ARTEFACT case - solutions

Diego Medrano Jiménez
Rua Gama Cerqueira
466, apto 920
Cambuci, 01539-010
São Paulo, Brasil

diego.medranojimenez@gmail.com
+55(11)95109-9379

Problem 1A: Recommendation engine

How can you use Monoprix's data to build the recommendation algorith?

The idea here would be to look for similar purchase patterns between the actual customer and the customers of the loyalty program using the 10 years' history.

Probably the products in the catalog vary from one year to another, so first of all I would group them using generic names and getting rid of the brands or additional information[1]

$$\textit{leite} \quad \leftrightarrow \quad \begin{cases} \text{Paulista leite 1.0\%} \\ \text{Leite Shefa integral 1L} \\ \dots \end{cases}$$

so that I can perfectly compare as well the actual customer with any other from 10 years ago.

First of all I would list the average of purchases of the data history for each customer per product and per month so that I have normalized quantities to look for similarities between customers; and I would look for the k-nearest neighbors of the actual customer using a cosine distance between the lists of those averaged purchases. From the similar neighbors I could now ponderate them and predict a possible shopping list for the actual customer.

Finally, to make a good recomendation, I just have to rank the obtained shopping list and offer the actual customer the higher ranked products if he has not already bought them.

Problem 1B: Natural language

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?

Again, using NLP algorithms[2], we can easily solve any spelling problem of the product names. On one side, *Novig's method* uses deletion, transposition, replacement and insertion of characters to recursively find the correct form of a word; alternatively, the *SymSpell algorithm* has a faster performance despite being based only on deletion of characters.

[1] This can be done for example through natural language processing (NLP) with dictionaries and "lemmatization & stemming" algorithms.

[2] Python NLTK library works perfectly for this kind of problems.

Problem 2A:
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.

In this case we have a history of sales of that specific product over the last two years with an stable demand. This means we can compute the total amount of sales per month to keep track of the time evolution of this quantity.

The first thing to do would be to make a linear (or polinomial, depending on the data) regression[3] of the data before the campaign so that we can predict what would be the expected amount of sales today without any campaign. After that, with the real amount of sales today once the campaign was launched, it is possible to compare both predicted and real values with their errors and estimate the benefit (or loss) of the campaign. Schematically:

- Before the campaign:

- History of sales → Regression with MLE[4] → Predicted amount of sales today ≡ Sales(pred)

- After the campaign:

- Actual amount of sales today ≡ Sales(act)

- Success rate of the campaign:

$$\Rightarrow \boxed{\text{Sales increased by } \left(\frac{\text{Sales(act)}}{\text{Sales(pred)}} - 1\right) \cdot 100\% \text{ thanks to the campaign.}}$$

For the regression we do not have to consider many assumptions. We just need that the price of the product does not change over that period of time. Apart from that, we can perform the regression directly to predict the total amount of sales during the first month after the campaign is launched.

Problem 2B:
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?

Since we do not have a history of sales we cannot perform the same kind of regression. However we still have tools in hand to measure the effect of the campaign. In this case we can use the results of the campaign to make a logistic regression, estimate how many sales there would have been on the control region after a campaign and thus, compare with the actual results of the control region. Again, schematically:

- Region where the marketing campaign was launched:

- Number of purchases → Logistic regression → Estimate on the control region ≡ Purch(pred)

- Control region:

[3] It can be done maximizing a likelihood function with the parameters of the model.

[4] Maximum Likelihood estimator.

- Actual number of purchases on the control region ≡ Purch(act)

- Success rate of the campaign:

$$\Rightarrow \boxed{\text{Sales increased by } \left(\frac{\text{Purch(pred)}}{\text{Purch(act)}} - 1\right) \cdot 100\% \text{ thanks to the campaign.}}$$

For the logistic regression to be reliable, I assume that for every potential purchase, the campaign collected enough data and basic information about the customer (gender, age, education, neighborhood status. . .) as well. The prediction ('purchase' / 'no purchase') is then performed for every potential customer on the control region so that we can finally compute the total amount of purchases to compare with.

If the input data of the logistic regression shows any kind of complex correlation, the prediction can also be performed with decision trees via random forests.

Problem 3A:
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)

I am going to work with a contrast hypothesis with significance level of 5% using the values in *LOCAL* and *SALARIO_MENSAL* in order to check whether the supermarket company is following the new policy described in the exercise. In particular, I am going to define the following Null-Hypothesis

$$\begin{cases} \bar{X}_{\text{salary}}^{\text{INTERIOR}} = \bar{X}_{\text{salary}}^{\text{CAPITAL}} & H_0 : \text{Null-Hypothesis} \\ \bar{X}_{\text{salary}}^{\text{INTERIOR}} \neq \bar{X}_{\text{salary}}^{\text{CAPITAL}} & H_1 : \text{Alternative-Hypothesis} \end{cases}$$

where each sample is characterized by the mean and variance of the salaries according to

$$\bar{X} = \frac{\sum_{i=1}^{|X|} x_i}{|X|} \qquad s^2 = \frac{\sum_{i=1}^{|X|} \left(x_i - \bar{X}\right)^2}{|X| - 1}$$

and follow these steps:

- Extract the values in *LOCAL* and *SALARIO_MENSAL* into a new dataset.

- Split the dataset into two different samples corresponding to the categories in *LOCAL* = ["INTERIOR", "CAPITAL"].

- Compute the quantity t_{stat} associated to the Student's t-test[5] where each sample has a different variance

$$t_{\text{stat}} = \frac{\bar{X}_{\text{salary}}^{\text{INTERIOR}} - \bar{X}_{\text{salary}}^{\text{CAPITAL}}}{\sqrt{\frac{\left(s_{\text{salary}}^{\text{INTERIOR}}\right)^2}{\left|X_{\text{salary}}^{\text{INTERIOR}}\right|} + \frac{\left(s_{\text{salary}}^{\text{CAPITAL}}\right)^2}{\left|X_{\text{salary}}^{\text{CAPITAL}}\right|}}}$$

[5] Also called *Welch's t-test* when variances of the samples are different.

- Compute the critical value z_α that corresponds to a significance level of $\alpha = 0.05$ according to the equation

  $$\alpha = 0.05 = 1 - \int_{-z_\alpha}^{z_\alpha} f_X(x)\,\mathrm{d}x = 1 - \int_{-z_\alpha}^{z_\alpha} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathrm{d}x\,.$$

  The value z_α can also be found on the internet, consulting integral tables for the t-Student distribution as a function of the number of degrees of freedom ν.

- Compare $|t_{stat}| \overset{?}{>} z_\alpha$.

- Accept or reject the Null-Hypothesis.

Check all the details in the attached file `salary_analysis.ipynb`.

Problem 3B:
Implement the approach you described in Python or R.

Check the attached file `salary_analysis.ipynb`.