Data Science

Problem 1: Predictive shopping list for Monoprix

Monoprix is a french retailer whose activity is primarily focused on cities. In order to reward its loyal customers, Monoprix started to deploy new innovative services. The goal is to provide new ways for the customer to interact with Monoprix, and help him in his daily life.

In this situation has Monoprix decided to launch a new vocal based experience, using smart speakers like Google Home. The purpose of the service is to build shopping lists by simply talking to the smart speaker. For example, you can ask « Remind me to buy egg » and the speaker will add eggs to the shopping list.

Monoprix wants this service to be intelligent, and asked us to create an algorithm that could learn from the customers habits and suggest them the products they might have forgotten to add on the list.

Therefore, Monoprix provides you with 10 years of purchase history, for the customers of their loyalty program. Also, you have at your disposal the full product catalog.

A. **First step**: Recommendation engine
   Question: How can you use Monoprix's data to build the recommendation algorithm ?

In view of the problem of building the recommendation algorithm for Monoprix, I would follow the pipeline steps to develop the recommendation algorithm:

1. I would clean and process the data (cleaning sales history, product reviews, etc.).

2. To gain insights, I would conduct a thorough investigative analysis exploratory and explanatory of the information to discover underlying patterns and relationships.

3. The data would best be partitioned with 70% dedicated to model training, 20% for testing purposes, and the final ten percent allocated towards evaluation of results.

4. **I would use a hybrid recommendation model approach based on user-based collaborative filtering and content-based filtering, due this approach can better capture user behavior, resulting in more accurate and personalized recommendations for users and also can generalize better than only one type of model recommendation (model approach based on user-based collaborative filtering). This hybrid model use cosine similarity and k-nearest neighbors.**

5. Would train the recommendation hydro model

6. After rigorously evaluating the model's performance in training and testing, I would consider implementing it in production more broadly if it proves sufficiently effective and generalist on the evaluation set.

B. **Second step**: Natural Language
Question: During the project, we realize that the product catalog is really dirty because products are wrongly named. Which solutions can you propound to correct products' names?

Given that the product catalog is "very dirty", a manual approach may not be viable, in this case I would follow a pipeline to clean the catalog data using NLP.

1. Text Preprocessing (Removing punctuations, special characters)

2. Tokenization (Division of product names into words)

3. Removal of Stopwords (Elimination of common words, "is", "the", "of", etc.)

4. Lemmatization and Stemming (Reduction of inflected or derived words to their base form)

. Correction of Specific Product Names in the catalog (For example: "Playstation XYZ Pro" which was actually "Playstation XYZ")

Problem 2: Measuring the effect of a marketing campaign

A Pharmaceutical company is trying to measure how much their last marketing campaign on a specific product has helped increase its sales and asks for your advice on how to do it. Previewing this, the Brand/Product manager already built a marketing campaign set up in which he separated one control region (specific region where he didn't roll out the campaign) from the others.

A. Assume the product has been on market in the last 2 years with a stable demand**. Explain a model you would advise the company to use and its main assumptions.**

For this problem, I recommend using difference-in-differences. If you include other factors on the stated result beyond the changes that were observed across the analyzed time period, your findings may be strengthened by using a difference-in-differences model in your analysis. One may assess the effect of the marketing initiative on sales between the experimental and control localities by comparing and contrasting the revenue changes between the locations that received the promotional push and the places that did not receive such exposure before and after the launch. The primary premise is that the treatment and control groups would

have eventually followed comparable paths in the absence of the campaign. This enables us to credit the marketing campaign's influence for any variations in sales growth. The DiD.

B. Assume now that the product is new, so that the campaign was a launching one. **In this scenario is it possible to measure the effect of the campaign on sales? If yes, what model would you suggest and why?**

Even in this case, campaign impact can still be determined by a controlled experiment known as a randomized controlled trial (RCT). In order to do this, regions are assigned at random to either get the campaign (treatment group) or not (control group). Any differences in the sales results between the treatment and control groups can be ascribed to the campaign's impact. Because RCTs help demonstrate causality and control for potential confounding factors, they are regarded as the gold standard for assessing the efficacy of therapies. Thus, in order to determine how the campaign affected sales in this particular circumstance, I would advise running an RCT.

Problem 3: Regression Analysis

A supermarket company has a new internal policy to not discriminate **significantly** salary according to the location of their employees. They gathered the data from all of their employees and want you to verify if they are already following the new policy.

