Market Basket Analysis – Instacart

STEPS to run notebook

1-Create a directory or use existing directory

2-Place Following Notebooks In that directory

- Instacart Market Basket Analysis – Which has Data Analysis (EDA)
- Apriori Implementation - Apriori Algorithm Implementation which is used to find frequvancy of the product
- Instacart FeaturesXG – all the models are implemented
- AutoML - Implemented H2O and Tpot
- Kaggle Best – Which has features from the best kernal available on Kaggle

3-Download all csv files from https://www.kaggle.com/c/instacart-market-basket-analysis/data

4-Copy these files in the same folder as your jupyter notebook

5- You need to install a few packages for XGB, LGBM,Tpot using pip or conda install

6- Installation guide can be found here

https://xgboost.readthedocs.io/en/latest/build.html

https://pypi.org/project/lightgbm/

https://catboost.ai/docs/installation/python-installation-method-pip-install.html

https://pypi.org/project/imbalanced-learn/

7- Run jupyter server from command prompt

Importing libraries and reading csv's

The libraries that we will be using are:

Numpy : NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas : Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Matplotlib : Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

Scikit-learn : Scikit-learn is a machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN.

Keras : Keras is a deep learning library that: Allows for easy and fast prototyping (through user friendliness, modularity, and extensibility). Supports both convolutional networks and recurrent networks, as well as combinations of the two.

We shall be loading all the above libraries and several of their features which we will be using.

List of files imported and loaded

- Aisles.csv – This contains the names of the aisles based on the products in them.
- Departments.csv – It has the names of department categorized by products types.
- Order_Product_Prior – It has details of all the previous customer orders.
- Order_Product_Train.csv – This is the dataset which will be used to train the test dataset explained next.
- Orders.csv – It is the main table containing details about the customer orders and also tells which record belongs to which table, train, prior or test.
- Products.csv – This contain detail of all the products sold by Instakart along with their ProductID.