516 Case 3 Loan

Kun Qian

2/12/2021

In this project, we use publicly available data from LendingClub, a website that connects borrowers and investors over the internet. There are 9,578 observations, each representing a 3-year loan that was funded through the LendingClub.com platform between May 2007 and February 2010. There are 14 variables in the dataset. We will be trying to predict **NotFullyPaid**, using all of the other variables as independent variables.

Setup

```
set.seed(1234)
# load data
loan <- read.csv("Loans.csv")
# inspect data
# head(loan)
str(loan)
```

```
## 'data.frame':    9578 obs. of  14 variables:
##  $ CreditPolicy  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Purpose       : chr  "debt_consolidation" "credit_card"
"debt_consolidation" "debt_consolidation" "debt_consolidation" ...
##  $ IntRate       : num  0.119 0.107 0.136 0.101 0.143 ...
##  $ Installment   : num  829 228 367 162 103 ...
##  $ LogAnnualInc  : num  11.4 11.1 10.4 11.4 11.3 ...
##  $ Dti           : num  19.5 14.3 11.6 8.1 15 ...
##  $ Fico          : int  737 707 682 712 667 727 667 722 682 707 ...
##  $ DaysWithCrLine: num  5640 2760 4710 2700 4066 ...
##  $ RevolBal      : int  28854 33623 3511 33667 4740 50807 3839 24220 69909
5630 ...
##  $ RevolUtil     : num  52.1 76.7 25.6 73.2 39.5 51 76.8 68.6 51.1 23 ...
##  $ InqLast6mths  : int  0 0 1 1 0 0 0 0 1 1 ...
##  $ Delinq2yrs    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ PubRec        : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ NotFullyPaid  : int  0 0 0 0 0 0 1 1 0 0 ...
```

a) Let us start by building a logistic regression model

i. Split the data, report the accuracy on the test set of a simple baseline model.

```
library(caTools)
# create a vector to split the dataset into train and test
split <- sample.split(loan$NotFullyPaid, SplitRatio = 0.7)
table(split)
```

```
## split
## FALSE  TRUE
##  2873  6705

train <- loan[split==TRUE,]
test <- loan[split==FALSE,]

# check the distribution of ture and false in the objective variable
sum(train$NotFullyPaid)/nrow(train)

## [1] 0.1600298

# Since 0 is majority, i.e. loans are fully paid back, we will use predicting
all 0 as the baseline model
accuracy_baseline <- nrow(test[test$NotFullyPaid==0,])/nrow(test)
accuracy_baseline

## [1] 0.8398886
```

Accuracy of the baseline model is 83.99%

ii. Build a logistic regression model that predicts NotFullyPaid

```
# Build the logistic regression model
loanLog <- glm(NotFullyPaid~., data=train, family=binomial)
summary(loanLog)

##
## Call:
## glm(formula = NotFullyPaid ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4651  -0.6141  -0.4914  -0.3670   2.5102
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 7.343e+00  1.548e+00   4.744 2.09e-06 ***
## CreditPolicy               -3.034e-01  1.009e-01  -3.008 0.002629 **
## Purposecredit_card         -5.774e-01  1.311e-01  -4.403 1.07e-05 ***
## Purposedebt_consolidation -4.260e-01  9.219e-02  -4.621 3.81e-06 ***
## Purposeeducational          1.278e-03  1.870e-01   0.007 0.994547
## Purposehome_improvement     1.044e-01  1.485e-01   0.703 0.481923
## Purposemajor_purchase      -3.583e-01  1.978e-01  -1.812 0.070044 .
## Purposesmall_business       4.947e-01  1.381e-01   3.582 0.000341 ***
## IntRate                     3.212e+00  2.080e+00   1.544 0.122508
## Installment                 1.144e-03  2.098e-04   5.453 4.95e-08 ***
## LogAnnualInc               -4.237e-01  7.141e-02  -5.932 2.99e-09 ***
## Dti                         3.990e-03  5.448e-03   0.732 0.463943
## Fico                       -7.357e-03  1.690e-03  -4.354 1.34e-05 ***
## DaysWithCrLine              1.088e-05  1.560e-05   0.698 0.485358
## RevolBal                    3.477e-06  1.143e-06   3.042 0.002350 **
```

```
## RevolUtil                   2.159e-03  1.529e-03   1.412 0.157951
## InqLast6mths                9.678e-02  1.632e-02   5.930 3.03e-09 ***
## Delinq2yrs                 -2.784e-02  6.436e-02  -0.433 0.665374
## PubRec                      2.541e-01  1.146e-01   2.217 0.026607 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5896.6  on 6704  degrees of freedom
## Residual deviance: 5476.2  on 6686  degrees of freedom
## AIC: 5514.2
##
## Number of Fisher Scoring iterations: 5
```

Significant variables:

