CE888: Data Science and Decision Making

Lab 3: Modelling

Ana Matran-Fernandez

28 January 2019

Institute for Analytics and Data Science
University of Essex

Table of contents

Setting up

CE888 repository

- If you haven't done it yet, send me (amatra@essex.ac.uk) an email NOW with subject "CE888 github" and your GitHub username. (e.g., "CE888 github amatra").
- You don't need to email me when you finish the practice.
- If you have changed anything in your repository since the last time you were in this computer, make sure you do: `git pull` from the repository folder.
- This will download all the changes you did into your local folder.

Downloading the lab 3 materials

- ☐ Go to the Moodle page for this week:
- ☐ `https://moodle.essex.ac.uk/course/view.php?id=6683&section=9`
- ☐ Download the slides and code for today's practice into your local Github directory (e.g., `/labs/lab3`).
- ☐ Unzip the code, commit and push it before you make any changes.
- ☐ Start with the instructions from `project_start.pdf`
- ☐ This might take a while. In the meantime, come and get a USB drive!

Regression

Lab structure

- Inside `lab3` you will see three ipython notebooks
- Open them and see what is inside:
  - facebook_regression.ipynb
  - facebook_classification.ipynb
  - Breast_Cancer_dataset.ipynb
- Start with `Breast_Cancer_dataset.ipynb` to get used to `scikit-learn`.
- Push your changes to GitHub!
- After this, you will create your own notebook and work on a new dataset (see next slide).

Modelling

Bank Marketing dataset

- ☐ Create a new ipython notebook
- ☐ Check the dataset
  - ☐ `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`: here you have a description of each attribute.
  - ☐ It's saved as `bank-additional-full.csv` in your lab directory

Your tasks

In the Ipython notebook you created

- ☐ Load the data from `bank-additional-full.csv`
- ☐ Use a classifier (anything, but `ExtraTreesClassifier` with 100 estimators is the easiest option) on the data with outcome/output variable `y`
  - ☐ Convert to dummies using `df_dummies = pd.get_dummies(df)`
  - ☐ Columns `y_no` and `duration` must be deleted — use something like del df_copy[”attribute”] for this
  - ☐ Plot histogram of the label `y_yes`
  - ☐ Get the values and run a classifier (with outcome `y_yes`)
  - ☐ Report the results of 10-Kfold stratified cross-validation
  - ☐ Get feature importances and a confusion matrix
- ☐ Once you are done, save your changes in github