UCLA Extension Data Science Intensive

Instructor: William Yu

Project 5

10/23/2018

A. **Predicting Blood Donation**

- Import blood_traindata.xlsx dataset. In this train dataset, we want to train the model to predict whether the person will donate blood this month with four possible predictors: "Months since last donation", "Number of donations", "Total volume donated (c.c.)", "Months since first donation".
- First, use a logistic model to predict the probability of those people to donate their blood in the blood_testdata.xlsx. To determine the choice (variables) of the logistic model, do a simple train-test validation from the blood_traindata.xlsx.
- Second, run all the classification machine learning models you learned in the class with the 10-fold cross validation. Choose a best model and predict whether the people in the blood_testdata.xlsx will donate or not (donate: 1; not: 0).
- Submit your result filling two columns (Col F: Made donation this month; Col G: Probability from logistic model) with your R script.
- Hint: if your model cannot run in the beginning, try to transform one variable to the form of logarithm.
- After your submission, I will calculate the accuracy of your prediction based on the real donation decision.

B. **Bonus Question -- Predicting Bank Client's Subscription**

- This is a big dataset, including 32,951 observations in bank_traindata.csv. The goal is to use this dataset to train and find the best model to predict if a client of this Portugal bank will subscribe the term deposit (Col U: y (yes, or no)) in the bank_testdata.csv. The descriptions of variables as in bank_readme.
- There seems to be some data collinearity issue so some classification models cannot work properly. In addition to the logistic model, see if you can come out a good model to predict.