Homework Assignment 3 - Classification Challenge

Dataset

This dataset is related to direct marketing campaigns of a banking institution. The dataset includes the following features:

1. **age**: (numeric)
2. **job**: type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3. **marital**: marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4. **education**: (categorical: "unknown","secondary","primary","tertiary")
5. **default**: has credit in default? (binary: "yes","no")
6. **balance**: average yearly balance, in euros (numeric)
7. **housing**: has housing loan? (binary: "yes","no")
8. **loan**: has personal loan? (binary: "yes","no")
9. **contact**: contact communication type (categorical: "unknown","telephone","cellular")
10. **day**: last contact day of the month (numeric)
11. **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
14. **previous**: number of contacts performed before this campaign and for this client (numeric)
15. **poutcome**: outcome of the previous marketing campaign

(categorical: "unknown","other","failure","success")

Output variable (desired target):

1. **y**: has the client subscribed a term deposit? (binary: "yes","no")

Assignment

Your goal is to use classification to predict the ‘y’ column.

Obtain predictions for the `hw-3-test-data.csv`. Save your predictions to a file called `answers.csv`, which should contain a single column of predictions. For the test data, impute any missing data using the transformers that have been trained on your training data before testing.

Submit both your notebook and your answers.csv file

Rubric

Points will given for each of the following sections present in your notebook. Each section must be named with a heading. Code must be documented where necessary, and markdown should be present to explain summary findings at the end of each section. Points will be awarded based on the thoroughness and quality of code in each section.

1. **Exploratory Data Analysis** - Did you:
   - Explore your data
   - Examine distributions
   - Look for correlations
   - Identify nulls
2. **Preprocessing** - As necessary, did you correctly:
   - Handle outliers
   - Encode features
   - Standardize features

   - Handle nulls (whether by imputation or some other method)
   - Handle imbalanced classes
3. **Modeling & Evaluation** - Did you:
   - Try more than one classifier
   - Use a cross-validation procedure
   - Avoid data leakage
   - Correctly evaluate F1-Score
4. **Test data** - Did you:
   - Train your model appropriately
   - Impute data correctly
   - Achieve at least 50% F1 on the test data
   - Save your answers to an `answers.csv` file