ADVERTISEMENT FOR *PREDICTIVE MODEL BUILDING*

Hasthika S. Rupasinghe Arachchige Don, Lasanthi C. R. Pelawa Watagoda, Alan T. Arnholt
Appalachian State University
Mathematical Sciences; 121 Bodenheimer Dr.; Boone, NC 28608
hasthika@appstate.edu

Knowing how to build a predictive model is an important skill for anyone working with data (De Veaux et al. 2017; Hicks and Irizarry 2018; Hardin et al. 2015). Unfortunately, detailed explanations of many algorithms used by researchers to create predictive models along with directions on how to use software to implement the algorithms are not commonly found in undergraduate textbooks. The fifth edition of the excellent text *STATS Data & Models* by De Veaux, Velleman, and Bock (2019) devotes chapter 27 to statistical learning and data science. The text does a great job of introducing and describing the standard types of problems in the statistical learning/data science fields; however, the instruction is not sufficiently detailed for the general reader of such a text to implement the ideas. The fact that statistical learning is even mentioned in an introductory text is a testament to the importance of statistical learning algorithms and their importance in the toolbox of practitioners and undergraduates. One text that does cover both theory and implementation of statistical learning algorithms (appropriate for advanced undergraduates or master's students according to its preface) is *An Introduction to Statistical Learning with Applications in R* by James et al. (2017).

Hicks and Irizarry (2018) suggest the following five guiding principles when teaching a data science course: "1) Organize the course around a set of diverse case studies; 2) Integrate computing into every aspect of the course; 3) Teach abstraction, but minimize reliance on mathematical notation; 4) Structure course activities to realistically mimic a data scientist's experience; 5) Demonstrate the importance of critical thinking/skepticism through examples."

One of the challenges instructors face when using a standard text is providing activities that mimic a data scientist's experience since data sets that accompany standard texts are generally clean and ready to be analyzed. The challenge and the need to expose students in applied statistics courses to unclean data is detailed in Zhu et al. (2013). A second challenge is the plethora of `R` packages and differing syntax among `R` packages one may choose to implement the numerous statistical learning algorithms.

The **caret** package solves the problem of differing syntax among `R` packages by providing a unifying interface with consistent arguments for 238 models. For details on one of the 238 models, one should consult the `caret` vignette at http://topepo.github.io/caret/available-models.html. The **caret** package has many functions of which only a small subset are needed to start building predictive models. Three of the principle functions used for building predictive models from the `caret` package

are `createDataPartition()`, `trainControl()`, and `train()`. The `trainControl()` function controls the resampling method used to train a predictive model. In this document, the method is either cross validation or repeated cross validation; however, there are thirteen different resampling techniques one could specify. For complete details, see the `R` help file for `trainControl()` by typing `?trainControl` at the `R` prompt. The ability to have a unified resampling method applied to any one of the 238 different models in **caret** permits the user to build predictive models that optimize the bias variance trade-off. While a resampling method such as cross validation is not conceptually difficult to understand, implementing a resampling method for a particular algorithm can be challenging.

The links to the guided labs that accompany this document (constantly being updated) introduce a case study where students integrate computing into a dynamic document using `R Markdown` as advocated by B. Baumer et al. (2014), Hardin et al. (2015), De Veaux et al. (2017), and Hicks and Irizarry (2018), create statistical models with a consistent interface using the **caret** package, mimic the workflow of a data scientist by using data that is of questionable integrity, and practice critical thinking/skepticism by working with questionable data. The ability to create a dynamic document involving predictive models with undergraduates that incorporates the full data science process is innovative simply because the computational tools required to do so in a reasonable amount of time have not previously been available. Transformations and model validation are two aspects of model building that were particularly onerous that the **caret** package simplifies immeasurably.

The guided labs have been used with both graduate and undergraduate classes. The computational preparation for both classes was similar coming into the course where the project was used. For the graduate students, the guided labs were implemented in a course where the students were already using `R`, `R Markdown`, and had been exposed to the grammar used in the **ggplot2** package. The prerequisite for the graduate class was a standard undergraduate (non-calculus based) introductory statistics course. The undergraduate course (STT 3860) where the students used the guided labs also has as a prerequisite a standard undergraduate (non-calculus based) introductory statistics course in addition to a data visualization and management course (STT 2860). The STT 2860 course introduces students to `R` (both base and tidyverse), `R Markdown`, Git, and data management.

