FinTech_Assignment1

Jiaqiu Wang

February 13, 2017

FNCE 385/885 Assignment 1: Credit Modeling

Initialization & Data Preparation

```
rm(list = ls()) # Clear the memory
library("pROC") # The package needed for plotting ROC curves
```

```
## Warning: package 'pROC' was built under R version 3.3.2
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## Load data
# Load the data from the data file. The first row is variable names.
# To avoid trouble we do not use "string as factor" option
Data_set <- read.csv("File1_IS_data.csv",header = TRUE)

#check the data type of each column in the data set
str(Data_set)
```

```
## 'data.frame':    10000 obs. of  14 variables:
##  $ id                 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ loan_amnt          : int  17625 2800 5375 20000 10000 20000 9500 4000 15000 6000 ...
##  $ int_rate           : num  18.49 7.62 15.31 14.09 12.12 ...
##  $ grade              : Factor w/ 6 levels "A","B","C","D",..: 4 1 3 2 2 4 3 3 3 2 ...
##  $ sub_grade          : Factor w/ 33 levels "A1","A2","A3",..: 17 3 12 10 8 17 12 11 13 10 ...
##  $ emp_length         : Factor w/ 11 levels "< 1 year","1 year",..: 5 6 5 1 3 3 3 3 3 9 ...
##  $ revol_bal          : int  12002 3897 6070 12174 13547 23178 10647 16904 9635 24963 ...
##  $ revol_util         : num  88.9 73.5 38.4 49.9 88.6 87.8 56.9 77.5 86.8 57.3 ...
##  $ fico               : num  672 727 682 702 707 677 667 687 702 702 ...
##  $ home_ownership     : Factor w/ 5 levels "MORTGAGE","NONE",..: 5 5 4 5 5 1 1 1 1 1 ...
##  $ annual_inc         : num  45000 44500 22880 95000 68000 ...
##  $ verification_status: Factor w/ 3 levels "Not Verified",..: 3 1 3 3 1 3 1 2 2 1 ...
##  $ loan_status        : Factor w/ 6 levels "Charged Off",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ default            : Factor w/ 2 levels "Defaulted","Paid": 2 2 2 2 2 2 2 2 2 2 ...
```

```r
# Transform "defaulted" into a binomial response, which gives 1 if defaulted or zero otherwise.
Data_set$default <- Data_set$default == "Defaulted"
```

Q1. Basic model estimate

a. Before estimating the model, I would expect to see the coefficient for fico score is negative, since people with a higher fico score tends to have less probability of default.

```r
#b. First logistic regression
model1 <- glm(default ~ fico, family = "binomial", data = Data_set)
# family = "binominal" tells R to run a logistic regression.
summary(model1)
```

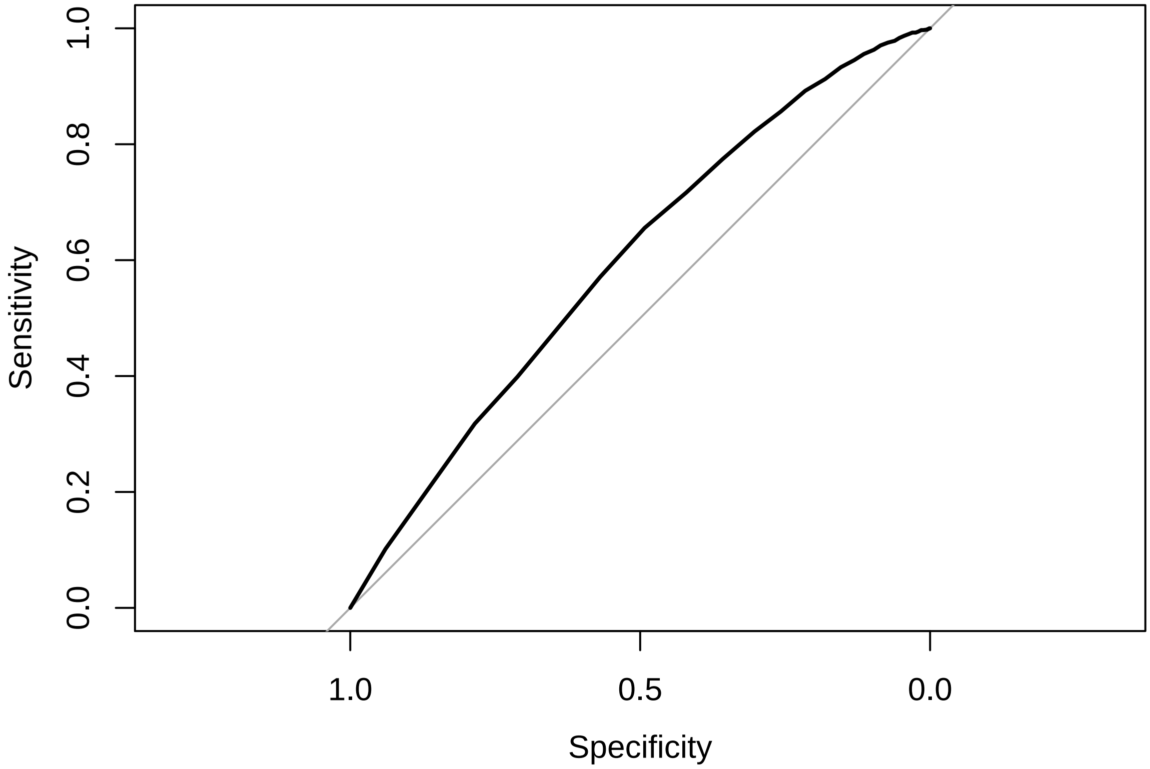
```
##
## Call:
## glm(formula = default ~ fico, family = "binomial", data = Data_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7259  -0.6441  -0.5698  -0.4430   2.6000
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.681073   0.755259   10.17   <2e-16 ***
## fico        -0.013414   0.001092  -12.29   <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: 8869.3  on 9999  degrees of freedom
## Residual deviance: 8694.0  on 9998  degrees of freedom
## AIC: 8698
##
## Number of Fisher Scoring iterations: 5
```

