Empowerment Labs
Machine Learning EngineerTest

The main objective of this test is to measure your abilities and knowledge in Python, as well as your ability to solve problems.

Rules:

- The problems presented in the test must be solved using only Python.
- It is allowed to use any Python library, as well as research information on web pages.
- Each point must be resolved in a different python file.
- Once the test is resolved send the python files with the csv compressed by email.

1. Given the file *astronauts.csv* you'll need to create a function that determines and writes in a csv file. The function must have one parameter, the name of the file

    a) Average missions and average days in space by country
    b) % of astronauts in space by country.
    c) Company that has sent the most missions by country.
    d) Main achievement by country, taking into account that the main achievement is the achievement that is less repeated.

2. The file *bbc_news.csv* has data extracted from news during 2022. For this exercise it's required to create a function that takes the data from the csv file and prints x most repeated words per month. The function must have two parameters: the name of the file to be read and the number of words to show. To define the most frequent words must remove stopwords and applied stemming techniques.

3. The file *fraud.csv* has data about transactions in credit cards. This table has the following columns:

    trans*date*trans_time: The date and time of the transaction.
    cc_num: credit card number.
    merchant: Merchant who was getting paid.
    category: In what area does that merchant deal.
    amt: Amount of money in American Dollars.
    first: first name of the card holder.
    last: last name of the card holder.
    gender: Gender of the cardholder.Just male and female!
    street:Street of card holder residence
    city:city of card holder residence
    state:state of card holder residence
    zip:ZIP code of card holder residence
    lat:latitude of card holder

long:longitude of card holder
city_pop:Population of the city
job:trade of the card holder
dob:Date of birth of the card holder
trans_num: Transaction ID
unix_time: Unix time which is the time calculated since 1970 to today.
merch_lat: latitude of the merchant
merch_long:longitude of the merchant
is_fraud: Whether the transaction is fraud(1) or not(0)

You will need to train a classification model to detect fraud in transactions. The model must include the next variables: Month, day of week, age , state, amount, gender and category.

a. EDA about the dataset
b. What do you think about the unbalance in the sample for the is_fraud variable? How can you correct it?
c. Which model did you choose? Why?
d. Define the confusion matrix and its metrics
e. Define the ROC curve and AUC metric
f. Which is your opinion about the model performance?
g. Include 2 new variables to retrain the model and define the previous metrics for the new model.