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?

Answer: To build a simple recommendation algorithm it's necessary to create a pre-processing stage, where at first you filter all the 10 years of data for each client ID. After that you'll have the purchase history separated by client and the next stages of processing will be the same for all of them.

What you should've do next it's to order the data of a specific client by the day that the shopping happened, assuming the database contains the items purchased and the day it occurred. By observing the distribution of data over the months it'll be possible to verify some patterns of shopping.

I believe that the weekly pattern will be useful to generate some good recommendation insights and from the assumption that the month contains 4 weeks (W1, W2, W3 and W4) we can group the data by this period. During this grouping we can count the number of times an item appeared in a specific week and order the items by the sum of appearances, so we'll have the customer data grouped by weeks of the month and for each week there will be an ordered list of the most common items that this client usually buys.

Finally, when a customer's shopping list reach the historic average of that week of the month, the algorithm of recommendation will compare the current items of the shopping list with the N most common items that he usually buys in that week, if an element isn't in his shopping list the algorithm will suggest that element.

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?

Answer: To correct products' names I propose the following, an iterative algorithm that runs over each element in the product catalogue, for each element we could've apply 2 functions.

The first function will be responsible to clean and standardize string format containing the product's name. This can happen using some python built in functions like *lower()* and *strip()*, generating an output.

So, the second function can use the Levenshtein Distance (or other metric to compare strings like Jaccard Distance) to compare the output of the previous function with the python's list of correctly spelled words found in the nltk library, generating a value of distance to transform a string into another (the smaller the distance value, the more similar they are). After compare the product's catalogue element with the nltk words, the correctly spelled word will have the lowest distance, and the output of this second function will be the autocorrection of the misspelled word in the catalogue.

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.**

Answer: In this case I would suggest an A/B test, where in the region B, more specific the control region, they don't run the new marketing campaign and it's expected that the demand of the product still stable the same. But in the region A they keep running the campaign for a period of test. At the end of the test period, let's say 3 months, they constate if the demand of the product in the A region have increased, decreased or still the same. If the demand has increased in the region A and has maintained stable in region B it's safe to say that the marketing campaign was effective by the percent of demand's growth in region A over region B.

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?**

Answer: It's hard to suggest the effect of the campaign in this case, because there isn't a past value of demand of this new product to verify if it's affected by the marketing campaign. What

they can verify is if a launching marketing campaign is better than none, by doing an A/B test where on the A region they apply the campaign and on the B region they don't apply it. After the product's demand stabilize in both regions it would be possible to constate if the launching campaign is worth or not.

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)

Answer: I would use the data to train and evaluate different ML models, and after I would select the best model by a specific metric.

From the selected model I would verify the predictions for employees but using the 2 different locations and by the values predicted I would verify if the salaries were significantly different for different locations.

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

Answer: The implementation of this question is located at the following repository: https://github.com/PortoSamuel/Processo-seletivo-Artefact. As you can note, the last commit it's on deadline.