uc3m

BIG DATA INTELLIGENCE / MACHINE LEARNING I

ASSIGNMENT #1:
PREDICTING BANK PRODUCT SUBSCRIPTION

2025-26

SOURCE: MICROSOFT COPILOT USING DALL-E 3

INDEX OF CONTENTS

NOTES:
We need to use all available data from our dataset selected for training KNN, TREES, RANDOM FOREST and GRADIENT BOOSTING. We need to perform the outer evaluation (estimation of future performance) with train-test. The test dataset is ignored until the final step of estimation of future performance. The inner evaluation is used inside train for Hiper Parameter Tunning (cross-validation or train-test) (selecting the best parameters for training and testing on the outer part)

INTRODUCTION

The purpose of the first assignment is to practice with machine learning methods, both basic and advanced, including hyper-parameter tuning and preprocessing the data to adapt the dataset to the ML methods (encodings, imputation, constant features, etc.).

The topic of this assignment is predicting whether a bank customer will subscribe to a term deposit. A bank would like to build a predictive model based on customer demographic, financial and interaction data, in order to identify clients more likely to accept the product when contacted. The target variable is *deposit*.

GENERAL CONSIDERATIONS

1. Results **must be reproducible**. Therefore, set the seed at the appropriate places. But instead of using seed 42, use your **Student ID number**.

2. There are **two datasets**: the available data set (for model training, hyper-parameter tuning, and model evaluation) and the competition dataset (for using the model: making predictions for future instances). At the bottom of this document, you will find the names and meaning of each of the variables in the datasets.

3. Each group must use a different available data set. The supplied datasets have the names **bank_xx.pkl** (in **pickle** format), where **xx** is the last two digits of the NIA of one of the members of the group, and **bank_competition.pkl.** The competition dataset is common for all students.

4. The model evaluation method for this assignment will be **Holdout** (train/test). Decide on the most appropriate metric for this problem. Report also confidence intervals.

5. **Execution time** of the training process for all methods (fit) should also be reported.

6. **Preprocessing** should be conducted using **pipelines when appropriate** and using the required preprocessing steps for each of the chosen methods.

STEPS TO FOLLOW

1. SIMPLIFIED EDA (0.6 POINTS)

Do a **simplified EDA**, mainly to determine how many features and how many instances there are, which variables are categorical/numerical, what categorical variables have high cardinality, which features have missing values and how many, whether there are constant columns, whether there are ID columns, and whether it is a regression or classification problem. If the latter, is it imbalanced? Analyze particularly the variable ***pdays***, and describe how you intend to preprocess this variable (and preprocess it already, if possible at this stage).

2. DEFINITION OF THE OUTER AND INNER EVALUATION (0.4 POINTS)

- Decide on some appropriate train set size / test set size (the latter, for the estimation of future performance / outer evaluation), justify your answer, and split your data into the train and test partitions. Important: most of the assignment will be carried out using only the train partition. The test partition will only be used when you have concluded what your final model is going to be, only then you will use the test partition to evaluate it.

- Decide on how the inner evaluation is going to be carried out. The inner evaluation is used when doing hyper-parameter optimization, but it is also used to evaluate and compare different alternatives (please, refer to the CASH problem in the theory class (Combined Algorithm Selection and HPO)). Therefore, most of the assignment will use the inner evaluation, except at the end, where the test partition (outer) will be use to evaluate the final model.

3. BASIC METHODS: KNN AND TREES (1.4 POINTS)

1. Train, evaluate and **compare the two basic methods** with default hyperparameters, and also a **dummy method**.
2. **Use shallow trees for interpretability to get some understanding of the problem** (what are the important features and the main decisions of a tree).
3. Do **hyperparameter tuning for KNN and Trees** using on the one hand GridSearch or Random Search, and on the other, Optuna. Does HPO improves results over default hyperparameter values? At what computational cost? Which HPO technique obtains the best results?
4. Use the results of HPO t**o understand the relation between hyper-parameter values and performance**, for KNN.
5. At this stage, summarize your results and draw some conclusions. Based on your findings, decide on one of the HPO methods for the remainder of the assignment and justify your answer.

4. ADVANCED METHODS (1.5 POINTS)

- **Choose two advanced methods**: one bagging technique (out of **Random Forests** and **Extra Trees**) and one gradient boosting technique (out of **gradientboosting**, **histgradientboosting**, **lightgbm**, **xgboost**, or **catboost**). Peruse the documentation of the methods and justify your selection. You can also compare the results of two of them to support your choice. Try default values and hyper-parameter optimization and compare both cases to determine whether HPO improves results, and at what cost.

5. RESULTS AND FINAL MODEL (0.5 POINTS)

- **Report your results**: report the inner evaluation of all alternatives tested (use a table), select the best one according to the inner evaluation and evaluate it on the test set (outer evaluation **== estimation of future performance**), estimating also a **confidence interval**.

- **Using the best method**, train the **final model** and use it to **make predictions on the competition dataset**. Save both the **final model** in an appropriate ML format and the **competition predictions** in a csv file.

6. PROBABILITY CALIBRATION (0.8 POINTS)

- Select the best method of the previous section among the ones which are able to predict probabilities, determine visually whether it is calibrated, and use one of the scikit-learn methods for improving the calibration. Check visually if the calibration was improved without worsening the original metric on the test set.

7. OPEN CHOICE TASK (0.8 POINTS)

- Decide on your own some additional task, either because it could improve results or because you find it particularly interesting. Justify your selection. Some possibilities are: doing feature selection, using Streamlit or other library to deploy your best model, etc. (beware, using Streamlit can be time consuming).

WHAT TO HAND IN

- A **jupyter notebook** with the code in the proposed order of steps. Please **use some of the cells to comment** about what you are doing and your results. In particular, emphasize your conclusions after each step with short arguments based on your results.
  - If it is more convenient, you can also hand in a file with Python code instead (i.e. a script) and a separate report.
  - **If you decide to use any AI chatbot**, **briefly explain in those commented cells what purpose you used it for and how you used it (for instance, you can quote the prompt and the output used**).
  - Please **write the names of the components of your group at the beginning of the notebook**.

- A file containing your **final trained model** in an appropriate ML format.

- A pickle or .csv file containing **your final model's predictions (values of your model's predictions** on the competition set).

- Any additional script/notebook if you require it (for instance, if you use Streamlit)

APPENDIX: VARIABLE DESCRIPTION

Variable	Short Description
age	age in years
job	type of job
marital	marital status
education	education level
default	has credit in default?
balance	average yearly balance
housing	has a housing loan?
loan	has a personal loan?
contact	contact communication type
day_of_week	last contact day
month	last contact month
duration	last contact duration, in seconds

campaign	number of contacts performed during this campaign and for this client
pdays	number of days that passed by after the client was last contacted from a previous campaign. The value is -1 if no/unknown contact was produced.
previous	number of contacts performed before this campaign and for this client
poutcome	outcome of the previous marketing campaign
deposit	has the client subscribed a term deposit? TARGET VARIABLE