PR2: Regression

Overview and Assignment Goals:

The objectives of this assignment are the following:

- Figure out how to best represent features numerically.
- Experiment with different feature selection and/or dimensionality reduction techniques.
- Experiment with various regression models.
- Think about dealing with imbalanced data.

Detailed Description:
Develop predictive models that can determine, given a large number of attributes describing a donor, how much money they are likely to donate in the next campaign.

The dataset you are provided describes a number of profile features for donors in a not-for profit donation campaign. The field names are somewhat cryptic. The description of each field can be found in the *dictionary.txt* file. The goal is to predict the donation amount for each potential donor, which is coded in the TARGET field.

You have been provided with a training set (train.dat.bz2) and a test set (test.dat.bz2) consisting of donor profiles, one per line in the file. The files are compressed with the bz2 library and contain train.dat and test.dat, respectively. Both files are comma-separated value formatted files and contain a header. The train.dat file has 76329 samples and 480 features, the last of which is the TARGET variable. The test.dat file has 19083 samples and 479 features, omitting the TARGET variable.

Note that the dataset is imbalanced. Many of the samples have a TARGET of 0.0. The test set was selected to have similar distribution of TARGET values as the training set. We will use root mean square error (RMSE) as the evaluation metric for this task.

Caveats:

+ Remember that not all features will be good for predicting the target. Think of feature selection, engineering, reduction (anything that works).
+ Use the data mining/machine learning knowledge you have gained until now, wisely, to optimize your results.

Rules:

- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes will result in an honor code violation.
- You are allowed 5 submissions per day.
- After the submission deadline, only your chosen or last submission is considered for

the leaderboard.

Deliverables:

- Valid submissions to the Leader Board website: https://clp.engr.scu.edu (username is your SCU username and your password is your SCU password).

Canvas Submission for the report:

- Include a 2-page, single-spaced report describing details regarding the steps you followed for feature extraction, feature selection, and classifier model development. The report should be in PDF format and the file should be called **<SCU_ID>.pdf**. Be sure to include the following in the report:
    1. Name and SCU ID.
    2. Rank & RMSE for your submission (at the time of writing the report). If you chose not to see the leaderboard, state so.
    3. Your approach.
    4. Your methodology of choosing the approach and associated parameters.
- Ensure you submitted the correct code on CLP that matches your output. Code does not need to be submitted on Canvas.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (30%).

Files: available on Canvas.