**README: Replication of *To Bag is to Prune***

## **Overview**

This repository, named **"replication\_codes\_final"** contains the necessary data and code to replicate Figures 1 through 7 and the Tables 2 and 3 from the paper **"To Bag is to Prune"** by Philippe Goulet Coulombe. This README will guide you through the replication process for each figure.

## **Repository Structure**

The repository is organized into subfolders, each corresponding to a figure or set of figures. Each folder contains the specific data and code needed to replicate the respective figure(s):

### **Folder Descriptions**

1. **Empirical/**
   * **data/:** contains the data that are used in the exercice
   * **results/:** output of empirical\_results\_loop.R
   * **tools/**:
     + tools/keras\_mlp\_cv.R – Keras function for a feed foreward neural network
     + tools/ booging\_functions\_pack4.R
     + tools/dataset-nocut.R– Generate the needed dataset
   * **Code**:
     + empirical\_results\_loop.R – The prediction out of sample code for generating Figures 1 and 5.
     + intro\_bp\_code\_figure1.R – Replicate Figure 1
     + R2\_barplots\_figure5.R – Replicate plots from Figure 5
     + tables2\_3\_demo.R – Replicate results from Tables 2 and 3
2. **Section2\_examples/**
   * **results/:** contains the output and of simulations\_ols\_vs\_greedyls.R and simulations\_R2\_test.R
   * **tools/ :** 
     + tools/datasets\_nocut.R – Generate the needed dataset
     + tools/keras\_mlp\_cv.R – Keras function for a feed foreward neural network
     + tools/dataset-nocut.R– Generate the needed dataset
   * **Code :**
     + simulations\_ols\_vs\_greedyls.R – Script for running the necessary simulations to generate Figure 2.
     + ols\_vs\_greedyls\_plot\_figure2.R – Script for generating the Figure 2.
     + simulations\_R2\_test.R – This script is used to perform the exercise needed to generate Figure 3.
     + R2\_test\_plot\_figure3.R – Script for generating the Figure 3.
3. **Simulation/**
   * **results**/ :
     + **figure7**/ : contains the output of main\_simu\_fig7.R
     + **figure4\_6**/ : contains the output of main\_simu\_fig4\_6.R
   * **tools/ :**
     + booging\_functions\_pack3.R – Contains the model functions required for the simulations.
   * **Code:**
     + main\_simu\_figure4\_6.R – Script for running simulations related to Figures 4 and 6.
     + main\_simu\_figure7.R – Script for running simulations related to Figure 7.
     + graphs\_simu\_figure7.R – Script to generate figure 7
     + graphs\_simu\_figure4\_6.R – Script to generate figures 4 and 6

## **Data**

|  |  |
| --- | --- |
| **Abbreviation** | **Data Source** |
| Abalone | http://archive.ics.uci.edu/ml/datasets/Abalone |
| Boston Housing | http://lib.stat.cmu.edu/datasets/boston |
| Auto | https://archive.ics.uci.edu/ml/datasets/Auto+MPG |
| Bike Sharing | https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset |
| White Wine | https://archive.ics.uci.edu/ml/datasets/Wine+Quality |
| Red Wine | https://archive.ics.uci.edu/ml/datasets/Wine+Quality |
| Concrete | https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength |
| Fish Toxicity | https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity |
| Forest Fire | https://archive.ics.uci.edu/ml/datasets/Forest+Fires |
| NBA Salary | https://www.kaggle.com/aishjun/nba-salaries-prediction-in-20172018-season/data |
| CA Housing | https://www.kaggle.com/camnugent/california-housing-prices |
| Crime Florida | https://www.census.gov/data/datasets/1990/dec/summary-file-3.html |
| Friedman 1 R2=0.7 | https://cran.r-project.org/web/packages/tgp/vignettes/tgp.pdf |
| Friedman 1 R2=0.4 | https://cran.r-project.org/web/packages/tgp/vignettes/tgp.pdf |
| GDP H = 1 | https://drive.google.com/file/d/1NxDKvr-gyf2hhZS7oJ2h12vFk-N3SZAJ/view?usp=sharing |
| GDP H = 2 | https://drive.google.com/file/d/1NxDKvr-gyf2hhZS7oJ2h12vFk-N3SZAJ/view?usp=sharing |
| UNRATE H = 1 | https://drive.google.com/file/d/1NxDKvr-gyf2hhZS7oJ2h12vFk-N3SZAJ/view?usp=sharing |
| UNRATE H = 2 | https://drive.google.com/file/d/1NxDKvr-gyf2hhZS7oJ2h12vFk-N3SZAJ/view?usp=sharing |
| INF H = 1 | https://drive.google.com/file/d/1NxDKvr-gyf2hhZS7oJ2h12vFk-N3SZAJ/view?usp=sharing |
| INF H = 2 | https://drive.google.com/file/d/1NxDKvr-gyf2hhZS7oJ2h12vFk-N3SZAJ/view?usp=sharing |

