Capstone project report

Marco Zenere July 9th 2021

1 Description of the problem/context

Würth is a world leader in the trade of assembly and fastening materials. Their catalogue contains a wide range of products and some of them are sold only during specific periods of the year and therefore are considered seasonal. The company would like to understand how to identify these type of products starting from customers purchases.

2 Description of the data

2.1 Invoices Table

	customer_id	customer_sector	customer_size_code	order_date	product_id	product_quantity	turnover_euros
0	994820	Dri	H	2018-08-09 00:00:00	511	1.0	22.00
1	645923	Fil	I	2017-11-02 00:00:00	511	1.0	19.00
2	169929	Fot	P	2019-05-17 00:00:00	511	1.0	21.90
3	113298	Fil	H	2017-03-09 00:00:00	511	1.0	18.85
4	93722	Fil	H	2016-12-05 00:00:00	511	3.0	64.95

Figure 1: Invoices Table

Figure 1 shows how the table containing the customers purchases is structured. The table contains the following information:

- customer_id: Customer identification
- customer_sector: Customer's working sector code. There are six different codes: 'Dri', 'Fil', 'Dip', 'Foo' and 'Mus'
- customer_size_code: Customer identification code. There are three different codes: 'H', 'I', 'P'
- order_date: Date of the purchase
- product_id: Id of the product purchased by the customer
- product_quantity: Quantity of products purchased by the customer
- turnover_euros: Total cost of the purchase made by the customer

2.2 County Table

	customer_id	province
0	3	BZ
1	4	LE
2	5	BZ
3	6	BS
4	7	BS

Figure 2: County Table

Figure 2 shows how the table containing the province of origin of each customer is structured. The table contains the following information:

- customer_id: Customer identification
- province: Province abbreviation

3 Description of the goals

The goal of the capstone project is to determine a rule that identifies seasonal products starting from customers purchases. Seasonal products are normally influenced by the weather, but one of the main problems is that the weather is not always the same year after year. Periods of heat and cold don't always begin in the same months.

4 Data cleaning

The table 'County' and some of the columns of 'Invoices' table were not used during the development of the code since there are not relevant for the type of analysis. In particular, the columns excluded from 'Invoices' table are: 'customer_id', 'customer_sector', 'customer_size_code' and 'turnover_euros'. To achieve the project goal, the idea was to use the clustering method called K-Means, but the table had to be restructured. Looking at the table, a product ID may appear in many rows and if data processing is performed, the same product may appear in multiple clusters and is an ambiguous result.
The table was structured considering a product per row and each row is structured as sort of time series with a monthly granularity . The time series is composed by 1s and 0s depending on if at least one product was sold in that period or not. It is a simplification because the optimal approach would consider a threshold computed considering the product sales trend but the data provider didn't provide this type of information. During the development, a quarter-time granularity was tested but the monthly option was chosen because the final result seemed more accurate.
An additional adjustment to the table was done to manage the non-seasonal products that seem just inserted into their catalogue. The time series of these products contain more 0s than 1s due to the long period considered (from 2015 to 2020) and the clustering method used for this project considers them as seasonal products. The problem was addressed by adjusting the time series so that they look close to a non-seasonal product' time series. The adjustment consists in converting all the values of the time series to 1 if the product was sold for at least 10 months during the last year. It's important to mention that the adjustment was kept only for the training of the model. The clustering output has been associated with the dataset without this adjusted so that in the test phase similar products find the match without having to adjust their time series.

5 Methodology

The goal of the capstone project was achieved using K-means clustering algorithm with 'n_cluster' parameter set to 2 (seasonal and non-seasonal clusters). The decision of using this approach was taken for the following reasons:

1. Data provided was unlabeled
2. After the data cleaning process, the table's columns were only of numerical type
3. The algorithm is relatively scalable and efficient in processing large data sets

In case the company would like to discover if a product is seasonal, they don't need to re-run the model because the results of the clustering were saved into two tables: one for seasonal product and one for non-seasonal products. The rule decided to classify the product is based on the Jaccard similarity measure: if the product is similar to at least one of the products in the 'seasonal' table for at least 70% (similarity by comparing the time series of the two products), the product is considered 'seasonal'.

6 Discussion of the results

By analyzing the output of the clustering algorithm, the results meet expectations. As mentioned in section 4, the result is the desired one after correcting the time series of non-seasonal products that appear to have been recently added to the catalog. Without this precaution, these products resulted in a category that in my opinion was ambiguous. Finally, the Jaccard coefficient rule seems to work well for the purpose of this project, and the decided threshold looks like a good compromise. It would be interesting to quantify the accuracy of the developed model through labeled sample data. During development, I asked if it was possible to obtain this data, but the company did not provide it to me so as not to influence my choices on the development and tuning of the model.