Both the graduate and undergraduate students had access to a University maintained RStudio Pro server. While R, RStudio, and Git can be installed by students on their personal machines, we found using a server virtually eliminates the time instructors were spending providing technical support despite providing students with detailed installation instructions. For instructors wishing to use a server without institutional hardware or support, RStudio Cloud is a fantastic and free resource provided by RStudio. In both the graduate and undergraduate courses where the project was used, Git was used for version control. Specifically, GitHub Classroom was used to distribute starter code for each class. While Git is not needed for students to complete the

project, the instructors use Git issues to provide feedback which leaves a permanent trail of comments the instructor provides to the students.

Predictive Model Building at https://stat-ata-asu.github.io/PredictiveModelBuilding/ provides an overview of linear regression and introduces a few algorithms generally not found in undergraduate texts. After a brief explanation of the algorithms, the statistical programming language `R` is used to build predictive models using a well-known data set. *Predictive Model Building* includes parametric and nonparametric estimation of f, visualizing and checking data, preprocessing data, k-fold cross validation, and detailed instructions for cleaning a data set.

To practice the material discussed in *Predictive Model Building*, three (more to come) guided R Markdown labs are provided. The guided labs all have public GitHub repositories with additional files that may be cloned to the user's machine. Instructors can also use the GitHub repositories to set up assignments with Github Classroom. For the Git averse, the guided R Markdown labs are also provided as RStudio Cloud projects where the `R` packages needed to answer the guided lab have been installed. To use the projects on RStudio Cloud, the user will need to create an account on RStudio Cloud. Since there are numerous `R` packages, the project will take a few minutes to deploy the first time the user opens the project. To save a permanent copy of the project, click on the red "Save a permanent copy" text in the upper right of the RStudio title bar. Regardless of the approach one uses to work with the `R Markdown` file, the end result is that the reader/student is creating a dynamic document that records the steps used to solve the questions as advocated by B. Baumer et al. (2014), Hardin et al. (2015), De Veaux et al. (2017), and Hicks and Irizarry (2018).

Guided Labs

1. Questioning and Cleaning the bodyfat data Lab:
    - GitHub repository - https://github.com/STAT-ATA-ASU/AHL-GL-BFcleaningSC
    - rstudio.cloud project - https://rstudio.cloud/project/1164604
2. Linear models with the bodyfat data Lab:
    - GitHub repository - https://github.com/STAT-ATA-ASU/AHL-GL-BFlinearSC
    - rstudio.cloud project - https://rstudio.cloud/project/1164829
3. Non-linear models with the bodyfat data Lab:
    - GitHub repository - https://github.com/STAT-ATA-ASU/AHL-GL-BFnonlinearSC
    - rstudio.cloud project - https://rstudio.cloud/project/1169242

References

Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J. Horton. 2014. "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations in Statistics Education* 8 (1). https://escholarship.org/uc/item/90b2f5xh.

De Veaux, Richard D., Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, et al. 2017. "Curriculum Guidelines for Undergraduate Programs in Data Science." *Annual Review of Statistics and Its Application* 4 (1): 15–30. doi:10.1146/annurev-statistics-060116-053930.

De Veaux, Richard D., Paul F. Velleman, and David E. Bock. 2019. *Stats: Data and Models*. 5 edition. Hoboken, NJ: Pearson.

Hardin, J., R. Hoerl, Nicholas J. Horton, D. Nolan, B. Baumer, O. Hall-Holt, P. Murrell, et al. 2015. "Data Science in Statistics Curricula: Preparing Students to "Think with Data"." *The American Statistician* 69 (4): 343–53. doi:10.1080/00031305.2015.1077729.

Hicks, Stephanie C., and Rafael A. Irizarry. 2018. "A Guide to Teaching Data Science." *The American Statistician* 72 (4): 382–91. doi:10.1080/00031305.2017.1356747.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2017. *An Introduction to Statistical Learning: With Applications in R*. 1st ed. 2013, Corr. 7th printing 2017 edition. New York: Springer.

Zhu, Yeyi, Ladia M. Hernandez, Peter Mueller, Yongquan Dong, and Michele R. Forman. 2013. "Data Acquisition and Preprocessing in Studies on Humans: What Is Not Taught in Statistics Classes?" *The American Statistician* 67 (4): 235–41. doi:10.1080/00031305.2013.842498.