Big Data Analytics I - Project 2019

Business Analytics in Banking

Task

The data set used in this project is related to direct marketing campaigns of a banking institution which was based on multiple phone calls to prospective clients. The purpose of this project is to produce the best model to predict the probability that a client will subscribe to a bank term deposit on the basis of various predictors. The competition is hosted on the following Kaggle website: https://www.kaggle.com/t/e192b9f0f11045238da5abffd0a74825.

The data has been splitted into training and test sets. The full training set is available to you. But only the predictors are provided for the test set. You can evaluate your predictions by submitting them to the Kaggle website. Note that only 50% of the test set is used to compute your *public score*. Your final score using the full test set will be provided at the end of the competition.

Here is a list of the predictors:

1. age
2. job: type of job
3. marital: marital status
4. edu: education
5. default: has credit in default?
6. housing: has housing loan?
7. loan: has personal loan?

Variables related to the last contact of the current campaign:

8. contact: communication type of the last contact
9. month: month of the year of the last contact
10. day_of_week: day of the week of the last contact
11. campaign: number of contacts performed for this client during this campaign (includes last contact)
12. pdays: number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
13. previous: number of contacts performed for this client before this campaign
14. poutcome: outcome of the previous marketing campaign

Important: You are not allowed to use any other external information/data to build your model and produce your predictions.

The evaluation metric for this competition is the log loss:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i log(\hat{p}_i) + (1-y_i) log(1-\hat{p}_i)],$$

where

- n is the number of data points in the test set
- $\hat{p}_i$ is the predicted probability
- $y_i \in \{0,1\}$
- $log()$ is the natural (base e) logarithm

Note: the actual submitted predicted probabilities are replaced with $\max(\min(\hat{p}, 1-10^{-15}), 10^{-15})$. A smaller log loss is better.

1. Your first task is to form a team with three people.

2. Each team member should create a Kaggle account (using your UMONS email address)
3. Form a team on Kaggle.
4. Do some basic exploration of the dataset
5. Build your first model. Predict the test set, and upload your predictions to Kaggle.
6. Try, and try again to improve your model. You can submit one prediction per day.

Project report and presentation

The data analysis report can be a maximum of 5 pages, and must abide by the section structure described below.

- Section 1: Introduction

The introduction will describe the data set and motivate the problem. It should be brief.

- Section 2: Methodology

This section describes the models and methods you have used, including a justification of your choices. You should also present your model fitting, diagnostics, etc.

- Section 3: Results and Discussion

This includes for example graphs and tables, as well as a discussion of the results.

- Section 4: Conclusion

This includes summary of the findings.

You should clearly explain what you have done, using figures to supplement your explanation. Your figures must be of proper size with labeled, readable axes. In general, you should take pride in making your report readable and clear. You will be graded both on *stastical/computational content* and *quality of presentation.*

Finally, each team will make a presentation of their work for the class (max 10 slides, in PDF format). Each team will have 10 minutes for presentation, and few minutes for Q & A. All members of the team must participate by speaking in the presentation. Score will be given by other members of the class. All students must be present to evaluate the presentations, and if not points will be deducted from the absent individual's score.

Grading

- Total points: 20
- Accuracy of classifier on Kaggle: 6
- Report: 7
- Presentation: 7

Deadlines

Do not wait until the last minute. Late submissions will not be allowed.

- April 24, 11:55pm: Submit the names of each member of the team on Moodle.
- May 1, 11:55pm: At least one Kaggle submission needs to have been made.
- May 12, 11:55pm: The Kaggle competition closes.
- May. 15, 11:55pm: Upload to Moodle (i) your project report, one per group, and (ii) the slides for your presentation.
- May 17, 1:15pm: Give your presentation.