Assignment – Data Science Associate

Objective:

The objective of this assignment is to analyse a dataset regarding the number of goals scored in soccer matches. Candidates are expected to apply statistical methods to derive insights and fit appropriate probability distributions.

Dataset:

You are provided with data containing the number of goals scored in 2000 soccer matches. The dataset includes the following information:
Total number of matches = 2000
Number of goals: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Number of matches: 150, 400, 600, 500, 250, 75, 15, 4, 1, 5

Tasks:

- 1. Calculate the probabilities for each value of X (number of goals) based on the provided data.
- 2. Compare the observed probabilities with those predicted by the Poisson distribution. Estimate the parameter λ from the data as the average number of goals per match.
- 3. Fit the Negative Binomial distribution to the data and determine its parameters (r and p). Justify your choice of distribution based on the shape of the probability distribution.
- 4. Create visualisations to represent the observed frequencies and fitted distributions using a Python library (e.g., Matplotlib or Seaborn).
- 5. Write a brief report summarising your findings, including:
  - The calculated probabilities
  - The comparison with the Poisson distribution
  - The results of the Negative Binomial fit
  - Any conclusions and recommendations based on your analysis.

Coding Requirements:

1. Use Python for your analysis and provide well-commented code to ensure readability.
2. Submit your code in a Jupyter Notebook(.ipynb) or as a Python script(.py).

Question 2

1. Exploratory Data Analysis (EDA)

- **Task**: Perform EDA on the leads dataset. Include visualizations and descriptive statistics to analyze key aspects of the data.

2. Lead Scoring Logic Development

- **Task**: Develop a scoring model to identify leads with a higher probability of conversion.
    - **Considerations**:
        - Use relevant features from the dataset to calculate a lead score.
        - Define a scoring formula or logic based on observed patterns and insights gained from EDA.
    - **Questions to Address**:
        - What factors significantly contribute to a higher lead score?
        - How will you test and validate the scoring model for effectiveness?

3. Top Contributors/Features

- **Task**: Identify and rank features that contribute most significantly to lead conversion.
    - **Questions to Address**:

        - Which features (e.g., UTM-Source, CountOfClickEvents, WebTimeSpent) have the highest impact on successful conversions?
        - How can these insights be used to prioritize leads and optimize efforts?

4. Differentiation Features for Positive and Negative Sale Results

- **Task**: Analyze features that differentiate successful leads from those that did not convert.
    - **Questions to Address**:
        - What are the common characteristics of leads that converted successfully compared to those that did not?
        - Are there specific behavioral patterns that correlate with positive or negative outcomes?
        - Can you propose any recommendations for improving lead conversion based on your findings?

Coding Requirements:

1. Use Python for your analysis and provide well-commented code to ensure readability.
2. Submit your code in a Jupyter Notebook(.ipynb) or as a Python script(.py).

Deliverables

- A report summarizing findings from the EDA, including visualizations and key insights.
- A lead scoring model and rationale for the scoring logic used.
- A presentation highlighting the top contributors/features and differentiation features for sale results.