# Capstone Term I (AIDI 1003-01)

# MODULE - 4

# Instacart Market Basket Analysis

# Instructor: Uzair Ahmed

*Team Members:*

*Elton Lionel Monteiro - 100799050*

*Sandeep Kour - 100801133*

*Sharvil Solanki – 100799051*

**1. Introduction**

**Problem Statement:**

Instacart’s data science team plays a big part in providing a delightful shopping experience. Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session. Product suggestions can be used to boost sales using promotional offers [1].

**Dataset Information:**

The dataset consists of a relational set of files describing customers' orders over time. The goal is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, between 4 and 100 of their orders’ data is available, with the sequence of products purchased in each order. We also provide the week and hour of day the order was placed, and a relative measure of time between orders [1]. Details about recency and frequency is available but monetary data is not. You can find the dataset [here](https://www.kaggle.com/c/instacart-market-basket-analysis/data).

**2. Executive Summary:**

Product suggestions for instacart based on data can significantly improve the business dynamics by using relevant information to predict the baskets and provide insights into the online transactions of customers, which can be analyzed to determine customized promotional offers to boost sales numbers [1].

Association Rules are widely used to analyze retail basket or transaction data and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules [2].

XGBoost is a supervised machine learning algorithm which is used both in regression as well as classification. It is an application of gradient boosted decision trees designed for good speed and performance. It stands for eXtreme Gradient Boosting [3]. Based on the prediction results, the model performance can be tweaked to get desired results

Models have been trained on various predictor features (x) that will describe the characteristics of a product and behaviour of a user regarding one or multiple products. This is achieved by analyzing the prior orders of the dataset. These results will then be fed to the models (XGBoost and Association Rule mining) to see their performance and getting desired output.

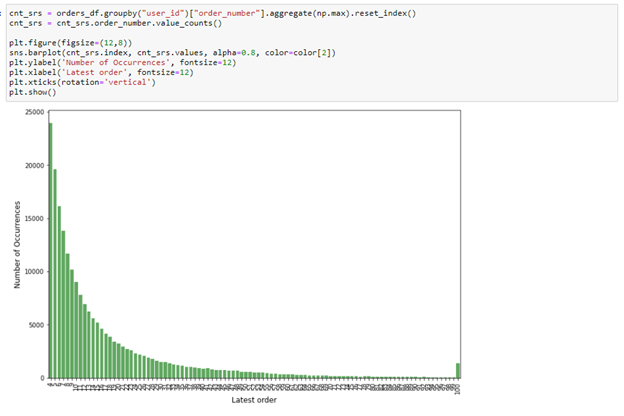
Models have to be trained for predictions, based on initial data exploration and analysis.The models under consideration here are XGBoost and Association rules mining as they are robust enough to handle complex tasks such as market basket analysis, which is the scope of this project.

**3. Evaluative Report**

**Exploratory Analysis:**

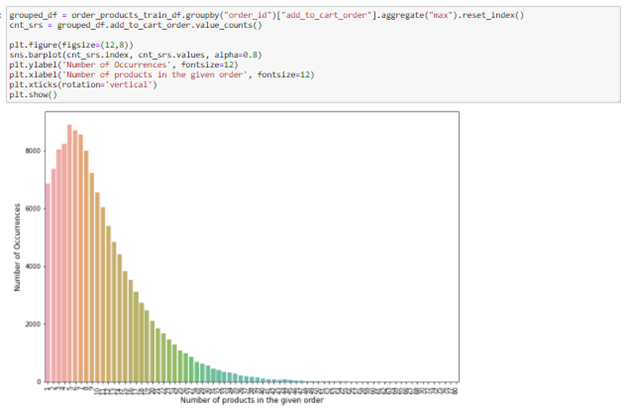
Exploratory Data Analysis (EDA) is an approach to analyse the data to discover patterns, summarize characteristics, spot anomalies and also to check hypotheses based on statistics and graphical representation [4].

