Data Scientist Take-Home Exercise

Instructions:

Thank you for your interest in joining BMC Software as a data scientist. This take-home exercise will assess your skills and problem-solving abilities in data science. Please follow the instructions below to complete the exercise:

Problem Statement:

Your Marketing department would like to contextualize companies website to show viable products your customers might purchase based on past orders/transactions. To help with this effort, the Marketing department contacted you to build an ML model. Analyze the provided dataset and build a predictive model to predict the probability that a given customer will buy a "***printer-related product."*** To achieve the following

- Share your findings with the marketing team (non-technical or Business stakeholders)
- Build a production-ready ML Model for deployment

Dataset:

Download the JSON datasets from the link.

Data Dictionary

Data File Name	Description	Joining Keys
customer_info	Provides information of customers who have purchased atleast one product	Customer ID
product_info	Provides list of products your company sells	Product ID
customer_transaction_info	Provides information of orders, product & relevant information that customers bought	Order ID, Customer ID, Product ID
orders_returned_info	Orders that the customers have returned either they didn't like the product or damaged or other reason unknown	Order ID
region_seller_info	Customers region covered by your sales team	Region

"***printer-related product"*** can be identified if a customer has ordered a product if the *Product Name* field contains the word “printer" and the product sub-category is *'Machines.'*

Requirements:

1. This exercise should be completed individually. See the Disqualification criteria below.
2. The time limit for completion is seven days from when you received this from BMC HR or Hiring Manager.
3. Environment Set-up:
   - Use Python as a programming language; specify libraries and versions used for the exercise.
   - Provide detailed instructions to set up the required environment to ensure reproducibility.
   - Include a list of any additional dependencies or packages necessary to run your code.
4. Documentation and Communication:
   - Create a technical report (Jupyter Notebook etc.) documenting your approach, methodologies, and findings.
   - Explain the steps taken during analysis, modeling, and evaluation.
   - Include visualizations and supporting materials.

- Present findings and conclusions concisely and clearly.

Submission:

Submit your completed exercise as a Jupyter Notebook (exported as HTML, Slides, etc.). Include your code, visualizations, and any additional materials used.

Guidance for Reproducibility:

- Include detailed comments in your code to enhance readability.
- Use relative file paths and ensure all necessary files are included with your submission.
- Provide clear instructions for reproducing your results, including the required commands or steps.

Evaluation:

Your submission will be evaluated based on your ability to demonstrate data science life-cycle following criteria:

- Data cleaning and preprocessing techniques
- Exploratory data analysis and insights gained
- Feature engineering choices and justification
- Model selection and performance evaluation
- Clarity and organization of the technical report
- Demonstrate prediction workflow

Disqualification:

- Plagiarism or using others' work without proper attribution will result in disqualification.
- Failing to follow the instructions, unauthorized sharing of exercise in any online forms, or submitting the exercise within the specified time frame will also lead to disqualification.

References:

Please specify any references, external resources, or libraries utilized during the exercise.

Note: Contact [provide contact information] for any questions or clarifications.

Best of luck! We look forward to reviewing your submission.

BMC