Homework 6

MSCS 6520 Business Analytics
Spring 2018
Assigned: April 16, 2018
Due: April 23, 2017 (by beginning of class)

Exercises

For this assignment, you're going to build a model to predict ad clicks. We're going to use a subset of a data set released by Criteo, a prominent digital advertising company.

1. Download the data set from D2L. The zip file contains a single file (criteo_ad_click.csv). All of the variables have been anonymized so you'll have to use the col_names = FALSE parameter for the read_csv() function. The resulting variables names will be X1 through X22 with X1 being the label.
2. Transform your data. For example, you will need to convert X1 and X15 – X22 into factors.
3. Perform exploratory data analysis. Is this data set balanced? (How many clicks vs not clicks?) Use box and jitter / heat map plots to identify which variables you think will be most predictive.
4. Build a model to predict the ad clicks. You may use any combination of forward or backward feature selection you wish – just make sure to report the outcomes. You may also consider advanced feature engineering such as: taking the logarithm (using log1p()) of continuous features, binning continuous features, and trying interactions between factors.
5. What were the accuracy and confusion matrix for your best model?
6. Which features were most predictive?

Prepare a document containing the answers to the above questions and plots. Submit the document as a PDF to D2L. You may work in pairs, in which case, you should only submit one PDF per group.