Latest Order vs Occurrence:



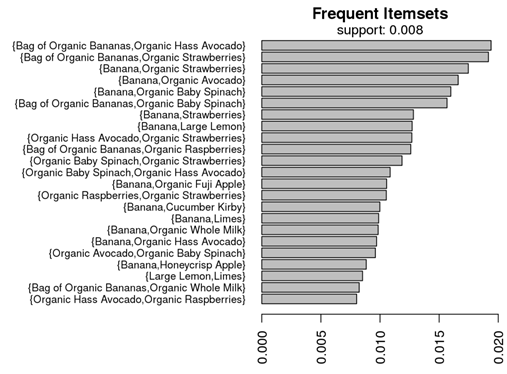
We can clearly see based on the latest order that most orders are small and large orders are less frequent, which makes the job of predicting slightly easy for the most part.

Number of products vs Occurrence:



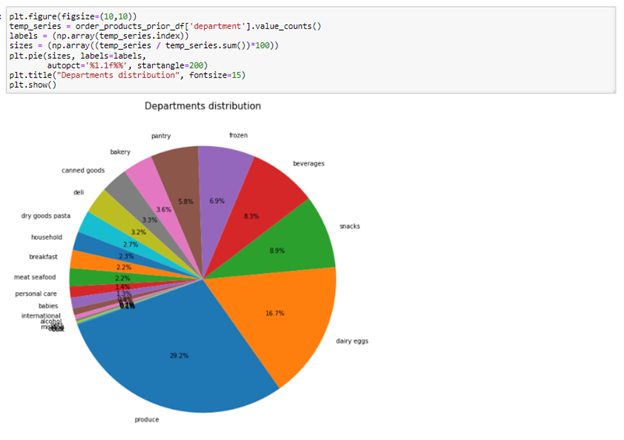
Overall size of an order and the number of products in each order are shown above, we can see that the most frequent size of the order is between 3-10 items per order.

Most frequently ordered items:



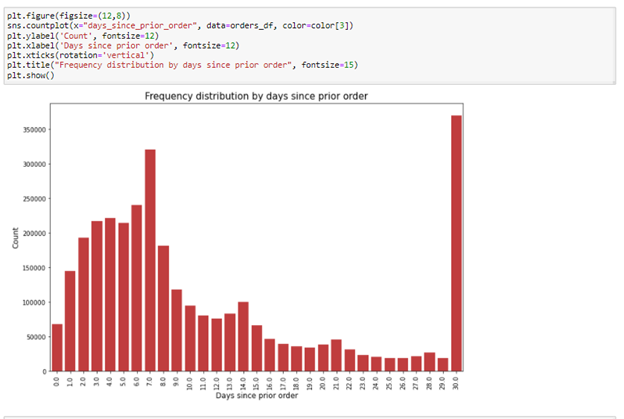
We can use this information to train the model based on the popularity of the items being ordered to improve the performance.

Department-wise popularity of products:



The most popular products by department are shown above and we can see that FMCG items dominate the chart with more than two thirds of the ordered quantity.

Days since Prior Order:



We can figure out the frequency of orders by customers. But caution has to be advised as this column has missing values, also all the orders made past the 30 day mark are being shown along with the orders made on the 30 day mark.

**Machine Learning Model:**

The goal of developing a predictive model is to develop a model that is accurate on unseen data. This can be achieved using statistical techniques where the training dataset is carefully used to estimate the performance of the model on new and unseen data.

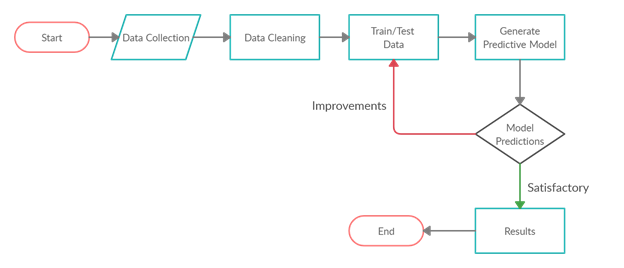
The simplest method that we can use to evaluate the performance of a machine learning algorithm is to use different training and testing datasets. We can take our original dataset and split it into two parts. Train the algorithm on the first part, then make predictions on the second part and evaluate the predictions against the expected results. This algorithm evaluation technique is fast. A downside of this technique is that it can have a high variance. This means that differences in the training and test dataset can result in meaningful differences in the estimate of model accuracy.

