

1. CodingTest

Introduction and Objective

The objective of the test is to assess familiarity with basic data manipulation, coding skills, code quality.

You can use either Python, Java, or Scala to complete the test. You can also use SQL for querying the data. You can use open-source third-party packages, but please clearly indicate the dependencies as part of your build script (requirements.txt/maven/sbt, etc).

You can create a zip or tar archive of your code and related files and submit them by email. Please make sure that the code is working and well tested and has unit tests to prove the functionality of any subroutines which are part of the code package. Please also include a readme file to explain code organization, main modules, and explanation of any tricky data structures or algorithms that you have used. The readme should also include steps to build and run the package.

Problem

For this test, we will use the dataset from the UCI Machine learning repository about marketing <https://archive.ics.uci.edu/ml/datasets/bank+marketing> which is trying to predict if the client will subscribe (yes/no) a term deposit (variable y).

Download the bank-additional.zip file from this link
<https://archive.ics.uci.edu/ml/machine-learning-databases/00222/>

When you extract the link, it will contain bank-additional-full.csv. All questions below are based on the bank-additional-full.csv file.

As mentioned above you can use Python/Java/Scala and/or SQL for completing this task.

1. Write a program to load the file into an SQLite (or Derby if using Java/Scala) table. Create a DDL script for table creation. Optimize the DDL for storage space. Create indexes if necessary.
2. Print a table showing the percentage of married/single/divorced/unknown people for all job types as shown below. You just need to print to console, no need to beautify the output

Job	Married	Single	Divorced	Others/Unknown
Services	20%	15%	35%	30%
Entrepreneur	10%	40%	40%	20%
Admin	...			
Technician	...			
...				

3. Compute histograms of age for people with a university degree. This should be done in either Python or Java
 - a. Histogram of the age of all people with a university degree
 - b. Histogram of the age of all people with university degree and label = yes. (Value of column y = yes)
 - c. Histogram of the age of all people with university degree and label = no. (Value of column y = no)
4. Plot the histograms generated in step 3 and save them to a file (.jpg or .png) submit plotting code along with the image.

2. SoftwareEngineering System Design and DS-ML

Introduction and Objective

The objective of the test is to assess familiarity with the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements and your understanding of developing software in a systematic method

5. For FP online website. When a user clicks on a product to open its details page, you would like to suggest 5 more items that the user may be interested in, based on item features and the user's purchase history, and display them at the bottom of the page. What services and database tables would you need to support this behavior?
6. What data would you like to collect from an FP-online to measure user engagement and product popularity?
7. You want to generate a live visualization of what portion of webpage users are currently viewing and clicking, sort of like a heat map. What components/services/APIs do you need in place to enable, on the client and server end?
8. Let's say you are asked to predict whether a person has diabetes or not and you applied logistic regression (since the dependent variable is yes/no). Which method you will use to evaluate the model and why?
9. How many models will be required for making the final prediction if you are training one vs all classifier models on a given dataset where the target variable has categories as 0,1,2,3,4,5,6,7?