ISE/DSA 5103 Intelligent Data Analytics
Homework #6

Instructor: Charles Nicholson

See course website for due date

Learning objective: Data wrangling, regression modeling, and analysis.

Submission notes:

1. **There are TWO components to this assignment**: a *Canvas* submission and a *Kaggle* submission.
2. Team assignment! Teams of up to 3 people allowed! Include all team member names on the submitted work. You will form teams on Canvas *and* those same teams will need to be formed on Kaggle.
3. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader *may* view your R code, but should never *have* to in order to find your solutions.
    - In the PDF, clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.) Also, note that only *relevant* and informative computer output should be provided.
    - Make sure to *provide comments* on what your R code is doing. Keep it clean and clear!
    - You will submit your complete R or Rmd script. Note: include `library` commands to load *all* packages that are used in the completion of the assignment. Place these statements at the top of your script.
    - Do not zip your files for submission. Submit exactly two files. Name the files "LastName-HW1" with the appropriate file extension (that is, .pdf for the write-up and .R or .Rmd file for the code)
4. A significant portion of your submission requirement is contribution to an in-class Kaggle competition. See notes below!

Kaggle competition notes:

1. In order to join the competition, you need to create a Kaggle account. Only one account per student is allowed.
2. Join the competition URL in the Canvas homework description page.
3. Once you join Kaggle and the competition, to create a team:
    (a) Have one person click on "Team"
    (b) Request a merge by searching for one of the other team members user names and "Request Merge"
    (c) Create a team name as stated in the "Rules" section of the Kaggle competition web page.
4. Grades will, in part, be based on the quality of your predictions as compared to the other teams in the class. It is your responsibility to read the rules and information on the competition website.

1 Online retail sales prediction

In many businesses, identifying which customers will make a purchase (and when), is a critical exercise. This is true for both brick-and-mortar outlets and online stores.

The data provided in this assignment is website traffic data acquired from an online retailer and provides information on customer's website site visit behavior. Customers may visit the store multiple times, on multiple days, with or without making a purchase.

Your goal is to predict how much sales revenue can be expected from each customer. The variable `revenue` lists the amount of money that a customer spends on a given visit. Your goal is to predict how much money a customer will spend, *in total*, across all visits to the website, during the allotted one-year time frame (August 2016 to August 2017).

More speciffically, you will need to predict a transformation of the aggregrate customer-level sales value based on the natural log. That is, if customer i has k_i revenue transactions, then you should compute:

$$custRevenue_i = \sum_{j=1}^{k_i} revenue_{ij} \quad \forall i \in customers$$

And then transform this variable as follows:

$$targetRevenue_i = \ln(custRevenue_i + 1) \quad \forall i \in customers$$

You will be evaluated on how well you can predict the target revenue on a test data set available at the Kaggle.com website (see the Canvas assignment page for the private competition URL)

(a) (50 points) **Preparation and modeling.**

i. (10 points) *Data understanding.* Generate a *Data Quality Report.* Also, choose at least two meaningful visualizations and/or analyses and explain their relevance.

ii. (10 points) *Data preparation.* Choose two of the most critical data preparation actions you took and explain the reasoning for these actions.

iii. (20 points) *Modeling.* Build an OLS model and 3 or more regression variant models (these may include robust regression, PLS, PCR, ridge regression, LASSO, elasticnet, MARS, or SVR) and summarize their performance in a table (as shown in Table 1). Clearly state your resampling approach. Note: You may combine models, techniques, etc.

iv. (10 points) *Debrief.* For your best predictions, describe your approach, e.g., did you examine interactions? did you use any type of model stacking? what was your secret sauce? Did you have any problems during the modeling process? If so, how did you overcome those?

(b) (50 points) **Competition modeling.**

- Upload your predictions to the Kaggle website and check the predictive performance on the "Public Leaderboard"
- You may submit multiple times throughout the competition, however, there is a limit to the number of submissions per day.
- Score is based on ranked performance on the "Private Leaderboard"; extra-credit is possible.
- All modeling approaches *covered in lecture* can be used (OLS, robust methods, dimension reduction methods, penalized methods, MARS, SVR, PCA, LDA, k-nn, t-SNE, transformations, missing value imputations, etc.) To be fair, approaches not yet discussed in detail are not allowed (e.g., tree-based models, neural network based models, clustering, are not allowed at this time.)
- You must outperform the *benchmark* model to receive any credit.
- May the odds be ever in your favor.

Table 1: Summary of Model Performance

Model	Method	Package	Hyperparameter	Value	CV performance RMSE	R^2
OLS	`lm`	`stats`	N/A	N/A	3192.4	0.5770
lasso	`glmnet`	`glmnet`	λ	1.21	4548.1	0.4910
PLS	`plsr`	`pls`	`ncomp`	18	2999.6	0.5913
Huber loss	`rlm`	`MASS`	N/A	N/A	1771.3	0.6831
elasticNet	`enet`	`elasticnet`	fraction lambda	0.72 0.1	979.9	0.7188
MARS	`earth`	`earth`	degree	2	1370.1	0.6222
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Team naming convention

Give your team a proper name!

Please prefix your team name according to the following protocol:

(O) if you are online students only (enrolled in sections 994, 995, 996, or 997);

(C) if you are on-campus students (enrolled in section 001);

Additionally, you are in group in Canvas. Please note your group number and post-pend that to your Kaggle team name, e.g.: (O) Yeet-13, (C) Data Maniacs-21, (O) Awesomesauciness-32. When submitting your PDF to Canvas, **leave a comment with your Kaggle team name.**