Cross validation is an approach that you can use to estimate the performance of a machine learning algorithm with less variance than a single train-test set split. It works by splitting the dataset into k-parts (e.g. k=5 or k=10). Each split of the data is called a fold. The algorithm is trained on k-1 folds with one held back and tested on the held back fold. This is repeated so that each fold of the dataset is given a chance to be the held back test set. After running cross validation you end up with k different performance scores that you can summarize using a mean and a standard deviation [5]

**Product Prediction:**

This is a classification problem because we need to predict whether each pair of user and product is a reorder or not. As a result, we need to come up and calculate various predictor variables (X) that will describe the characteristics of a product and the behavior of a user regarding one or multiple products. We will do so by analyzing the prior orders of the dataset. We will then use the train users to create a predictive model and the test users to make our actual prediction [6].

**4. Methodology**

Architecture:

**Technologies Used:**

AI projects differ from traditional software projects. The differences lie in the technology stack, the skills required for an AI-based project, and the necessity of deep research. To implement your AI aspirations, you should use a programming language that is stable, flexible, and has tools available. Python offers all of this, which is why we see lots of Python AI projects today.

Python offers concise and readable code. While complex algorithms and versatile workflows stand behind machine learning and AI, Python’s simplicity allows developers to write reliable systems. Developers get to put all their effort into solving an ML problem instead of focusing on the technical nuances of the language [7].

Implementing Association Model based on Market Basket Analysis:

Association Rule Mining is one of the ways to find patterns in data. It finds:

● features (dimensions) which occur together

● features (dimensions) which are “correlated”

We can measure effectiveness of the rule are as Follows:

● Support

● Confidence

● Lift

● Others: Affinity, Leverage [8]

Implementing XGboost for prediction:

1. Import the ready features from EDA notebooks and reshape data: This step includes loading files into pandas DataFrames and some data cleaning.

2. Create the test and train DataFrames: In this step we create two distinct DataFrames that will be used in the creation and the use of the predictive model.

3. Create the predictive model: In this step we employ XGBoost algorithm to create the predictive model through the train dataset.

4. Apply the model: This step includes applying the model to predict the 'reordered' variable for the test dataset.

**5. Modelling Results:**

**XGBoost**

XGBoost is a supervised machine learning algorithm which is used both in regression as well as classification. It is an application of gradient boosted decision trees designed for good speed and performance. It stands for eXtreme Gradient Boosting [3]. XGBoost is an algorithm that has recently been dominating applied machine learning

The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model. Some key algorithm implementation features include:

* Sparse Aware implementation with automatic handling of missing data values.
* Block Structure to support the parallelization of tree construction.
* Continued Training so that you can further boost an already fitted model on new data.

The two reasons to use XGBoost are also the two goals of the project:

1. Execution Speed.
2. Model Performance.

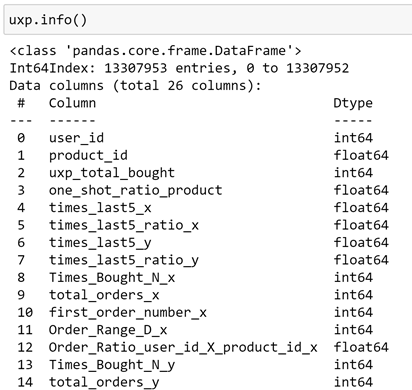
The XGBoost library implements the gradient boosting decision tree algorithm. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

This approach supports both regression and classification predictive modeling problems [2].

**Results**

The objective of this part is to predict the future behavior (which products they will buy) based on the features that we have created in our trained model.

