Predictive task description

You've been asked to help the marketing team in increasing performance of their marketing campaign. The team wants to laverage Advanced Analytics to improve campaign targeting. In other words they want to identify customers for which the gain from being contacted is the highest. To establish a proof of concept they provided data from previous campaign for both control and targeted groups, which were selected at random from non-users before the campaign start. Aim of the campaign was to persuade customers to subscribe to the term deposit. The product (term deposit) was available also to the control group (but not marketed).

Assuming the total campaign budget is fixed calculate expected lift from using model predictions vs random selection (as done before).

We would like you to laverage relevant analytical methods to develop proof of concept for the next wave of the campaign (same message) and it's expected lift. For results robustness please propose appropriate evaluation and validation framework.

Dataset

Please find the dataset here: link (https://empik-my.sharepoint.com/:f:/p/kmarmolowska/EoqAeRiB0-tFn1qWqeMX8j0BArtjUJGC_z29R_WliNhsUA?e=wmgV51) (password: DataScience42)

The dataset contains information about direct marketing campaign (phone calls) of a Portuguese banking institution.

Disclamer: Dataset is based on publicly available Bank Marketing dataset (link (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#))

Attributes

Clients' data

- **age** (numeric)
- **job:** type of job (categorical: "admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
- **marital:** marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
- **education** (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school","illiterate", "professional.course", "university.degree", "unknown")
- **default:** has credit in default? (categorical: "no","yes","unknown")
- **housing:** has housing loan? (categorical: "no","yes","unknown")
- **loan:** has personal loan? (categorical: "no","yes","unknown")

Target data

- **test_control_flag**: contains information if the person was part of the campaign ("campaign group") or control group ("control group"). Campaign group was called (details on the contact below) and offered term deposit. All customers could subscribe to the deposit (also control group)
- **y**: has the client subscribed a term deposit? (binary: "yes","no")

Data related with the last contact of the current campaign

- **contact:** contact communication type (categorical: "cellular","telephone")
- **month:** last contact month of year (categorical: "jan", "feb", "mar", …, "nov", "dec")
- **day_of_week:** last contact day of the week (categorical: "mon","tue","wed","thu","fri")
- **duration:** last contact duration, in seconds (numeric)

- **campaign:** total number of contacts performed during this campaign and for this client (numeric, includes last contact)

Data about previous campaigns

- **pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- **previous:** number of contacts performed before this campaign and for this client (numeric)
- **poutcome:** outcome of the previous marketing campaign (categorical: “failure”,“nonexistent”,“success”)

Social and economic context attributes

Data gathered for the day the day of subscription to the term deposit.

- **emp.var.rate:** employment variation rate - quarterly indicator (numeric)
- **cons.price.idx:** consumer price index - monthly indicator (numeric)
- **cons.conf.idx:** consumer confidence index - monthly indicator (numeric)
- **euribor3m:** euribor 3 month rate - daily indicator (numeric)
- **nr.employed:** number of employees - quarterly indicator (numeric)

So…

After you prepare the solution please send us:

- **programming code** in open-source language (preferably Python or R) used in the process
- **report / presentation with the results** (e.g. notebook, markdown, pptx - can contain also codes if convenient) which should include:
    - description of proposed approach
    - appropriate evaluation metrics for model/models (including ROC graphs and AUC for comparison purposes)
    - expected campaign lift

Your ideas on what some next steps could be are also welcomed.

Through your solution we would like to learn more about how you approach a Data Science problem, you thought process and decisions. The lift from the model and model scores are important, but be aware this is not a kaggle competition :).

Good luck and have fun!