Máster de Formación Permanente en BIG DATA e Inteligencia Artificial

Universitat d'Alacant
Universidad de Alicante

Modulo 5
Analítica de datos y extracción de conocimiento mediante técnicas de IA

Task 4

Description:

Your company wishes to better understand the evolution of our sales globally.

More specifically, it is unclear whether we should create a VIP program for our clients, focus on particular customer segments or which markets, if any, we should pay more attention to. We also do not know if higher category shipments are being profitable or we are losing money instead and should drop them.

Try to help your company improve its situation through a data science process that answers the questions posed in this task.

Submit a pdf file that covers the process followed to process the dataset and answering the questions. **For each question** you should include (i) **a screenshot with the value obtained** that answers the question, (ii) **a brief answer that interprets the value** obtained.

Remember that the file submitted must **include the full name(s) of the student(s) involved**.

To do:

We start from a defined objective and a set of identified data. From here, **first of all**, you must **load the sales data into Jupyter**. You can do this **from the ORACLE database or** alternatively you can load it directly **from the CSV**. The database will give you the flexibility to execute SQL statements that will obtain the results directly. In the case of CSV, you will have to transform and operate with the initial dataframe to obtain the answers you are looking for.

Secondly, we are going to process the data, starting with a **basic understanding** of it **and detecting errors**. Perform **data profiling** and **answer the following questions**:

- Is each row an individual and complete shipping order?
- How many customers are in the data set?
- How many orders does each customer place on average?
- Are there columns with missing values? Which ones? Are they critical to the analysis we are doing?
- Are there columns highly correlated? Which ones? If there are correlations, do they all make sense?

After the initial data profiling task, we are going to carry out a brief **Exploratory Data Analysis (EDA**) to **understand the data at a deeper level**. Perform the necessary operations to answer the following questions:

- How much is the average shipping time of packages according to their ship mode?
- How much is the average profit that the company obtains per package ship mode? Is there any case where packages incur in loses?
- How much profit do the 10 best clients (profit-wise) generate compared to the rest?
- How much is the total profit per market?

Focus now on the market that provides the greatest benefit:

- How has profit evolved over time in this market?
- What are the best-selling categories (by quantity) in the region of that market that has placed the most orders?
- Are the best-selling categories those that generate the highest total profits?

Based on all the answers you have obtained, what recommendations would you make to the managers of your company in order to improve results?