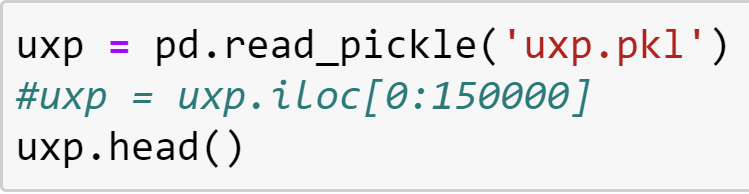
* Features



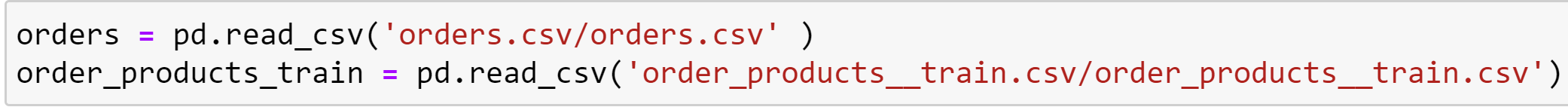
* Pickle and CSV file



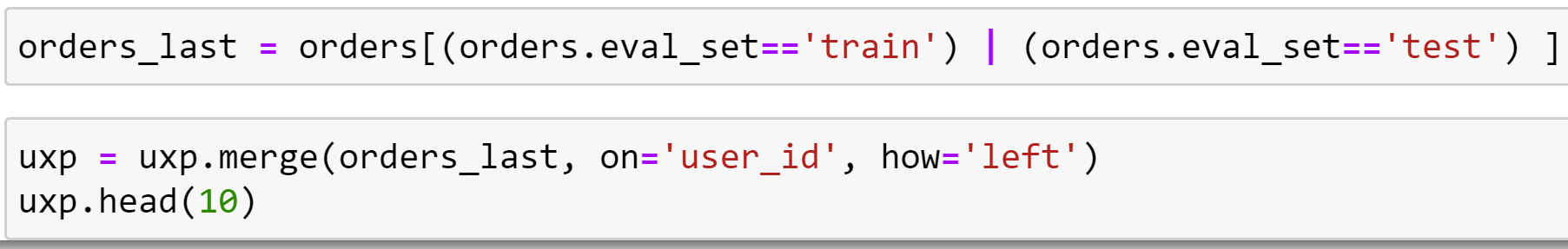
* Load the pickle file that contains the features that we created



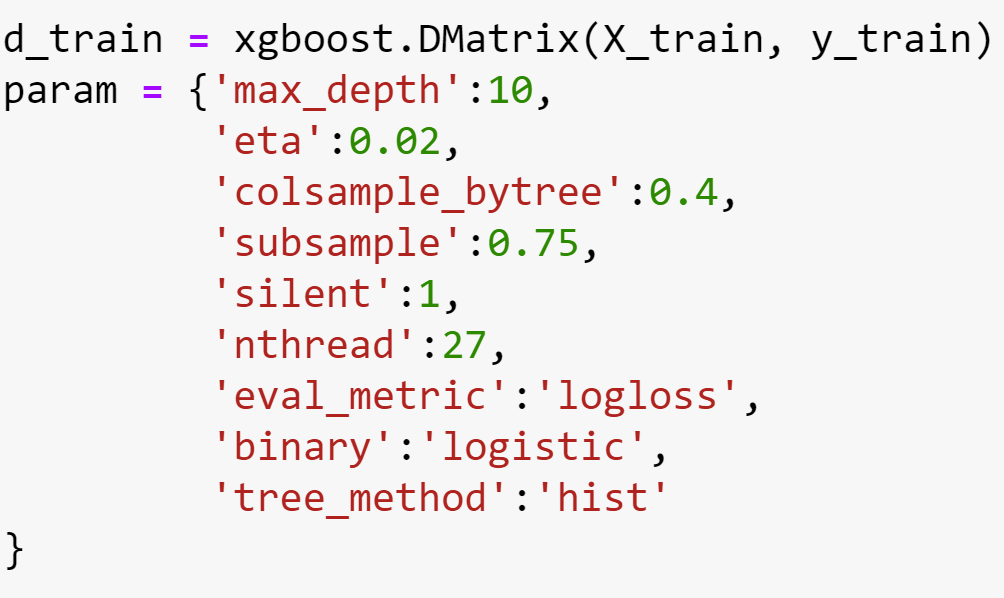
* Load the .csv files of orders and the products that have been purchased



