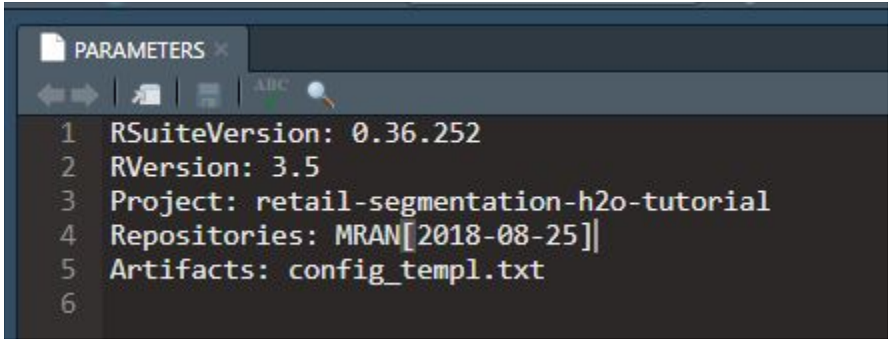
# Retail Segmentation with *h2o*

# Initial steps

- Create the project
- Add the packages
- Set the repository reference date
- Set the debug level
- Add the data folder and copy the raw data
- Quick test running *master.R*

# PARAMETERS

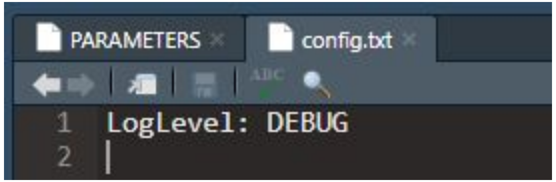
Change the MRAN date



```
RSuiteVersion: 0.36.252
RVersion: 3.5
Project: retail-segmentation-h2o-tutorial
Repositories: MRAN[2018-08-25]
Artifacts: config_templ.txt
```

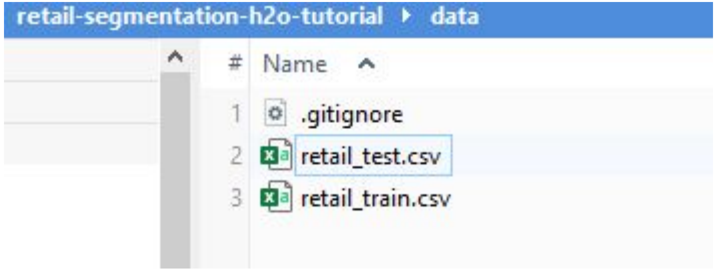
# config.txt

Set the log level to DEBUG



```
LogLevel: DEBUG

```

# Add data files

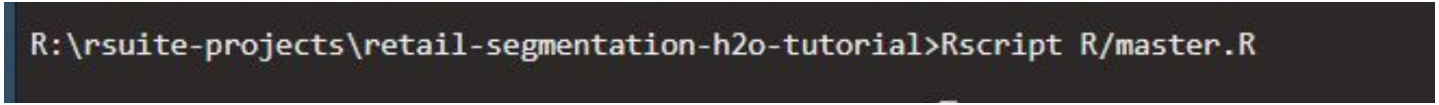


retail-segmentation-h2o-tutorial ▸ data

| # | Name |
|---|---|
| 1 | .gitignore |
| 2 | retail_test.csv |
| 3 | retail_train.csv |

# Run master.R

- We run the scripts from the terminal



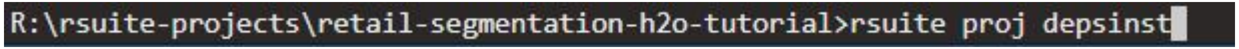
```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>Rscript R/master.R
```

# Next steps

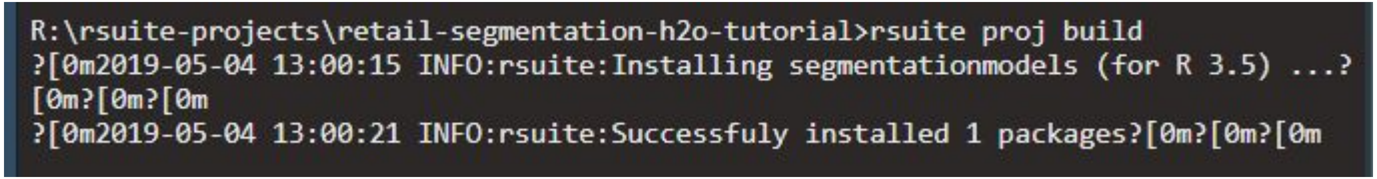
- Package
  - Add imports to `DESCRIPTION`
  - Add imports to `packages_import.R`
  - Add 1st script `find_best_model.R`
- Project
  - add 1st script `build_p2b_nosegmentation_model.R`
- Terminal
  - Install dependencies and build project
  - Run script

# terminal: Install and build

- Install dependencies
- Build project

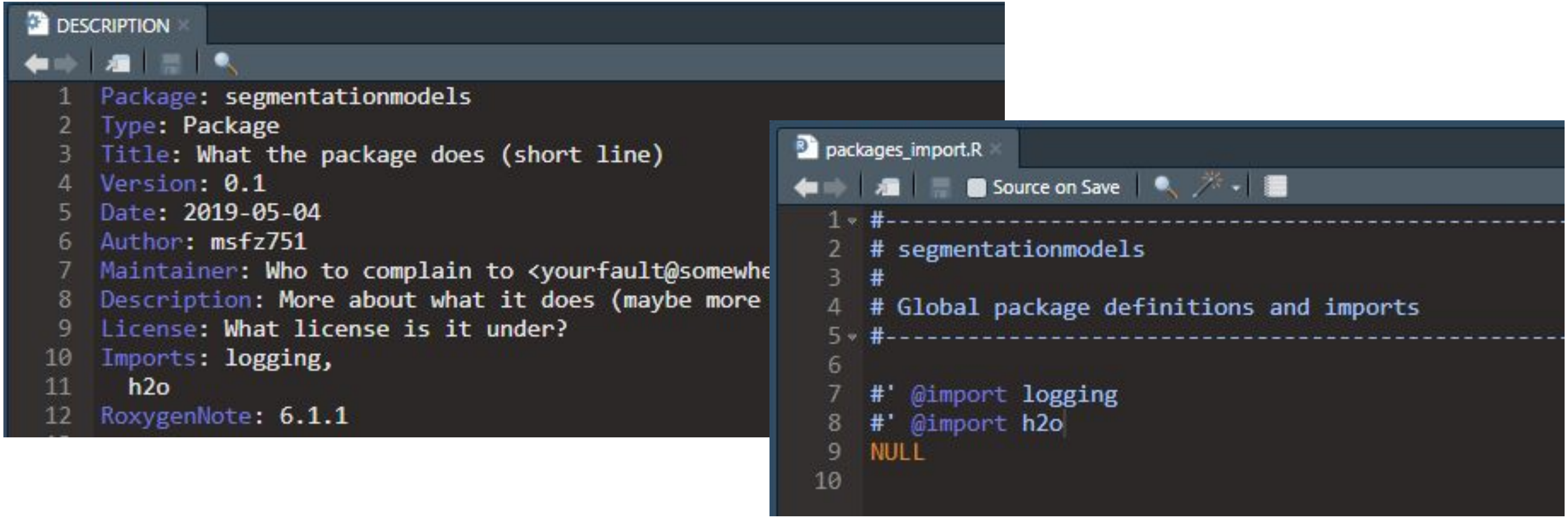


```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj depsinst
```



```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj build
?[0m2019-05-04 13:00:15 INFO:rsuite:Installing segmentationmodels (for R 3.5) ...?
[0m?[0m?[0m
?[0m2019-05-04 13:00:21 INFO:rsuite:Successfuly installed 1 packages?[0m?[0m?[0m
```

