ISE 5103 Intelligent Data Analytics
Homework #6

Instructor: Charles Nicholson

See course website for due date

Learning objective: Advanced regression modeling.

Submission notes:

1. Teams of 1 or 2 (optional) – 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) 10 page max – 10% penalty per page over the allowance.
   (c) You may include an appendix with supplementary information if desired. The appendix does not count against your page limit.
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, I do not want to see "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. Name the files LastName-HW1 with the appropriate file extension (that is, .pdf for the write-up and .R or .Rmd for the 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/t/495cc1d2f46647dabe8af1d04a194e87`
- 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 instructions in the "rules" tab. 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/t/e5aa719be3584ef091d9f61fc931fb36`.

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 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.

(a) (50 points) You must build at least 5 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.

The deliverable for this part is three-fold:

- Choose one model with hyper-parameters and justify your choice on how you tuned the model. Please support with one or more visualizations.
- From your work, choose two different model class instances and compare/contrast the results in detail, e.g., you may discuss differences in regression coefficients, model complexity, residual diagnostics, etc.
- Summarize all 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. Include results from your Homework #5 OLS model.

					CV performance	
Model	Method	Package	Hyperparamter	Selection	R^2	RMSE
OLS HM-5	`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.						

(b) (50 points) Build the best possible regression model(s) to predict the target value.

i. (15 points) For your best model, report the variables, coefficient estimates, and p-values (you may do this in the appendix if it is a large model) Additionally, report the re-sampled RMSE and R^2 values as well as any tuning parameter values. Describe your modeling approach, e.g., did you examine interactions? did you create use more than one model? what was your secret sauce?

ii. (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.