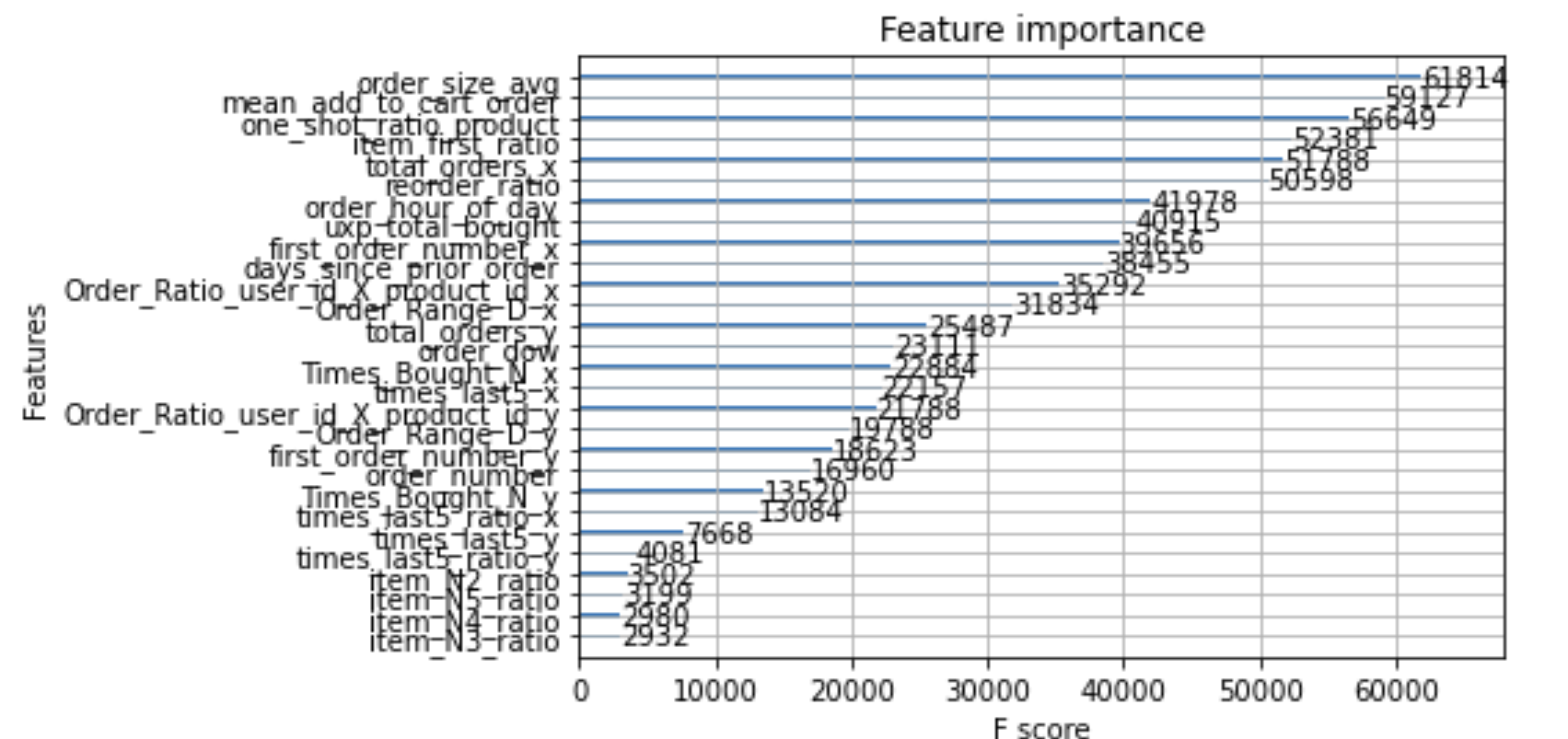
* Keeping only train and test orders, excluding the prior orders



