Name: \_\_\_Sri Vamshi Polela\_\_\_\_

CSDA 5430 Predictive Analytics

Week 7 Assignment

1. **Justify why both association rules and clustering analysis are among unsupervised learning algorithms in predictive analytics and how their unsupervised nature serve their goals in analytics?**

* Association rules and clustering analysis are both categorized as unsupervised learning algorithms in predictive analytics due to their common goal of extracting patterns and insights from data without the need for pre-defined labels or outcomes.
* In both association rules and clustering analysis, the absence of outcome variables implies that they are not employed for predicting specific outcomes.
* Association rules, like in market basket analysis, unveil relationships among items without predefined class labels, making them advantageous for exploring datasets. Their unsupervised nature is beneficial for discovering patterns, as seen in {charcoal, lighter, chicken wings} → {barbecue sauce}. This is especially valuable in biomedical research for revealing patterns in DNA or medical claims without prior outcome knowledge, serving the goal of knowledge discovery.
* Clustering analysis groups similar things together based on their features, finding patterns without predefined categories. These groups are called clusters. It helps organize data into natural groups, revealing possible classes when we don't have specific labels in advance. Clustering's unsupervised approach uncovers hidden structures, making it useful for identifying meaningful groups without predefining what those groups should be. **(Okay. In association rules, the goal is to identify items in transaction-type databases to discover which groups of products, or items, tend to be purchased together. In clustering analysis, the goal is to segment the data into a set of homogeneous clusters of observations for the purpose of generating insight. In both algorithms, unsupervised learning is implemented to learn from the data with no outcome variable.)**

1. Given the following transactions dataset, calculate support, confidence and lift for the two association rules*.* Show your calculation and interpret the result.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Transaction | charcoal | lighter | chicken wings | barbecue sauce |
| 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 0 | 1 | 1 | 1 |
| 5 | 0 | 1 | 1 | 0 |
| 6 | 0 | 0 | 1 | 1 |
| 7 | 0 | 1 | 0 | 1 |

1. **Rule 1: {charcoal} 🡪 {barbecue sauce}**

**Support**: Measures how frequently it appears in the transaction data.

* Support (charcoal) = = = 0.4285
* Support (barbecue sauce) = = = 0.7142
* Support (charcoal, barbecue sauce) = = = 0.2857

**Confidence**: Measures the joint occurrence of charcoal and barbecue sauce over the charcoal

* Confidence (charcoal 🡪 barbecue sauce) = = = 0.6667
* **Lift** (charcoal 🡪 barbecue sauce) = = = 0.9334

1. **Rule 2: {barbecue sauce} 🡪 {charcoal}**

**Support:**

* Support (charcoal) = = = 0.4285
* Support (barbecue sauce) = = = 0.7142
* Support (charcoal, barbecue sauce) = = = 0.2857

**Confidence**:

* Confidence (barbecue sauce 🡪 charcoal) = = = 0.4
* **Lift** (barbecue sauce 🡪 charcoal) = = = 0.9334

**Interpret the results:**

Rule 1: Says that if a customer buy charcoal, then he also buy barbecue sauce.

Rule 2: Says that if a customer buy barbecue sauce, then he also buy charcoal.

**Confidence:**

* Confidence tells us how often one product is bought when another is purchased.
* For example, in Rule 1, there's a 66.67% chance that if someone buys charcoal, they'll also buy barbecue sauce.
* For Rule 2: There is a 40% chance that if someone buys barbecue sauce, they’ll also buy charcoal.
* In general, there is a lower likelihood that someone who buys barbecue sauce will also purchase charcoal, as indicated by the 40% confidence in Rule 2.

**Lift:**

* Lift helps us understand if there's a real connection between the products. In both rules, the lift is 0.9334, suggesting a weak link. It means the chance of buying both items together is only slightly higher than if the purchases were unrelated. **(It is not higher than random. When lift is less than 1, the confidence is less than benchmark.)**

**9/10 points**

1. Exam grocery transaction dataset in file *MarketBasket.csv* that contains data of one week duration of South France from a well-known chain of Marts. You can use the following lines of R code to read the data as transaction.

library(arules)

basket <- read.transactions('MarketBasket.csv', sep = ',', rm.duplicates = TRUE)

1. What are number of transactions and number of items in the dataset?

A close-up of black text

Description automatically generated

Total number of transactions = 7501 and number of items = 119 **(good)**

1. Show the first five transactions.

A white background with black text

Description automatically generated

**(Look good)**

1. Find out the top frequent grocery items that have minimum support 0.1 and plot them. Which item has the highest support?

