Individual Project Description

I. Introduction

This project is actually the extension of the lab quiz on Week 12, which encourages you to improve the performance of models by any means discussed in the lecture or external resources. Unlike the group project, the problem here is well formulated, the dataset is given and the performance evaluation metrics is also determined. In addition, you are supposed to conduct this project by yourself, and the grading **mainly relies on the performance** of your proposed model evaluated on the test dataset.

II. Your Role

You are required build a predictive model with Python that can predict if a bank client will subscribe a term deposit after being targeted during the direct marketing campaigns (phone calls) from the bank. In particular, you are supposed **rank** the clients by their probability of subscribing a term deposit. And a client is labeled as "positive" if he/she has subscribed a bank term deposit.

Therefore, the performance will be evaluated with the **AUC score of the ROC curve.** And you are supposed to build a well-trained binary classification model with ranking scores of being positive. In general, you can enhance the performance of the model proposed and built by you based on following methods discussed in the course.

1) Data clean and preprocess
2) Class imbalance reduction
3) Feature engineering
4) Model selection and parameter tuning
5) Ensemble learning

You are free to use any methods of above categories (not all required), or any other methods to enhance the model performance. You are provided with one training dataset (clients with true labels) for you to develop and optimize your proposed models. Based on the best trained model selected by you, you are then required to predict the label and corresponding ranking scores of being positive for clients in the test dataset (without true labels).

On the other hand, the teacher holds the true labels of the clients in the test dataset, and the AUC-ROC score will be computed by comparing the predicted ranking scores of being positive provided by you and the true labels held by the teacher. You will be mainly graded on the performance evaluated in the test dataset as above. In addition, the novelty of operations you take to enhance the model performance will also be taken into consideration.

After evaluation, **top 3 students** will be invited to **share** their experience during the last lecture on May 8, 2020 with for 5-10 minutes, and each student will be awarded with maximum 2% overall grade based the quality of sharing. This sharing is not compulsory but encouraged.

III. Provided dataset

You are provided with two dataset for this project as follows:

1) **bank_marketing_train.csv** contains the information of 26,246 clients. The information of each client includes 19 different features and 1 target label that indicates if the client has subscribed a bank term deposit;
2) **bank_marketing_test.csv** contains the information of 8000 clients. The information of each client includes 19 different features.

The definitions and explanations of 19 features and labels are listed in the appendix.

IV. Deliverables and Due Dates

Electronic submission will be required on following materials with on/before **Sat. May 2, 2020 at 23:59.**

1) Filled **"Individual Project Report.xlsx"** with the main operations you take to enhance model performance, and the performance of your built model evaluated by 5-fold CV on the training dataset;
2) "**bank_marketing_test_scores.csv**" to store the predicted ranking score for the 8,000 clients in the "**bank_marketing_test.csv**". This file is supposed to have 8,000 rows, and each row records corresponding ranking score of being positive for the corresponding client in the same order in "**bank_ marketing_test.csv**". You may refer to the "bank_marketing_test_scores(example),csv" for example;
3) A **packed Python project file** which includes your Python code and any data required by your Python code.

In addition, the invitation for sharing will be sent before Tue. May 5, 2020 at 23:59 and the acceptance of invitation should be confirmed before Wed. May 6, 2020 at 23:59**.**

V. Notes

- You need to make sure that the submitted Python code can generate and save the ranking scores are consistent with the scores in the file submitted by you separately. Otherwise, the scores generated by the Python code will be adopted for grading and will incur a penalty on the grade. If there is any error in your submitted Python code such that it cannot generate valid ranking scores, the whole project will be skipped for grading.
- You are NOT allowed to use any extra data in addition to the provided data listed as above.
- You are free to modify add/delete the categories/operations in "Individual Project Report.xlsx" at your convenience.
- Understand the data itself is very important, you are suggested to read and think about the defined features twice before you take any movement.
- There are several missing values in some categorical attributes, all coded with the "unknown" label.

Appendix: Dataset Description

This dataset is based on "Bank Marketing" UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The definitions of features in this dataset are listed as following table. For more information, read [Moro et al., 2014].

#	Feature	Data type	Explanation
#Clients' personal features:			
1	age	numeric	Age in years
2	job	categorical:	Type of job("admin.","blue-collar", "entrepreneur", "housemaid","management", "retired","self-employed","services", "student","technician", "unemployed","unknown")
3	marital	categorical	Marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
4	education	categorical	Education level, ("basic.4y","basic.6y","basic.9y", "high.school","illiterate", "professional.course", "university.degree","unknown")
5	default	categorical	Has credit in default? ("no","yes","unknown")
6	housing	categorical	Has housing loan? ("no","yes","unknown")
7	loan	categorical	Has personal loan? ("no","yes","unknown")
#Features related to the last contact of the current campaign:			
8	contact	categorical	Contact via ("cellular","telephone")
9	month	categorical	Last contact month of year ("jan", "feb, ... , "dec")
10	day_of_week	categorical	Last contact day of the week ("mon","tue", "wed","thu","fri")
# Other attributes:			
11	campaign	numeric	Number of contacts performed during this campaign and for this client (includes last contact)
12	pdays	numeric	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	numeric	Number of contacts performed before this campaign and for this client
14	poutcome	categorical	Outcome of the previous marketing campaign ("failure","nonexistent","success")
# Social and economic context attributes:			
15	emp.var.rate	numeric	Employment variation rate - quarterly indicator
16	cons.price.idx	numeric	Consumer price index - monthly indicator
17	cons.conf.idx	numeric	Consumer confidence index - monthly indicator
18	euribor3m	numeric	Euribor 3 month rate - daily indicator
19	nr.employed	numeric	Number of employees - quarterly indicator
# Target variable:			
20	y	categorical	Has the client subscribed a term deposit? ("yes","no")

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.0013