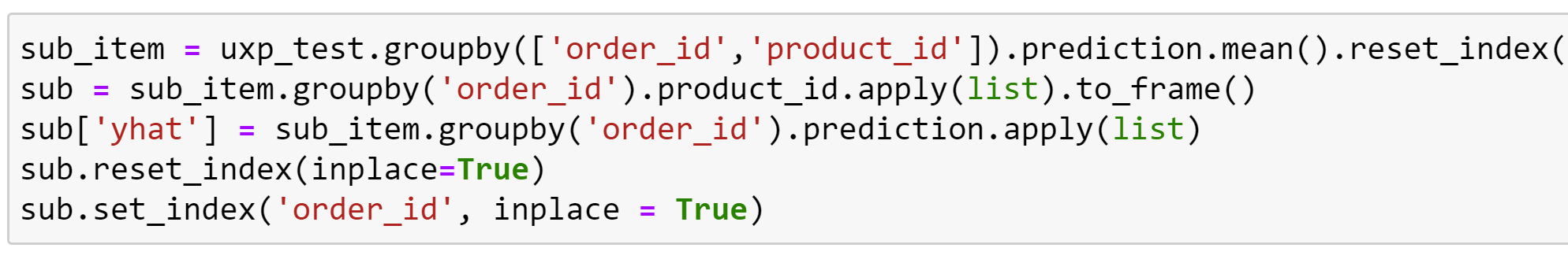
* Employ XGBoost algorithm to create the predictive model through trained dataset

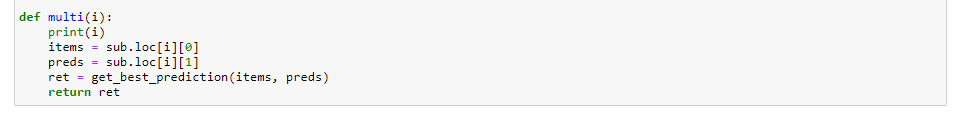


* Features

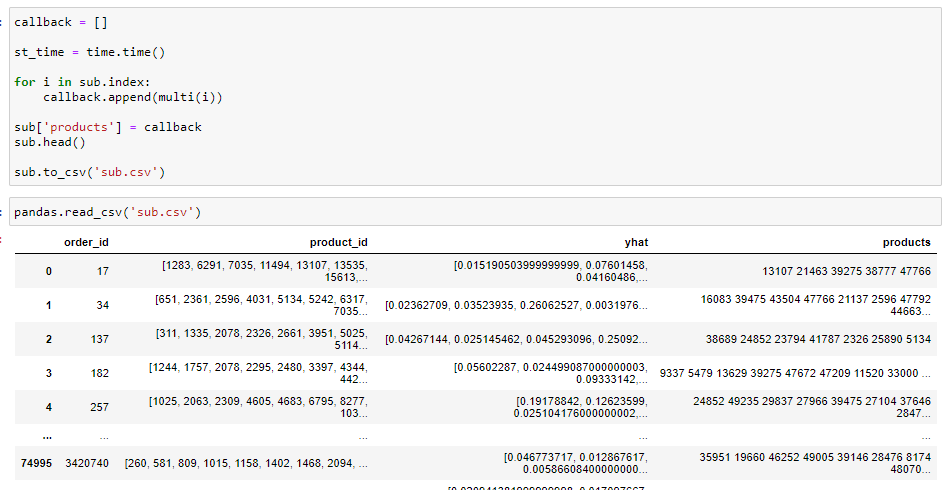


* For each order\_id predict what products could be purchased based on previous orders and predicted values.





* Snapshot of final prediction of products



**Association Rules Mining**

With the rapid growth of e-commerce websites and general trend to turn towards data for answers across industries (especially retail), every organization is trying to find more opportunities for best product bundles to run discounts and promotions on.

In return for these decisions is the expectation is the growth in sales and reduction in inventory levels. Performing the analysis on “what is bought together” can often yield very interesting results.

