Threat Fabric Data Science Challenge

The goal of this data science assignment is to assess if you are comfortable with handling a data science case end to end. Hence, we would like to understand the steps you follow during the process how you transform raw data into a business objective result.

Description: In this assignment you are requested to analyze keystroke data. Keystroke recognition is a behavioural biometric which utilizes the unique manner in which a person types to verify the identity of an individual. Typing patterns are predominantly extracted from computer keyboards, but the information can potentially be gathered from any input device having traditional keys with tactile response (i.e., cellular phones, PDA's, etc). Although other measurements are conceivable, patterns used in keystroke dynamics are derived mainly from the two events that make up a keystroke: the Key-Down and Key-Up. The Key-Down event takes place at the initial depression of a key and the Key-Up occurs at the subsequent release of that key.

In this challenge, we provide you two datasets:

- Train_keystroke.csv: in this dataset the keystroke data from 110 users are collected. All users are asked to type a 13-length constant string multiple times (from 6 to 7) and the keystroke data (key-press time and key-release time for each key) are collected. The data set contains 750 rows and 27 columns. The last column indicates UserID, and the rest shows the press and release time for first to 13th character.
- Test_Keystroke.csv: This dataset contains 130 rows, and the structure is same as Train dataset except that the UserID is not included.

Objective: Determine UserID for each row in "Test_Keystroke.csv" dataset based on collected keystrokes data.

Deliverable: We ask you the following deliverables:

1. A CSV file containing UserID for "Test_Keystroke.csv"
2. The Code in Python (preferably in Jupyter Notebook with enough description). The code should have following elements:
    - Data Understanding part
    - Feature Engineering part
    - EDA part
    - Modeling
    - Evaluation
3. Explains to your colleague data scientist(s) (format not important) your code, e.g., techniques used, findings and data science recommendations.

If you need anu clarifications (e.g., definition), please do not hesitate to contact us on jobs@threatfabric.com.

You have maximum 10 days for this assignment, starting day after you receive this document as well as data.
Please send your output CSV file (#1), code (#2) and presentation (#3) to jobs@threatfabric.com.

We wish you a lot of fun with this challenge!