Please answer the following questions using github/gitlab/ bitbucket repo (Select the one that is a good fit for you). Finally, please provide a public repo link so the hiring manager can read your answers.

Task 1: General Questions

1. What is your preferred language when building predictive models and why?

I prefer using Python and SQL when building predictive models. Python is my go-to language due to its extensive libraries and frameworks, such as scikit-learn and TensorFlow, which simplify the development of complex machine learning models. Additionally, its readability and ease of use make it an efficient choice for rapid prototyping and iterative development. SQL is invaluable for data extraction and manipulation. It allows me to efficiently query large datasets and perform complex joins and aggregations, ensuring that I can work with the most relevant and clean data for analysis.

2. Provide an example of when you used SQL to extract data.

In a previous project, I used SQL to extract patient records from a large healthcare database. I wrote queries to filter data based on specific criteria, such as diagnosis codes and admission dates, which allowed me to build a dataset that was relevant for analyzing trends in patient outcomes. For instance, I extracted data on patients with osteomyelitis and their comorbid conditions to assess the impact of these factors on treatment outcomes. This data extraction was crucial for developing a predictive model that informed our clinical strategies.

3. Give an example of a situation where you disagreed upon an idea or solution design with a co-worker. How did you handle the case?

During the Family and Social Services Administration project, I disagreed with the proposed data cleaning process. The initial regression results were suboptimal, suggesting that the cleaning process might not be addressing the underlying issues effectively. To resolve this, I scheduled a meeting with the team to thoroughly review the original data and the proposed cleaning steps. By discussing our observations and exploring alternative approaches collaboratively, we identified a more effective cleaning strategy that improved the regression results.

4. What are your greatest strengths and weaknesses and how will these affect your performance here?

My greatest strength is my communication skill. I make it a point to consult with experts when I encounter uncertainties, ensuring that the solutions I implement are well-informed and effective. This approach helps me deliver high-quality results rather than just completing tasks. My weakness is that I can become stressed when projects are incomplete or when facing roadblocks. This sometimes leads me to push my team to expedite progress. However, I am actively working on managing this stress and balancing urgency with a collaborative approach to maintain a positive team dynamic.

Task 2: Python model development

Objective: Given the dataset, **train.csv** - the training dataset; **Exited** is the binary target and **test.csv** - the test dataset; your objective is to predict the probability of Exited, write a python script (main.py) that when run (e.g. python main.py) will output:

- a CSV file containing the following:
    - Predicted Exited from test.csv
    - Evaluation of the predictive ability of the model.(e.g. F1 Score, Confusion Matrix…etc)
- Plots that can help us visualize the classification data and the prediction curves.
- Please submit codes, explanations, and plots when finished. Try to be more **specific**, a README might be helpful.