ISE 5103 Intelligent Data Analytics
Homeworks #4 and #5

Instructor: Charles Nicholson

See course website for due date

Learning objective: Data wrangling, regression modeling and analysis.

Submission notes:

1. Teams of 1, 2 or 3 – make sure to set this up correctly in Canvas *and* on the Kaggle.com competition site. Ask if you don't know!
2. 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.
   (a) I expect high-quality, clear, concise yet complete, easy to read PDFs.
   (b) 15 page max – penalties will be incurred per page over the allowance.
3. 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. For example, do not include "warning" messages, or the results of "library" commands, etc.
4. Make sure to provide comments on what your R code is doing. Keep it clean and clear!
5. You will submit your complete R 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/code.
6. Do not zip your files for submission. Submit exactly two files. The PDF, like usual, is your primary submission. You must also submit your R code.

Kaggle competition notes:

- In order to join the competition, you need to create a Kaggle account. Only one account per student is allowed.
- Join the competition here: `https://www.kaggle.com/c/2020-5013-hw4`
- Once you join Kaggle and the competition, to create a team:
  - Have one person click on "Team"
  - Then, request a merge by searching for one of the other team members user name and "Request Merge"
  - Create a team name as stated in the "Rules" section of the Kaggle competition web page (and summarized at the end of this document.) Even if you want to work alone, you need to form a team so that we can recognize if you are online or on-campus student.
- 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 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. You will be predicting customer sales.

The data and many details about the problem can be found here:
`https://www.kaggle.com/c/2020-5013-hw4/`.

The data 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 ultimate 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 specifically, 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:

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

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.

Parts (a) and (b) constitute Homework #4. Parts (c) and (d) constitute Homework #5.

(a) (50 points) Conduct a thorough exploratory data analysis (EDA), using any and all techniques you prefer, to help understand the data and prepare you for the *data wrangling* step next.

While you may do as much you like, make sure to choose *at least* 5 EDA analyses of value and provide commentary as to what you have found. Please note: Each visualization, table of statistics, etc., that you include must be described thoughtfully. Your "insight" represents half of the points for this problem – make it good.

(b) (50 points) Data preparation. Choose five categories of data preparation tasks for the data. For each, state the method and describe/visualize the results. The categories you may choose from for include:

- missing value imputation
- resolution of outliers
- aggregations
- transformations
- collapsing categories
- creating new categories or binning
- feature engineering or feature extraction

You may wish to consider using the `dplyr`, `lubridate`, and/or `forcats` packages, among others, to help you manage this process.

(c) (20 points) Continue preparing the data for modeling. This requires you to consider missing values, outliers, transformations, aggregations, and/or any other data preparation technique you find useful as in the previous step, however, now it should be an iterative process with the subsequent modeling step. The deliverable for this step is a concise summary of the choices you have made and an explanation for your choices for your *best* model.

(d) (80 points) Modeling.

i. (40 points) You must build an ordinary least squares (OLS) model and at least 3 different classes of models from the following list: robust regression, lasso, ridge, elasticnet, PLS, multiadaptive regression splines, and/or SVM-regression. Each of your models with hyper-parameters should be tuned using a re-sampling method of your choice. Summarize model performances in a table that identifies: R method and underlying library (not `caret`), specifics with respect to tuning parameters, and re-sampled performance metrics.

					CV performance	
Model	Method	Package	Hyperparamter	Selection	R^2	RMSE
OLS	`lm`	`stats`	NA	NA	0.417	1.012
lasso (large)	`lasso`	`elasticnet`	fraction	0.84	0.618	0.741
lasso (small)	`lasso`	`elasticnet`	fraction	0.27	0.559	0.814
Huber loss	`rlm`	`MASS`	NA	NA	0.633	0.739
MARS	`earth`	`earth`	degree	3	0.701	0.719
etc.						

ii. (5 points) For your best model, describe your modeling 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?

iii. (35 points) Submit your model predictions to the Kaggle.com competition website and outperform your peers in high quality predictions on the test data. You can submit multiple times each day to get feedback on the "public leaderboard". The final competition placement will be based on the "private leaderboard" standings. See the competition website for more details. You must outperform the baseline model to earn any points.

Team naming convention

You have to give your group 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); (M) if you are mixed team.

Additionally, you will need to sign-up for a group in Canvas. Please note your group # and post-pend that to your Kaggle team name, e.g.: (O) Yeet-13, (C) Data Maniacs-21, (M) Awesomesauciness-32.

When submitting your PDF to Canvas, please leave a comment with your Kaggle team name. Only one team member has to submit the final work on Canvas.