- Credit Policy
- Purpose is either: credit_card, debt_consolidation, major_purchase, or small_business
- installment
- LogAnnualInc
- Fico
- InqLast6mths

iii. Application A has a FICO credit socre of 700 while Application B has a FICO score of 710. What's Logit(A)-Logit(B)?

$$\begin{aligned} Logit(A) - Logit(B) &= \beta_{FICO} * X^{A}_{FICO} - \beta_{FICO} * X^{B}_{FICO} \\ &= -0.00736 * (700 - 710) \\ &= 0.0736 \end{aligned}$$

iv. Predict the probabilityu of the test set loans not being paid back in full.

```
# predict the risk probability
PredictedRisk <- predict(loanLog, newdata = test, type="response")
# add the predicted probability as a column to the test data
test$PredictedRisk <- PredictedRisk
# report accuracy
sum(round(test$PredictedRisk) == test$NotFullyPaid)/nrow(test)
```

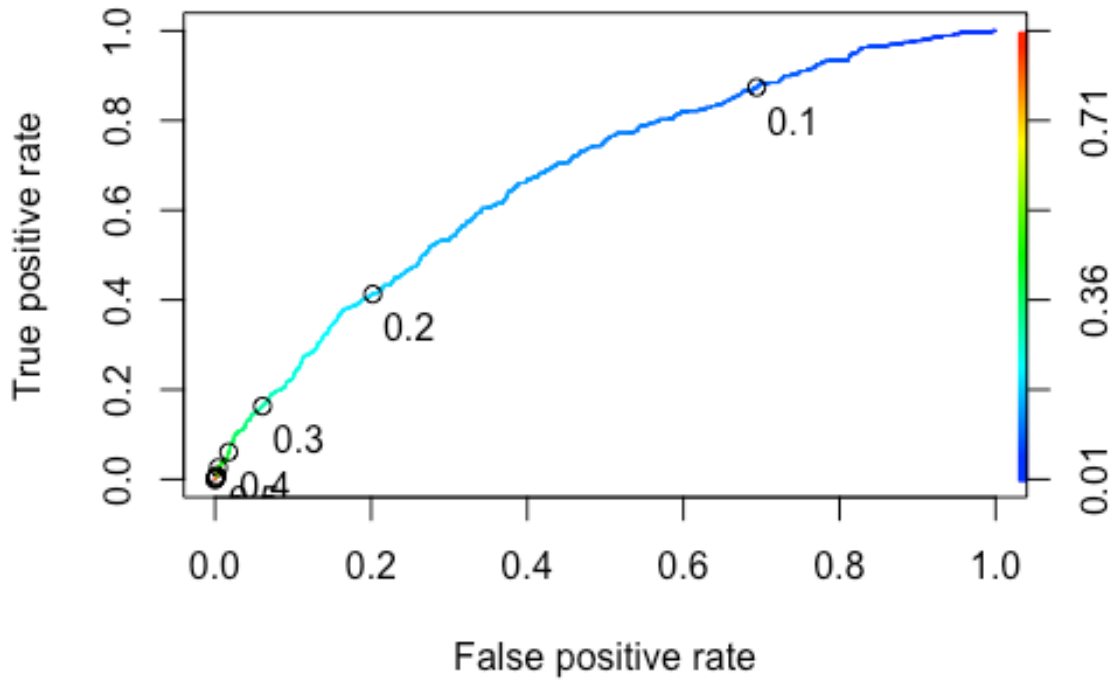
```
## [1] 0.8402367
```

The baseline model has an accuracy rate of 83.6% The logistic regression model performs a little worse than the baseline model.

v. What's the AUC?

```
library(ROCR)
ROCRPred <- prediction(test$PredictedRisk, test$NotFullyPaid)
ROCRperf <- performance(ROCRPred,"tpr","fpr")
```

```
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0.1,1,by=0.1),
text.adj=c(-0.2,1.7))
```




```
# report AUC
ROCRAUC <- performance(ROCRPred, "auc")
AUC <- ROCRAUC@y.values[[1]]
AUC

## [1] 0.6752311

# get accuracy at a better cut-off of 0.2
# sum((test$PredictedRisk>0.2) == test$NotFullyPaid)/nrow(test)
```

