Showcase: on segmentation importance for marketing campaign in retail using R and H2O

Wit Jakuczun, WLOG Solutions

Introduction

Agenda

- "At the corner" business case
- What is segmentation?
- How to build (optimal) segmentation models in H2O?
- How to combine segmentation and predictive modelling?
- Summary

Who am I?

- Job
  - owner of company WLOG Solutions
- Education
  - Mathematician (Warsaw University)
- My expertise:
  - Solving business problems with analytical solutions
  - Implementing and delivering optimization and predictive models
- Contact details:
  - email: w.jakuczun@wlogsolutions.com
  - WWW: www.wlogsolutions.com

If you want to follow me…

I have used:

- R version 3.3.2
- H2O version 3.10.0.8

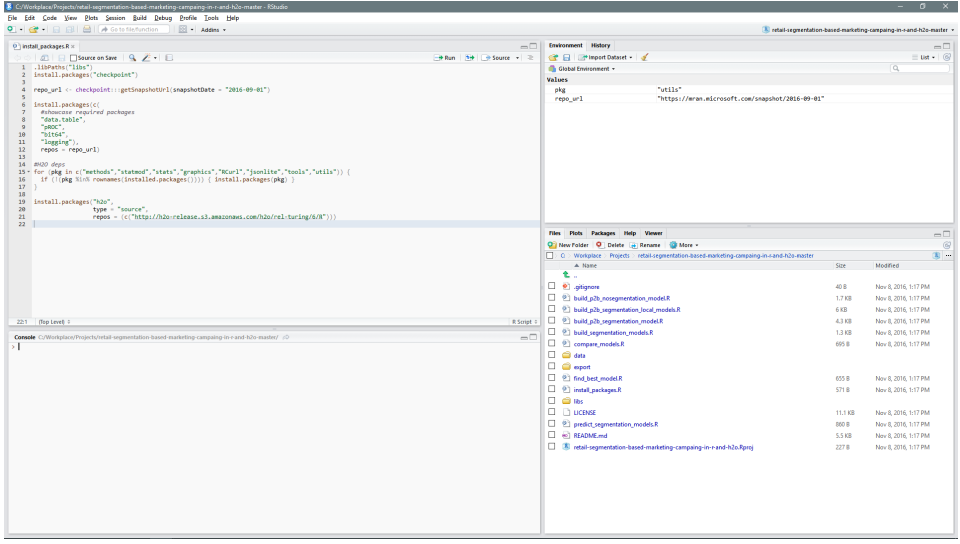
Code available at WLOG's github space

- Direct link

First step

- Download repository using link and extract into any folder
- Open `retail-segmentation-based-marketing-campaign-in-r-and-h2o.Rproj` in RStudio

The result


C:/Workplace/Projects/retail-segmentation-based-marketing-campaing-in-r-and-h2o-master - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
retail-segmentation-based-marketing-campaing-in-r-and-h2o-master
install_packages.R
Source on Save
Run
Source
.libPaths("libs")
install.packages("checkpoint")
repo_url <- checkpoint:::getSnapshotUrl(snapshotDate = "2016-09-01")
install.packages(c(
"h2oconsole required packages
"data.table",
"pROC",
"bitops",
"logging",
repos = repo_url)
#H2O deps
for (pkg in c("methods","statmod","stats","graphics","RCurl","jsonlite","tools","utils")) {
if (!(pkg %in% rownames(installed.packages()))) { install.packages(pkg) }
}
install.packages("h2o",
type = "source",
repos = (c("http://h2o-release.s3.amazonaws.com/h2o/rel-turing/6/R")))
22:1 (Top Level)
R Script
Console C:/Workplace/Projects/retail-segmentation-based-marketing-campaing-in-r-and-h2o-master/
Environment History
Import Dataset
List
Global Environment
Values
pkg "utils"
repo_url "https://mran.microsoft.com/snapshot/2016-09-01"
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Workplace Projects retail-segmentation-based-marketing-campaing-in-r-and-h2o-master
Name Size Modified
.gitignore 40 B Nov 8, 2016, 1:17 PM
build_p2b_nosegmentation_model.R 1.7 KB Nov 8, 2016, 1:17 PM
build_p2b_segmentation_local_models.R 6 KB Nov 8, 2016, 1:17 PM
build_p2b_segmentation_model.R 4.3 KB Nov 8, 2016, 1:17 PM
build_segmentation_models.R 1.3 KB Nov 8, 2016, 1:17 PM
compare_models.R 695 B Nov 8, 2016, 1:17 PM
data
export
find_best_model.R 655 B Nov 8, 2016, 1:17 PM
install_packages.R 571 B Nov 8, 2016, 1:17 PM
libs
LICENSE 11.1 KB Nov 8, 2016, 1:17 PM
predict_segmentation_models.R 860 B Nov 8, 2016, 1:17 PM
README.md 5.5 KB Nov 8, 2016, 1:17 PM
retail-segmentation-based-marketing-campaing-in-r-and-h2o.Rproj 227 B Nov 8, 2016, 1:17 PM

Packages installation

```
source("install_packages.R")
```

Important: all installation is done locally in `libs` folder. Your R environment is not messed up!

At the corner's business case

What is *At the corner*?

At the corner *is an* ***analytical driven*** *retail chain selling a wide variety of products.*

What is *At the corner*'s business challenge?

At the corner *would like to introduce a new product.*

What is their business approach?

