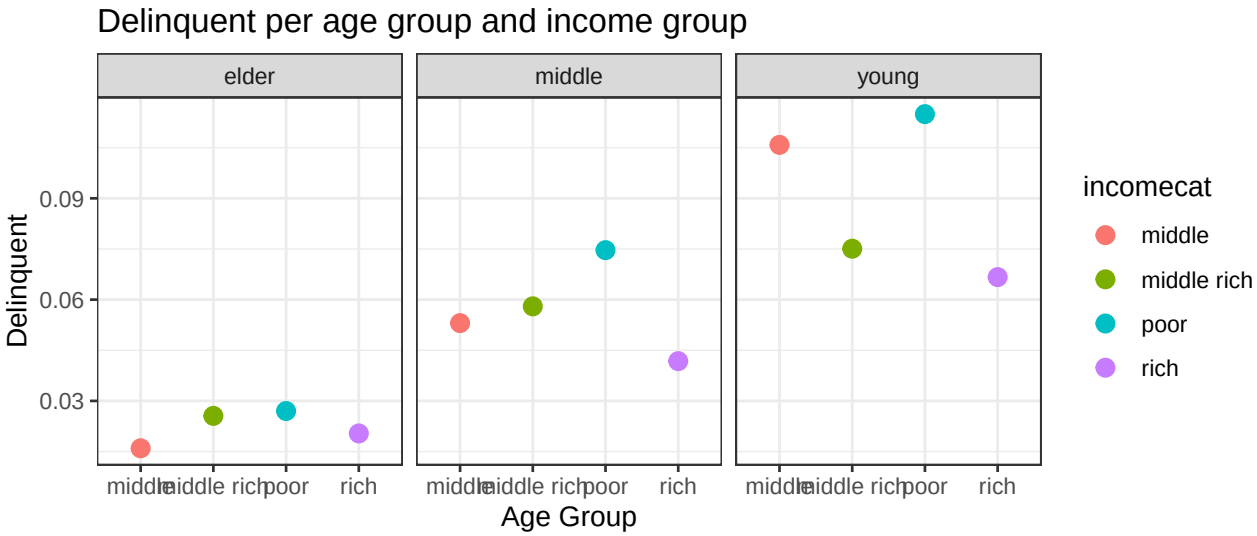
Model Choice



From the plot, it seems that the the younger age group has higher prob of Delinquent.



From the plot, it seems that the the richer age group has lower prob of Delinquent.



This plot shows the ratio of delinquent in separated groups of people of different income and ages. The plot shows that in each income group, people of different age group has a trend that the younger age group has smaller ratio of delinquent. The effect of income is not so apparent. The ratio of delinquent in elder people has a smallest to largest order in the income group: middle, rich, middle rich, poor. While the ratio of delinquent in young people has a smallest to largest order in the income group: rich, middle rich, middle, poor. The order in middle-age group is rich, middle, middle rich, poor.

As the outcome is binary for whether or not the person has overdue return in 2 years, so I consider about logistic regression. Also as the plots show, the age group might has a random effect so I try to use the generalized multilevel models. (To make it take less time to run, I make a sample subset of 8000 samples.)