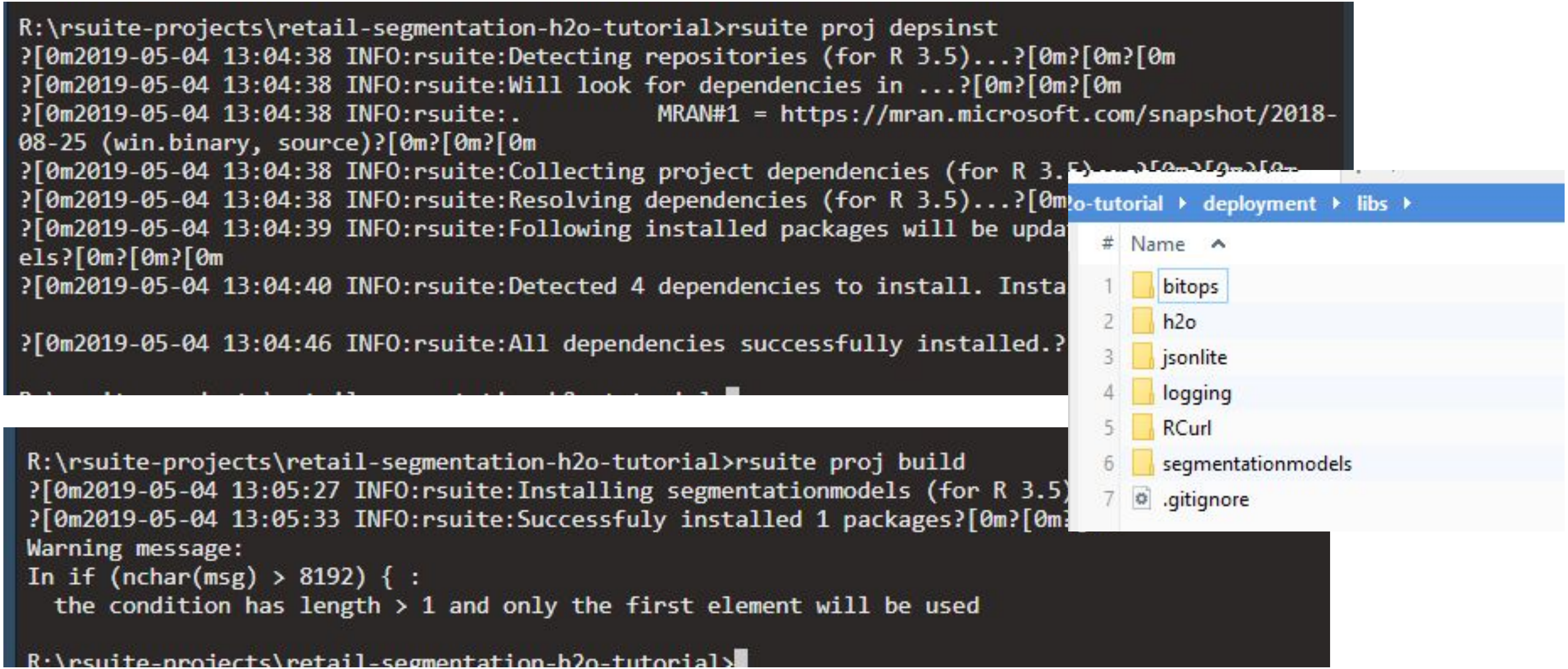
# Add package **h2o**



```
Package: segmentationmodels
Type: Package
Title: What the package does (short line)
Version: 0.1
Date: 2019-05-04
Author: msfz751
Maintainer: Who to complain to <yourfault@somewhe
Description: More about what it does (maybe more
License: What license is it under?
Imports: logging,
  h2o
RoxygenNote: 6.1.1
```

```
#--------------------------------------------------
# segmentationmodels
#
# Global package definitions and imports
#--------------------------------------------------

#' @import logging
#' @import h2o
NULL

```

# terminal: Install dependencies and build project



```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj depsinst
?[0m2019-05-04 13:04:38 INFO:rsuite:Detecting repositories (for R 3.5)...?[0m?[0m?[0m
?[0m2019-05-04 13:04:38 INFO:rsuite:Will look for dependencies in ...?[0m?[0m?[0m
?[0m2019-05-04 13:04:38 INFO:rsuite:.          MRAN#1 = https://mran.microsoft.com/snapshot/2018-
08-25 (win.binary, source)?[0m?[0m?[0m
?[0m2019-05-04 13:04:38 INFO:rsuite:Collecting project dependencies (for R 3.
?[0m2019-05-04 13:04:38 INFO:rsuite:Resolving dependencies (for R 3.5)...?[0m
?[0m2019-05-04 13:04:39 INFO:rsuite:Following installed packages will be upda
els?[0m?[0m?[0m
?[0m2019-05-04 13:04:40 INFO:rsuite:Detected 4 dependencies to install. Insta

?[0m2019-05-04 13:04:46 INFO:rsuite:All dependencies successfully installed.?

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```

```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj build
?[0m2019-05-04 13:05:27 INFO:rsuite:Installing segmentationmodels (for R 3.5)
?[0m2019-05-04 13:05:33 INFO:rsuite:Successfuly installed 1 packages?[0m?[0m?
Warning message:
In if (nchar(msg) > 8192) { :
  the condition has length > 1 and only the first element will be used

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```

o-tutorial ▸ deployment ▸ libs ▸

| # | Name |
|---|---|
| 1 | bitops |
| 2 | h2o |
| 3 | jsonlite |
| 4 | logging |
| 5 | RCurl |
| 6 | segmentationmodels |
| 7 | .gitignore |

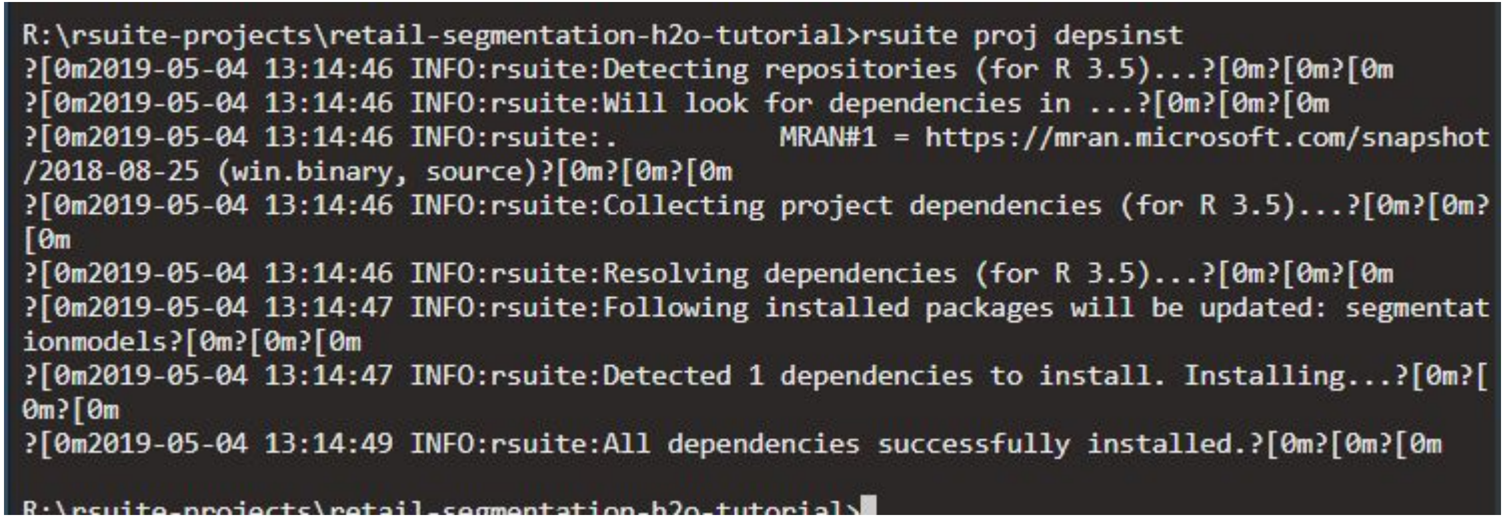
# project: Add the first script to the project

