Supermarket Dataset for Predictive Marketing 2023

Dataset:

The dataset includes more than 2 million records of purchases made at the widely recognised Hunter's supermarket. Hunter's E-grocery is a well-known and rapidly expanding new-generation lifestyle brand. They are popular in 10 counties and are constantly seeking novel approaches to advance and foresee the needs of their clients. The way people shop in stores has undoubtedly been impacted by black swan events like Covid-19, the Ukraine crisis, and gas shortages. The dataset is a perfect sample to propose a business value for informative-based decision-making. The dataset contains information on about 2019501 unique purchases based on 12 attributes. Further details of these attributes are as follows:

- order_id – (A unique number to identify the order)
- user_id - (A unique number to identify the user)
- order_number – (Number of the order)
- order_dow – (Day of the Week the order was made)
- order_hour_of_day – (Time of the order)
- days_since_prior_order - (History of the order)
- product_id – (Id of the product)
- add_to_cart_order – (Number of items added to cart)
- reordered – (If the reorder took place)
- department_id - (Unique number allocated to each department)
- department – (Names of the departments)
- product_name – (Name of the products)

Data Pre-processing:

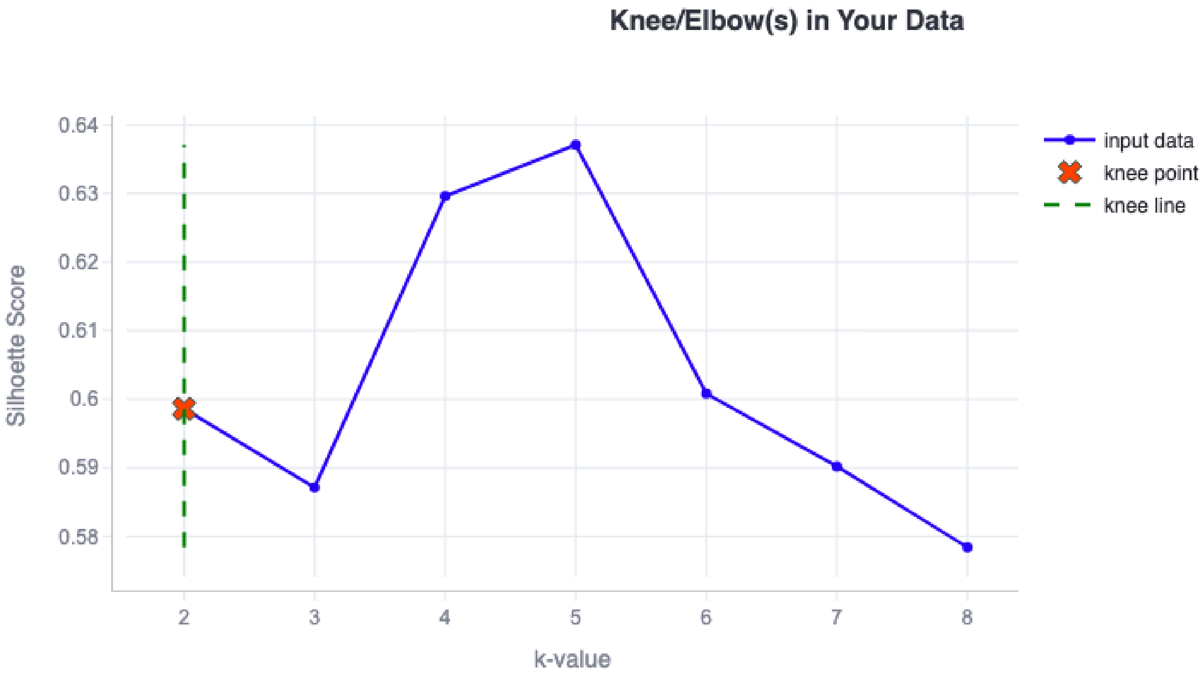
The dataset is first cleaned by removing all the NaN values and replacing them with the median values of that column. Furthermore, unique features and indexing are also cropped out. Since the model only reads data in numerical form, the categorical data are transformed into numerical labels for the model's convenience in processing the data.

