BA222 Project 3

Prof. Leder-Luis

Due: Thursday, Dec 7, 12PM

Regression Prediction Problem

In this project, we will work with data from a bank that contacted customers trying to get them to subscribe to a "term deposit", i.e. to put in money into an account for a fixed length of time.

The goal of this project is to build the best regression model to predict whether a customer subscribed to the term deposit. However, there are 21 variables to this dataset, and so you must choose which variables to include. There are more than 500,000 possible models in this dataset; you cannot try all of them manually.

A good model will have the following properties:

- Reasonable, well-motivated variables
- Good statistical significance on at least some included variables
- A strong R^2
- A strong Out of Sample R^2 (to be discussed in class)

Steps you must take

- Download the zip file bankdata.zip
- **Carefully read** bankdata_dictionary.txt. Please note that this specifies the "target" variable you are trying to predict, and also which variable(s) are inappropriate to include.
- Run your regression models on bankdata_training.csv
- Compute out of sample R^2 using bankdata_full.csv

See next page

Teams: Please work in teams of 4-5. Submit your teams by Friday, Nov 25 here: `https://forms.gle/DnMS9sFTnbBzZ6WJ8`

You may work with anyone, including people you worked with during previous projects. Email Nicholas Antonelli (nantonel@bu.edu) immediately if you need help finding a partner.

Deliverables – you will turn in the following:

- A PDF with **one page** of text describing the variables in your final model, and why you chose to include them
- Within that one page report, please state your fitted equation with coefficients.
- Within that one page report, please state your R^2 and out of sample R^2
- As part of the same PDF, include a screenshot of your regression output using that model as fit to the training data (not counted toward page length)
- Please include a paragraph at the end of the report (not counted toward page length) describing how each team member participated.
- A jupyter notebook that reproduces your results, and only includes the final regression and computations (and does not include other models you tested)