b. Interest Rate and Loan

i. Use a logistic regression model to predict NotFullyPaid using only IntRate

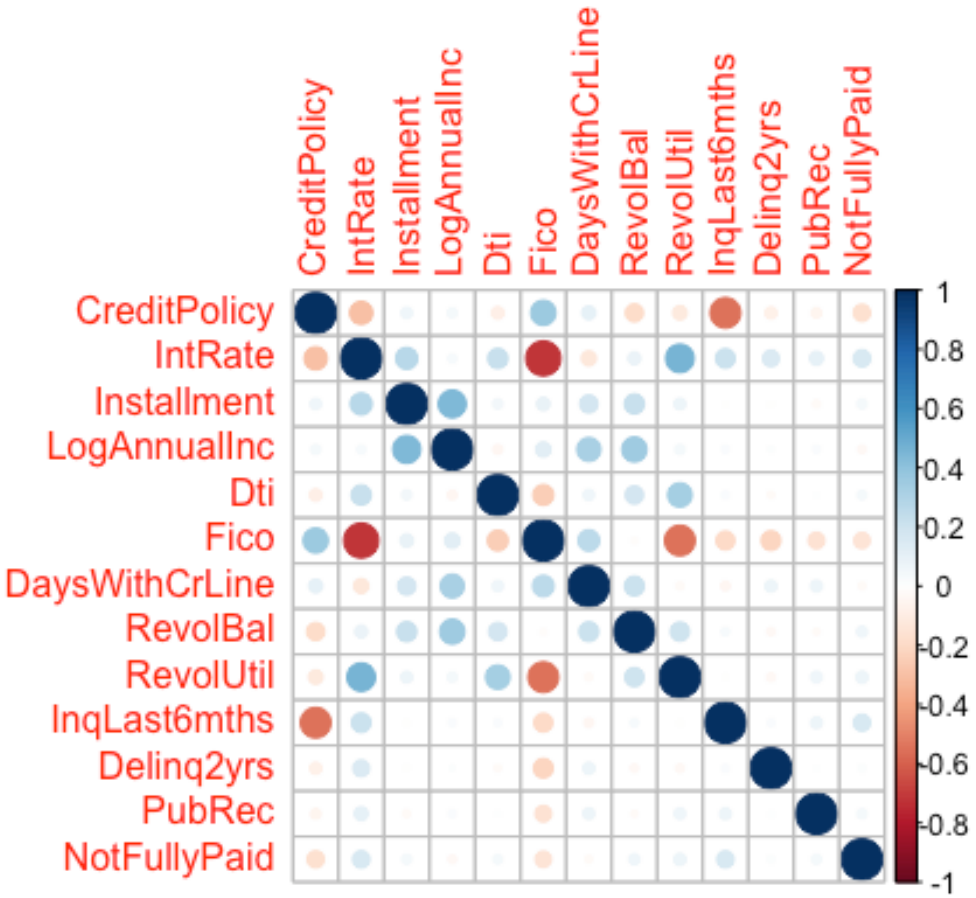
```
intLog <- glm(NotFullyPaid~IntRate, data=train, family = binomial)
summary(intLog)

##
## Call:
## glm(formula = NotFullyPaid ~ IntRate, family = binomial, data = train)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.0963  -0.6294  -0.5377  -0.4255   2.3197
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.8282     0.1704  -22.47   <2e-16 ***
## IntRate      17.1353     1.2795   13.39   <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: 5896.6  on 6704  degrees of freedom
## Residual deviance: 5710.9  on 6703  degrees of freedom
## AIC: 5714.9
##
## Number of Fisher Scoring iterations: 4
```

Interest rate is significant in the simple model, but insignificant in the previous full model. The difference might be caused by multicollinearity.

```
library(corrplot)
corrplot(cor(train[,-2]))
```



Interest rate is strongly (negatively) correlated with FICO score.

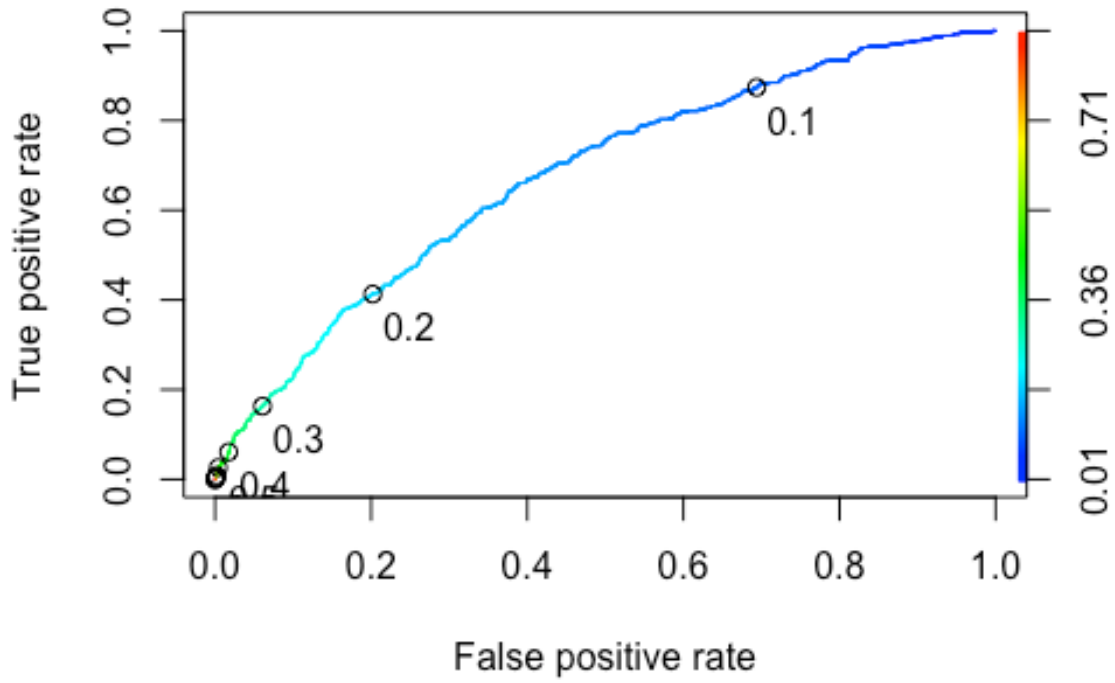
ii. use the interest rate model to predict probability of NotFullyPaid on the test set

```
PredictedRiskInt <- predict(intLog, test, type="response")
summary(PredictedRiskInt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05732 0.11238 0.14982 0.16079 0.19617 0.45170
```

