

Data Science & LLM Technical Assessment

Thank you for your interest in the role. As part of the process, we'd like you to complete the following technical assessment. The task is designed to evaluate your practical skills across the data science lifecycle: predictive modelling, feature engineering, working with language models, and communicating your results clearly.

Assignment Overview:

You've been asked to help a clinical data team develop an AI assistant that:

1. Predicts whether a patient will be readmitted to hospital within 30 days
2. Extracts and categorises key information from free-text discharge notes.

This assignment combines a binary classification problem with a lightweight NLP/LLM-based task, both based on the same dataset.

Dataset

Download the dataset:

👉 Assignment_Data.csv (attached)

The dataset includes:

- Structured features (age, gender, diagnosis code, medication type, etc.)
- A binary label: readmitted_30_days (0 = No, 1 = Yes)
- Free-text discharge notes for each patient

Your Tasks

1. Predictive Modelling (Binary Classification)

- Perform data cleaning and exploratory analysis
- Apply feature engineering: e.g. convert categorical variables to numerical, handle imbalances, derive new features
- Train a classification model (e.g. logistic regression, random forest, XGBoost)
- Evaluate using ROC AUC, F1-score, and confusion matrix
- Explain which features were most influential

2. Named Entity Recognition from Discharge Notes (LLM/NLP)

- Use an open-source NLP or LLM model (e.g. SpaCy, Flan-T5, BART from Hugging Face) to extract clinical entities from discharge notes
- Identify and label relevant medical terms: diagnosis, treatment, symptoms, medications, or follow-up actions
- Optionally use prompt engineering to guide the LLM for consistent output
- Briefly discuss risks such as hallucination, entity ambiguity, or limitations of general-purpose models in clinical contexts

3. Evaluation & Reporting

- Submit a well-documented Jupyter notebook or Python script
- Include a short report (max 1 page) explaining:
 - Your approach
 - Key results
 - Practical implications
 - What you'd do with more time or more data

4. Submission Guidelines

- Upload everything to a public or private GitHub repo
- Include:
 - Your code
 - README with instructions
 - Final report (PDF or Markdown)
 - Any dependency files (e.g. requirements.txt)

We're more interested in your thought process and clarity than in perfection. We look forward to reviewing your work!