Results

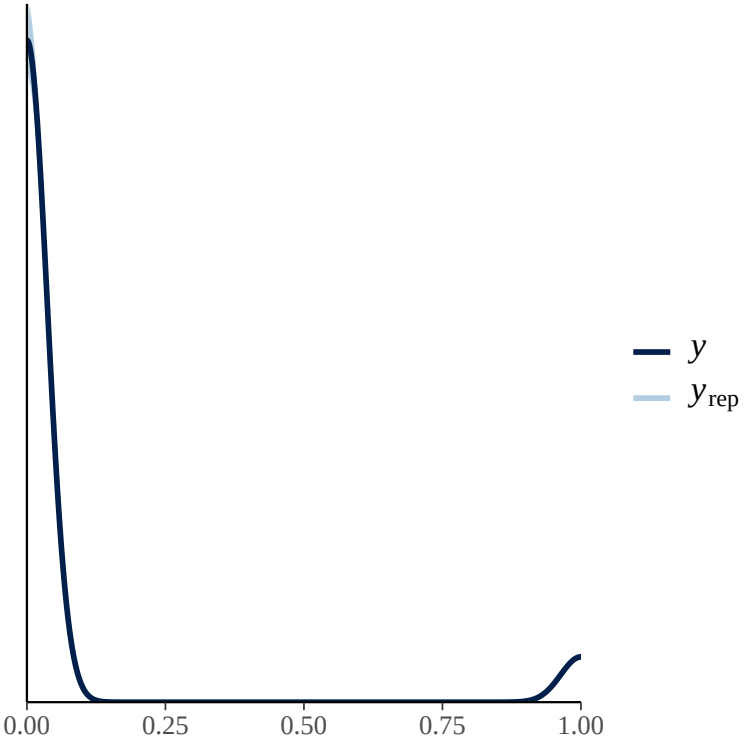
Multilevel mixed effect model

```
##
## Model Info:
##  function:     stan_glmer
##  family:       binomial [logit]
##  formula:      SeriousDlqin2yrs ~ agecat + (1 | incomecat)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 8000
##  groups:       incomecat (4)
##
## Estimates:
##                                              mean   sd   10%   50%   90%
## (Intercept)                                 -4.3    0.5 -4.9  -4.3  -3.8
## agecatmiddle                                 1.4    0.4  0.9   1.4   1.9
## agecatyoung                                  2.0    0.4  1.5   2.0   2.5
## b[(Intercept) incomecat:middle]              0.0    0.2 -0.3   0.0   0.2
## b[(Intercept) incomecat:middle_rich]         0.0    0.2 -0.3   0.0   0.2
## b[(Intercept) incomecat:poor]                0.3    0.2  0.0   0.3   0.5
## b[(Intercept) incomecat:rich]               -0.3    0.2 -0.6  -0.3  -0.1
## Sigma[incomecat:(Intercept),(Intercept)]     0.2    0.4  0.0   0.1   0.5
##
```

```
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.1    0.0  0.1   0.1   0.1
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                                          mcse Rhat n_eff
## (Intercept)                              0.0  1.0  1219
## agecatmiddle                             0.0  1.0  1800
## agecatyoung                              0.0  1.0  1743
## b[(Intercept) incomecat:middle]          0.0  1.0   793
## b[(Intercept) incomecat:middle_rich]     0.0  1.0   902
## b[(Intercept) incomecat:poor]            0.0  1.0   779
## b[(Intercept) incomecat:rich]            0.0  1.0   862
## Sigma[incomecat:(Intercept),(Intercept)] 0.0  1.0   479
## mean_PPD                                 0.0  1.0  3675
## log-posterior                            0.1  1.0  1071
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

From the multilevel mixed effect model fitted, comparing to the old people group, people in the middle-age group has 35% more likely to be seriously delinquent in 2 years. Comparing to the old people group, people of young age has 50% more likely to be seriously delinquent. It is an interesting result to see that the difference of delinquent possibility between different age groups. For people within the same age group, when they are poor(having monthly income less than 3812), it is more likely that they would be delinquent. When they are in the rich income groups(having monthly income more than 7500), they have smaller possibilities of being delinquent.

Model Check



The plot of the posterior prediction and estimates shows that the overlapping of dots is quite good, which shows that the model used is a good fit.

Discussion

1.By grouping people by ages, there is an interesting finding of the difference of the possibility of delinquent between the groups. Age is seldom considered in the credit scoring calculations, we don't directly know the causation of the relationship. What is making difference between different age groups. Future analysis could be targeted at the correlation between credit defaulting possibility and age groups.

2. Although the GGLM model seems to fit well, for the time limit, the result is iterated from random sampling, which might not be able to tell the real story of all. Also, the grouping standard is by the percentage distribution of the data, I took each 1/4 boundary line as the grouping boundary. It works for this selected dataset but is not permitted for the others. Different approach of making group might lead to different result, which needs more examination.

Reference

[1]Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01. [2]Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A (2019). "Visualization in Bayesian workflow." *J. R. Stat. Soc. A*, *182*, 389-402. doi: 10.1111/rssa.12378 (URL: https://doi.org/10.1111/rssa.12378). [3]McKeon, C. S., Stier, A., McIlroy, S., & Bolker, B. (2012). Multiple defender effects: Synergistic coral defense by mutualist crustaceans. Oecologia, 169(4), 1095–1103. http://doi.org/10.1007/s00442-012-2275-2 [4]Kaggle.Give me some credit.(2011). Retrieved Dec,2020 from https://www.kaggle.com/c/GiveMeSomeCredit/data.
[5]The EDA process gets inspirations from:Nicolas.Exploratory Data Analysis - Preparing for one of the top performing models.(2020). Retrieved Dec,2020 from https://www.kaggle.com/nicholasgah/eda-credit-scoring-top-100-on-leaderboard.

Appendix