The highest probability of a loan not being paid back in full on the test set is 45.2%. No loan would we predict would not be paid back in full if we used a threshold of 0.5 to make prediction.

```
ROCRPred2 <- prediction(PredictedRiskInt, test$NotFullyPaid)
ROCRperf2 <- performance(ROCRPred,"tpr","fpr")
plot(ROCRperf2, colorize=TRUE, print.cutoffs.at=seq(0.1,1,by=0.1),
text.adj=c(-0.2,1.7))
```




```
# report AUC
ROCRAUC2 <- performance(ROCRPred2, "auc")
AUC2 <- ROCRAUC2@y.values[[1]]
AUC2
```

```
## [1] 0.6133313
```

c. How our logistic regression model can be used to identify loans that are expected to be profitable

How much does a $10 investment with an annual interest rate of 6% pay back after 3 years, using continuous compounding of interest?

$$\begin{aligned} P &= c * e^{rt} \\ &= 10 * e^{0.06*3} \\ &= 11.97 \end{aligned}$$

ii. What is the profit?

Profit when investment is paid back in full

$$\begin{aligned} FullProfit &= c * e^{rt} - c \\ &= c * (e^{rt} - 1) \end{aligned}$$

Profit when investment is not paid back in full, assume the pay-back is short of $b dollars, where $0 <= b <= c * e^{rt}$

$$NotFullProfit = c * e^{rt} - c - b$$

Notice that if $b = c * e^{rt}$, which means the borrower did not pay back any money, the loss will equal to the investment(c).

iii. Compute the profit of a $1 investment in each loan

Assume that if the loan is not paid in full, the borrower pay nothing. Therefore, assume the profit will be -c.

```
investment = 1
test$Profit[test$NotFullyPaid==0] <-
investment*(exp(1)^(test[test$NotFullyPaid==0,"IntRate"] * 3)-1)
test$Profit[test$NotFullyPaid==1] <- -investment
max(test$Profit)
```

```
## [1] 0.8894769
```

Maximum profit of a $1 investment is $0.889

iv. alternative investment strategy

```
HighInterest <- test[test$IntRate>=0.15,]

mean(HighInterest$Profit)
```

```
## [1] 0.2691649
```

```
# rate of not paid in full
sum(HighInterest$NotFullyPaid)/nrow(HighInterest)

## [1] 0.2271715
```

- The average profit of a $1 investment is $0.269
- The proportion of the high-interest loans were not paid back in full is 22.7%.

v. What is the profit of an investor who invested $1 in each of these 100 lonas? How many of the 100 selected loans were not paid in full? How does this compare to the simple strategy?

```
# sort by predicted risk
HighInterest <- HighInterest[order(HighInterest$PredictedRisk),]

# new data frame with the top100 loans with the least risk
SelectedLoans <- HighInterest[1:100,]
mean(SelectedLoans$Profit)

## [1] 0.3635442

sum(SelectedLoans$NotFullyPaid==1)

## [1] 16

sum(SelectedLoans$Profit)

## [1] 36.35442
```

- The average profit of an investor who invested $1 in each of these 100 loans is $0.364
- 16 of the 100 loans are not paid back in full
- The total profit increased from $20.94 to $36.4!

d. assumptions in financial situations

The situation described in the question is completely possible and need to be careful of. For example, the models are built on historical data; If the behavior and patterns change in the future, the model will lose its effectiveness. In this case, the prediced risk might no longer hold effective. One possible solution is to implement some protective mechanism. For example, cross check the model performance with the reality periodically, and if the evaluation metric fall out of a pre-defined range, analyst will pause and re-examine the model to prevent false prediction.