Saemi Ramirez

011926418

D212 PA 3 Association Rules and Lift Analysis

9/24/2024

WGU

1. Research Question
   1. Research question and clustering technique  
      What are the top 3 medications that were prescribed the most along with abilify in the medical dataset?
   2. Define a goal  
      Apply market basket analysis to determine the top highest 3 association rules for abilify sorted by confidence
2. Reasons of chosen market basket analysis
   1. How market basket analyzes the selected dataset & Expected outcome
      1. Market basket analysis is a data mining technique employed to retailers to boost sales by gaining deeper insights into customer buying behaviors. It involves examining extensive datasets, like purchase histories, to identify product groupings and determine which items are frequently bought together. (TechTarget)
      2. An association rule for market basket analysis should be applied to extract insights from large transaction datasets. This technique can reveal relationships between items in the basket using machine learning or statistical methods.   
         For example, there are 5 items in the shop including the bread, butter, jam, milk, and egg. And there are 5 baskets or transactions with mixed items of those 5.   
         A chart of different types of bread

         Description automatically generated  
         Once the computer translates the data as 0 and 1, the metrics can be calculated that can lead for the business action using support, confidence, and lift.
      3. Support is a frequency/entire transaction. The support value helps determining which rules are worth exploring further. For example, if the bread was in 4 transactions out of 5 as total, then it can be 4/5 which is 0.8.
      4. Confidence is a “if then” condition which indicates the likelihood of the subsequent item(s) appearing in the same transaction, given that certain item(s) are already present. The formula is confidence (A→B) = frequency (A,B) / Frequency (A). For example, the confidence (Bread → milk) = 2/4 = 0.5 and confidence (Milk→Jam) = 1/3 = 0.33.
      5. Lift measures the ratio between the probability of B occurring when A is present and the probability of B occurring without knowledge of A. The formula is Lift (A→B) = Support(A,B)/(Support(A) \* Support (B)) which is Confidence(A→ B)/Support(B). For example, Lift(Bread→Milk) is 0.5/(6/10) = 0.83.  
         (Kimnaruk)
      6. The final list that are the most related with abilify sorted by confidence will be the expected outcome
   2. Provide 1 example of transactions in the dataset  
      Following screenshot is the transactions of the medical dataset after the clean up  
      A screenshot of a computer screen

      Description automatically generated
   3. One assumption of market basket analysis  
      One of the assumptions of market basket analysis is a representative sample. It is assumed that the dataset is a representative sample of all customer transactions. Having a representative sample is essential for drawing valid conclusions about overall shopping patterns. If the dataset is biased or unbalanced, the insight of gained from the analysis may not accurately reflect the behavior of the broader customer base. (Deniran)
3. Perform data preparation
   1. Transform the dataset for market basket analysis. Include a copy of cleaned dataset
      1. Read the medical\_market\_basket.csv file and check the information using info()  
         A screen shot of a computer

         Description automatically generated
      2. There are total 15002 entries but the column with the most non-nulls is Presc01 with 7501 rows. Filtering the data with column Presc01 using notna(). Data is now filtered with 7501 rows and 20 columns  
         A screenshot of a computer

         Description automatically generated
      3. Convert the DataFrame data to the list of lists using for loop, then feed the list to TransactionEncoder with the columns  
         A screenshot of a computer

         Description automatically generated
      4. Here are the first few columns captured  
         A screenshot of a computer

         Description automatically generated
      5. Remove the NaN columns from the transformed dataset using drop function  
         A screenshot of a computer

         Description automatically generated
      6. The cleaned dataset is attached as ‘cleaned\_dataset.csv’  
         A close-up of a white background

         Description automatically generated
   2. Execute the code used to generate association rules with Apriori algorithm
      1. Create Apriori object first with not showing if the support is less than 0.02 with column names shown  
         A screenshot of a computer

         Description automatically generated
   3. Provide values of the support, lift, and confidence of the association rules table
      1. Association\_rules function was used to create a rule\_table filtered by lift with min\_threshold 1 which is considered to be a strong rules  
         A screenshot of a computer

         Description automatically generated
   4. Explain the top 3 relevant rules generated by the Apriori algorithm
      1. Here is a screenshot of the top 3 rules when filtered by confidence   
         A screenshot of a computer

         Description automatically generated
4. Summarize your data analysis
   1. Summarize the significance of support, lift, and confidence from the results of the analysis  
      As the presented screenshot above on C4, the top 3 rules that are related to abilify when the antecedents are metformin, glipizide and lisinopril. Talking about metformin and abilify first, their support value, which is calculated by number of transactions contained both metformin and abilify divide by entire transaction, is 0.023. The confidence, which is calculated by the transaction containing metformin and abilify divide by all the transaction that contains metformin, which is 0.456. The lift value, which is Confidence(A→ B)/Support(B) gives 0.456/(number of transactions with abilify/entire number of transactions), resulted in 1.915. In the same way, the second and the third rule with glipizide and lisinopril, their support values were 0.028 and 0.041, confidence values with 0.419 and 0.417, and the lift values with 1.758 and 1.748.
   2. Discuss the practical significance of your findings from the analysis  
      The support values for top 3 were varied from 2.3% to 4.1% which is considered to be low. However, the confidence values were over 40% for all 3 which is high likelihood.
   3. Recommend a course of action  
      Because the lift values for all top 3 medications are greater than 1, which indicates the positive correlation between the prescriptions and abilify. The hospitals should foster collaboration between departments because these 4 medicines can be prescribed from doctors from different departments.
5. Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2a116ef5-b3dc-4969-bce9-b1f600a9076d>
6. Third-party code to support the analysis
7. Third-party in-text citations and references
   1. Deniran, Oluwakemi Helen. *Boosting Sales with Data: The Power of Market Basket Analysis in Retail*. Medium. (November 27, 2023). <https://medium.com/@chemistry8526/boosting-sales-with-data-the-power-of-market-basket-analysis-in-retail-c79cc10a14df>.
   2. Kimnaruk, Yannawut. *What are market basket analysis and the apriori algorithm?* Medium. (September 18, 2022). <https://yannawut.medium.com/what-are-market-basket-analysis-and-the-apriori-algorithm-fe0e8e6e34d>.
   3. TechTarget Contributor. *Market basket analysis*. TechTarget. <https://www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis>.