ML 4375 – Intro to Machine Learning

Project 2 (200 points)

- Select 2 large data sets (minimum: 30K rows). Good sources are listed at the end of this document. One should be suitable for regression and the other for classification.

For each data set, your project will be evaluated as follows:

- (10 pts) You will get more points for larger/messier data sets:
  - 0-5 pts <30K
  - 6-10 pts >=30K
- (10 pts) Data cleaning:
  - provide a link where you found the data
  - describe what steps you had to do for data cleaning (more points for messier data that needed cleaning)
- (20 pts) Data exploration:
  - use at least 5 R functions for data exploration
  - create at least 2 informative R graphs for data exploration
- (30 pts) Run at least 3 ML algorithms on each data set, **using at least 5 algorithms** in all.
  - this portion of your R script should include:
    - code to run the algorithms
    - commentary on feature selection you performed and why
    - code to compute your metrics for evaluation as well as commentary discussing the results
- (10 points) Run at least one ensemble method such as Random Forest, XGBoost
  - this portion of your R script should include:
    - code to run the algorithms
    - commentary on feature selection you performed and why
    - code to compute your metrics for evaluation as well as commentary discussing the results
- (10 pts) Results analysis
  - rank the algorithms from best to worst performing on your data
  - add commentary on the performance of the algorithms
  - your analysis concerning why the best performing algorithm worked best on that data
  - commentary on what your script was able to learn from the data (big picture) and if this is likely to be useful
- (10 pts) Project depth
  - 0-3 project minimally meets requirements
  - 4-6 project exceeds minimum requirements
  - 7-10 project went well above the requirements

Turn in:

- Upload the Rmd script and knit pdfs to eLearning, zipped together. Use one script for each data set.
- Send the knit pdfs to your 3 reviewers
- No need to upload data
- Project 2 presentations will be at the end of the semester

Regression algorithms: Linear regression, kNN, Decision Trees, SVM

Classification algorithms: Logistic regression, Naïve Bayes, kNN, Decision Trees, SVM

Ensemble methods: Random Forest/Boosting, XGBoost

Good Sources for Data Sets

- Kaggle, free but you need to register: https://www.kaggle.com/
- UCI machine learning repository: https://archive.ics.uci.edu/ml/index.php
- Open data on AWS: https://registry.opendata.aws/
- Google data sets: https://toolbox.google.com/datasetsearch
- Microsoft research open data sets: https://msropendata.com/
- US government data: https://www.data.gov/
- Data sets by topic: https://github.com/awesomedata/awesome-public-datasets