Michelle Rice

DSC630-T301 Predictive Analytics

Term Project Milestone 3: Preliminary Analysis

Predicting Grocery Purchases

7/4/2021

**Abstract**

This project will analyze grocery purchases to find correlations between purchased products and predict what a customer might buy based on their shopping history and other products that they purchase. This information can then be used for marketing, pricing and product placement. The dataset being used was obtained from Kaggle and contains approximately 1500 orders and includes the items in the order, the day of the week, time of day and days since last order. I will be using the Apriori algorithm and association rules to determine correlation between products.

**Background**

A key component to retail success is understanding the customers and their buying preferences and patterns. If we know what a customer is likely to want to buy before they come to the store, or shop online, we can use that information to help manage inventory, promotions and targeted marketing. The best way to understand what customers are likely to buy in the future is to look at what they have purchased in the past. Some of this is simple average sales: for example, if we know that we generally sell 500 gallons of milk every day, we would plan our inventory to accommodate for that. For this type of planning, in addition to knowing what products a specific customer purchases, it is valuable to know how frequently the product is purchased and when it is purchased.

This study will look at an additional component to purchase history that can give us even more information. Rather than looking at a product individually, looking at what products customers purchase together can give us even greater insight. In addition to what is purchased and when, there is great value in knowing what other products are purchased at the same time. This is referred to as basket analysis and can guide us even further in sales and inventory planning and provide information needed for successful marketing and promotions.

**Problem Statement**

What products is a customer likely to buy based on their other purchases? How can we use purchase history along with basket analysis to predict future sales and use that in marketing, inventory, pricing and product placement strategies?

**Methods**

The analysis was done using the Apriori algorithm, which identifies three key results:

**Support:** the percentage of transactions that contain all of the items in an itemset (e.g., bread, peanut butter, jelly). The higher the support, the more frequently the itemset occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future transactions.

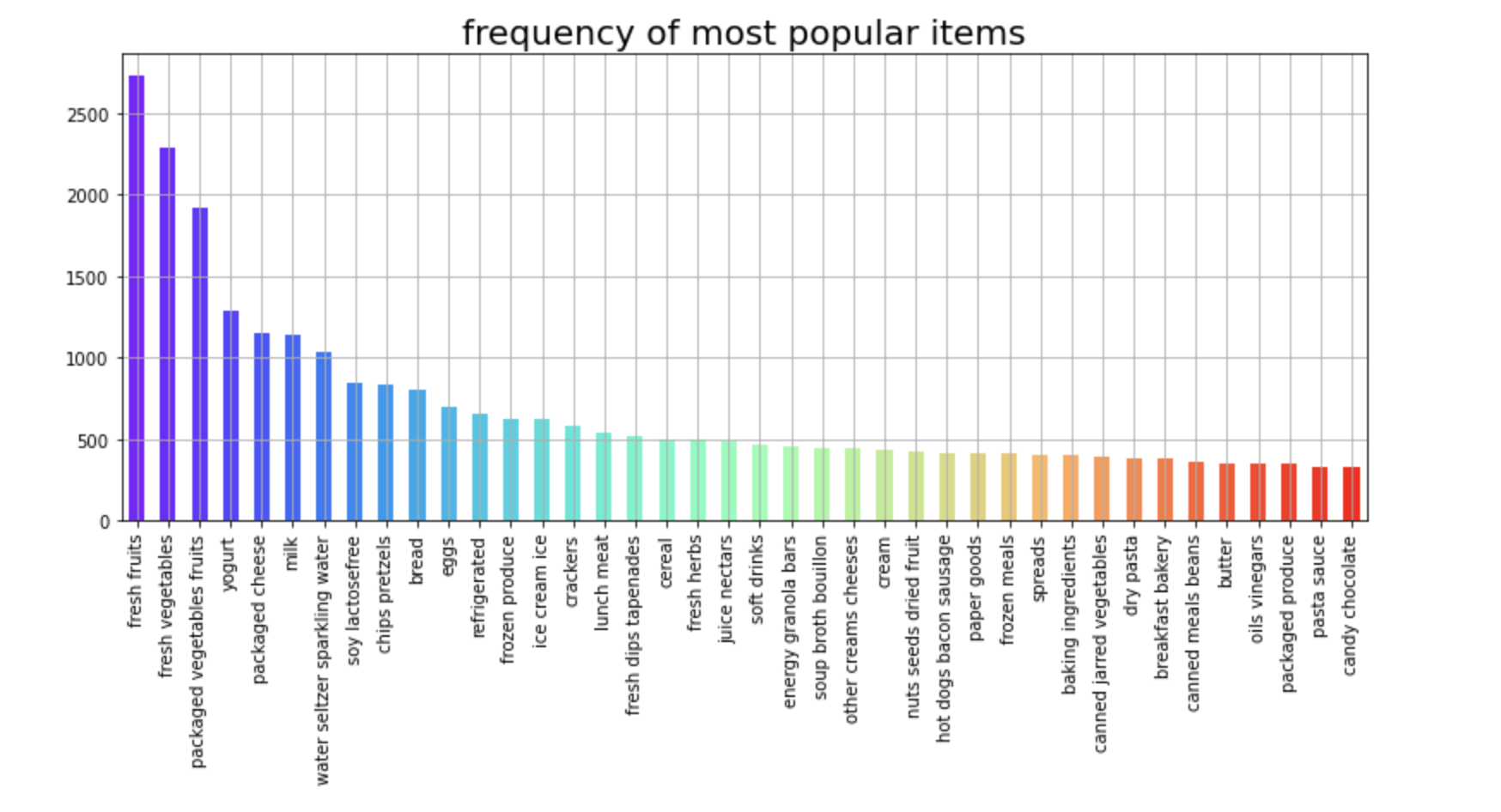
**Confidence:** the probability that a transaction that contains the items on the left hand side of the rule (bread and peanut butter) also contains the item on the right hand side (jelly). The higher the confidence, the greater the likelihood that the item on the right hand side will be purchased or, in other words, the greater the return rate you can expect for a given rule.

