Project 10: Market Basket Insights

Selecting a machine learning algorithm :

The Apriori algorithm is commonly cited by data scientists in research articles about market basket analysis and is used to identify frequent items in the database, then evaluate their frequency as the datasets are expanded to larger sizes.

Advantages :

Finding items that buyers desire to buy is the major goal of market basket analysis. Market basket analysis may help sales and marketing teams develop more effective product placement, pricing, cross-sell, and up-sell tactics.

Disadvantages :

It means that the businesses may not fully understand why certain products are frequently purchased together or how changes in one product may affect the sales of other products.

Applications :

Market basket analysis can increase sales and customer satisfaction. Using data to determine that products are often purchased together, retailers can optimize product placement, offer special deals and create new product bundles to encourage further sales of these combinations.

Training the model :

Data collection :

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

Data Mining :

It concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.

Data Preprocessing :

The Apriori algorithm was used for processing the input data and result was produced as the list of rules that are strongly associated with each other.

Data Extraction :

Data extraction is the first step that can be considered part of the data transformation process. The collection of data sometimes can be a hard work due each client has it own way to store the data. Oftenly, data is distributed among different resources and have different formats. Other times, data is poorly structured or even unstructured. All these aspects makes data extraction a hard task.

Data Cleaning :

Cleaning data is used in the process of removing data that is not relevant or needed as well. Part of the work is to know which information is relevant or can apport value to the algorithm and treat it for each specific case. Another common situation is data duplicated. Due databases are from big companies and comes from different sources sometimes the information is repeated. This provoke an overlap of information absolutely useless.

Apriori algorithm :

There are three components in APRIORI ALGORITHM:

* SUPPORT
* CONFIDENCE
* LIFT

**SUPPORT:**It is been calculated with the number of transactions divided by the total number of transactions made,

support(pen) = transactions related to pen/total transactions

i.e support -> 500/5000=10 percent

**CONFIDENCE:**It is been calculated for whether the product sales are popular on individual sales or through combined sales. That is calculated with combined transactions/individual transactions.

Confidence =   combine transactions/individual transactions

i.e confidence-> 1000/500=20 percent

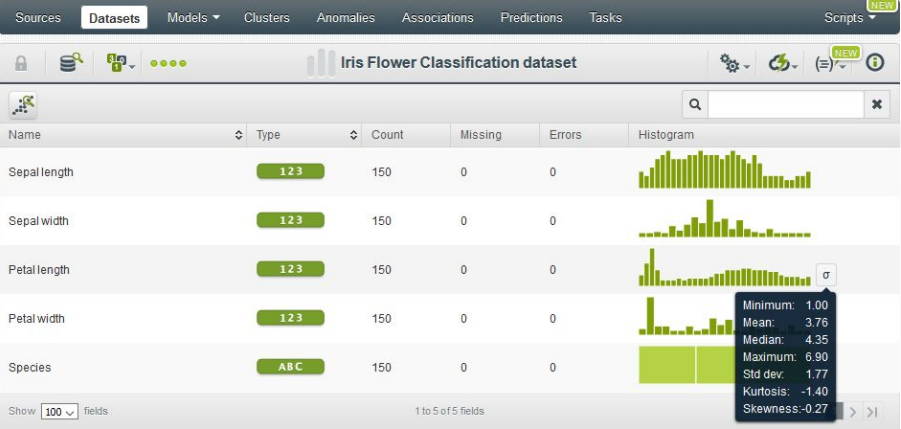
**LIFT:**Lift is calculated for knowing the ratio for the sales.

Lift-> 20/10=2

When the Lift value is below 1 means the combination is not so frequently bought by consumers.

But in this case, it shows that the probability of buying both the things together is high when compared to the transaction for the individual items sold.

With this, we come to an overall view of the Market Basket Analysis in Data Mining and how to calculate the sales for combination products.



Types of Market Basket Analysis:

1. It involves identifying frequent item sets and generating association rules that express the likelihood of one item being purchased with the purchase of another item. It is used to identify the relationships or associations between items in a transactional dataset.
2. This type of market basket analysis focuses on the order in which items are purchased in a transaction. It identifies frequent item sequences and generates sequential association rules describing the likelihood of one item sequence being followed by another.
3. This type of market basket analysis involves grouping similar items or transactions into clusters or segments based on their attributes. It helps to identify customer segments with similar purchasing behaviors, which can inform product recommendations and marketing strategies.

Tickets dataset :

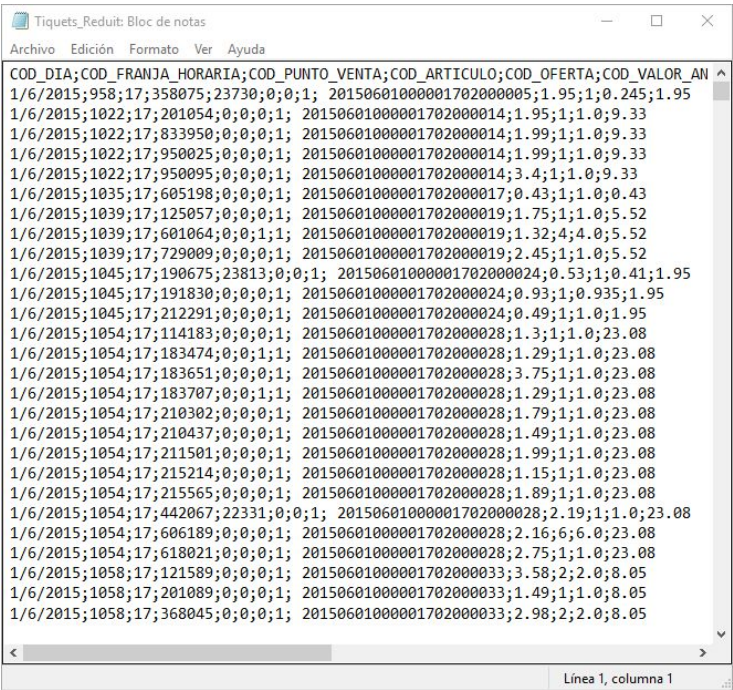
* Number of Instances: 214,712,174

● Number of Attributes: 13

● Missing Values? Yes

● Size of Dataset: 21.4 GB

* The first dataset used in this project was the historical tickets record. It was composed by 36,76,526 tickets from 203 different stores.
* Those tickets were expended during the last year, exactly from the June 2015 until May 2016.
* This is because we were interested in train our algorithms with newer data as possible.
* A setback we had to face with this dataset was the size of it. The file was considerably large and the client had problem to send it to us.
* Normally, when we work on project at CleverData, we have problems at the moment to pass the data from the client to our servers.
* The are tools and methods used for this types of problems, are common in BigData companies.



Puntos Venta dataset

● Number of Instances: 273

● Number of Attributes: 73

● Missing Values? Yes

● Size of Dataset: 226 KB

* The last dataset used contains information of each store the client posses.
* This information belongs to structural and un-structural variables like the size of the store, the location of it or the shop category it belongs .
* Our hypothesis was that shops can be similars due structural and un-structural variables. For instance, a shop with parking will have a higher mean ticket due the people go there on car and can take with him more products.
* Or the category of the shop can lead to have some specific products that others don’t have. All this information Influenced the behaviour of the shop.
* Our client tried in past occasions classify its stores using some metrics. However, due we didn’t know how were they created or based on, we decided to not use them in the clustering to not add noising information.
* In addition, the client was interested to see how the algorithm classified the stores, so we didn’t use those features.