FNCE 385/885 Assignment 1: Credit Modeling

Due: 3:30 p.m., Nov. 14, 2016

You are an asset manager who had been asked to allocate $1 million worth into a portfolio of MPL. In this assignment you will analyze a database of consumer loans originated on an online platform.

You should start with a sample database of 10,000 three-year consumer loans, *File1_IS_data.csv*. The database contains the following details for each borrower: ID, Loan Amount, Interest Rate (Annual in Percentage), Credit Grade, Detailed Credit Grade, Length of Employment, Revolving Credit Balance, Revolving Credit Utility, FICO Score, Home Ownership, Annual Income, Employment Verification, the Statues of the Loan, Default Statues.

1 Basic model estimate

In this part you are expected to estimate a *logistic regression* model introduced in class to explain the probability of default using a consumer's *fico* score *only*.

(a) Before estimating the model, what sign do you expect to see for β_{fico}, the coefficient for fico score? Give a brief explanation. One to two sentences should suffice.

(b) Write an R script to estimate the model. You can look at the class notes or session note 2 for details. Report your estimates for the intercept and the coefficient for *fico*. Does your result suggest that *fico* score has significant explanatory power?

2 Model evaluation

Now that you have the first fitted model, you are expected to evaluate its usefulness.

(a) Use the model you have, estimate the *probability* of each consumer's default.

(b) Use the probability you estimated, create an *ROC* curve plot. Again you can look at session note 2 for how to do that. You should not calculate the curve manually.

(c) Is the area below your *ROC* curve greater than 0.5? Is this consistent with your results in Part 1(b)? Report the area. You should have the number when you plot the *ROC* curve in R.

(d) Suppose you develop a strategy that you reject all the loan applicants whose predicted probability of default is greater than 0.1. What is the proportion of consumers you mistakenly reject (false positives)? (That is, the number of consumers who are rejected but do not default, over the total number of consumers who do not default). What is proportion of consumer you correctly reject (true positives)? (The number of consumers who are rejected and actually default, over the total number of consumers who default.)

3 An out-of-sample analysis

An important goal of statistical modeling is that we want to use the models to make predictions, and a model which can explain existing well does not necessarily lead to great predictive power.

In this part you will conduct an out-of-sample analysis and get an impression on the issue addressed above. We will keep working on the original data set.

(a) To start with, create a sub-sample data set with the first 9,000 loans. This will be your training set. A good command of doing so is *data_small <- data_set[1:9000,]*, where *data_set* should be the name you use for your full data set. Then you should estimate the model in Part 1 again, using the training set. Report the parameter estimates.

(b) Based on the estimate you get from the training set, *predict* the probability of default for the remaining 1,000 loans. Then use the predicted probability to draw an *ROC* curve. Specifically, you should put the new ROC curve together with the one you obtain in Part 2. Additional argument *add = TRUE* in R command will help.

(c) Does the area below the new ROC line get smaller or larger? Do you have an explanation?

(d) As a manager, you might want to take advantage of the data you have and conduct a multi-variate logistic regression analysis. Do you want to use all variables available? Why or why not?

(e) Propose three variables in the data set that you think might help improving the performance of the model, and run a logistic regression analysis with them and *fico*. Report the model estimates. Are your new variables significant? Do they improve the model performance? Will you keep any of them if you want to report a final model?

4 (BONUS) A business decision to make

Note: This part gives bonus points if you finish it successfully. However you should be able to obtain full credit for this homework assignment with Part 1 to 3 ONLY. There will be a competition for those who submit their results.

The online platform received additional 1,000 applications for loans. Their analysts calculated the interest rate they want to charge based on their model. The data of the new applicants are stored in file *File2_OOS_predictor_data.csv*.

Now as an asset manager you have the right to pick 100 of the applications and form a portfolio. Suppose that you allocate your assets *equally* on the 100 applications you pick. Suppose that in an event of default, you get zero payment.

Based on the data you have, pick 100 applications. You should work with the models you obtained in Part 1 to 3, taking into consideration the probability that one loan can default.

In your assignment submission you should argue for your decision. The probability of default should be one good criterion, but you should explore more.

You results should be stored in a .csv file, with first column being the ID of the applicants (10,001 to 11,000), and the second column being the indicators (TRUE for being picked, and FALSE for not being picked). The teaching team will calculate the realized return of your portfolio, and the competition will be based on that. Again, please look at session 2 notes for details of outputting your results. (If there are more than 100 loans picked, the teaching team will only use the first 100.)