Market Basket Analysis

Apriori Algorithm

What is Market Basket Analysis?

The retailer wants to target customers with suggestions on itemset that a customer is most likely to purchase .I was given dataset contains data of a retailer; the transaction data provides data around all the transactions that have happened over a period of time. Retailer will use result to grove in his industry and provide for customer suggestions on itemset, we be able increase customer engagement and improve customer experience and identify customer behavior. I will solve this problem with use Association Rules type of unsupervised learning technique that checks for the dependency of one data item on another data item.

Strategy

- Data Import
- Data Understanding and Exploration
- Transformation of the data – so that is ready to be consumed by the association rules algorithm
- Running association rules
- Exploring the rules generated
- Filtering the generated rules
- Visualization of Rule

Algorithm Used

The Apriori algorithm in data mining is a popular algorithm used for finding frequent itemsets in a dataset. It is widely used in association rule mining to discover relationships between items in a dataset. The Apriori algorithm was developed by R. Agrawal and R. Srikant in 1994.

Add a little bit of body textThe Apriori property is a fundamental property of frequent itemsets used in the Apriori algorithm. In other words, if an itemset appears frequently enough in the dataset to be considered significant, then all of its subsets must also appear frequently enough to be significant. For example, if the itemset {A, B, C} frequently appears in a dataset, then the subsets {A, B}, {A, C}, {B, C}, {A}, {B}, and {C} must also appear frequently in the dataset

Steps in Apriori Algorithm

- Define minimum support threshold - This is the minimum number of times an item set must appear in the dataset to be considered as frequent. The support threshold is usually set by the user based on the size of the dataset and the domain knowledge.
- Generate a list of frequent 1-item sets - Scan the entire dataset to identify the items that meet the minimum support threshold. These item sets are known as frequent 1-item sets.
- Generate candidate item sets - In this step, the algorithm generates a list of candidate item sets of length k+1 from the frequent k-item sets identified in the previous step.

- Count the support of each candidate item set - Scan the dataset again to count the number of times each candidate item set appears in the dataset.
- Prune the candidate item sets - Remove the item sets that do not meet the minimum support threshold.
- Repeat steps 3-5 until no more frequent item sets can be generated.
- Generate association rules - Once the frequent item sets have been identified, the algorithm generates association rules from them. Association rules are rules of form A -> B, where A and B are item sets. The rule indicates that if a transaction contains A, it is also likely to contain B.
- Evaluate the association rules - Finally, the association rules are evaluated based on metrics such as confidence and lift.

Required Libraries

- **NumPy**

  NumPy is a Python library used for working with arrays. It also has functions forworking in domain of linear algebra, fourier transform, and matrices.

- **Mlxtend**

  Create counterfactual records, draw PCA correlation graphs and decision boundaries, perform bias-variance decomposition, bootstrapping, and much more. MLxtend library (Machine Learning extensions) has many interesting functions for everyday data analysis and machine learning tasks.

- **Pandas**

  Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

- **Matplotlib**

  Matplotlib is a cross-platform, data visualization and graphical plotting library (histograms, scatter plots, bar charts, etc) for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB.