Take home coding exercise (Data Science)

Overview

There are two parts to this exercise:

1> The first part of the coding exercise is to generate synthetic data using ChatGPT
2> The second part is to implement EDA & ML/DL models on the generated data

Synthetic Data Generation

Generate synthetic data to capture hourly pay rates for nurses in US across the major metros

1. Locations: Dallas, Atlanta, New York, Philadelphia, Washington, San Francisco, Los Angeles, Seattle, Chicago, San Diego, Miami, Boston, Detroit, Phoenix, Houston).
2. Each row of the generated data should contain the Job Title, Location (State & City), Hospital Name, Contract start date, Contract end date, Hourly Pay rate.
3. The hourly pay rate should show seasonal uptick during flu & Christmas holiday season
4. Generate Hospital name using City name as prefix followed by a suffix that could be one of

“Corporate”, "NonProfit", "Community”,”Veterans”, “Govt”

5. Assume the following Job Titles for nurses:
    a. 1) RegisteredNurse_ICU 2) RegisteredNurse_MedSurg 3) RegisteredNurse_Telemetry 4) RegisteredNurse_Oncology 5) RegisteredNurse_Pediatric 6) PhysioTherapist 7) LabTechnician 8) RegisteredNurse_CriticalCare 9) RegisteredNurse_Cardiology 10) RegisteredNurse_Surgery
6. Generate the data for the years 2023 & 2024
7. The contract duration cannot exceed 13 weeks.
8. Generate a total of 250,000 rows

EDA & ML/DL models

1. Using EDA, show
    a. variations of the hourly pay rates across the major metros
    b. show the uptick in pay rates during flu & holiday season
    c. show hourly pay rates against the desirability of a city (cost of living, schools, crime rates, public transport etc…)
    d. show specialization (oncology, cardiology, surgery) getting higher pay vs other job titles
2. Implement any two ML/DL models to predict the hourly rate. You will need to justify why you picked those models

3. For both models, show
    a. how you are handling the high cardinality of hospitals & propose/implement alternate methods (with pros & cons)
    b. metrics used for measuring the accuracy of the model (what metrics did you use & why did you choose them?)
4. Implement a Streamlit application to demo the model. The Streamlit app should allow the user to input a requirement (Job Title, Location, Hospital, Contract Start Date & Contract End Date) and show the predicted Hourly Rate
5. For bonus points: Implement one timeseries-based forecasting model using Prophet or NeuralProphet or StackedLSTM

Submission

1. For the first part, submit your link to ChatGPT session along with the generated data
2. For the second part, submit your google Colab notebook

Notes

1. It is important that you do the work yourself. Submitting other people's work will lead to immediate disqualification
2. During the demo, you will be asked to make changes to demonstrate your understanding of the problem. Hence it is even more important to follow above guideline