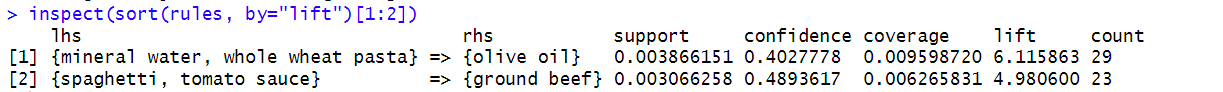
A graph of a number of objects

Description automatically generated

* Mineral Water has the highest support frequency.

**(good)**

1. Use support = 0.003, confidence = 0.4, minlen = 2 to generate the grocery association rules. Sort the rules with highest lift. Translate two of the rules into “if-then” statements. Interpret the various measures (such as support, confidence, and lift).



**Rule 1:** If a customer buys mineral water and whole wheat pasta, then they are likely to buy olive oil as well.

**Interpretation:**

* Support (0.0039): Approximately 0.39% of all transactions include mineral water and whole wheat pasta. **(This is incorrect. The support is for all items of the rule, including mineral water, whole wheat pasta, and olive oil.)**
* Confidence (0.40): If a customer purchases mineral water and whole wheat pasta, there is a 40% chance that they will also buy olive oil.
* Lift (6.12): The chance of buying olive oil when mineral water and whole wheat pasta are purchased together is 6.12 times higher than if these purchases were unrelated.

**Rule 2:** If a customer buys spaghetti and tomato sauce, then they are likely to buy ground beef as well.

**Interpretation:**

* Support (0.0031): About 0.31% of all transactions include spaghetti and tomato sauce. **(Same issue here.)**
* Confidence (0.49): If a customer buys spaghetti and tomato sauce, there is a 49% chance that they will also buy ground beef.
* Lift (4.98): The chance of buying ground beef when spaghetti and tomato sauce are purchased together is 4.98 times higher than if these purchases were unrelated.

**(Others are good)**

**9/10 points**

1. In the hotdogs clustering example on chapter 12 slide 14, we created 3 clusters of hotdogs. By using the following lines of R code, a dendrogram of average linkage clustering on the data can be generated. You can try different types of linkage for hierarchical clustering.

d.norm <- dist(hotdog.df.norm, method = "euclidean")

hc <- hclust(d.norm, method = "average")

plot(hc, hang = -1, ann = FALSE)

* **Compare the two different approaches in clustering analysis: k-means clustering and hierarchical clustering.**
* **K-means clustering** is a partitioning method that divides a dataset into a pre-specified number of clusters, known as k, based on the similarity of data points. It iteratively assigns points to the nearest cluster centroid and updates centroids until convergence. While computationally efficient and suitable for well-separated, spherical clusters, K-means requires the prior knowledge of the desired number of clusters and may yield different results based on initial centroid selection.
* **Hierarchical clustering** creates a hierarchical structure of clusters, often represented as a dendrogram. This approach does not require specifying the number of clusters in advance and can be either agglomerative (bottom-up) or divisive (top-down). Agglomerative hierarchical clustering starts with individual data points and merges them, forming a tree-like structure. While offering flexibility in exploring clusters at different levels of granularity, hierarchical clustering tends to be more computationally intensive, especially for large datasets, and is sensitive to noise and outliers. **(okay)**
* If we choose to create 3 clusters on the average linkage dendrogram, what is the value of cutoff height? What are the sizes for each cluster?

* By seeing the dendrogram the cut-off height is 2 to create 3 clusters.
* Also, you can calculate the cut-off height by following formula.

A close-up of a text

Description automatically generated

* I took 2 as cutoff height.

