Columbia GSB: Machine Learning
Homework 2

Dr. George A. Lentzas

This Homework is due by 11:59pm on Tuesday Feb 28th. Sumbit your results in .pdf format in Canvas. Good Luck!

This homework is based on the paper (and associated dataset) by S. Moro, P. Cortez and P. Rita, ”A Data-Driven Approach to Predict the Success of Bank Telemarketing” Decision Support Systems, Elsevier, 62:22-31, June 2014. We will use machine learning methods to identify customers that have high purchase intent and direct the marketing resources to them.

”A Portuguese retail bank uses its own contact-center to do direct marketing campaigns, mainly through phone calls (tele-marketing). Each campaign is managed in an integrated fashion and the results for all calls and clients within the campaign are gathered together, in a flat file report concerning only the data used to do the phone call. ... In this context, in September of 2010, a research project was conducted to evaluate the efficiency and effectiveness of the telemarketing campaigns to sell long-term deposits. The primary goal was to achieve previously undiscovered valuable knowledge in order to redirect managers efforts to improve campaign results. ... Since this project started being analyzed in detail in September of 2010, it meant that there were available reports for about three years of telemarketing campaigns ... ”

1. Go to the UCI ML Repository (click on the hyperlink) and download the data. There you will find two .zip files, you should use the one called ”bank-additional.zip” (ignore the ”bank.zip” file which is an older version of the same data). In the .zip file you dowloaded you should find and use the ”bank-additional-full” .csv file. Import the `bank-additional-full.csv` in R Studio. You can do that either using the R Studio dataset GUI or running the command `read.table()`. Remove the variables `duration`, `date_of_week`, `month` and `nr.employed` and explain why removing these variables makes sense. Print `summary()` for your data and briefly discuss the imported dataset. (10 Points)

2. Examine the ”bank-additional-names.txt” file and read about the attributes in the data frame. You will use the input variables (minus the variables which we deleted) to predict the output variable (whether the client subscribed for a term deposit).

Notice that there are missing values in some features, which are coded with the ”unknown” label. Remove any rows with ”unknown” observations. An easy way to do this is to replace the ”unknown” field with NAs and then use the `na.omit()` function. Recall that tree methods are not great at handling multiple unordered categorical predictors; hence change the ”job” attribute to have only two values (employed and unemployed) and marital to have only (”single” and ”married”). Similarly change the ”education variable” to a numeric ordered dummy variable taking 6 increasing values (hint: use `as.numeric()` to ensure this is not a character in R). (10 Points)

3. Now split the sample into two equal sub-samples, for training and testing. Use `set.seed(1)` and the `sample()` command like in the the R Lab to create a training set and a test set. Then recall that the `tree()` function in R takes either numeric or factor inputs; hence transform any character variables in the data frame into factors (use the `as.factor()` command to do this). (10 points)

4. Fit a simple classification tree to your training data to predict the output variable. Try using both ”gini” and ”deviance” as the splitting criteria; what do you observe? (hint: check out the `tree.control()` function). Print your trees. (10 points)

5. Fit a random forest to you training data and print the variable importance graph. (20 points)

6. Fit a boosted tree to your training data and print the variable importance graph. (20 points)

7. Compare the prediction accuracy of the simple tree, the random forest and the boosted tree on the test data. Which one does better? What are your conclusions from this analysis? (20 points)