build_p2b_nosegmentation_model.R

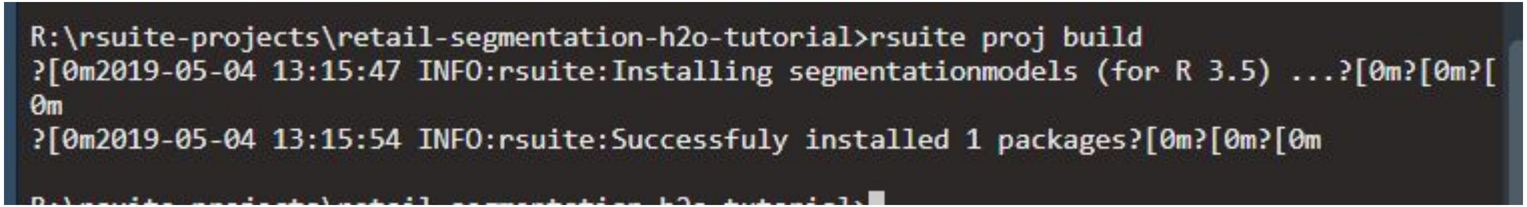
# package: Add 1st script - find best model

find_best_model.R

# terminal: Install dependencies and build project

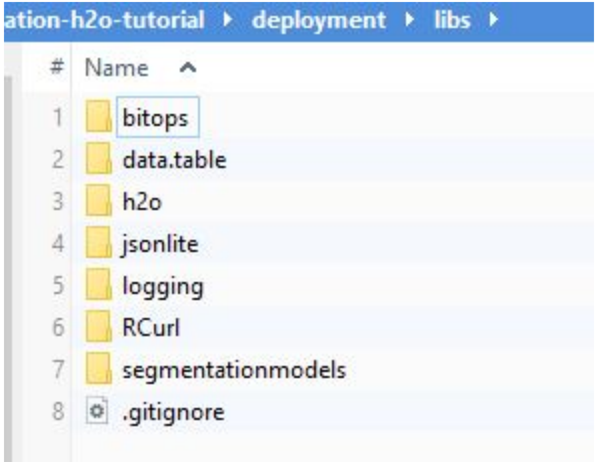


```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj depsinst
?[0m2019-05-04 13:14:46 INFO:rsuite:Detecting repositories (for R 3.5)...?[0m?[0m?[0m
?[0m2019-05-04 13:14:46 INFO:rsuite:Will look for dependencies in ...?[0m?[0m?[0m
?[0m2019-05-04 13:14:46 INFO:rsuite:.          MRAN#1 = https://mran.microsoft.com/snapshot
/2018-08-25 (win.binary, source)?[0m?[0m?[0m
?[0m2019-05-04 13:14:46 INFO:rsuite:Collecting project dependencies (for R 3.5)...?[0m?[0m?
[0m
?[0m2019-05-04 13:14:46 INFO:rsuite:Resolving dependencies (for R 3.5)...?[0m?[0m?[0m
?[0m2019-05-04 13:14:47 INFO:rsuite:Following installed packages will be updated: segmentat
ionmodels?[0m?[0m?[0m
?[0m2019-05-04 13:14:47 INFO:rsuite:Detected 1 dependencies to install. Installing...?[0m?[
0m?[0m
?[0m2019-05-04 13:14:49 INFO:rsuite:All dependencies successfully installed.?[0m?[0m?[0m

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```



```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj build
?[0m2019-05-04 13:15:47 INFO:rsuite:Installing segmentationmodels (for R 3.5) ...?[0m?[0m?[
0m
?[0m2019-05-04 13:15:54 INFO:rsuite:Successfuly installed 1 packages?[0m?[0m?[0m

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```

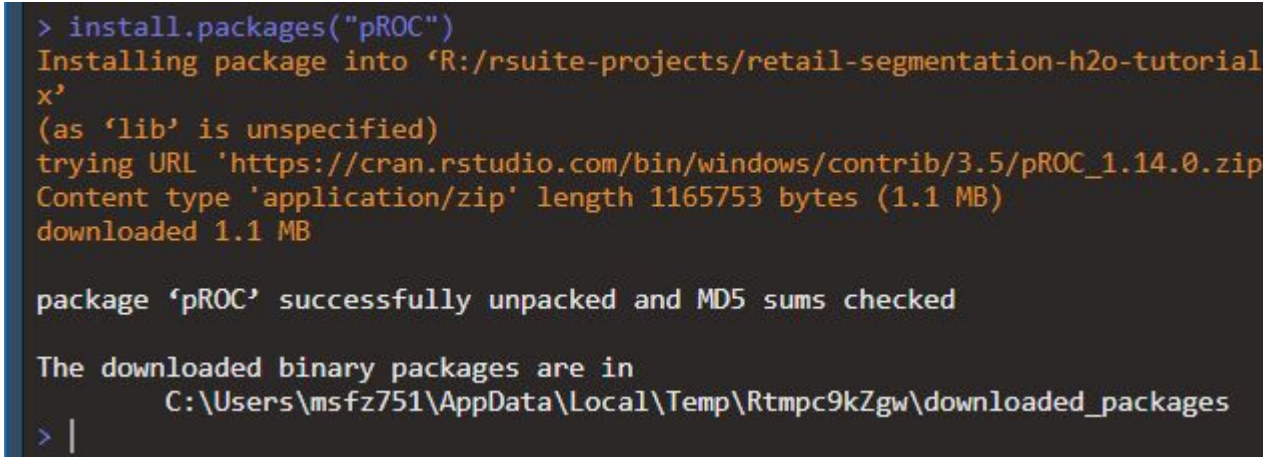
# new packages have been added



| # | Name |
|---|---|
| 1 | bitops |
| 2 | data.table |
| 3 | h2o |
| 4 | jsonlite |
| 5 | logging |
| 6 | RCurl |
| 7 | segmentationmodels |
| 8 | .gitignore |

# terminal: Run the *no-segmentation model*

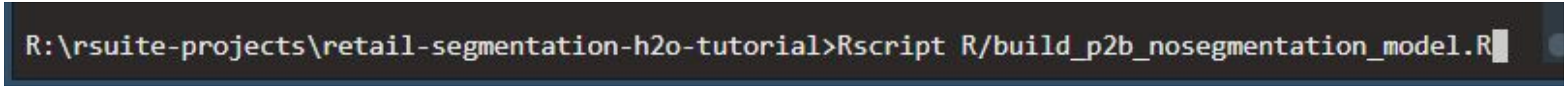
- It will run h2o but will stop on error because the package pROC is missing
- Install pROC outside the project (global) because pROC is called by the project not the package



