ISE 5103 Intelligent Data Analytics
Homework #7

Instructor: Charles Nicholson

See course website for due date

Learning objective: Classification modeling.

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) **12 page max** – 2 point 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 URL in the Canvas homework description page.
- Once you join Kaggle and the competition, to create a team:
  1. Have one person click on "Team"
  2. Request a merge by searching for one of the other team members user names and "Request Merge"
  3. Create a team name as stated in the "Rules" section of the Kaggle competition web page. FOLLOW THE NAMING CONVENTIONS FOR YOUR TEAM NAME. 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 Hospital Readmits

A hospital readmission is an episode when a patient who had been discharged from a hospital is admitted again within a specified time interval. Readmission rates have increasingly been used as an outcome measure in health services research and as a quality benchmark for health systems. Hospital readmission rates were formally included in reimbursement decisions for the Centers for Medicare and Medicaid Services (CMS) as part of the Patient Protection and Affordable Care Act (ACA) of 2010, which penalizes health systems with higher than expected readmission rates through the Hospital Readmission Reduction Program. The real-world clinical care data provided in this competition comes from multiple hospitals across the United States for several years.

The challenge: Readmissions

In this machine learning problem, you have the exacting task of trying to predict whether or not a patient will be readmitted to the hospital. The target takes on binary values where 0 implies that the patient was not readmitted, and 1 implies that the patient was readmitted. You are predicting the latter.

The data and many details about the problem can be found on the course Kaggle competition page – see the Canvas page for link.

Your classification model will be evaluated using *log loss*. Your predictions, therefore, will be probabilities of readmission.

$$\text{log loss } = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right]$$

where N is the number of observations; y_i is the actual observed value for the outcome i^{th} observation and equals either 1 or 0; and $p_{i,j} \in [0,1]$ is the predicted probability that observation i equals 1.

You will be evaluated on how well you can predict the hospital readmits on a test data set available on the Kaggle.com class competition page.

(a) (60 points) Build at least 5 different classes of classifiers from the following list: linear discriminant analysis, logistic regression, penalized logistic regression (e.g., using glmnet), MARS (for classification), kNN, decision tree, random forest, bagged trees, boosted trees, SVM, neural nets. Hyper-parameters should be tuned using a re-sampling method of your choice.

- (10 points) R code that successfully constructs and tunes at least 5 distinct model types. This code does *not* need to be in the PDF, but it should be easy to find in the R code.
- (15 points) Summarize all model performances in a table that identifies: R method and underlying library (not just `caret`), specifics with respect to tuning parameters, and re-sampled performance metrics of your choice.

					CV performance	
Model	Method	Package	Hyperparamter	Selection	Accuracy	Kappa
logreg	`glm`	`stats`	NA	NA	0.627	0.570
lasso (logreg)	`glmnet`	`glmnet`	lambda	0.84	0.618	0.741
decision tree	`rpart`	`rpart`	cp	0.27	0.559	0.814
MARS	`earth`	`earth`	degree	3	0.701	0.719
etc.						

- (20 points) Choose one model (of your choice) and provide at least 3 potential "insights" relating to hospital readmits that might be of some use to hospitals, insurance companies, doctors, patients, and/or government administration.
- (15 points) Using at least 3 different types of performance evaluation techniques, quantify (and/or visualize) the predictive quality of the above model. Performance evaluation techniques include (but are not limited to): accuracy, kappa, weighted accuracy, weighted kappa, log-loss, F_1-score, ROC, AUROC, Precision-Recall Curve, AUPRC, concordant pairs, confusion

matrices, difference in distributions of predicted probabilities, D statistic, cumulative gains, cumulative lift, Kolmogorov-Smirnov chart, etc.

Note: the `confusionMatrix` command in the `caret` package computes 13 separate metrics, you may count only 1 of these towards your total distinct analysis methods.

Second note: The performance metrics on training data are ok, but no one really cares about training performance; you need to compute measures based on your resampling holdout sets/folds.

(b) (40 points) Submit your best 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. To earn points, you must outperform the "benchmark" model. Extra-credit opportunity is available for the best predictive performance.

(c) (X points) 5 to 10 points of **extra-credit** for each individual who creates an informative Kaggle.com notebook on the competition site and earns at least 8 "upvotes" from the class.