Semi-Final Coding Problem

Duration : 120 minutes

Problem Statement
A junior data scientist, John, start a project call “Rating Young Employees” for fun. He found and crawled employee data from multiple sources in Internet. Although there are a lot of data with different features, John decided to select 4 features: Birthday, Gender, Degree, Salary and Level.

Fortunately, his dataset is around 200K samples and half of them with label (“Rating”). Now John wants to build prediction model to label the samples without “Rating” but he doesn’t want to do it by himself but with your support.

Note that:
- 1st that he must exploring the dataset to understand them.
- 2nd his senior told him that the accuracy will be more than 75%.

Data Description

- Birthday: The birthday of employees with format “DD/MM/YYYY”
- Degree: There are 4 level of degrees (N/A = 0, Bachelor = 1,Master = 2,Doctor = 3)
- Salary level: There are 4 level of salary (Low = 0, Average = 1,High = 2,Super High = 3)
- Gender: There are 2 value for gender (Male = 0, Female = 1)
- Rating: There are 2 value for rating (Bad = 0, Good = 1)

Note that:
- Besides Rating (cleaned), all features with above format are what John expects from dataset but the real dataset isn’t as beautiful as that.

Question:
1. Data cleaning
    a. Cleaning and Standardizing “Birthday” as “dd/mm/yyyy” or “mm/yyyy” (when no day number)?
    b. Cleaning and Standardizing “Degree”?
    c. Cleaning and Standardizing “Sex”?

2. Data summary
    a. Distribution of “Birthday”?
    b. Distribution of “Degrees”?
    c. Distribution of “Sex”?

3. Data exploration
    a. Do Males have higher education level (degree)? Prove your answer by an actual data.
    b. For people who has degree higher than ***Master***, figure out distribution of their ages.
    c. Distribution of people who has the higher salary (level = 3)?
    d. Show correlation between all column in Train dataset.
    e. Comparing the characteristics of Train and Test dataset
    Hint: In term of correlation and distribution.

4. Prediction
    a. Build predictive models on Test dataset to classify the “Rating” (0 or 1)