ML 4375 – Intro to Machine Learning – Summer 2018 – Mazidi

Project 2

Big data

Kaggle

* (20 pts) Select 2 large data sets (minimum: 100K rows). Kaggle is a good source. They can be either for regression or classification.
* For each data set:
  + 0-5 pts <500K
  + 6-10 pts >500K
* (30 pts) Try at least 3 ML algorithms on each data set, **using at least 5 algorithms** in all.
  + your R script should include:
    - link to the source of the data
    - code to run the algorithms,
    - code to compute your metrics for evaluation
    - commentary on the performance of the algorithms
    - see additional requirements below
* (30 pts) Prepare a 4-minute presentation (you can just use your Rmd or pdf file). The text in your Rmd file and your presentation should indicate for each data set:
  + which algorithms you tried with what results
  + which algorithm performed best
  + comment why the best performing algorithm worked best on that data
  + what feature selection you tried
  + what metric(s) you used to evaluate the models
  + what you learned from the data
* (10 points) Script runs correctly
* (10 points) Project depth
  + 0-3 project minimally meets requirements
  + 4-7 project exceeds minimum requirements
  + 8-10 project went well above the requirements
* Upload the Rmd script to eLearning
* Project 2 presentations will be at the end of the semester