**Lift:** the probability of all of the items in a rule occurring together (otherwise known as the support) divided by the product of the probabilities of the items on the left and right hand side occurring as if there was no association between them. For example, if bread, peanut butter and jelly occurred together in 2.5% of all transactions, bread and peanut butter in 10% of transactions and jelly in 8% of transactions, then the lift would be: 0.025/(0.1\*0.08) = 3.125. A lift of more than 1 suggests that the presence of bread and peanut butter increases the probability that jelly will also occur in the transaction. Overall, lift summarizes the strength of association between the products on the left and right hand side of the rule; the larger the lift the greater the link between the two products.

The set of items on the left-hand side is the antecedent of the rule, while the one to the right is the consequent.  The probability that the antecedent event will occur is the supportof the rule. That simply refers to the relative frequency that an itemset appears in transactions.

**Initial Data Exploration**

The initial data exploration provides some general information about the basket data, including the most frequently purchased items as shown in this graph.

****

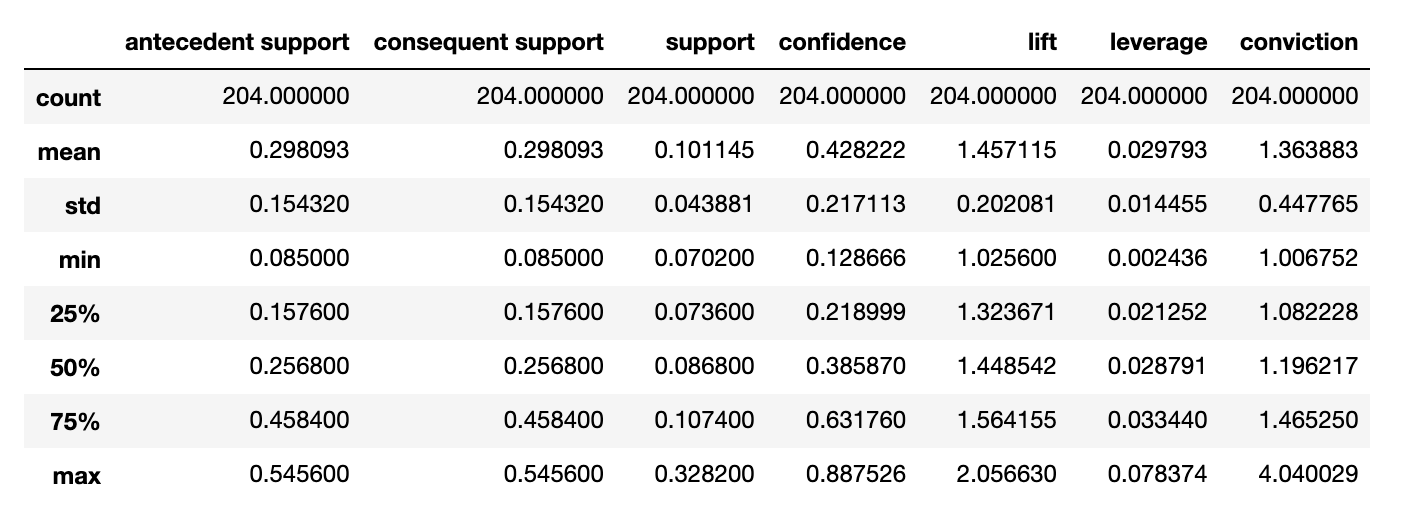
The data does not appear to point to anything unexpected initially with the most frequent purchases being common products such as fruits, vegetables, cheese, milk, bread, eggs, etc.

