# STAT 6240 - HW #4 Part 1 - Due 3/21

**Due Date:** March 21st, Thursday, 6:10 PM

Your homework submission should contain:

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

You should use *R Markdown* or *R Notebooks*.

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

### `kaggle` Competition - Avito Duplicate Ads Detection

In this homework assignment, you will be working on the "Avito Duplicate Ads Detection" competition, which ran between May-July 2016. The competition is available at https://www.kaggle.com/c/avito-duplicate-ads-detection. You are asked to create a powerful classification models to identify whether a pair of ads are duplicates. As you might expect, this process will involve building and diagonising numerous (at least hundreds) of models.

This assignment is worth 250 points and has two parts. In the first part of the homework, you will create your own models and will be evaluated on the kaggle leaderboard score of your personal model.

In the second part, which will be due 3/28, you will be asked to aggregate some of the best models from the class to create an even better classifier with higher AUC scores.

1. Sign up on kaggle.com.
2. Go to the website and read *Description* and *Evaluation* under the *Overview* tab, and the data explanations under the *Data* tab.
3. Join the competition as a late submission, and download all files from the *Data* tab, except for files that contain the images (these files all have titles with `Images_X.zip`).
4. **Your task is to come up with a classifier with the highest AUC possible *without using the images***. Try numerous classifiers and variable transformations to improve your model.
5. After choosing a final classifier, perform out-of-sample predictions on the test data (use `ItemInfo_test.csv` and `ItemPairs_test.csv`) and upload your predictions to kaggle. You can see how well your method worked on the test dataset under the Leaderboard tab.
6. At the very least, your report should contain:
    - A basic exploratory data analysis;
    - Fitting and evaluations of all of the methods we have seen in class for classification (this includes LDA, QDA, Logistic regression, SVM, Decision Trees, Random Forests, various forms of Boosting).

7. 40% of your grade will depend on your method's performance on the test AUC on kaggle. Please take a screen shot of your leaderboard score and upload it with your homework to Blackboard.
    - *We will randomly check some of the applications on kaggle to ensure that you submitted your true AUC. If you are found to be cheating, you will get a score of -200 from this assignment.*
8. The other 60% of your grade will depend on the quality of your report, the depth of your analysis and the effort you put into the project. Since this is the first such assignment, we will be lenient in the grading of the project report. You will get full points as long as you have properly done everything that was asked for *on your own.*

**Some hints:**

- If you have no idea where to start from, kaggle has numerous Kernels which are R Notebooks prepared by some of the competitors. These Kernels contain code and a decent approach to the problem. You can find them under the *Kernels* tab for the competition
    - At this link you can find an R Kernel that prepares the data, creates very basic features and uses logistic regression for prediction.
- If you are looking for code to implement Decision Trees and others, read the R-Session from the 8th Chapter of Introduction Statistical Learning with R. You can also find the code for this online. To fit, adaboost, you are recommended to use the `fastAdaboost` package. Click here for a quick tutorial on the package.
- If you are loooking for inspiration, checkout the interviews with the 1st, 2nd and 3rd place teams of this competition.

*This is not an easy assignment. Start early.*