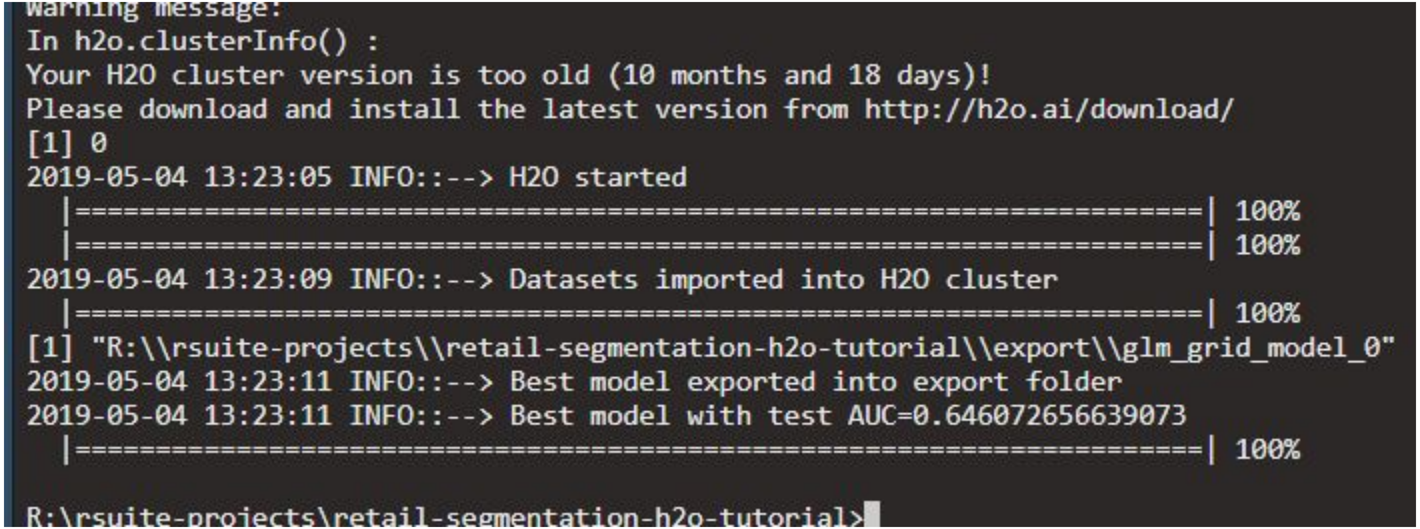
```
> install.packages("pROC")
Installing package into ‘R:/rsuite-projects/retail-segmentation-h2o-tutorial
x’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/pROC_1.14.0.zip
Content type 'application/zip' length 1165753 bytes (1.1 MB)
downloaded 1.1 MB

package ‘pROC’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\msfz751\AppData\Local\Temp\Rtmpc9kZgw\downloaded_packages
> |
```

# terminal: Run the same model again ...

- No it runs h2o and rPROC



```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>Rscript R/build_p2b_nosegmentation_model.R
```