Association rule learning is a rule-based method for discovering relations between variables in large datasets. In the case of retail POS (point-of-sale) transactions analytics, our variables are going to be the retail products. It essentially discovers strong associations (rules) with some “strongness” level, which is represented by several parameters.

We can measure effectiveness of the rule are as follows:

* Support
* Confidence
* Lift
* Others: Affinity, Leverage

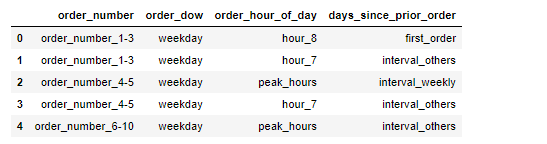
We will be looking at the Apriori algorithm, for the association between products, which is one of the most common techniques in Market Basket Analysis. It is used to analyze the frequent itemset in a transactional database, which then is used to generate association rules between the products [3].

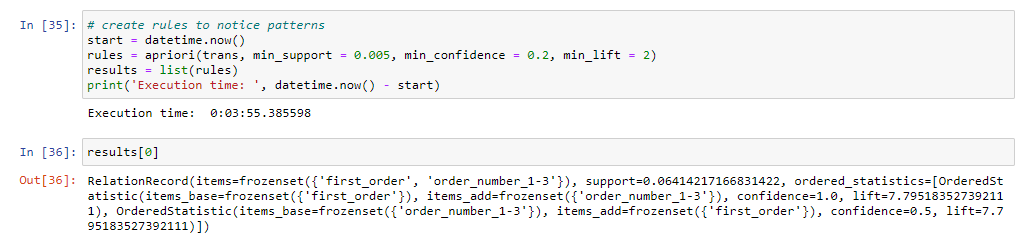
Once item pairs have been identified as having positive relationships, recommendations can be made to customers to increase sales.

**Results:**

The data sets have been loaded with all necessary columns needed to build the association rules. (Please Refer the Python notebooks included).

We can now build the association rules by converting the numerical values to categorical variables in the orders table.

