MF 850: Advanced Computational Methods

Professor Gustavo Schwenkler Fall 2016

Additional problem set

Due: Monday, December 5, by 10 am via email to Professor Schwenkler (gas@bu.edu)

(By solving this additional problem set you can make up for parts of your midterm grade. This problem set will be graded on a scale from 0 (worst) to 10 (best), and any points you receive will be added to your midterm grade. If you decide to work on this problem set and hand it in, the final grade for your midterm will therefore be $\max\{\textit{Grade from midterm} + \textit{Points from this problem set}, 60\}$*. You may still wish to work on this problem set even if you do not need extra points for your midterm as it will help you practice for the final project.)*

The file "mf850-loan-data.csv" contains data on different personal loan applications. For each loan, the file contains data characterizing the loan applicant, as well as the decision of the bank whether to approve the loan or not. In this project you will use these data to construct a statistical model that predicts whether a loan will be approved or not by a bank depending on the characteristics of the loan applicant. You may use any of the tools we studied in class. The model you construct may be as simple or as complex as you deem appropriate.

The data contains the following entries for each loan:

- "CHK_BAL": Balance on the applicant's checking account
- "DURATION": Loan duration
- "CRED_HIST": Credit history of the applicant
- "PURPOSE": Indicated purpose for the loan
- "CRED_AMT": Credit amount requested
- "SAV_BAL": Balance on the applicant's savings account
- "EMPL_LENGTH": Length of applicant's employment
- "INSTALLMENTRATE": Loan's installment rate as a percentage of the applicant's disposable income
- "GENDER": Applicant's gender
- "MARITALSTATUS" Applicant's marital status
- "GUARANTORS": Additional applicants or guarantors (co-signers) of loan

- “ATADDRESSSINCE”: Years lived in applicant’s current address
- “ASSETS”: Applicant’s most valuable asset
- “AGE”: Applicant’s age
- “CRED_OTHERS”: Does the applicant hold personal loans at other institutions? If so, where?
- “HOUSING”: Housing type
- “CRED_HERE”: Does the applicant hold other credit lines at the bank providing the loan? If so, how many?
- “JOBQUALITY”: Job type
- “DEPENDENTS”: Number of dependents supported by the applicant
- “HOMEPHONE”: Does the applicant have a home phone?
- “FOREIGNER”: Is the applicant a foreigner?
- “CRED_APPROVED”: Was the loan approved?

Rules

You may work on this project in a team consisting of **at most 4 people.** Your submission should consist of 2 components:

(i) A function that implements your model in R or Matlab. Your function should take as input all the variables in the data that characterize a loan applicant, and should return a “Yes or No” decision on whether to approve the loan or not.

(ii) A written summary of the analysis that you carried out in order to construct your model. Your written summary should contain all estimates, tables, and figures, and should not be longer than 10 pages.

The grade you will receive for this project will be constructed as follows:

- *Accuracy (4 points).* The prediction accuracy of your model will be tested on a special test data set selected by Professor Schwenkler. The test data is not included in the file “mf850-loan-data.csv”. Professor Schwenkler will rank all model submission from most to least accurate, and will assign a grade for accuracy according to a submission’s ranking.

- *Model justification (3 points).* Professor Schwenkler will judge whether your choice of model is justified based on the analysis that you describe in your written summary. Models that appear unjustified (or randomly selected) will receive a low model justification score.

- *Model interpretability (2 points).* Your model will also be judged along the interpretability dimension. A model that can be easily interpreted and is consistent with standard economic thinking and theory will receive a high model interpretability score.

- *Model uniqueness (1 points).* Your model will also be evaluated along the uniqueness dimension. If your predictions are based on model that multiple submissions use, then you will receive a low model uniqueness score.