Yugen.ai
Your Trusted AI Partner

Assignment – ML Engineer

Note:

1) You can choose scripting language of your choice (Python preferred)
2) There are 6 questions in the assignment.
3) We would expect a submission from you within 3-4 days' time after we share it with you. In case you foresee any issue, please inform us proactively.
4) To submit your response, you could either share a "jupyter notebook" having code blocks below their respective questions or paste your code in this document below each question.
5) Please save your assignment on your system until the interview round, so that we could discuss your approach etc. over a call.
6) Model Accuracy & Latency of prediction is not very important, as this assignment is to understand your approach, ability to code hands-on and creating an end-to-end pipeline.

Overall Objective

In this assignment, you will analyze an open dataset about a marketing campaign of a Portuguese bank in order to design strategies for improving future marketing campaigns. The object of this campaign is to pursue customers to subscribe the term deposit. The Marketing campaign was based on phone calls. The classification goal is to predict if the client will subscribe a term deposit (variable y). Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

The dataset name is "Bank-full.csv".
Number of Instances: 45211 for bank-full.csv
Number of Attributes: 16 + output attribute.

Attribute information:

\Input variables:
1 - age (numeric)
2 - job : type of job (categorical)
3 - marital : marital status (categorical; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary)
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary)
8 - loan: has personal loan? (binary)
# related with the last contact of the current campaign:
9 - contact: contact communication type (categorical)
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)
# other attributes:
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - 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)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical)

Output variable (desired target):
17 - y - has the client subscribed a term deposit? (binary: "yes","no")

Missing Attribute Values: None

<u>Questions:</u>

Import the data and then:

1) Perform data cleaning, missing value treatment, outlier treatment on it. (list down your assumptions in the script).
2) Create a classification model using a non-neural-network based classifier like logistic regression/ xgboost/ adaboost/ catboost etc. of your choice
3) Compare the accuracy/performance of the above model with another model made using a neural network. (free to choose your function)
4) Deploy the model on any cloud service of your choice (AWS, Azure, GCP, Heroku) and set up an inference pipeline (1 input at a time → 1 output at a time)
5) Make an API call from Postman or a python script to get real time predictions (1 input → 1 output)
6) Modify the inference script / API so that it can ingest and predict multiple payloads (n inputs → n outputs)