Tasks Analytics Lab

Dear Siemens applicant, in order to give an impression of our typical problems we need to solve daily, we provided you some tasks that resemble our day-to-day business. Please answer the tasks below, if you do not have time to do all of them, please focus on the first two questions. We expect to receive your commented code as well as a short text description on your observations about the problems, the results and why you did the way you did it.

You are free to solve the problem in R or Python, where a language is not explicitly mentioned. If you do not have time to solve all four tasks, please focus on the first and the second task.

Good luck!

The Siemens Analytics Lab

TASK 1: Binary Classification

You are requested to solve a binary classification problem, for which both a dataset is provided. The dataset is split into two tables, which should be combined first. The “id” column is present in both tables and should be used for matching. The “class” column is the target variable.

TASK 2: Identifying alternative materials

The csv file attached consists of material description and technical attributes for several fuses. Your job is to find an algorithm that shows for every material 5 alternatives by means of similar materials. To achieve this goal, please conduct the following steps:

1. Load the data an make a descriptive analysis of the various attributes. Which difficulties do you see with the data. Name at least 2 findings and show potential solutions to circumvent them
2. Build a model that can identify similar materials based on the attribute Part Description. Explain your approach.
3. Outline the steps you would undertake to expand the model under 2. to also account for all the other attributes within the dataset.

TASK 3: Functions for Categorical attributes

The first rows of the dataset are given only as an example:

TICKET ID	COUNTRY	DELAYED (Y/N)
T3656	DE	Y
T1598	EN	Y
T2105	CH	N
…	…	…

1. Write a function with 2 input arguments
    a. the values of a categorical attribute like the values of Country in the above dataset
    b. a percentage between 0%-100%. The output should be as the first input with the values that occur less frequent than the input percentage replaced by a generic categorical value
2. Why could this function be useful in building a logistic regression model predicting a binary target like the Delayed (y/n) in the above dataset?
3. Describe shortly another approach, which could be useful to treat a categorical attribute with too many different values. Alternatively, describe briefly an algorithm that addresses internally this issue and does not require such a treatment of the categorical attributes.

TASK 4: Date Functions

1. Write a function that takes as input two timestamps of the form 2017/05/13 12:00 and calculates their differences in hours
2. Expand the above function to only count the time difference between 09:00 – 17:00 and only on weekdays.

Examples:

2017/10/13 19:00 minus 2017/10/13 17:10 = 0 minutes / 0 hours

2017/10/13 17:00 minus 2017/10/13 16:10 = 50 minutes / with rounding 1 hour

Note: Only the difference in hours needs to be returned

TASK 5: SQL

Table name: Customer

CUSTOMER_ID	COUNTRY	INDUSTRY
C1	DE	Automotive
C2	EN	Healthcare
C3	CH	F&B
…	…	…

Table name: Sales

SALE_ID	INVOICE_TOTAL	CUSTOMER_ID	DATE
S1	10	C22	2016/04/20
S2	56	C100	2016/04/22
S3	15	C1	2016/04/23
…	…	…	…
…	…		

1. Write a SQL query that displays the amount of customers per industry.
2. Write a SQL query that displays the average invoice total per industry.
3. Write a SQL query that displays what each customer spent per month, if that value is bigger than 100.

NOTE: Solve these tasks using SQL (or MySQL) code.