RAISING
THE VILLAGE

DATA SCIENTIST- WRITTEN TECHNICAL INTERVIEW QUESTIONS 11-09-2024

INSTRUCTIONS:

i. Answer all questions.

ii. Duration: 4.5hrs.

iii. Answers should be provided in electronic form.

iv. Ensure that your answers are mailed back on **Wednesday September 11, 2024, by 04:30pm**.

v. Save your answer file(s) using both names in the format “**FirstName Surname – DATA SCIENTIST ANSWERS”**.

Background:

Raising the Village (RTV) works with last-mile communities to lift them out of ultra-poverty through a 24-month agricultural intervention program. Progress among households (HHs) varies significantly, with some advancing quickly while others struggle. RTV is now seeking to develop a data driven system that identifies early indicators of struggle among households, allowing for timely interventions. The system must be robust, scalable, and able to integrate seamlessly with RTV’s existing infrastructure.

Objective:

Develop a scalable, interpretable machine learning model that predicts the likelihood of household struggle. Deploy this model as part of an integrated system, including an inference API, a data processing pipeline, model development pipeline, inference pipeline and continuous monitoring. The final solution should not only predict outcomes but also provide actionable insights and support real-time data-driven decision-making.

Submission Requirements

- Code: Submit a Jupyter Notebook or Python scripts, ensuring the code is clean, modular, and adheres to SOLID and DRY principles. Include a README with instructions.
- Version Control: Submit the assessment via a GitHub repository link with a clear commit history.
- Report: Provide a summary report (1-2 pages) detailing your approach, model interpretability, and deployment strategy.

Time:4.5 Hours

- For any follow-up questions please contact- 0705945524 or 0775648275 or 0757005504

Section 1: Dataset Description

- The dataset comprises household demographics, program participation details, geographic data, and outcome metrics, spanning multiple villages and years.
- Data dictionary

Section 2: Tasks

1. Data Preprocessing

- Handle missing values and outliers (***2 marks***).
- Create new features, including interaction terms where appropriate. (***1 marks***)
- Convert categorical variables to numerical representations. (***1 mark***)
- Scale or normalize numerical features where necessary. (***1 mark***)
- Create a new variable called ‘ProgressStatus’ based on the ‘HHIncome+Consumption+Residues/Day’ variable, categorizing the values as follows: "On Track" for values >= 2.15, "At Risk" for values >= 1.77, "Struggling" for values >= 1.25, and "Severely Struggling" for values below 1.25. (***1 mark***)
- Implement techniques to handle the imbalance in ‘ProgressStatus’. (***2 marks***)
- Split the data into training, validation, and testing sets. (***1 marks***)
- Ensure data preparation follows DRY (Don't Repeat Yourself) principles to avoid redundancy. ***(1 mark***)

2. Exploratory Data Analysis (EDA)
   - Examine and visualize the distribution and relationships of features and target variables. ***(10 marks***)
   - Identify patterns, correlations, and potential issues. ***(5 marks***)
   - Discuss the implications of these insights for model development. ***(5 marks***)

3. Model Development
   - Implement at least two different models, you can start with a non ML approach as a baseline ***(12 marks)***
   - Use cross-validation to assess model performance and avoid overfitting ***(5 marks).***
   - Apply SOLID principles to structure the code, ensuring it is modular, maintainable, and adheres to good design practices, considering future model updates and maintenance. ***(7 marks)***

4. Model Evaluation
   - Evaluate models using relevant metrics including custom metrics if appropriate ***(4 marks).***
   - Analyze and justify the trade-offs for the different metrics especially in the context of ‘At Risk’ households. ***(3 marks)***
   - Include confusion matrices and classification reports ***(2 marks).***
   - Interpret results, discuss potential biases, and justify the choice of the final model***.(1 mark)***

5. Code Quality & Version Control
   - Ensure the code is clean, well-documented, and adheres to coding standards. ***(4 marks)***
   - Implement version control using Git, with clear commit messages and a well-structured repository. ***(3 marks)***
   - Apply CI (Continuous Integration) practices to automate testing and linting. ***(3 marks)***

6. Machine Learning Interpretability:

- Implement interpretability methods and evaluate the effectiveness in explaining model predictions. ***(5 marks)***
- Provide a report on how interpretability impacts decision-making for field officers. ***(5 marks)***

7. Inference Endpoint Implementation***: (10 marks)***
    - Set up an inference endpoint using a framework like FastAPI, Flask or django. ***(4 marks)***
    - Implement input/output validation. ***(3 marks)***
    - Set up logging to capture request details, errors, and predictions. ***(3 marks)***

8. Data Literacy & Communication ***(10 marks***):
    - Prepare a concise report summarizing the approach, findings, and recommendations. ***(4 marks)***
    - Visualize the model’s impact on household identification and provide actionable insights for field officers***. (6 marks)***

**********END**********