Final Exam

IDS 572, Spring 2020

May 08, 2020 Time allowed: 120 minutes

EXAM CONDITIONS

- You need to submit your R script (saved as .R file) in the link provided on Blackboard under the "Final Exam" content area by 3:15 pm. The submitted files through emails will not be graded. Hence, make sure you submit your codes much before the deadline to avoid connection issues.

- Please try to write your R codes as clear as possible. Comment each line if needed. .

- In case there are questions asked in the exam, please write your detailed answer in your R.script file. Make sure you write it as a comment so that you do not receive an error when running your code.

- Please include your full name at the top of your R script file. In addition, use the following naming convention for the file: Lastname_Firstname_FinalExam.R.

- You are not allowed to communicate with any other individuals while completing this exam in any way. Failure to abide by this condition will imply a violation of the Honor Code of University of Illinois at Chicago and will subject violators to the consequences stated under violations of the Honor Code in the UIC Student Handbook.

Good luck!

To minimize loss from the bank's perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant's demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The Bank.data contains data on 9 input variables and the classification target indicating whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

The variables in this dataset are:

- Age (numeric)
- Sex (categorical: male, female)
- Job (categorical: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- Housing (categorical: own, rent, or free)
- Saving accounts (categorical: little, moderate, quite rich, rich)
- Checking account (categorical: little, moderate, rich)
- Credit amount (numeric, in USD)
- Duration (numeric, in month)
- Purpose (categorical: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
- Target (categorical: 1 - Good, 0 - Bad)

Please answer the following questions.

(1) Provide a 5 number summary for the credit amount.

(2) Are the variables Duration and Credit amount highly correlated? Do we need to remove one of them?

(3) What is the distribution of Credit amount for "Good" and "Bad" instances. Draw a side-by-side boxplot to support your answer. Add a title to your plot and name your x and y axes.

(4) Provide a two-by-two table that contains the frequency of different housing types (free, own, rent) for "Good" and "Bad" instances.

(5) Are there any missing values in this dataset? If yes, handle these missing values.

(6) Are there any outliers in this dataset? If yes, take care of outliers.

(7) Divide you data into 70% training and 30% test data.

(8) Run a logistic regression model on your training examples to predict the target variable and use "forward" selection technique to come up with the best variable sets for this model.

(9) Explain what is the AIC measure.

(10) What are the significant variables using your logistic regression model? Explain how you select them.

(11) Run a C5.0 decision tree model using the rpart() function that provides the best outputs. Include all the input variables into your model. What are the parameters minbucket, minsplit and cp

values used in your tree function? Why do we use these parameters? How do you select the best values for these parameters.

(12) Write two best decision rules to predict the target variable? Explain how do you select these rules.

(13) What are the important variables suggested by your decision tree model?

(14) Run a SVM model. Include all the variable in your model. Use cross validation for picking the parameter gamma. Fix cost = 100.

(15) Write your own function in R that receives the predicted and true labels and computes the recall, precision, and accuracy of a model.

(16) Which of your models (among logistic regression, decision tree, and SVM) is the best model to predict the target variable? Explain your choice of evaluation measure(s).

(17) Draw a ROC curve for all your models and compare the performance of your models using their ROC curves.

Thanks for being a great class! Have fun and enjoy life!