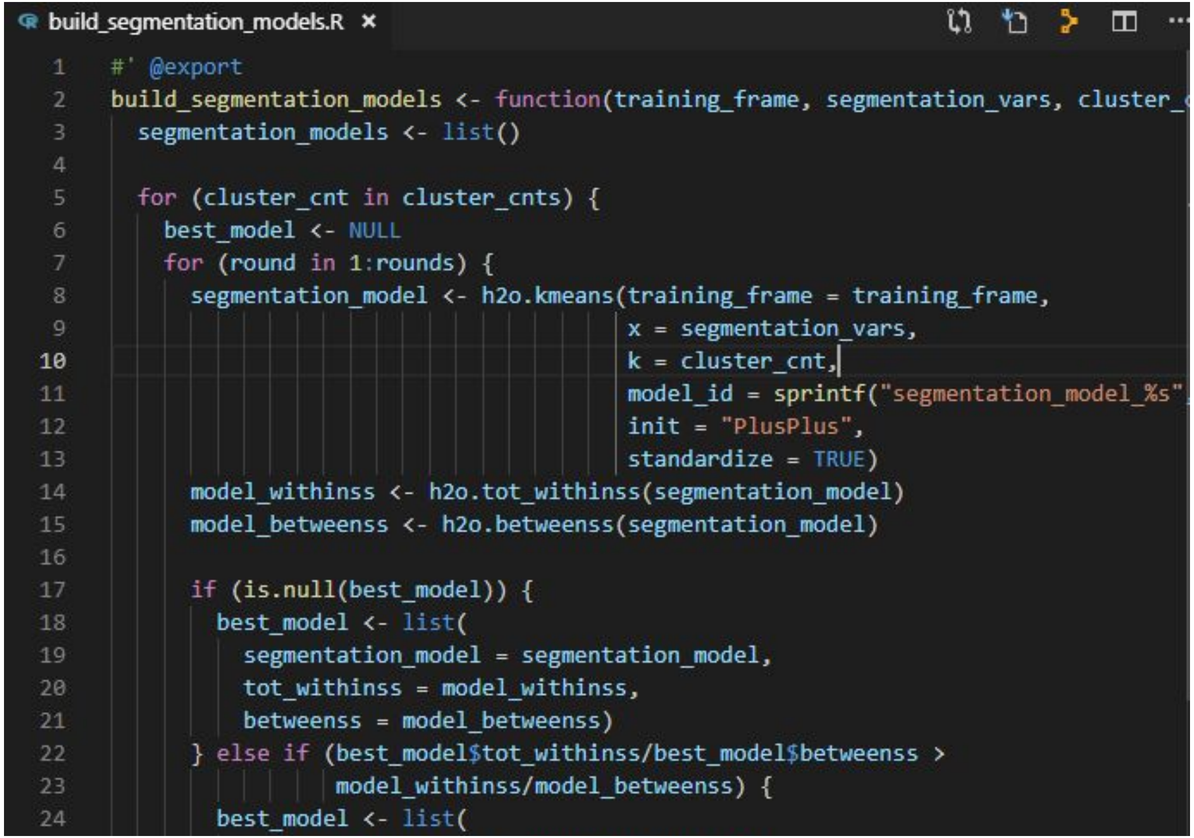


```
Warning message:
In h2o.clusterInfo() :
Your H2O cluster version is too old (10 months and 18 days)!
Please download and install the latest version from http://h2o.ai/download/
[1] 0
2019-05-04 13:23:05 INFO::--> H2O started
  |=====================================================================| 100%
  |=====================================================================| 100%
2019-05-04 13:23:09 INFO::--> Datasets imported into H2O cluster
  |=====================================================================| 100%
[1] "R:\\rsuite-projects\\retail-segmentation-h2o-tutorial\\export\\glm_grid_model_0"
2019-05-04 13:23:11 INFO::--> Best model exported into export folder
2019-05-04 13:23:11 INFO::--> Best model with test AUC=0.646072656639073
  |=====================================================================| 100%

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```

# Next steps

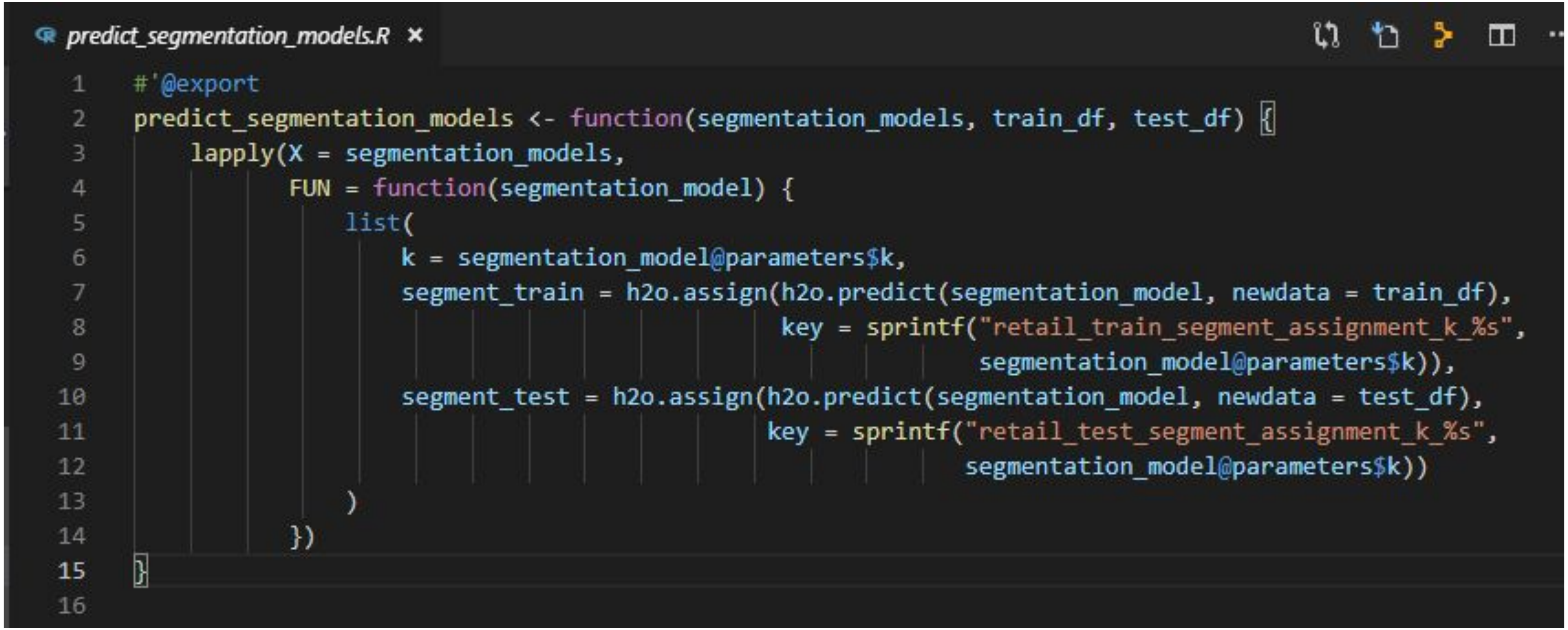
- Package
  - Add 2nd script `build_segmentation_models.R`
  - Add 3rd script `predict_segmentation_models.R`
- Project
  - add 2nd script `build_p2b_nosegmentation_local_models.R`
- Terminal
  - Install dependencies and build project
  - Run script `R/build_p2b_nosegmentation_local_models.R`

# package: Add the 2nd script - build models



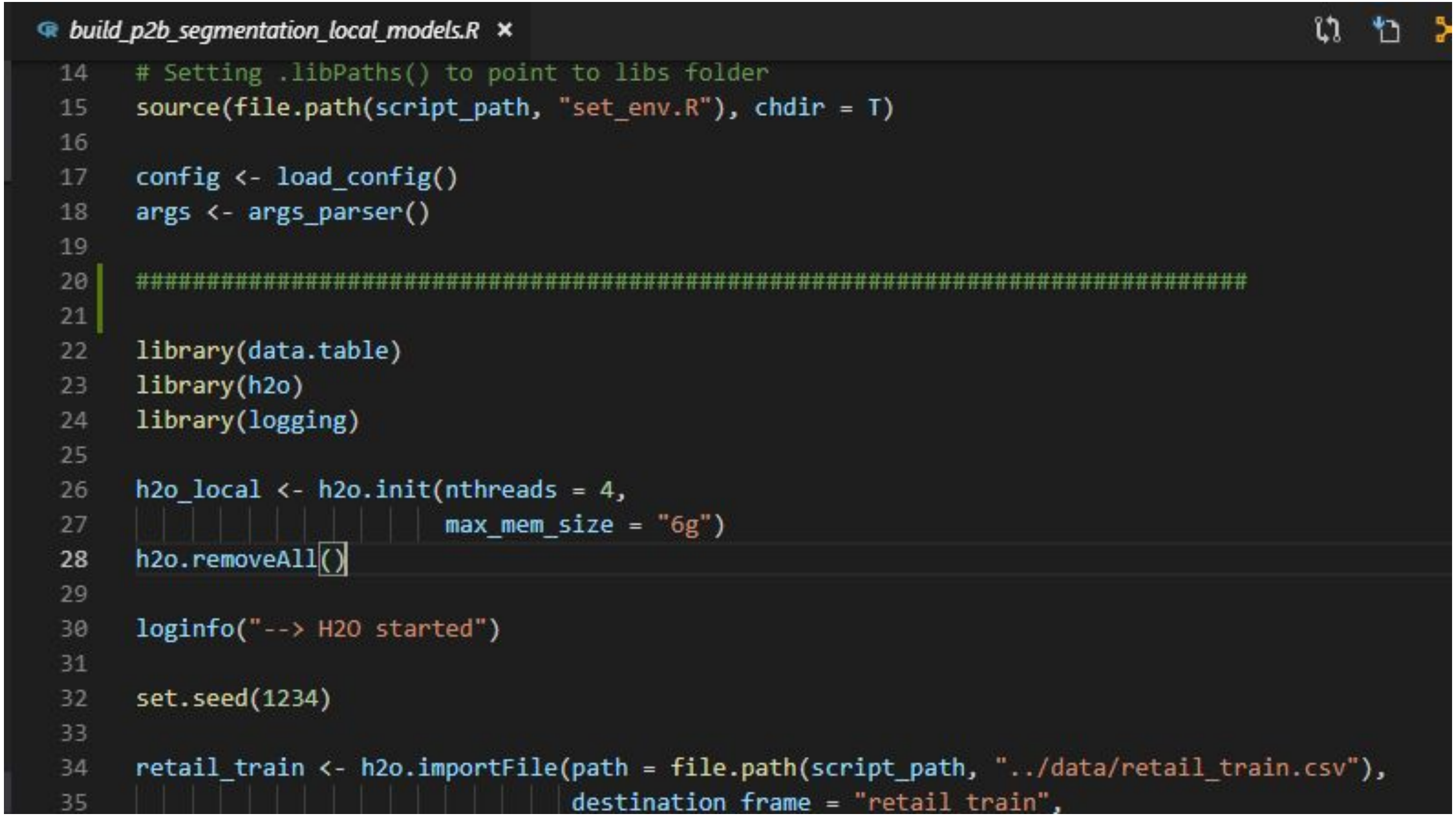
```r
#' @export
build_segmentation_models <- function(training_frame, segmentation_vars, cluster_c
  segmentation_models <- list()

  for (cluster_cnt in cluster_cnts) {
    best_model <- NULL
    for (round in 1:rounds) {
      segmentation_model <- h2o.kmeans(training_frame = training_frame,
                                       x = segmentation_vars,
                                       k = cluster_cnt,
                                       model_id = sprintf("segmentation_model_%s"
                                       init = "PlusPlus",
                                       standardize = TRUE)
      model_withinss <- h2o.tot_withinss(segmentation_model)
      model_betweenss <- h2o.betweenss(segmentation_model)

      if (is.null(best_model)) {
        best_model <- list(
          segmentation_model = segmentation_model,
          tot_withinss = model_withinss,
          betweenss = model_betweenss)
      } else if (best_model$tot_withinss/best_model$betweenss >
                 model_withinss/model_betweenss) {
        best_model <- list(
```

# package: Add the 3rd script - prediction



