UWEO StatR201 – Winter 2013: Final Project

<u>*Due:*</u> *Proposals - Thursday March 14, before Class, by email to assaf@uw.edu*

Project Reports – Friday March 29, 11:59 PM, in Dropbox

<u>Instructions:</u>

- **Submission format is .pdf. Upload code excerpts in a <u>separate</u> text file (*.txt,*.r,*.rmd).**
- **Grading is Pass/Fail, based on adherence to requirement, and demonstration of reasonable proficiency and effort in using the methods commensurate with experience level and time limitations. Pass/Fail <u>not</u> based on predictive performance of methods used.**

Choose one dataset that is of interest to you. This could be a publicly available one, or one from work. **Please make sure to cite the dataset's source if so required, and/or to obtain permission for using the dataset if it is from a business source.**

The dataset should have between about 1000 and 50000 observations (n), and between about 10 and 500 features (p). On this dataset, you will carry out a complete predictive-modeling process – either regression or classification, as befits the question of interest.

Report your work as a **single, readable and aesthetically-acceptable document** (except for code that should be attached as a separate file). **The report should be between 10 and 20 pages long, including figures** (but excluding references and appendices).

<u>The work steps:</u>

- Before starting out, separate **a test set comprising 25%-50% of the data** (less for smaller datasets, more for larger ones). This test set will be left out until the last stage. If you dataset is "naturally" grouped, then no group should be split between training and test set. For example, datasets like the smartphone-activity one which is grouped by person, have a test set composed of certain subjects' entire sets of observations – **and no data from persons belonging to the training set.**
- Carry out extensive exploratory analysis **on the training set only**, including justified decisions on variable transformations, and exclusion/filtering of observations or features.
- Choose **<u>at least two</u> prediction methods, and <u>at least one</u> of them should be more sophisticated/advanced** (e.g., if you chose regression then you need to attempt something **<u>beyond</u>** various forms of subset selection). For that more advanced method, explore usage beyond the function's *"factory defaults."*
- Tune the methods on the training set, using the appropriate tuning parameter(s), **and a context-appropriate (cross-) validation grouping.**
- Finally, predict the test set's observations, and score your predictions by method, using context-appropriate performance metrics.

You are strongly encouraged to ask me questions during the entire process.

The final report should include:

1. A brief introduction **(1-2 paragraphs max)** describing the dataset and the question of interest.
2. Exploratory analysis summary, including some graphics, and justified decisions on variable transformations and exclusion/filtering of observations or features, and on the main and secondary scoring metrics.
3. Cross-validation/Bootstrap-validation tuning, for 2-4 prediction methods on your dataset. The methods should be appropriate to the problem at hand. The tuning report should be both in table (tuning variable values vs. performance), and also scatterplot or confusion-matrix form at least for the "winning" tuning choice.
4. Test-set descriptives analogous to #2 above (they can be shorter), and then test-set prediction scores. **Beside the main scoring metric, be sure to also include the standard one in case you chose a different one (RMSE for regression, "flat" error-rate for classification). Show an "observed vs. predicted" scatterplot or confusion-matrix for every final test-set prediction.**
5. Final discussion **(1-2 paragraphs max). Please avoid lengthy prose discussions. Inside Sections 2-4, keep commentary to one-sentence blurbs** (e.g., "regression method X causes more outliers than regression method Y").

Notes:

Meaning of context-correct CV groups: For example, if your data are "naturally" grouped (say, by person or by location), then your "ultimate" cross-validation should exclude entire groups together (i.e., all of one person's observations should be in the same CV group). Otherwise, you tuning is guaranteed to be too optimistic, and in any case unrealistic.

Use of the test set: you cannot re-tune your prediction method on your test set. You have to use the tuning values chosen by your validation process. Since this is an exercise, it is OK to try out a couple of distinct "flavors" of the method as a sensitivity analysis, for the learning and curiosity value. For example, if `kknn` CV indicates that a triangular kernel with Euclidean distance is narrowly better than a rectangular one with absolute-value distance, it is ok to designate one as your main choice, but still check out the performance of the other one as a side note.

Please avoid fluff. For example, if you already show visual scatters or quantile plots indicating covariate distribution, there is no need to list out a table with each feature/s summary statistics.

Good luck! And don't hesitate to ask questions!