Project 2

Deadline: 03.06.2024

Goal

The goal is to build a model that combines two advantages: has high accuracy and is parsimonious, i.e., is based on a small number of variables. Imagine that the company you work for has commissioned you to build a predictive model whose purpose is to identify customers who use the bank's marketing offer.

1 Data

We have **5000 historical training data**. Each client is described with **500 variables** (variables are anonymized). Your task is to build a model that predicts which customers in the test set took advantage of the offer.

Training data
x_train.txt - variable matrix for training data, containing information about 5000 clients.
y_train.txt - labels on the training set (value 1 means that the customer took advantage of the offer, value 0 means that the customer did not take advantage of the offer).

Test data
x_test.txt - variable matrix containing information about 5000 clients.

2 Task

Your goal is to build a model on training data and then identify 1000 customers in the test set who you think will benefit from the offer (of course, it is best to send the offer to those customers who are likely to use the offer). The number 1000 is related to the fact that the company can send offers to max. 1000 customers during one campaign. Additionally, you should indicate the variables that were used to build the model. The effectiveness of your prediction will be assessed as follows:

- For each designated customer who actually took advantage of the offer, the company will pay you €10.
- For each variable used, you must pay €200 (the company bears the cost of obtaining information related to individual variables).

Example 1 Among the 1000 customers you indicated in the test set, all of them took advantage of the offer, for which you get: $€10 \cdot 1000 = €10,000$. Your model is based on 20 variables. The penalty for the variables used is: $€200 \cdot 20 = €4,000$. Your score: $€10,000 - €4,000 = €6,000$.

Example 2 Among the 1000 customers you indicated in the test set, only 400 people took advantage of the offer, for which you receive: $€10 \cdot 400 = €4,000$. Your model is based on 3 variables. The penalty for the variables used is: $€200 \cdot 3 = €600$. Your score: $€4,000 - €600 = €3,600$.

The higher the score, the better, because it means a higher reward.

3 General additional remarks

- The projects are implemented in teams of 3 students.
- You can choose any programming language (Python/R are preferred), as long as the resulting files are in the correct format.
- You should evaluate at least 5 strategies (model/feature selection).
- The experiments should be described in the report (max 4 pages).

4 Final grade

The total number of points to be scored is 40, including:

1. score (20 points)
2. report (10 points)
3. presentation* (10 points)

* All teams must record a presentation (maximum 5 minutes). During the classes 50% of the group will additionally present their results live (maximum 10 minutes).

Live presentation dates: **Group 1**: 13.06.2024, **Group 2**: 11.06.2024, **Group 3**: 04.06.2024.

5 Solution

Your solution should be contained in two files:

1. File *STUDENTID_obs.txt* should contain 1000 indexes of customers from testing data to whom you want to send the offer.
2. File *STUDENTID_vars.txt* should contain the indexes of variables used by the proposed model.

STUDENTID is a student id of the first student from the team. Please see example files: *123456_obs.txt* and *123456_vars.txt*.
The submitted files must be in the same format!

Please upload a directory `Surname1_Surname2_Surname3` to GitHub repository https://github.com/kozaka93/2024L-DSAdvancedML via Pull Request. The title of PR should be `[P2] Surname1, Surname2, Surname3`. Every team should upload files to the folder `/projects/project2/groupX`, depending on attending the project group.

The directory should include:

- file *STUDENTID_obs.txt*
- file *STUDENTID_vars.txt*
- `codes` directory include all code needed to reproduce results
- report **(.pdf file)**

The recorded presentation **(.mp4 file** entitled `Surname1_Surname2_Surname3`) should be uploaded to MS Teams to the folder `/Projects/Project 2/Presentation/Group X`.