```r
#'@export
predict_segmentation_models <- function(segmentation_models, train_df, test_df) {
    lapply(X = segmentation_models,
           FUN = function(segmentation_model) {
               list(
                   k = segmentation_model@parameters$k,
                   segment_train = h2o.assign(h2o.predict(segmentation_model, newdata = train_df),
                                              key = sprintf("retail_train_segment_assignment_k_%s",
                                                            segmentation_model@parameters$k)),
                   segment_test = h2o.assign(h2o.predict(segmentation_model, newdata = test_df),
                                             key = sprintf("retail_test_segment_assignment_k_%s",
                                                           segmentation_model@parameters$k))
               )
           })
}
```

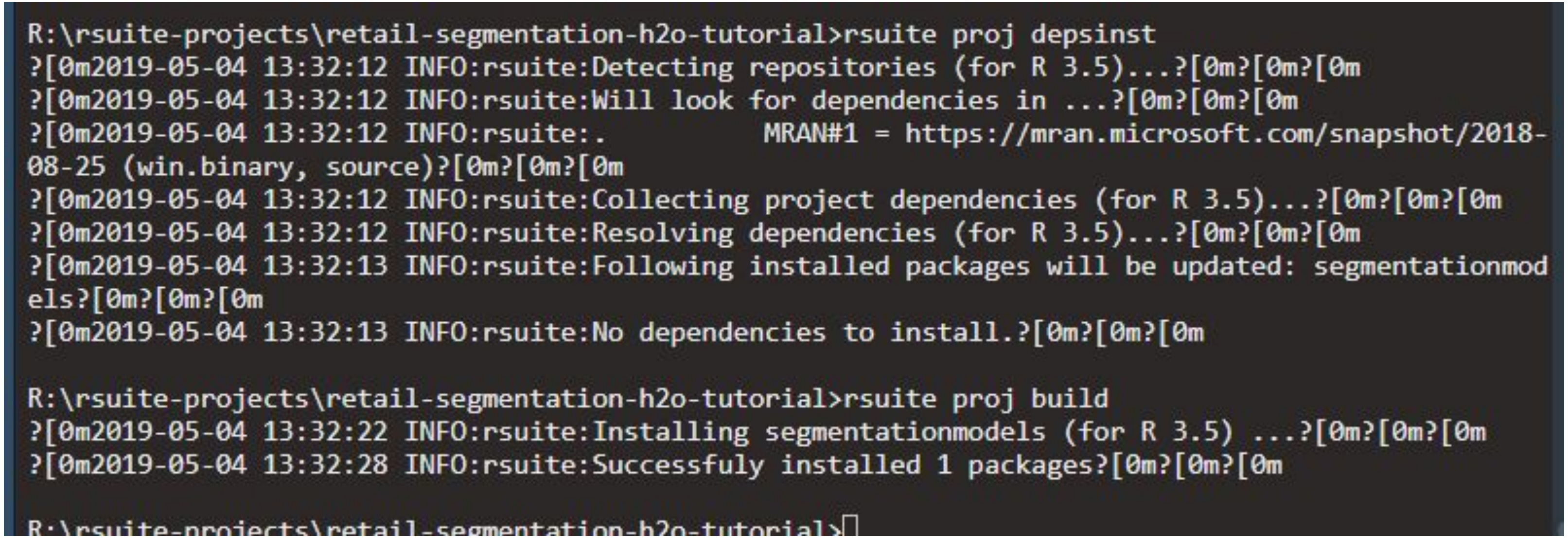
# project: Add script for segmentation of local models



```r
# Setting .libPaths() to point to libs folder
source(file.path(script_path, "set_env.R"), chdir = T)

config <- load_config()
args <- args_parser()

################################################################################

library(data.table)
library(h2o)
library(logging)

h2o_local <- h2o.init(nthreads = 4,
                      max_mem_size = "6g")
h2o.removeAll()

loginfo("--> H2O started")

set.seed(1234)

retail_train <- h2o.importFile(path = file.path(script_path, "../data/retail_train.csv"),
                               destination_frame = "retail_train",
```

# terminal: install dependencies and build project

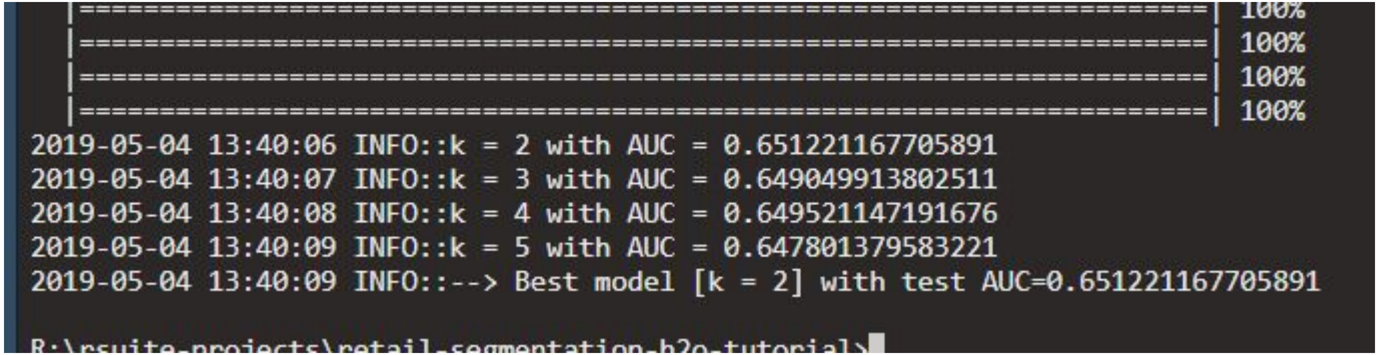
- the new package scripts will be added



```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj depsinst
?[0m2019-05-04 13:32:12 INFO:rsuite:Detecting repositories (for R 3.5)...?[0m?[0m?[0m
?[0m2019-05-04 13:32:12 INFO:rsuite:Will look for dependencies in ...?[0m?[0m?[0m
?[0m2019-05-04 13:32:12 INFO:rsuite:.          MRAN#1 = https://mran.microsoft.com/snapshot/2018-
08-25 (win.binary, source)?[0m?[0m?[0m
?[0m2019-05-04 13:32:12 INFO:rsuite:Collecting project dependencies (for R 3.5)...?[0m?[0m?[0m
?[0m2019-05-04 13:32:12 INFO:rsuite:Resolving dependencies (for R 3.5)...?[0m?[0m?[0m
?[0m2019-05-04 13:32:13 INFO:rsuite:Following installed packages will be updated: segmentationmod
els?[0m?[0m?[0m
?[0m2019-05-04 13:32:13 INFO:rsuite:No dependencies to install.?[0m?[0m?[0m

R:\rsuite-projects\retail-segmentation-h2o-tutorial>rsuite proj build
?[0m2019-05-04 13:32:22 INFO:rsuite:Installing segmentationmodels (for R 3.5) ...?[0m?[0m?[0m
?[0m2019-05-04 13:32:28 INFO:rsuite:Successfuly installed 1 packages?[0m?[0m?[0m