A diagram of a city

Description automatically generated A diagram of a city

Description automatically generated

* After cutting tree at height 2, you can see 3 clusters.
* **Size of Each cluster**

A black text with blue letters

Description automatically generated

**(Look good)**

* By comparing the clustering result of hotdogs between implementing k-means clustering and hierarchical clustering, which implementation is more useful to business owner? Discuss.

**K-means Clusters sizes:**



**Hierarchical Clusters:** Size of clusters is not appropriate when compared to k-means clustering.

In cluster 3 the size is 1.

* In the context of the hot-dog dataset, employing k-means clustering proves more advantageous for business owners pursuing various goals. K-means produces more balanced and evenly distributed clusters compared to hierarchical clustering, offering clearer insights for decision-making. Hierarchical clustering, with its fixed nature, may not adapt well to changes in data validity, potentially leading to misleading results.
* K-means clustering, with its ability to generate clusters of approximately equal size, enhances interpretability for businesses. On the contrary, hierarchical clustering's dendrograms can be intricate and challenging to understand. Considering these factors, particularly the balanced cluster sizes and adaptability to data changes, k-means clustering emerges as a more suitable and practical choice for businesses in this scenario. **(Okay. We can see that in hierarchical clustering for average distance of hotdog case, one of the clusters has only one observation – number 33. From business perspective, this one member “cluster” does not serve the clustering goal. Hierarchical clustering provides flexibility of generating any number of clusters. Therefore, it is important to validate that the clusters generated are meaningful before deploying the model. This is true for all unsupervised learning approaches.)**

1. The data in *WestRoxbury.csv* includes information on single family owner-occupied homes in West Roxbury, a neighborhood in southwest Boston, MA, in 2014. The data contain over 5,000 homes. This dataset has 14 variables, and a description of each variable is given at the end.

In R, perform the following steps as discussed in Ch12 on the first 1,000 homes and 6 variables: *TOTAL\_VALUE, LOT.SQFT, YR.BUILT, KITCHEN, FIREPLACE, REMODEL*. We would like to group similar homes based on provided features.

* Step 1: Collecting data (Discuss the business problem and how the data can support the analytics.)

**Business Problem:**

* The goal is to figure out what makes houses similar or different in West Roxbury, Boston.
* To understand the importance of each feature or factor that affects real estate. We want to understand how things like the total value of a house, its size, when it was built, whether it has a kitchen or fireplace, and if it has been remodeled, affect the real estate market in this area.
* This information can aid in understanding what factors influence property values and preferences in the neighborhood. The analysis may help in predicting trends, making informed decisions about investments, and providing useful information to individuals involved in the real estate market in West Roxbury. **(Okay. The goal is to explore and create clusters on the data based on the eight measurements on utilities.)**
* Step 2: Exploring data and preparing data (Use only the first 1000 rows of data and 6 variables.

A screenshot of a computer

Description automatically generated

* Explore data by creating a couple of plots of your interested measurements. Explain your findings.

**Distributions of *TOTAL\_VALUE, LOT.SQFT, YR.BUILT, KITCHEN, FIREPLACE, REMODEL:***

A graph of different sizes and numbers

Description automatically generated with medium confidence

**Data Exploration:**

The data exploration reveals interesting insights into key property features.

* The total value of properties exhibits a slight right skew, indicating a concentration around the 200-300 range.
* Lot sizes are right-skewed, peaking around 5000 sq.ft.
* The histogram for the year of construction is slightly left-skewed, concentrated between 1950 and 1970.
* Approximately 90% of houses feature only one kitchen.
* The barplot for fireplaces emphasizes a higher frequency of houses without fireplaces or with only one.
* Similarly, the majority of houses show no remodeling, suggesting a prevalent non-renovated status in the dataset.

A graph with a line and dots

Description automatically generated

* The scatter plot above highlights a crucial observation: as the lot size increases, there is a corresponding increase in the total value of the property. This suggests a positive correlation between lot size and property value.
* Create dummies for categorical variable and rescale the data so distance can be calculated properly.)
* Created dummies for remodel attributes and standardized data using scale function.

A screenshot of a table

Description automatically generated

**(Overall good)**

* Step 3: Training a model on the data (Use the function *kmeans()* for the k-means clustering algorithm to create 4 clusters. Set a seed so the results can be reproduced.)