**Initial Results**

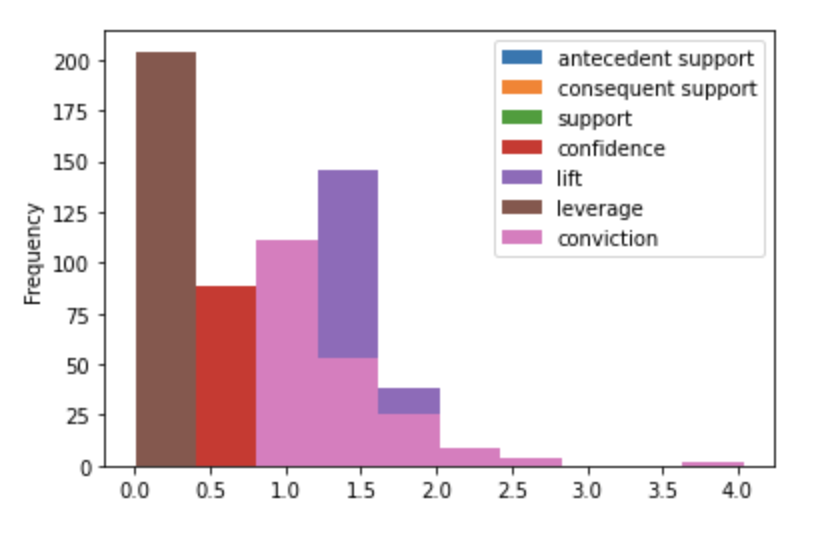
The first step in the analysis is to determine the association rules of the products. Using the Apriori model for this set of data when using a lift threshold of 1 returns 204 rules. The results tell us which item(s) are the antecedents and which are the consequents, along with the support, confidence and lift for each rule. Not surprising, many of the items we see here are those that were in the graphic of the most frequent purchases.

****

Looking at the descriptive stats for the rules, we can see the more details of the metrics for these rules. At first glance, I’m concerned that the support and confidence are low and this will require further analysis.

****

This chart shows us a graphical representation of those values:

****

**Discussion/conclusion**

The initial analysis indicates that this model will work well for this dataset. As mentioned previously, there needs to be more analysis done on the factors other than lift such as support and confidence to determine the value of the results. As the analysis continues, it is important to keep in mind that the real value of this is finding associations that we would not normally expect. Therefore, if most of the items that are returned in the rules are those very common, expected purchases, it may be necessary to adjust the lift threshold or find additional data to work with. The next steps will be to dig deeper into those rules and results and determine if that is necessary and then testing of the model.

**References**

McColl, Lynsey. Market Basket Analysis: Understanding Customer Behavior. *Select statistical services*. Retrieved from: <https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/>

Smartbridge. Market Basket Analysis 101: Anticipating Customer Behavior. Retrieved from: <https://smartbridge.com/market-basket-analysis-101/>

Simmons, Julian (July 27, 2020). Market Basket Analysis in Python. Retrieved from: <https://www.srose.biz/research-analysis/market-basket-analysis-in-python/>