CluePoints - Case

Introduction

As part of the recruitment process, two machine learning exercises have been selected to assess your skills in this field. First exercise is a classification use case with structured input data. Second exercise is a classification use case involving unstructured data, i.e. text.
To solve these exercises, you can use any tool or software that you like. The goal is not to reach the best performances but to show if you are able to properly apply machine learning techniques, write a clean code, understand the methods that you use, etc. These exercises are just a pretext to show your expertise in machine learning so don't hesitate to use complex techniques even if they are overkill.
You should be able to

- Shortly present your findings;
- Assess the performances of your solution;
- Describe the mathematical concepts behind your methods;
- Justify your implementation choices.

Exercise 1

Description

This exercise aims at predicting the length of stay in hospital for a patient. The prediction is done at the time of admission. For this, patient data is provided together with the target variable ("Stay") that indicates the range of the number of days that the patient spent in the hospital. Since some patients made several stays at the hospital, it's possible to have several records per patient.

Data

The labeled data is stored in the file `hospital_stay_data.csv`.
Here is an explanation of the features:

- `case_id`: Identification of a visit at hospital;
- `Hospital_code`: Unique code for the hospital;
- `Hospital_type_code`: Unique code for the type of hospital;
- `City_Code_Hospital`: City code of the hospital;
- `Hospital_region_code`: Region code of the hospital;

- `Available Extra Rooms in Hospital`: Number of extra rooms available in the hospital;
- `Department`: Department overlooking the case ward type;
- `Ward_Type`: Code for the ward type;
- `Ward_Facility_Code`: Code for the ward facility;
- `Bed Grade`: Condition of bed in the ward;
- `patientid`: Unique patient ID;
- `City_Code_Patient`: City code for the patient;
- `Type of admission`: Admission type registered by the hospital;
- `Severity of Illness`: Severity of the illness recorded at the time of admission;
- `Visitors with Patient`: Number of visitors with the patient;
- `Age`: Age of the patient;
- `Admission_Deposit`: Deposit at the admission time;
- `Stay`: Stay days by the patient.

Exercise 2

Description

This exercise aims at predicting the quality of Stack Overflow questions. For this, several features related to the question (like the title or the body of the question) posted on Stack Overflow are provided together with the target variable ("Type") that indicates the quality of the question (3 types of quality, see section Data).

Data

The training and validation data are stored in the files `stack_overflow_questions_train.csv` and `stack_overflow_questions_valid.csv`, respectively.
Here is an explanation of the features:

- `Id`: Question ID;
- `Title`: Title of the question;
- `Body`: Body of the question;
- `Tags`: Tags associated to the question;
- `CreationDate`: Date of the creation of the question;
- `Type`: Quality of the question with three possible values:
  - `HQ`: High-quality questions without a single edit;
  - `LQ_EDIT`: Low-quality questions with a negative score, and multiple community edits. However, they still remain open after those changes.
  - `LQ_CLOSE`: Low-quality questions that were closed by the community without a single edit.