ASSIGNMENT FOR TECHNICAL ASSESSMENT (DATA SCIENCE)

Analyse the data set with R or Python and their respective libraries. Then, answer the respective questions using proper visualization measures (graph, tables, etc.). Submit the report and the source code to the interviewer. Note that the questions are *open-ended* and the candidate should define the context when appropriate.

1. DATASET

The data set is collected in an office environment where a number of Bluetooth Low Energy (BLE) receivers are installed. A BLE beacon is then used as a transmitter to broadcast beacon signals while BLE receivers are receiving the broadcasted beacon signal. Based on the received signals, a wireless fingerprint is created and the fingerprint data can be used to train a model so that location of BLE beacon can be predicted. There are two datasets, Offline and Online dataset as explained below.

Offline Datasets: This dataset is collected for model training; it is also often called as calibration stage in literatures. In X_train.csv the header represents the MAC address of the BLE receivers and the rest of the data in it represents the correspond signal strength (in dbm) of the BLE beacon seen by that respective BLE receiver. Each row of X_train.csv representing a sample of data. Y_train.csv represents the corresponding location (or called Pin and labelled by PinId) when each sample data is collected. The actual X, Y coordinates (in meters) of each Pin can be found in pinInfo.csv. Offline data is collect from 35 locations with 45 samples at each location. Please refer to readme file in Offline folder for more detail explanations on the datasets.

Online Datasets: This dataset is collected for testing purpose. The data is collected in the same way as in offline phase, however, data collecting locations are random where it can be same as offline locations or falls in between them. The performance of the model will be measure using this dataset. The error of the model is calculated by measuring the Euclidian distance between the predicted location compared to ground truth location. Please refer to readme file in Offline folder for more detail explanations on the datasets.

You are required to optimise your prediction algorithm to get the best location prediction in X, Y coordinates that provide minimum distance error compared to ground truth.

- Please document your analysis and comment the code where ever necessary.
- It is required to explain your thought process and methodology used to achieve challenge metric.
- Visualize your interesting findings and parameter tuning if any.
- Analysis of the variables, models and other findings need to be presented in a report along with the coding. Please provide detail insight, recommendations and conclusions.
- Use your model to perform prediction using the data in ‘x_test_sumission.csv’ and save your result in the format as shown in ‘submission.csv’, both CSV files can be found in Online dataset folder. Your result will be compared to ground truths.

Marking Rubric

The assignment will be marked using the following rubric.

	Item	Description	Marks Allocation
1	Hypothesis	The data set may or may not link directly to the context of the questions and we expect a few assumptions to be made. The marks will be awarded based on the following criterions. - Invalid assumption – 0 ~ 3 marks - Less acceptable assumptions – 4 ~ 6 marks - Logic and comprehensible assumptions – 7 ~ 10 marks	10 marks
2	Methodology - Analysis	The candidate is expected to explain the analysis process in a comprehensive manner. - Invalid statistical analysis or mathematical model – 0 ~ 3 marks - Sounding analysis with minor ambiguity – 4 ~ 6 marks - Logic and comprehensive analysis – 7 ~ 10 marks	10 marks
3	Methodology – Machine Learning	The candidate is expected to explain the machine learning process in a comprehensive manner. - Invalid methodology – 0 ~ 3 marks - Sounding methodology – 4 ~ 6 marks - Logic and comprehensive methodology – 7 ~ 10 marks Please submit the Python or R code. Provide the flow chart if necessary.	10 marks
4	Result and Conclusion	The result and the conclusion must be aligned to the questions. - Result and conclusion do not align to the questions – 0 ~ 3 marks - Bad visualization (table or graph) but with satisfying result and conclusion or vice versa – 4 ~ 6 marks - Excellent visualization, result and conclusion – 7 ~ 10 marks	10 marks
5	Creativity	- The appropriateness of feature enrichment such as feature reduction, feature transformation and extra data pre-processing – 0 ~ 5 marks - The creativity and the application of external resource into the original dataset to improve the accuracy of model – 0 ~ 5 marks	10 marks
		Total Marks:	50 Marks