Deposit opening classification problem:

1. **Business problem:**

- The objective of the classification is to identify clients who will subscribe (yes/no) for a term deposit. (Variable y: Target function).
- The Bank wants us to conduct Exploratory Data Analysis (EDA) to identify relationships, trends in data. For example: correlations, bivariate analysis of target versus input variables, facts, univariate patterns, missing data,
- Develop and save a predictive model to roll out for future use. Explore different techniques and share your findings about the approach and benefits of the champion model.
- Prescriptive recommendations if any
- K-means Clustering is optional (Bonus point)
- If you are comparing more than four different supervised algorithms (bonus point). You can utilize the Pyspark or spark-Scala platform for this Mini project.

2. **Dataset information:**

- Data is about an XYZ bank's direct marketing campaign. Marketing campaigns were driven by telephone calls.
- Data Set In many cases, more than one contact for the same client was required., in order to access if the product (deposit) would be ('yes') or not ('no') subscribed
- The purpose of the classification is to forecast whether the customer will signup (yes/no) a term deposit (variable y).
- The dataset: XYZ_Bank_Deposit_Data_Classification.csv, 20 entries/columns, sorted by date between May 2008 and November 2010.

Attributes information:

1 - Age (Numeric)
2 - Job: type of job (categorical)
3 - Marital: marital status (categorical)
4 - Education (categorical)
5 - Default: has credit in default? (categorical)
6 - Housing: has housing loan? (categorical)
7 - Loan: has personal loan? (categorical)

regarding the latest contact in the ongoing campaign:

8 - Contact: contact communication type (categorical)
9 - Month: last contact month of year (categorical)
10 - Day_of_week: last contact day of the week (categorical)
11 - Duration: last contact duration, in seconds (numeric)

other attributes:

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; 999 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)

social and economic context attributes

16 - Emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - Cons.price.idx: consumer price index - monthly indicator (numeric)
18 - Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - Euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - Nr.employed: number of employees - quarterly indicator (numeric)

3. Students are required to submit their findings via GitHub. Our objective is to introduce students to Git AI/ML CI/CD industry practices. Don't be concerned about real-time deployment or integration. All we need to do is organize the following files in git.

   1. Data file
   2. Pickle file/saved model file (Refer the below short commands)
   3. Model file – py file
   4. Readme file describing project details.

5. PPT (with notes) or, Word report capturing the results

Saving model as serialized object:

```
lr = pipeline.fit(df) // Trained model
lr.save("/path")
pipelineModel = lr.load("/path")
df = pipelineModel.transform(df)
```