(Senior) Data Scientist Take Home Challenge

Dataset

This exercise is based on customer support data that can be downloaded at the following link:

https://drive.google.com/file/d/1pd4hR0U4Wm7sWtJp60-Del095WkqBveU/view?usp=sharing

The dataset includes customer support tickets for various tech products.

Features Description:

- Ticket ID: A unique identifier for each ticket.
- Customer Name: The name of the customer who raised the ticket.
- Customer Email: The email address of the customer (Domain name - @example.com is intentional for user data privacy concern).
- Customer Age: The age of the customer.
- Customer Gender: The gender of the customer.
- Product Purchased: The tech product purchased by the customer.
- Date of Purchase: The date when the product was purchased.
- Ticket Type: The type of ticket (e.g., technical issue, billing inquiry, product inquiry).
- Ticket Subject: The subject/topic of the ticket.
- Ticket Description: The description of the customer's issue or inquiry.
- Ticket Status: The status of the ticket (e.g., open, closed, pending customer response).
- Resolution: The resolution or solution provided for closed tickets.
- Ticket Priority: The priority level assigned to the ticket (e.g., low, medium, high, critical).
- Ticket Channel: The channel through which the ticket was raised (e.g., email, phone, chat, social media).
- First Response Time: The time taken to provide the first response to the customer.
- Time to Resolution: The time taken to resolve the ticket.
- Customer Satisfaction Rating: The customer's satisfaction rating for closed tickets (on a scale of 1 to 5).

Part I

1. What can you do with this dataset, what kind of insights can you generate? For example, top 10 issues based on the dataset.

2. From "Ticket Subject" and "Ticket Description" columns, how would you identify common queries from customers?
3. Can you explain how you would use the "Ticket Type" field to categorize customer inquiries into broad categories or high-level categories?
4. Given the "Customer Satisfaction Rating" in our dataset, how can we pre-processing this feature and how can we potentially use this feature in the models you are developing in Part II?

Part II

1. Design an AI or GenAI solution to categorize customer inquiries into different intent categories. How would you design an unsupervised learning model or generative AI model to discover new categories of customer inquiries, given that we don't have a clear categories list?
    - Given the "Customer Age" and "Customer Gender" fields, how might you personalize the AI model's responses to different customer segments?
2. How would you come up with new features to determine which queries should be escalated to human operators? For example, we can decide a ticket should be escalated if it has high or critical priority, can you think of something else?
    - Can we use this proxy 'Escalated' label to train a model that predicts whether a ticket is likely to need escalation? Or classifying which escalation level it is?

Part III (open questions)

1. Design an AI or GenAI approach to predict customer satisfaction (using the "Customer Satisfaction Rating" field) based on other features in the dataset.
4. Describe an approach to detect potential fraudulent activities, high-risk situations or complaint using the available data fields. How would you integrate this into the AI model's decision-making process for escalation to human operators?
5. Explain how you would implement a system that continuously learns and improves its responses over time, using the "Ticket Description", "Resolution", and "Customer Satisfaction Rating" fields.
5. How would you design a hybrid system that combines rule-based approaches for simple queries with more advanced machine learning models for complex tasks? Discuss the pros and cons of this approach.
6. Is there anything else you'd like to build and show us?

Additional notes for the candidate

The interview will include three sections: Introduction, Tech Deep Dive, and Take-home Solution Discussion.

- The Introduction session will take 15 minutes to discuss your work experience and projects.
- The Tech Deep Dive will take 20 minutes. The interviewers will ask a list of technical questions related to Gen AI, AI, and Data Science.
- Take-home Case Solution Demo:
    - You will have 10 minutes to demo your solutions.
    - Some follow-up technical questions related to your solution might be asked.
- For the take-home challenges: We are mainly looking for the thought process and approach followed by the candidate in solving open business problems using advanced AI, GenAI, ML, and Data Science.
- Please document your work and provide us with the results along with commented code.
- We will not use the candidates' work for commercial purposes outside the interview process.
- You will have one week to work on the data challenge using your tools and computing platform of choice.