Before answering the questions below take a look at the annexed dataset: (**1stPhase-SelectiveProcess-Data Science-Data Base.csv**)

a) Question: Describe how can you use the supermarket data to verify if employees from different locations have significantly different salaries ? (Include here how you are going to treat the variables before feeding into the model)

For this problem, if we only want to know if there is a statistically significant difference in salaries between the groups for the "locations" variable and the company is already following the new policy, we can take a simpler approach without needing to create a regression model. **We can simply use the t-test to check if the "locations" variable has a statistical difference in employee salaries (this method compares the distribution between the mean values in the groups).** If there is a significant statistical difference, we can then create a regression model to better understand the contribution of the "locations" variable to employee salaries.

Next, we will apply the student test:

```python
def eval_test_t_student(dataset, feature, target):
    """
    This method applies the t-test method.
    Args:
        dataset: DataFrame containing the data.
        feature: Name of the feature/column to compare groups.
        target: Name of the target variable.
    """
    # Splitting the DataFrame into two groups based on the feature
    unique_values = dataset[feature].unique()
    samples_by_group = [dataset[dataset[feature] == unique_value][target] for unique_value in unique_values]

    # Performing the t-test
    t_statistic, p_value = ttest_ind(*samples_by_group)

    # Printing the results
    print("t-test Statistic:", t_statistic)
    print("p-value:", p_value)

    # Conclusion based on the p-value
    if p_value < 0.05:
        print(f"There is a statistically significant difference between the groups of the variable {feature} and {target}.")
        return True
    else:
        print(f"There is no statistically significant difference between the groups of the variable {feature} and {target}.")
        return False

# Example usage:
# Assuming 'dataset' is the DataFrame containing the dataset
eval_test_t_student(dataset, "LOCAL", "SALARIO_MENSAL")
```

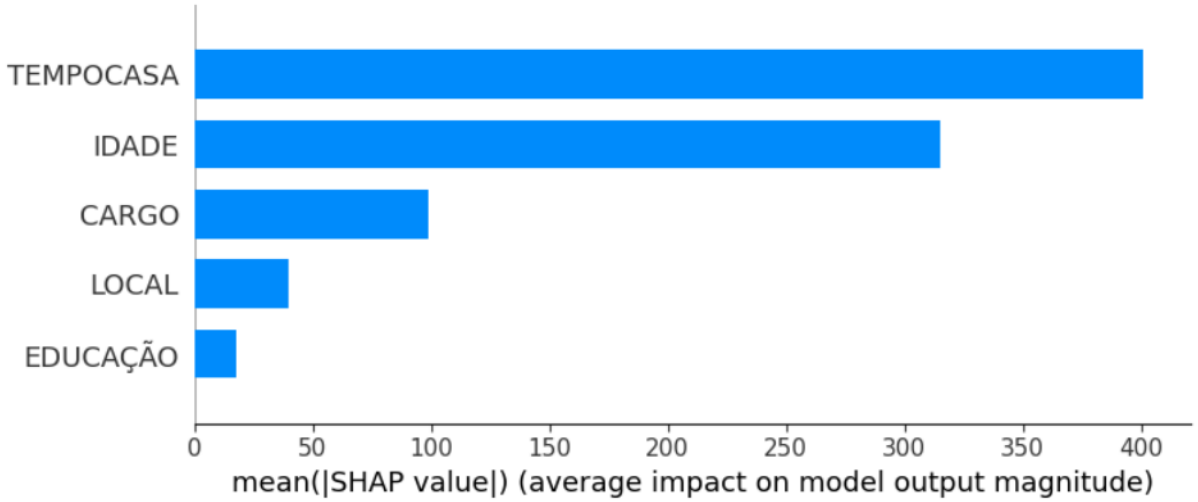
✓ 0.0s

```
t-test Statistic: -2.969717893005732
p-value: 0.0039592884372240028
There is a statistically significant difference between the groups of the variable LOCAL and SALARIO_MENSAL.
```

From the results of the p-value of the t-test student, we **can infer that there is a statistically significant difference between the "locations" variable and the employee variance variable.** Furthermore, we can already state that the company is not yet following the new policy. To perform feature engineering of the categorical variables within the dataset, I used target encoding, the target method avoids having sparse columns such as one hot encoding. In the next question I will implement a regression model with Random Forest to analyze the importance of the "locations" variable and **understand its variable importance with the shap values method and ajust R^2**.

Results:

Below is the final result of the shape values of the Regression Model, where we can see the result of the shape values of our trained model. The shape values help us understand how each variable contributes to the model's prediction for a given input



As we can see, the local variable is important for the model to predict employee salaries.

b) Question: Implement the approach you described in python or r:

Below is the code repository and its implementation description in Python:

https://github.com/GabrielFonsecaNunes/desafio_ds_artefact/blob/master/analise.ipynb