University of Kansas Business Analytics Competition

Challenge #3

**WheatFields Market Basket Analysis**

**Problem Statement**

Your problem for the last challenge in this competition will be to perform a market basket analysis on the 9,532 transactions provided in the Wheatfields.csv data set. Because there is no metric like an MSE for market basket analysis you will need to answer the following ten questions and will be graded on a score of 0-100 points. The questions and point break down are listed below. Also listed, is a description of the data provided as well as resources that will be a good starting point for your research into the apriori algorithm.

**Questions**

**Create visuals to answer the following questions**

1. What days of the week have the highest sales figures? **5 points**
2. What are our top 10 products in terms of highest sales? **5 points**
3. During which hours of the day does WheatFields have the highest sales figures? **5 points**

**Question 4:**

WheatFields bakers’ bread fresh every morning. To help the bakers better understand how much bread they should make each morning

1. Find the average quantity of bread sold for each day of the week.  **15 points**
2. Help them, with a visualization, to find out which days of the week they have their top sales. What do you see with the graph? **10 points**

**Apriori Questions Regarding Products**

1. What are the five products that have the highest frequency? **5 points**
2. Assuming a minimum Support of .02, if a customer purchases a Donut what product do we have the highest confidence they will purchase next? What is the rule’s; Support, Confidence, and Lift? **10 points**
3. Using support and confidence parameters of .02 and .25 respectively, what rule has the highest lift? What is the rule’s; support, confidence and lift? **10 points**
4. Using support and confidence parameters of .001 and .1 respectively, what are the 4 rules regarding a customer whose first purchase is Milk (lhs = {Milk}? What are the rules; support, confidence, and lift? **15 points**

**Questions continue on next page…**

**Apriori Questions Regarding Categories**

1. Assuming a minimum Support of .02, if a customer purchases an item from the Drink category what Item category do we have the highest confidence they will purchase from next? What is the rule’s; support, confidence, and lift? **10 points**
2. If a customer purchases an item from the food category then an item from the side order category, what is our confidence that they will make a third purchase from the drink category {Food,Side Order} => {Drink}? **10 points**

**Models for Challenge #3 will be due Nov. 20th at mid-night and results will be posted after Thanksgiving Break**

**Data**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| **Transaction\_Date** | Day – Month – Year that transaction occurred on |
| **Transaction\_Time** | Hour – Minute – Second that transaction occurred on |
| **Transaction\_ID** | Unique ID corresponding to a particular transaction |
| **Quantity** | Quantity sold |
| **Product** | Product name |
| **Price** | Price of product |
| **Category** | Category product belongs too |

**Resources - Apriori Algorithm**

The following videos will give you a good introduction to the apriori algorithm and should aid in questions 5 – 10.

[**https://www.youtube.com/watch?v=b5hgDPa7a2k**](https://www.youtube.com/watch?v=b5hgDPa7a2k)

[**https://www.youtube.com/watch?v=Gy\_nqzJMNrI&t=10s**](https://www.youtube.com/watch?v=Gy_nqzJMNrI&t=10s)

**\*While building an apriori model you can control what product/categories appear on the left-hand side and right-hand side, we recommend you research how this is done.**

**Resources – Code for importing data as sparse matrix**

You will want to install the package “arules” and use the following code to import the data as a sparse matrix.

trans <- read.transactions("Wheatfields.csv", format="single", cols=c(3,5), sep=",", rm.duplicates=TRUE)

trans2 <- read.transactions("Wheatfields.csv ", format="single", cols=c(3,7), sep=",", rm.duplicates=TRUE)

**\*trans will create the sparse matrix for products and trans2 will create the sparse matrix for categories.**

**Resources – Definitions and descriptions for association rules**

**Support**

Support is an indication of how frequently the item set appears in the data set.

supp(X⇒Y)=|X∪Y|n

In other words, it’s the number of transactions with both X and Y divided by the total number of transactions. The rules are not useful for low support values.

**Confidence**

For a rule X⇒Y, confidence shows the percentage in which Y is bought with X. It’s an indication of how often the rule has been found to be true.

conf(X⇒Y)=supp(X∪Y)supp(X)

**Lift**

The lift of a rule is the ratio of the observed support to that expected if X and Y were independent, and is defined as

lift(X⇒Y)=supp(X∪Y)supp(X)supp(Y)

Greater lift values indicate stronger associations

**Evaluation**

You will be evaluated on a 0 – 100 point scale, the breakdown of points is listed above with the questions. Models will be due Nov. 20th at midnight and will be graded by the end of Thanksgiving break. We will have an event the following week to announce the winners of the contest!