Methodology/Workflow

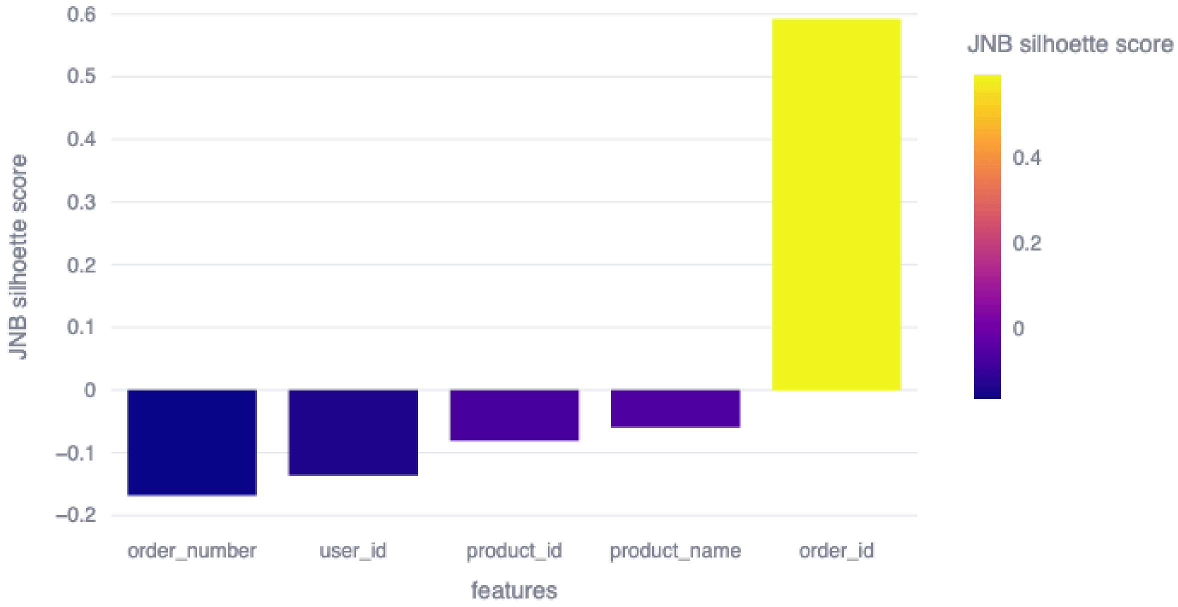
All the attributes are split into either numerical or categorical types based on their clustering threshold,i.e., if the number of unique values is < 40, then the feature is considered categorical. Among all the 12 features, 5 were tagged as numerical features, and the rest 7 as categorical in nature. The numerical features are then clustered into 2 (calculated through the elbow method) groups using JNB clustering [2]. The features are ordered according to their silhouette scores [3] and the top 75% are selected for further steps, since the total number of attributes are really less in this case. All possible subsets of these feature spaces are plotted into a feature matrix [4]. For this case study, we considered the combination of "product_id " and "user_id" with a silhouette score of 0.6758. The entropy of all the categorical features and the selected set is compared in Figure [5].

To create each cluster of the chosen numerical features, K-means clustering was employed. This is then the joint entropy calculated against all the categorical features. The maximum entropy is about 0.35 for 'days_since_prior_order', and the minimum is 0.05 for "order_hour_of_day". The user can select one or more categorical features from it, depending on the situation.

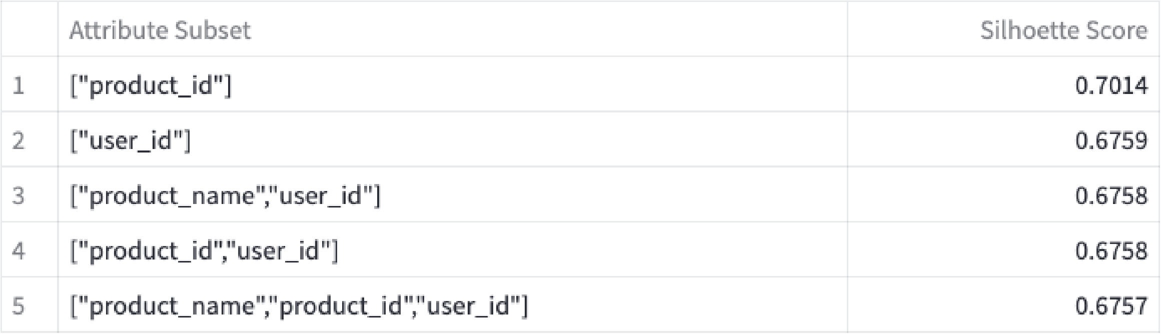
The k-values silhouette score graph, showcasing the elbow value [2]



