Read-Me and Instructions for Running the Final Project

CSC849 Spring 2017
Paul Previde

This document provides instructions on running the programs of the final project, as required by the instructions contained in the document “Demonstrating, Presenting, and Submitting Your Term Project.”

I. Requirements

My program, entitled “PaulPrevide_869_Project.Rmd”, is written in R, version 3.3 using the Rstudio integrated development environment. The program is in the form of an Rmarkdown file, with the file extension “.Rmd”. This file type opens in Rstudio. Rmarkdown is analogous to iPython, in that it allows code, output, and other text to be woven together into a presentation document. Also, Rmarkdown allows separating your program into discrete chunks, useful for development and testing.

The following libraries are required for all the data processing and analysis steps. The functionalities provided by each are listed in parentheses.

- dpylr (various R commands for manipulating data frames)
- ggplot2 (plotting functions)
- boot (bootstrapping and cross validation)
- glmnet (ridge, lasso, and elastic net regression)
- e1071 (naive Bayes classification)
- MASS (linear and quadratic discriminant analysis)
- randomForest (bagging and random forest classification)
- adabag (Adaboost classification)
- knitr (opening and executing an Rmarkdown file)

Instructions for installing the above tools and libraries are described in the next section.

II. Installation of the Required Tools and Libraries

1. Install R:

(a) Windows
Download and execute the installer, available at the following link:
https://cran.r-project.org/bin/windows/base/

(b) Linux/Ubuntu
For Ubuntu users, R is available through the Ubuntu Software Center repository. The following packages should be installed:
r-base
r-base-dev

(c) Mac

Download and install R for your version of Mac OS at the following:
https://cran.r-project.org/bin/macosx/

2. Install Rstudio IDE

The Rstudio IDE is available for download at the link below:
https://www.rstudio.com/products/rstudio/download/

3. Install required libraries:

The following commands can be copied into an R script and executed in order to install the required libraries:

```
install.packages(“dpylr”)
install.packages(“ggplot2”)
install.packages(“boot”)
install.packages(“glmnet”)
install.packages(“e1071”)
install.packages(“MASS”)
install.packages(“randomForest”)
install.packages(“adabag”)
install.packages(“knitr”)
```

III. Downloading and Using the Data

The College Scorecard data set used for this project is available at the following link:
https://www.kaggle.com/kaggle/college-scorecard

Because of the size of the dataset (~1 GB), it is not possible to upload the comma-separated values file to iLearn. To download the file and have it ready for use by the R program, follow these instructions:

1. Download the college-scorecard.zip file from the link above.
2. Move the Scorecard.csv file to a preferred directory that you wish R to use as the working directory (possibly the same directory that R is installed in).
3. In R, execute the command

```
setwd(<path_to_the_directory_you_chose>)
```

so that R will know where to find the csv file.

IV. Running the Program

Once R and Rstudio are installed, the programs associated with this project can be opened using File > Open File. The program will open in the upper-left window of Rstudio. To run the entire Rmarkdown file, find the “Run” drop-down menu just above the code, and choose “Run all” to run the entire program. You can also run just certain chunks of the code as well, using the other commands like “Run current chunk” or “Run all chunks below” from the same drop-down menu.