STAT 6240 - HW 6 - Due 4/18

Due Date: April 18th, Thursday, 6:10 PM

Your homework submission should contain:

1. Your R code,
2. Outputs from R (including figures),
3. Answers to the questions.

Do not print out your homework. Save it as a pdf file and upload it to Blackboard.

In this homework exercise, you will work on the Avito Duplicate Ads detection dataset.

- Randomly subsample some portion of the dataset (anywhere from 3% to 60% of the dataset, depending on your computer's computational capability). Then, subsample one third of this dataset, this will be your "training data". An another one third of the dataset will be your "validation data", and the final third will be your "test data".
- Based on your results from the previous homework (HW 4), fit 10 different models on the training data. All of these models can be of the same type, that is you can train 10 different xgboost models, if that was your best model. Keep in mind that it is nice to have some variety, so you might want to use 5 xgboost and 5 random forest models.
- Now, compute and store the probability predictions for each of these 10 models for both the "validation" and the "test data". Then, use your favorite classification method to model the responses in the validation data with the new 10 variables, which are the probability estimates. Finally, use this last "stacking" model to obtain classifications for the test data. Compute the AUC score of your final stacked model.

Bonus Points: For extra 20 bonus points, use a pre-trained neural network model on the images as one of your candidate models.