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```

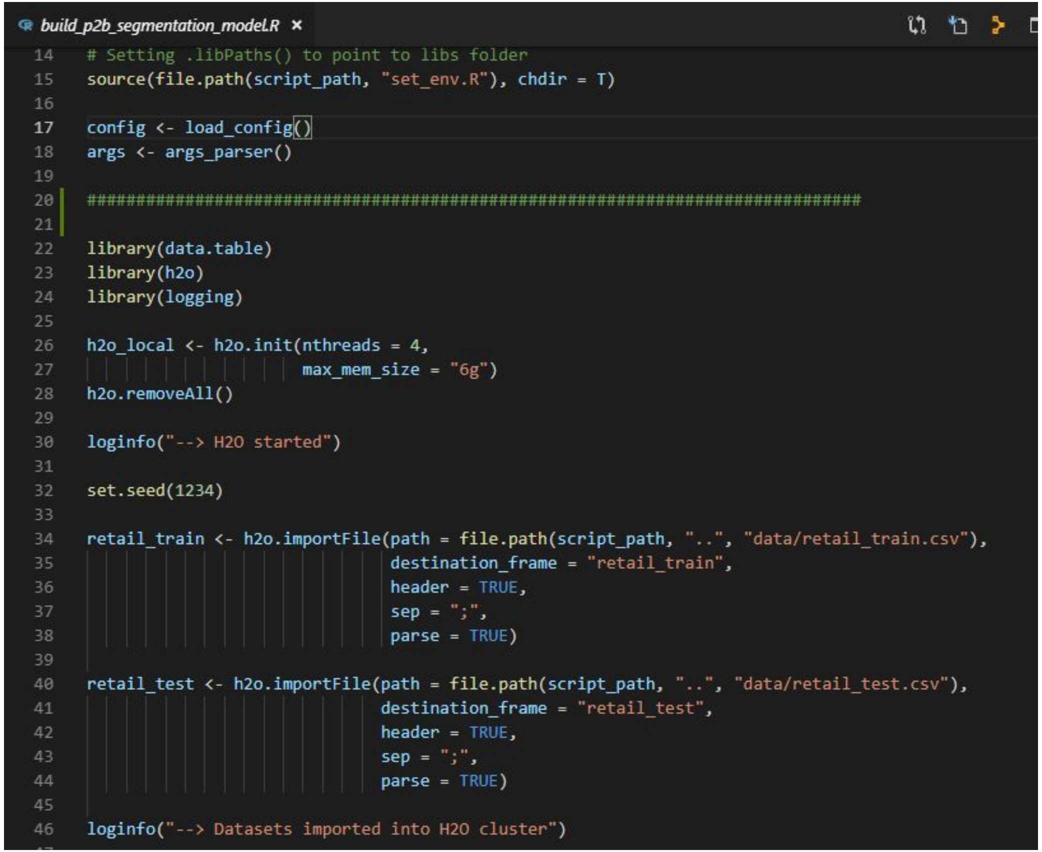
# terminal: Run segmentation for local models

- Results from h2o and pROC



```
  |=======================================================================| 100%
  |=======================================================================| 100%
  |=======================================================================| 100%
  |=======================================================================| 100%
2019-05-04 13:40:06 INFO::k = 2 with AUC = 0.651221167705891
2019-05-04 13:40:07 INFO::k = 3 with AUC = 0.649049913802511
2019-05-04 13:40:08 INFO::k = 4 with AUC = 0.649521147191676
2019-05-04 13:40:09 INFO::k = 5 with AUC = 0.647801379583221
2019-05-04 13:40:09 INFO::--> Best model [k = 2] with test AUC=0.651221167705891

R:\rsuite-projects\retail-segmentation-h2o-tutorial>
```

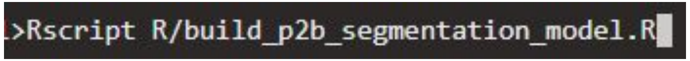
# project: Add script for segmentation of *standard* models



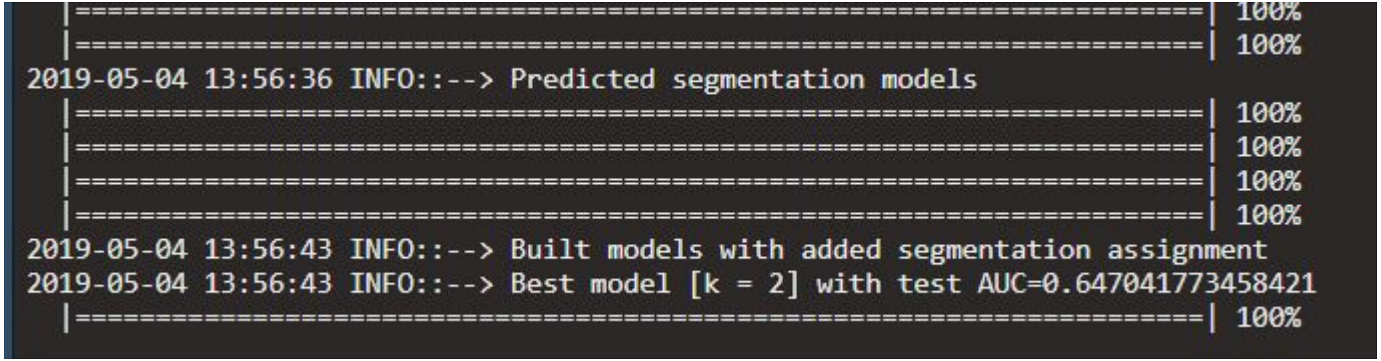
```r
# Setting .libPaths() to point to libs folder
source(file.path(script_path, "set_env.R"), chdir = T)

config <- load_config()
args <- args_parser()

###############################################################################

library(data.table)
library(h2o)
library(logging)

h2o_local <- h2o.init(nthreads = 4,
                      max_mem_size = "6g")
h2o.removeAll()

loginfo("--> H2O started")

set.seed(1234)

retail_train <- h2o.importFile(path = file.path(script_path, "..", "data/retail_train.csv"),
                               destination_frame = "retail_train",
                               header = TRUE,
                               sep = ";",
                               parse = TRUE)

retail_test <- h2o.importFile(path = file.path(script_path, "..", "data/retail_test.csv"),
                              destination_frame = "retail_test",
                              header = TRUE,
                              sep = ";",
                              parse = TRUE)

loginfo("--> Datasets imported into H2O cluster")
```

# terminal: Run the standard models

- In this case, we don’t need to install dependencies or build the project because we didn’t make any changes on the package, only on the project.



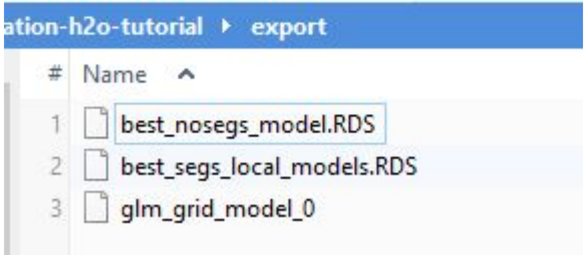
```
>Rscript R/build_p2b_segmentation_model.R
```



```
  |======================================================================| 100%
  |======================================================================| 100%
2019-05-04 13:56:36 INFO::--> Predicted segmentation models
  |======================================================================| 100%
  |======================================================================| 100%
  |======================================================================| 100%
  |======================================================================| 100%
2019-05-04 13:56:43 INFO::--> Built models with added segmentation assignment
2019-05-04 13:56:43 INFO::--> Best model [k = 2] with test AUC=0.647041773458421
  |======================================================================| 100%
```