(b) The estimates for the intercept and the coefficient for fico from the logistic model are 7.68 and -0.013 respectively. The result suggest that fico score has significant explanatory power with a near-zero p-value.

Q2. Model evaluation

```r
#a. Get the fitted default probability
Data_set$predicted_p <- predict(model1, type = "response")
#'type = "response"' tells R to give estimated probability of default directly

#b. Now we can draw a ROC curve.
ROC1 <- roc(default ~ predicted_p, data = Data_set) # Calculate the ROC curve
plot(ROC1) # Plot the ROC curve
```



```
#c. calculate the area under the ROC curve
# With package pROC you can use the function auc()
auc(default ~ predicted_p, data = Data_set)
```

```
## Area under the curve: 0.5981
```

The area under the ROC curve is 0.5981 which is greater than 0.5, this is consistent with the results in part1(b), since fico score has significant explanatory power.

```
#d. Calculating the percentage of false positives and true positive with cut-off 0.1:
#For any one who has estimated probability of default being 0.1, we label them as 'default'.
Data_set$Predicted_default <- Data_set$predicted_p > 0.1
Num_predicted_pos <- sum(Data_set$Predicted_default == TRUE)
Num_correct_pos_pred <- sum(Data_set$Predicted_default == TRUE & Data_set$default == TRUE)
true_positive <- Num_correct_pos_pred/Num_predicted_pos
false_positive <- 1-true_positive
#display the result for percentage of `correct' positive
cat("true positive rate:",true_positive)
```

```
## true positive rate: 0.1740018
```

```
cat("false positive rate:", false_positive)
```

```
## false positive rate: 0.8259982
```

The proportion of consumers you mistakenly reject (false positives) is 82.6%, the proportion of consumer you correctly reject (true positives) is 17.4%.

Q3.An out-of-sample analysis

```
#a. Create a subsample data set with first 9000 samples. The remaining data goes to test data
Data_set_training <- Data_set[1:9000,]
Data_set_test <- Data_set[9001:nrow(Data_set),]
names(Data_set_test)
```

```
##  [1] "id"                   "loan_amnt"            "int_rate"
##  [4] "grade"                "sub_grade"            "emp_length"
##  [7] "revol_bal"            "revol_util"           "fico"
## [10] "home_ownership"       "annual_inc"           "verification_status"
## [13] "loan_status"          "default"              "predicted_p"
## [16] "Predicted_default"
```

```
model2 <- glm(default ~ fico, family = "binomial", data = Data_set_training)
summary(model2)
```

