ECO 4444 Introduction to Business Analytics for FinTech
Fall 2021
Final Project
Due Date and Presentation: 1:00pm December 7

In this final project, you will use R in order to evaluate logistic regression models with the objective of building a mortgage application approval/denial classifier. Of interest is modeling the probability that an applicant will be approved or denied as a function observables in order to classify applicants as being ones who will be approved and who will not be approved. As with the mid-term project, you will perform k-fold cross validation in the process of model selection.

From our class page in Webcourses, download the dataset hmda_sw.csv. The dataset contains records on 2,380 applicants for mortgages in Boston. Also at our class page is a description of the variables and a paper published in the *American Economic Review* that first evaluated the data in order to test for the presence of racial discrimination in the mortgage approval process. You should first read the paper as it provides a detailed discussion of the data and insight on the determinants of the probability of being approved. After reading the paper and importing the dataset into R, do the following:

A. The variables in the dataset do not have intuitive names (e.g., the meaning of S3 is unclear). Referencing the data description and the *AER* paper, identify the qualitative dependent that you will be modeling and the set of co-variates that you intend to include in your various models, and rename the variables so that they have (somewhat) intuitive names. Be certain that the debt-to-income ratio and the race, self-employed, marital status, and education indicator variables are included, among other variables.

B. Generate summary statistics on the set of variables selected in A, and explain the composition of the sample and of the characteristics of an average (representative) applicant. In the process, you can (should) also generate and histograms and frequency counts on particular variables of interest, which can be referenced in your explanation of the composition of the sample and of a representative applicant.

C. With the full sample, estimate the logistic regression model, where the deny/approve dummy variable is the response variable and the debt-to-income ratio and the race, self-employed, marital status, and education indicator variables are the co-variates. Graph the ROC curve and calculate the AUC. Also, compute the confusion matrix at alternative cut-off levels, and calculate the classifier sensitivity, specificity, the false-positive rate, the false-negative rate, the model accuracy and error rate to confirm they are the same as those produced by R. Provide a written explanation summarizing the findings.

D. Next, using 10-fold cross validation, estimate a variety of logistic regression models with the training samples and evaluate their predictive performance with the test samples across a range of threshold values in each case. The models can (should) include interaction variables and polynomial terms (e.g., quadratic and cubic variables). The models may be evaluated with different performance measures, such as the test error, accuracy, and AUC. Document in a table the performance of the various models using the chosen performance measure(s).

E. Of the competing models that you estimated and thresholds that you evaluated, identify the superior model for classification purposes. Re-estimate the model with the full sample of data. Then, graph the ROC, calculate the AUC, and compute the confusion matrix at the chosen threshold level. Calculate the classifier sensitivity and specificity, the false-positive rate, the false negative rate, the accuracy, and overall error rate. How well does your superior model perform relative to the model estimated in C? Thoroughly explain. Note that to do so you may need to re-calculate the confusion matrix from the estimated model in C at the chosen threshold level.

F. Submit clean and well-organized hard-copies of your program, the output and written discussion/explanation of your findings from B, C, D, and E, your estimation results and tables for review in the format of a report. Be certain to be complete and thorough.

G. Have fun with the project and learn a lot in the process.