# Observe an `export` folder has been created

- These are the outputs of the scripts



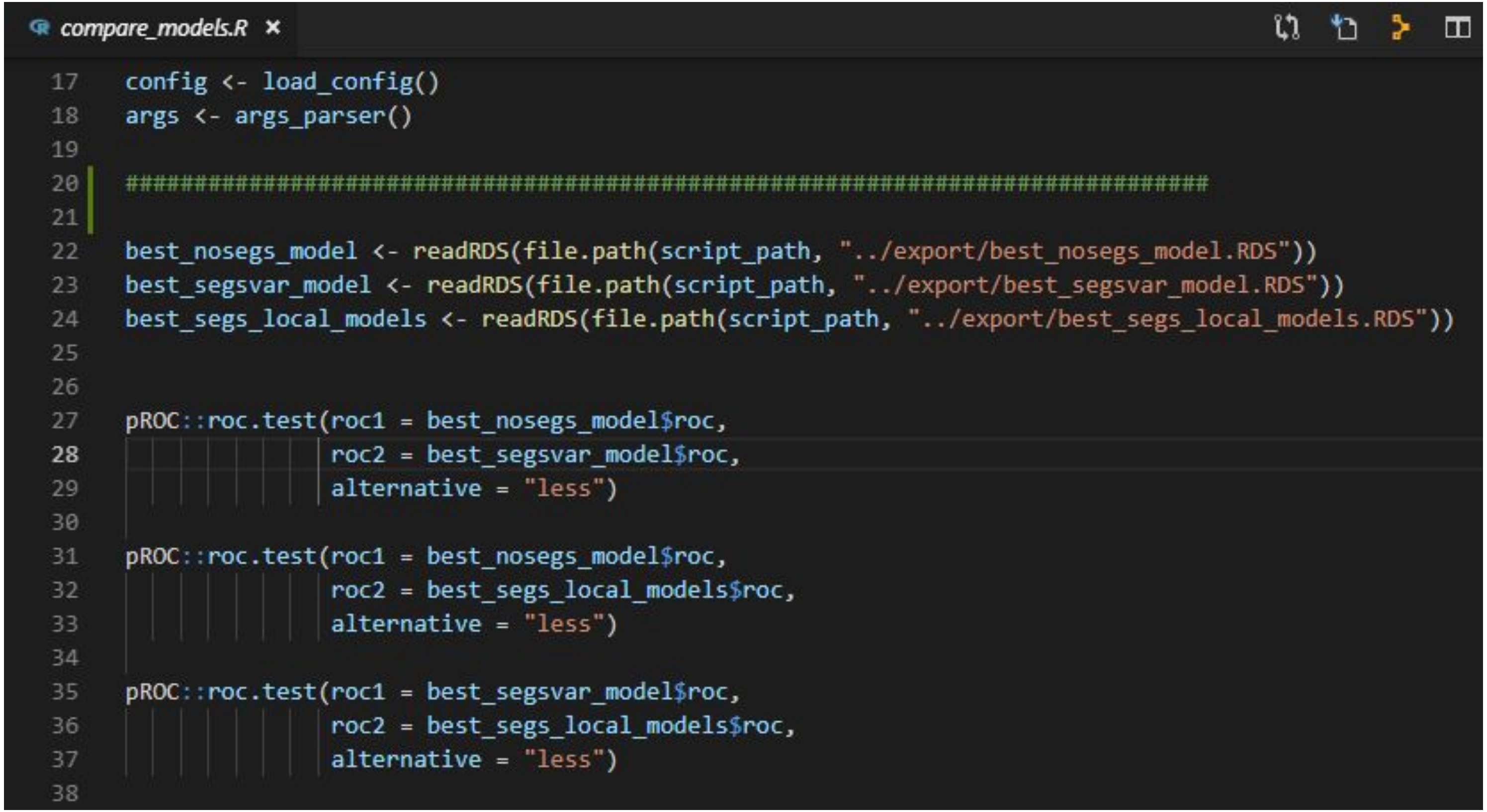
ation-h2o-tutorial ▸ export

| # | Name |
|---|---|
| 1 | best_nosegs_model.RDS |
| 2 | best_segs_local_models.RDS |
| 3 | glm_grid_model_0 |

# Next steps

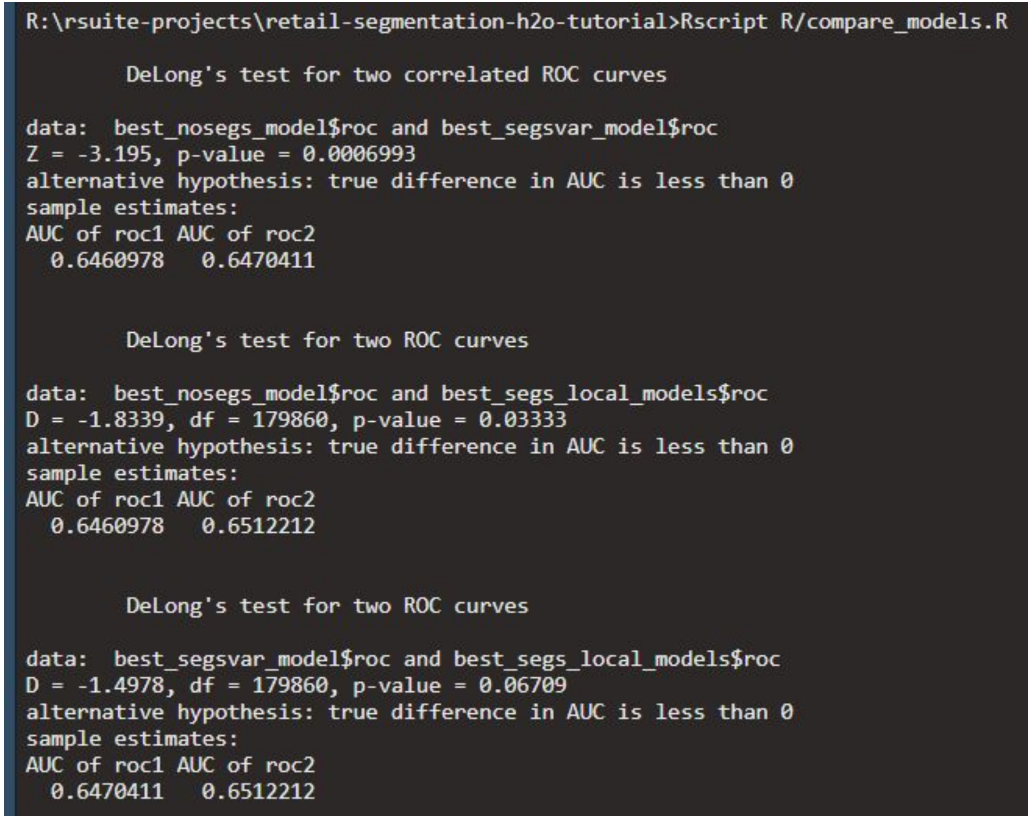
- Package
  - None
- Project
  - add 4th script `compare_models.R`
- Terminal
  - Run script `R/compare_models.R`

# project: add the last script for model comparison



```r
config <- load_config()
args <- args_parser()

###############################################################################

best_nosegs_model <- readRDS(file.path(script_path, "../export/best_nosegs_model.RDS"))
best_segsvar_model <- readRDS(file.path(script_path, "../export/best_segsvar_model.RDS"))
best_segs_local_models <- readRDS(file.path(script_path, "../export/best_segs_local_models.RDS"))


pROC::roc.test(roc1 = best_nosegs_model$roc,
               roc2 = best_segsvar_model$roc,
               alternative = "less")

pROC::roc.test(roc1 = best_nosegs_model$roc,
               roc2 = best_segs_local_models$roc,
               alternative = "less")

pROC::roc.test(roc1 = best_segsvar_model$roc,
               roc2 = best_segs_local_models$roc,
               alternative = "less")

```

# terminal: Run model comparison



```
R:\rsuite-projects\retail-segmentation-h2o-tutorial>Rscript R/compare_models.R

        DeLong's test for two correlated ROC curves

data:  best_nosegs_model$roc and best_segsvar_model$roc
Z = -3.195, p-value = 0.0006993
alternative hypothesis: true difference in AUC is less than 0
sample estimates:
AUC of roc1 AUC of roc2
  0.6460978   0.6470411


        DeLong's test for two ROC curves

data:  best_nosegs_model$roc and best_segs_local_models$roc
D = -1.8339, df = 179860, p-value = 0.03333
alternative hypothesis: true difference in AUC is less than 0
sample estimates:
AUC of roc1 AUC of roc2
  0.6460978   0.6512212


        DeLong's test for two ROC curves

data:  best_segsvar_model$roc and best_segs_local_models$roc
D = -1.4978, df = 179860, p-value = 0.06709
alternative hypothesis: true difference in AUC is less than 0
sample estimates:
AUC of roc1 AUC of roc2
  0.6470411   0.6512212
```