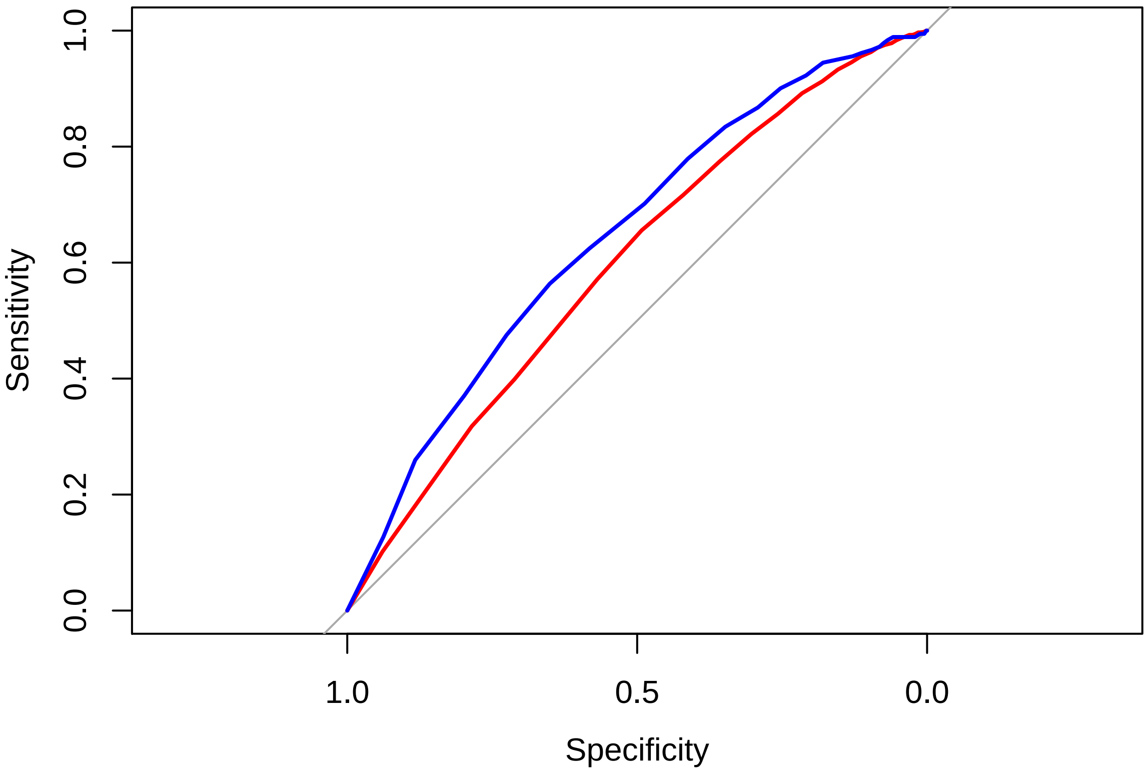
```
##
## Call:
## glm(formula = default ~ fico, family = "binomial", data = Data_set_training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7145  -0.6378  -0.5680  -0.4476   2.5702
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.164309   0.790718   9.061   <2e-16 ***
## fico        -0.012688   0.001142 -11.108   <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: 7920.7  on 8999  degrees of freedom
## Residual deviance: 7778.8  on 8998  degrees of freedom
## AIC: 7782.8
##
## Number of Fisher Scoring iterations: 4
```

The estimates for the intercept and the coefficient for fico using only the training dataset are 7.16 and -0.013 respectively. The estimated model is close to the model we get for Q1.

```
#b. Compare the performance
Data_set_test$predicted_p = predict(model2, newdata = Data_set_test, type = "response")
# Calculate the ROC curve
ROC1 <- roc(default ~ predicted_p, data = Data_set)
```

```
ROC2 <- roc(default ~ predicted_p, data = Data_set_test)

plot(ROC1, col = "red") # Plot the ROC curve, 'col = "red"' sets the color of the
# curve to be red.
plot(ROC2, add = TRUE, col = "blue") # The argument 'add = TURE' makes sure that the curve is added
```



c. The area below the new ROC line get larger compared with what we get earlier. Here since we just took the first 9000 records as training dataset instead of randomly sampling the data, the dataset may have some bias itself and the model just happen to fit better on the remaining 1000 data which we used as teating data.

d. As a manager, with the opportunity to conduct a multi-variate logistic regression analysis, I'm not going to use all variables available, since some of the variables are correlated themselves and some of the variables don't have much explainatory power. If adding those variables to the model the coefficients for the variables with actual explainatory power may be biased.

