PROJECT WORK – INSTRUCTIONS

05/12/2023

Welcome to the final SML project work!

As a reminder, recall that your final SML grade will be a weighted sum of this project work (60%) plus the closed-book exam on theory (40%)

Note for PhD students: only project work (100%), no closed-book exam on theory

Let's begin!!

PROJECT WORK RULES

1. You can work alone or in team (max 3 people)
2. Deadline: December 19th 23:59
   For PhD students the deadline is January 19th 23:59
3. This pdf, together with the datasets, will be uploaded by the end of the day in the cloud folder "Project work"

If you never downloaded the lectures material, the cloud link is here: https://bit.ly/3rbsk9r and the password to access is: sml_gdp23!

EXPECTED OUTPUT

The expected output is a **jupyter notebook (.ipynb)**

Inside the notebook write the list of people in the team. **For each person in the team make sure to write:**

1. First name
2. Last name
3. Registration number (matricola in italiano)
4. Degree program (that is: SDS, QFI, PhD, Erasmus, …)
5. Email

If there is more than 1 person in the team, it's sufficient that only one person from the team sends me the notebook file.

Important notes:

1. The notebook has to be sent via email at the address gardino.gb@gdpanalytics.com before the deadline

2. To make sure that I've received your file, inside the cloud folder "Project work" there will be an excel file "received projects.xlsx" with all the registration numbers (matricola) of the students from which I have received the file. **If you are not in this list you'll not be admitted to the final closed-book exam on jan/feb.** Once you send me the file, you should see the registration number of the people in the team within 24 hours, **if not write me again!**

THE CHALLENGE

For this challenge, you will be the data scientist of a multinational retail company.

Company description: known for its extensive network of hypermarkets, supermarkets, and minimarkets, it has woven itself into the fabric of daily life for countless consumers worldwide. With a commitment to providing a diverse array of products, ranging from groceries to electronics, household goods, and beyond, this retail company caters to the diverse needs of its customer base.

THE DATASETS

The company gives you the access to two datasets: sales and market

sales.csv

This file contains the daily sales for each market. Here it's the structure:

Column name	Data type	Notes
market_id	string	Reference the column "id" in the "market.csv" file
date	date	Format DD/MM/YYYY Dates range from 01/01/2021 to 31/12/2022
is_open	string	YES: the market was open on that day NO: the market was closed on that day. If a market is closed then sales_amount has to be zero. No strings different than YES/NO should be present
sales_amount	float	Daily sales in €. It contains only actual sales, no refunds. This implies that only values >= 0 should be present

market.csv

This file contains the market registry (that is, one row per market). Here it's the structure:

Column name	Data type	Notes
id	string	Unique identifier of the market
country	string	Location (country) of the market
market_type	string	Type of the market. Can be MINI/SUPER/HYPER
square_feet	integer	Square feet of the market
avg_customers	integer	Daily average of customers in the market
competitor_distance	integer	Distance in meters from nearest competitor
has_promotion	string	YES: the market may have promotions during the year NO: the market cannot apply any promotion

OBJECTIVES OF THE COMPANY

The company is interested in 2 objectives. Below are the details.

Objective 1

Considering a linear relationship between target and predictors, the company is interested in understanding the delta changes (that is, a 1-unit increase in the predictor leads to a XX variation in the target).

Objective 2

The company in 2024 wants to open 3 new markets: 1 MINI, 1 SUPER and 1 HYPER.

There are multiple options to choose from.

Options for MINI markets

Country	Square feet	Estimated average customers	Nearest competitor distance	Will it run promotions?
SPAIN	1850	190	4500	YES
FRANCE	2100	215	1850	YES
ITALY	1920	220	1450	YES

Options for SUPER markets

Country	Square feet	Estimated average customers	Nearest competitor distance	Will it run promotions?
SPAIN	5880	420	580	YES
FRANCE	5120	390	2560	YES
ITALY	4970	410	3520	YES

Options for HYPER markets

Country	Square feet	Estimated average customers	Nearest competitor distance	Will it run promotions?
SPAIN	10560	860	8940	YES
FRANCE	12570	880	7580	YES
ITALY	11980	790	11560	YES

For this second objective the goal is to find for each market type which is the best one to open in 2024.

WHAT YOU SHOULD DO

First thing first, which is the target for this task? It's "sales_amount". Clearly, we want to open markets with the highest revenues!

Below you'll find some guided steps to solve both objectives.

Objective 1

- Data cleaning
  1. Read the both .csv file using pandas
  2. Print both files and see if they have been read properly: sometimes there are empty columns
  3. Check for NaN values and/or errors: be careful here! Read the above structure of the files and check column by column if you see something that shouldn't be there
- Generate the dataset – part 1: you may have noticed that "sales.csv" contains temporal information. To capture the main relationships between target and predictors you could collapse the temporal dimension
- Generate the dataset – part 2: now that you have a new dataset without the dates, you can merge it with "market.csv". Let's call this new dataset "merged_df"
- Data visualization: use "merged_df" to explore the data. Do you see any predictors that could help in explaining the target? Do you see any outliers?
- Encode categorical variables
- Finally, fit a linear model and answer to "objective 1". When answering, think about "feature selection" and "dimension reduction methods": do you think they can be useful? Explain why.

Objective 2

Starting from the latest dataset generated in the previous point, test all the models that we studied throughout this course. Pick the best model (explain clearly what it means "best model") and then apply it to the above options to select the most profitable market per each

type. Recall that the company is interested in opening 3 new markets: 1 MINI, 1 SUPER and 1 HYPER.

Important note: inside the notebook, use markdown to clearly write your comments. Make sure to always explain what you are doing and why. Interpret results as best as you can, show me that you are really understanding what you are doing… show me that you're not doing copy-paste from chatGPT!

ONLY FOR PhD STUDENTS

If you are a PhD student there is an extra objective: use historical data to predict **weekly** sales for each market. The predictions has to be done for the first 8 weeks of 2023 (so, for each market make prediction for: week 1 2023, week 2 2023, … week 8 2023)

GOOD LUCK!