ODH DATA ANALYST TEST1

PROBLEM STATEMENT

Scenario:
We are launching a new Wi-Fi optimization feature called ‘band steering’ which moves Wi-Fi connected devices (e.g., iPhones, laptops, tablets etc) between two different frequency bands (freq_band), 2.4 GHz and 5 GHz. This happens by moving devices to the different frequency bands at specified RSSI levels.

We need to model the impact of the new feature.

Step 1 - Evaluate Phy Rates at Different RSSI levels
Using the client stats sample data provided to you:

1. Look at the RSSI levels (rssi_percs_25) and look at the spread of Tx and Rx weighted Phy Rate
2. Produce a histogram (graph and CSV) for the average Tx and Rx Weighed Phy Rate between -85 and -65 dB in 1 dB steps (hint: this should produce 21 outputs)
3. Create a pipeline design and spec for a data engineer, so it is possible to produce this as a graph in Grafana
    - The source database should Elasticsearch or Graphite

Step 2 - Evaluate RSSI threshold

1. Using the client stats sample data provided to you calculate what % of devices are connected to 2.4 GHz and 5 GHz

2. Now model the a future hypothetical scenario that could be caused after the new Band Steering feature is activated. In this future scenario the following four conditions will be met:
    - Every Device connected to 2.4 GHz with an RSSI >= -60 dB connects to 5 GHz
    - Every Device connected to 5 GHz with an RSSI <= -75 dB connects to 2.4 GHz
    - Every Device connected to 5 GHz with an RSSI > -75 dB stays on 5GHz
    - Every Device connected to 2.4 GHz with an RSSI < -60 dB stays on 2.4GHz

    Calculate what % of devices that will be on 2.4GHz and 5GHz using above condition. Will we gain more devices on 5GHz?

EXPECTED OUTPUT

1. A short presentation, presenting your analysis results

2. A pipeline design document including pseudo code describing how a data engineer should create pipeline

3. Your Pyspark code

DATA SCIENCE PROBLEM

CUSTOMER CHURN RISK PREDICTION

Problem statement:

As a business we would like to understand what's the risk of our customers churning. We would like to create a predictive model that as an output per customer where we have a KPI that indicates the churn risk. How would you approach this task?

Notes, in this hypothetical senecio you should consider the following parameters:

- You have access to technical data (e.g. Wi-Fi Quality, network performance, …) of customers, you don't have access to You do not have access personal data such as age, contract detail and etc.
- Access to 30 days of data before customers Churn. You don't have any long term metric for these customers.
- Your model should fit for 6M customers using TB of daily data.

Desired output:

- A presentation with following information (Max 5 pages):
  - How would you approach this modelling?
  - How can the possible model look like?
  - What should we considering such problem?
  - How will you deploy this model? What can be a possible deployment of this?

Note: you are not expected to build a model/design using data, instead we would like a presentation explaining how you would approach this.

EXPECTED OUTPUT

1. A short presentation, presenting your analysis results

2. A pipeline design document including pseudo code describing how a data engineer should create pipeline

3. Your Pyspark code