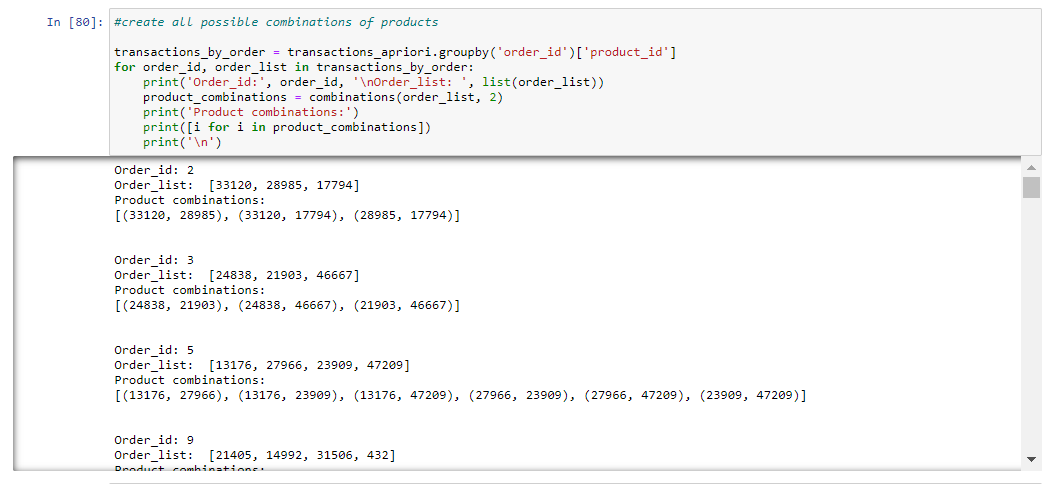


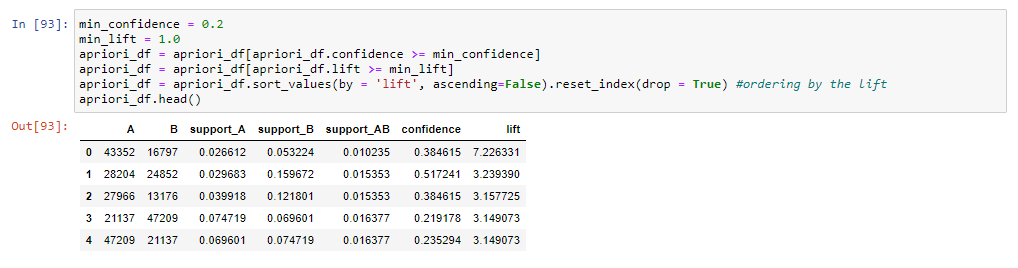


To use the Apriori algorithm it will be necessary to transform the data frame into a list.

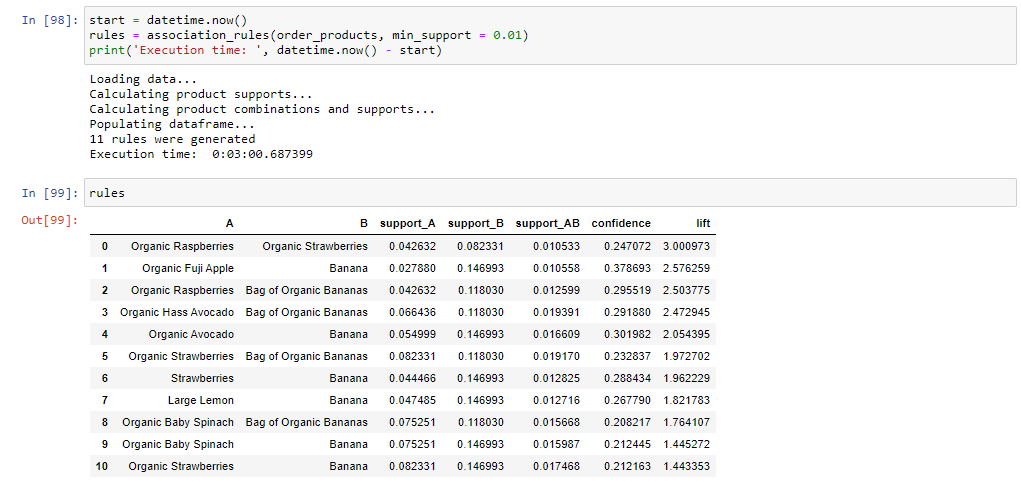
Apriori is an algorithm for frequent item set mining and association rule learning over large datasets. It proceeds by identifying the frequent individual items in the dataset and extending them to larger and larger item sets. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the data.

Create all possible combinations of products that will likely be ordered together, which will be used to define rules.





After we calculate Support, lift and confidence, we can now define the rules to predict the list of items.



We were able to see the presence of Banana and organic foods strongly. We can conclude that customers tend to take organic food accompanied by Banana in their purchase. Here the lift and the confidence already show higher values. If we analyze the first few rules, we see that the proportion of one occurring, and the other also occurring is very high, showing a preferable relationship between such products. It shows the habit of occurrence among the products, where one is taken, the chances of taking the other is high.

**6. References**

1) <https://www.kaggle.com/c/instacart-market-basket-analysis/data>

2) <https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce>

3) <http://dataanalyticsedge.com/2019/11/23/xgboost-using-python/>

4) <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

5) <https://machinelearningmastery.com/evaluate-gradient-boosting-models-xgboost-python/>

6) <https://www.kaggle.com/kokovidis/ml-instacart-f1-0-38-part-two-xgboost-f1-max>

7)  [https://steelkiwi.com/blog/python-for-ai-and-machine-learning/#:~:text=Python%20](https://steelkiwi.com/blog/python-for-ai-and-machine-learning/#:~:text=Python%20offers%20concise%20and%20readable,developers%20to%20write%20reliable%20systems.&text=Python%20code%20is%20understandable%20by,build%20models%20for%20machine%20learning)

8) <https://towardsdatascience.com/association-rule-mining-be4122fc1793>