Documentation of the fourth phase of the data mining project

Faridreza Mumtazzandi 9812762601

Alireza Nurbaksh 9812762496

1 –Introduction:

As we have seen in the last three phases, after recognizing the data, checking the quality of the data, performing the necessary pre-processing on it, and finally answering some real questions from the dataset, in the end, it is necessary to obtain the information related to three different issues. We can analyze or express our opinions in a specialized, complete, and correct way.

In this phase, our main focus is on the products that have been entered into the database, and we review these products according to time and organization. Our attention in this phase is on the PRODUCTINSTANCE dataset and we can find almost all the required features in this dataset.

In this phase, we would like to comment on the following three issues:

1. The ratio of the number of goods purchased to return to the warehouse by organization
2. Analyzing asset entry to asset exit by time and deputy
3. Status of similar products and price prediction of each product according to its cluster

We examine each case in order:

2 –The ratio of the number of purchased goods to the return to the warehouse by organization:

Contrary to the previous phase, where we focused on the input goods by holding, in this section, we want to do the necessary separation based on the organization, as we can see in the PDF file of the project, three characteristics have been selected, among which the appropriate characteristics are found to find the organization. We select the products that feature ID_REF_ORG_AD from the PRODUCTINSTANCE dataset.

In this section, returning to the warehouse means scrapping the goods, and for this purpose, it is a feature.

We use RETURNAMVALTOANBAR, which has four values and is fully explained in the main file of the project, and we will only say that the values 3 and 4 mean abort.

Finally, it is enough to obtain the number of scrapped goods for each organization, which is also fully mentioned in the main file codes.

3 –Analyzing asset entry to asset exit by time and deputy:

For this section, we will check the incoming goods whose arrival time is specified in the dataset with the CREATED attribute, with the UPDATED attribute, which refers to its latest changes, including their departure .

Finally, with a creative method, we will get a list of the date and time of all incoming goods and the time and date of their update, and we will also have their organization in hand. You can also refer to the main file for the codes of this section.

4 –The status of similar products and the price prediction of each product according to its cluster:

Finally, in the last part, it is necessary to first understand what is meant by similar goods, there are different definitions, such as goods that have the same type or goods that have the same price at the same time, which we consider this definition.

First, we separate the products of the PRODUCTINSTANCE dataset based on the input year and sort them according to their value obtained from the PRIMALVALUE attribute in each category. Now it is enough to creatively cluster all these products in such a way that if we did not know a product or wanted to predict it, we could easily predict it using this clustering. The code and description of this section are also available in the main file.

5 –Conclusion:

Finally, we reached the end of this huge project and saw how, despite the extremely messed up but real datasets, it is still possible to apply the practical methods of the data mining course and finally draw conclusions

and even provide answers to the important questions of the employer. Accurate and effective