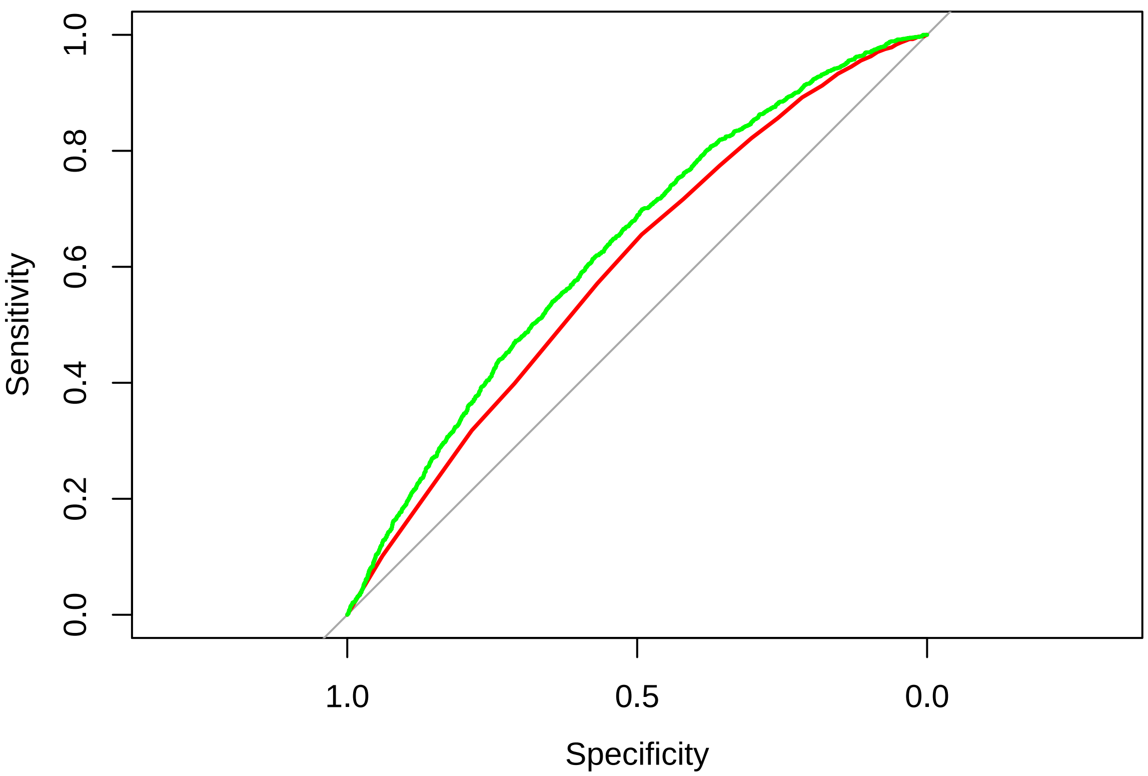
```
#e. A more complicated model
model3 <- glm(default ~fico + loan_amnt + int_rate + verification_status, family = "binomial", data = Da
summary(model3)
```

```
##
## Call:
## glm(formula = default ~ fico + loan_amnt + int_rate + verification_status,
##     family = "binomial", data = Data_set)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0318  -0.6393  -0.5416  -0.4135   2.4123
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -5.427e-02  1.009e+00  -0.054  0.95710
## fico                               -4.264e-03  1.343e-03  -3.176  0.00149
## loan_amnt                          -8.120e-07  3.991e-06  -0.203  0.83880
## int_rate                            9.929e-02  9.079e-03  10.937  < 2e-16
## verification_statusSource Verified  1.000e-02  7.527e-02   0.133  0.89427
## verification_statusVerified         3.301e-02  6.520e-02   0.506  0.61264
##
## (Intercept)
## fico                               **
## loan_amnt
## int_rate                           ***
## verification_statusSource Verified
## verification_statusVerified
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8869.3  on 9999  degrees of freedom
## Residual deviance: 8573.8  on 9994  degrees of freedom
## AIC: 8585.8
##
## Number of Fisher Scoring iterations: 5
```

```r
Data_set$predicted_p_new <- predict(model3, type = "response")

# Now we can draw a new ROC curve.
ROC3 <- roc(default ~ predicted_p_new, data = Data_set) # Calculate the ROC curve
plot(ROC1, col = "red") # Plot the ROC curve
plot(ROC3, add = TRUE, col = "green")
```



I added loan amount, interest rate and the employment varification status to the logistic regression model, the coefficients are only siginificant for fico socre and interest rate. Drawing the ROC curve, we can see the model performance get improved with the new variable since the area under ROC curve is larger. When adding interest rate to the model, the coeffcient estimate for fico score decrease to almost zero, which means interest rate includes almost all the infomation we can get from the fico score but also some other information. However, when reporting the final model, I will not keep any of the new variables, since the interest rate information will not be available when assessing the loan in real life.