merimen online

Data Scientist Pre-Qualification Assessment

Block D, Level 1,
UPM-MTDC Technology Center 3,
Universiti Putra Malaysia
43400 Serdang Malaysia

SCENARIO

A bank collected behavioral data for approximately 1000 customers. These behavioral data can be classified as behavior1, behavior2 until behavior11. That is, a customer can be described through 11 behavior types. All these behavior data are transformed and represented as real values corresponding to their behavior and the bank is confident that these behavior data are sufficient to predict whether a customer is fraudulent.

Referring to the given excel data *FraudDetection_Dataset.xlsx*, data in sheet 1 ("Data") contain 1000 customer records with their behavior values from behavior1 until behavior11. It is not known if there is any fraudster in the list, but the bank knows that majority of their customers are genuine law-abiding customers, and not fraudulent.

Data in sheet 2 ("KnownFraud") contains 7 customer records that are confirmed to be fraudulent. They can be observed from the rows that are marked with red "FRAUD" in column M.

The goal of this challenge is to assist the bank in identify fraudsters from 94 customers. Detect those customers having high probability of fraudulent and flag them as "FRAUD".

In specific, attempt the following challenges and provide solutions to the bank:

Question 1

Identify those customers in sheet 3 that are to be labelled as "FRAUD".

Question 2

Name your methods that led you to the conclusion in Question 1.

Question 3

How many different groups of customers can you identify from sheet 1? How do you identify the groups?

Question 4

How do you evaluate the performances of the models you built?

Question 5

How would you react if the human expert cannot agree with the recommendation made by the model you built?