## **Instructions for Replication**

The replication codes for this paper are organized into three sections: **Empirical**, **Simulation**, and **Section 2 examples**. The specific instructions for each part are detailed below. The needed packages are installed within the codes.

**Empirical**

This section is for replicating the empirical part of the paper, which includes Figures 1 and 5, as well as Tables 2 and 3.

1. **Main Script:** empirical\_results\_loop.R
   * Only modify the relative path variable ("path") in the code.
   * The results will be saved in the results/ folder under the name boog23nn\_newcode\_DS\*.RData (where "\*" is the dataset number).
2. **Figure 1:**
   * Use the script Figure1\_intro\_bp\_code.R to reproduce Figure 1 from the paper.
   * Only modify the relative path variable ("path") in the code.
   * Output: figure1\_bp\_intro\_abalone7.png
3. **Figure 5:**
   * Use the script Figure5\_R2\_barplots.R to generate the bar plots for Figure 5.
   * Only modify the relative path variable ("path") in the code.
   * Output: DS\*\_results\_barplots.png (where "\*" is the dataset number)
4. **Tables 2 and 3:**
   * Use the script tables2\_3\_demo.R to generate the results from Tables 2 and 3.
   * Only modify the relative path variable ("path") in the code.

**Section 2 - Examples**

This section is for replicating the simulation exercise, including Figures 2 and 3 from the paper.

1. **Figure 2**
   * Run the script simulations\_ols\_vs\_greedyls.R to replicate the simulation results. Only modify the path variable ("path") in the code. The simulation results will be stored in the results/ folder. Output : NeuralNet\_out\_DS\*.RData
   * Run the script ols\_vs\_greedyls\_plot\_figure2.R to generate Figure 2. Only modify the relative path variable ("path") in the code. Output: figure2\_ols\_vs\_Greedy2.png
2. **Figure 3**
   * Run the script simulations\_R2\_test.R to replicate the simulation results.
   * Run the script R2\_test\_plot\_figure3.R to generate Figure 3. Only modify the relative path variable ("path") in the code. Output : ols\_vs\_greedyls\_results.RData
   * Output: figure3\_NN\_vs\_RF.png

**Simulation**

This section is for replicating the simulation exercise, including Figures 4, 6, and 7 from the paper.

1. **Figures 4 and 6:**
   * Run the script main\_simu\_figure4\_6.R to replicate the simulation results. Only modify the relative path variable ("path") in the code. The simulation results will be stored in the results/figure4\_6/ folder.
   * Use the script graphs\_simu\_figure4\_6.R to generate Figure 2. Only modify the path variable ("path") in the code.
   * Output: simulplot\_mtry90\_figure4.png and simulplot\_mtry90\_figure6.png
2. **Figure 7:** 
   * Run the script main\_simu\_figure7.R to replicate the simulation results. Only modify the relative path variable ("path") in the code. The simulation results will be stored in the results/figure7/ folder.
   * Use the script graphs\_simu\_fig7.R to generate Figure 2. Only modify the path variable ("path") in the code.
   * Output: simulplot\_mtry50\_figure7.png

**Software Requirements**

To replicate the figures, the following software and packages are required:

* **R** version 3.6.3 has been used on Mac OS

### **R Packages:**

* glmnet
* rpart
* ranger
* mboost
* gbm
* ISLR
* caret
* dplyr
* mlbench
* AppliedPredictiveModeling
* knockoff
* earth
* pracma
* tgp
* ggplot2
* RColorBrewer
* reshape2
* readr
* GA
* e1071
* doParallel
* kernlab
* xts
* neverhpfilter
* grf
* EZtune
* tuneRanger
* party
* readxl
* forecast
* ggthemes
* extrafont
* scales
* wesanderson
* grid
* keras
* tensorflow
* bagofprunes (from <https://github.com/philgoucou/bagofprunes>)

**Troubleshooting**

If you encounter any issues, please ensure the following:

* All necessary R packages are installed.
* The data files are correctly placed in the specified folders.
* The relative path variable “path” is correct.