At the corner *decided to go with an e-mail marketing campaign. To optimize campaign costs and customers' comfort they decided to carefully select customers that would be contacted in the campaign.*

What has already been done?

- Conducted a pilot campaign and gathered customers responses
- Analytical table has been prepared

What is in the analytical table?

```
data_train <- fread("data/retail_train.csv")
colnames(data_train)
```

We have following categories of variables:

- marketing - did we contact a customer before?
- purchased - did the customer purchase in a result of pilot campaign
- demographic: sex, age, income,
- behavioural - what is customer's buying pattern?
  - basket of products
  - basket value
  - purchases in a nearest shop
  - mean distance to shops

What is our goal?

Score customers in `data/retail_test.csv`.

Our approach

We will build three types of models:

- (M1) Logistic regression with variables from analytical table.
- (M2) Logistic regression with variables from analytical table and a variable from a segmentation model based on behavioural variables.
- (M3) Local logistic regression models with variables from analytical table for segments calculated by segmentation model based on behavioural variables.

We will select the best one using *AUC* measure.

First model - no segmentation

Most important parts:

- `source("build_p2b_nosegmentation_model.R")`
- What is interesting?
  - `?h2o.grid` - model meta parameters fitting
  - `find_best_model.R` - find best model according to **AUC** measure
  - `?h2o.auc` - internal H2O function for calculating **AUC**
  - `pROC` package
    - `?pROC::roc` - ROC curve calculation
    - `?pROC::auc` - AUC measure for ROC curve calculation

First model - results

- Our baseline is `AUC = 0.6460`
- And what does Flow say?
  - `http://localhost:54321/flow/index.html`

What is segmentation?

WLOG SOLUTIONS

What is your definition?

Naive definition

Unsupervised approach for discovering groups of similar objects according to some distance/similarity measure.

My (our) definition

Discovering latent variables, that are strongly non-linear transformations of the input space. The transformation, being based on metric on input space, are too difficult for standard supervised algorithms to be discovered.

Why is segmentation difficult?

- Business perspective
  - It is almost impossible to formalize requirements for being good segmentation in general.
  - But **it is possible** (next slides) to formalize requirement for being good segmentation in predictive modelling.
- Technical perspective
  - Final segments depends on both variables and **the distance**.
  - Number of segments is unknown and must be calculated from data or given by the oracle.

Things to be considered

- Popular algorithms (like *kmeans*) are randomized
  - Repeat segmentation *N* times.
  - Select best segments using e.g. *within sum of squares metric*
- and iterative
  - Give enough number of iterations to be sure the algorithm has converged.
- Sometimes segment centres cannot be a mean
  - Can use more expensive medoid approaches

How to measure goodness of segmentation methods?

- A very informative method is silhouette
  - Only useful if we have the same distance.
    - For example choosing number of segments.
- But we are in predictive modelling
  - **Use predictive power of the final models!**

What is good segmentation for predictive model?

Good segmentation is a segmentation that significantly improves predictive model quality measure (e.g. AUC).

How to build segmentation models in H2O and R?

What is available?

- H2O provides *k-means* algorithm
- Tutorial is here

Let's analyse the code (1)

- Check `build_segmentation_models.R`
  - For given range of segments `cluster_cnts`
  - Generate `rounds` segmentations and select the best one

Let's analyse the code (2)

Main part - fitting the model:

```
segmentation_model <- h2o.kmeans(
    training_frame = training_frame,
    x = segmentation_vars,
    k = cluster_cnt,
    model_id = sprintf("segmentation_model_%s", cluster_cnt),
    init = "PlusPlus",
    standardize = TRUE)
```

Let's analyse the code (2)

And scoring segmentation model (check file `predict_segmentation_models.R`)

```
h2o.predict(segmentation_model, newdata = train_df)
```

How to combine segmentation and predictive modelling?

Two approaches

- Use segments assignment as another predictor.
- Build local models for segments.

Segment as another predictor

- Check `build_p2b_segmentation_model.R`
- Most important parts:
  - Lines 45-49: building segmentation models
  - Lines 53-55: predict segmentation models
  - Lines 59-77: build predictive models with segments
    - Lines 61-62: assign segments to customers
    - Lines 64-74: select best model for given number of segments
  - Lines 81-93: select best model

Segment as another predictor - results

- Best number of segments is 2.
- We obtained `AUC = 0.6470`

Local models for segments

- Check `build_p2b_segmentation_local_models.R`
- Most important parts:
  - Lines 48-52: building segmentation models
  - Lines 55-57: predict segmentation models
  - Lines 61-83: build local models for segments
    - Lines 63: assign segments to customers
    - Lines 67-79: build models for segments for different number of segments
  - Lines 87-105: predict local models for test data
  - Lines 107-132: select best models

Local models for segments - results

- Best number of segments is 2.
- We obtained `AUC = 0.6512`

Summary

Summary of results

- No segmentation was worst with `AUC = 0.6460`
- Segmentation as a predictor was second best with `AUC = 0.6470`
- Local models were best with `AUC = 0.6512`

Are the differences significant?

- Check `compare_models.R`
- Most important parts
  - One can compare significance differences for ROC curves
    - We used DeLong's test
- Conclusions
  - Adding segmentation as predictor is significant.
  - Local models give significant improvement to segment as predictor.

WLOG SOLUTIONS

Thank you for you attention!