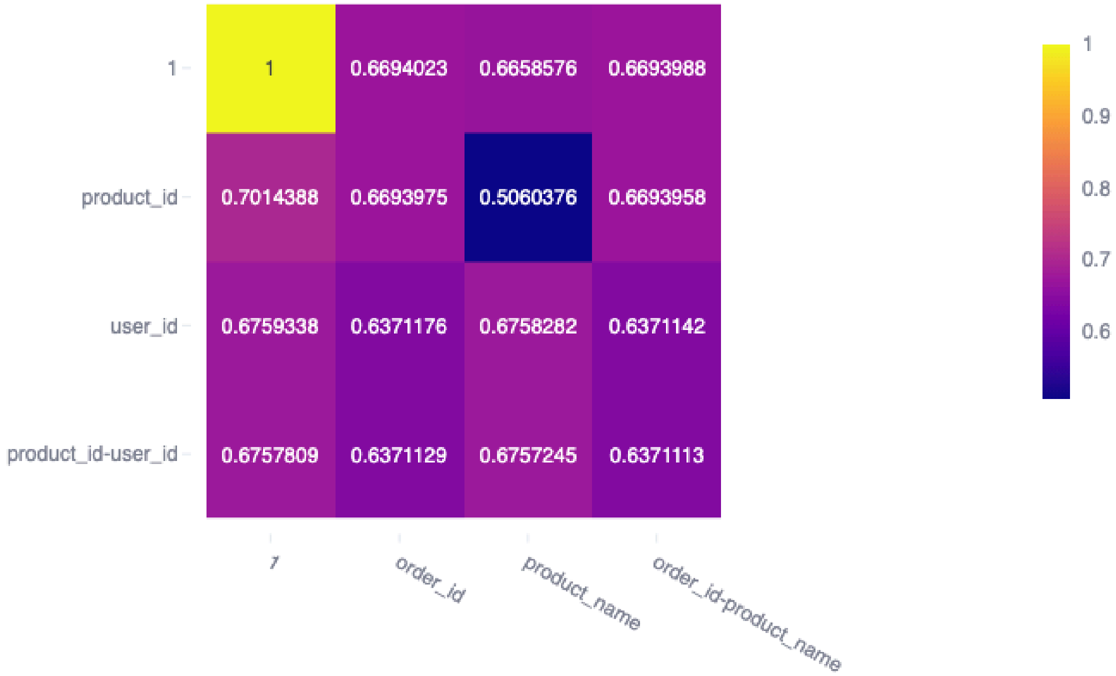
The Silhouette scores of JNB Clustering are based on the features [3]:



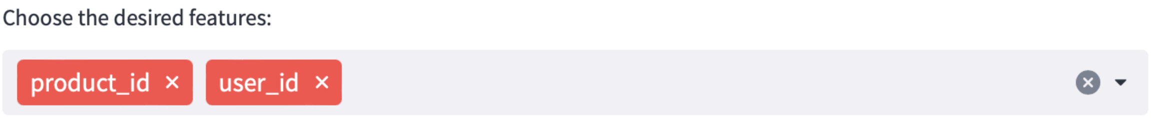
Feature matrix for case study 3 [4]:



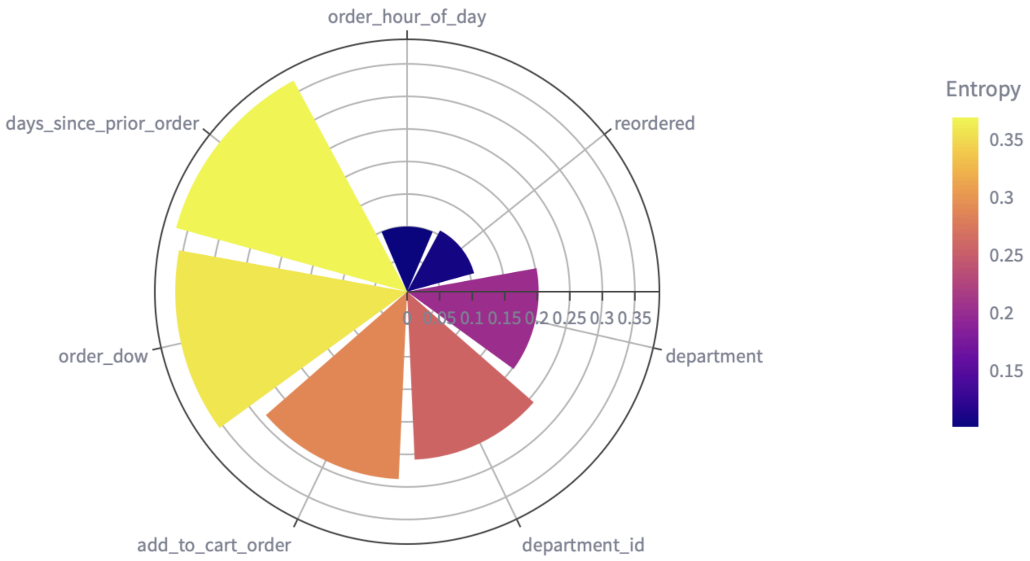
	Attribute Subset	Silhoette Score
1	["product_id"]	0.7014
2	["user_id"]	0.6759
3	["product_name","user_id"]	0.6758
4	["product_id","user_id"]	0.6758
5	["product_name","product_id","user_id"]	0.6757



The entropy of each categorical feature w.r.t. The selected numeric feature/s[5]


Choose the desired features:

product_id × user_id ×



Reference

[1]https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023