A close-up of a logo

Description automatically generated

**(okay)**

* Step 4: Evaluating model performance (Show the centers and members of each cluster. Plot visual presentation (profile line plot or bar plot) of cluster centroids. Generate silhouette plot. What is the value of average silhouette?)

**Centers of each cluster:**

A group of numbers on a white background

Description automatically generated

* The below bar plot illustrates the centers of each variable with 4 clusters.

A group of graphs with different colored squares

Description automatically generated with medium confidence

* **Silhouette Plot:**
* A graph of a number of objects

  Description automatically generated with medium confidenceA graph of different colored shapes

  Description automatically generated**Average Silhouette Value is 0.3**
* A high silhouette value suggests strong conformity within its cluster, indicating good clustering performance.
* A high silhouette for cluster 1, 3, 4 value means data fits well in its group, showing good clustering. But negative values in cluster 2 suggest poor fit with nearby groups, hinting that the clustering approach or the number of clusters might need adjustment.

**(Okay)**

* Step 5: Usage of cluster information (Interpret each cluster based on evaluation results.)

**Usage of cluster information A close-up of numbers

Description automatically generated**

**Cluster-1 (Poor Home Condition):** Consisting of 374 homes, this cluster represents properties in unfavorable conditions. These homes have notably low TOTAL\_VALUE, indicating a diminished market value. The LOT.SQFT is limited, suggesting smaller land areas. These properties are characterized by old constructions (low YR.BUILT), a scarcity of kitchens, the absence of fireplaces, and no significant remodeling, reflecting a need for potential improvements or renovations.

* From a business standpoint, these homes may not be considered favorable for investment due to their overall poor condition and potential renovation costs.

**Cluster-2 (Home Condition is favorable):** This cluster comprises 450 homes with characteristics indicating favorable conditions. The homes in Cluster-2 have relatively higher TOTAL\_VALUE, LOT.SQFT representing larger land areas, newer constructions (higher YR.BUILT), sufficient kitchens, fireplaces, and a mix of remodeling types.

* From a business perspective, homes in Cluster-2 are considered favorable for investment due to their overall good condition, larger land areas, and modern features.

**Cluster-3 (Mixed Home Conditions):** With 77 homes, Cluster-3 presents a mixed scenario. The TOTAL\_VALUE, LOT.SQFT, and YR.BUILT vary, indicating diversity in home conditions. KITCHEN and FIREPLACE features also show variability, and a range of remodeling types is observed.

* Understanding the mixed conditions in Cluster-3 is vital for businesses aiming to navigate the real estate market effectively and cater to a broad range of customer demands.

**Cluster-4 (Homes with Recent Remodeling):** This cluster encompasses 99 homes that have recently undergone remodeling. These homes exhibit a mix of characteristics, including moderate TOTAL\_VALUE and LOT.SQFT, varying YR.BUILT, and a blend of kitchen and fireplace features. The commonality among them is the recent remodeling, reflecting potential modernization or upgrades.

* Understanding and capitalizing on the unique characteristics of this cluster can be instrumental in maximizing returns on investment and attracting a specific segment of the real estate market.

**The below bar plot gives in-depth information of each variable behavior with their clusters.**

**A graph of different colored bars

Description automatically generated**

**A chart with many colored circles

Description automatically generatedThe below scatter plot reveals four distinct clusters b/w TOTAL\_VALUE and LOT.SQFT**

* Cluster-1 exhibits low total value and limited land area, indicating properties with lower market value and smaller land sizes.
* Cluster-2 represents properties with higher total value and larger land areas, reflecting favorable conditions.
* Cluster-3 displays mixed characteristics, with varying total value, land area, and construction year.
* Cluster-4 showcases homes with recent remodeling, featuring moderate total value and land area, and a blend of kitchen and fireplace features, suggesting modernization or upgrades.

**The below table gives you an of overview of each cluster for each variable:**

| **Variable** | **Cluster-1 (Poor Home Condition)** | **Cluster-2 (Favorable Conditions)** | **Cluster-3 (Mixed Conditions)** | **Cluster-4 (Recent Remodeling)** |
| --- | --- | --- | --- | --- |
| **Total Value** | Very low value | Relatively higher total value | Moderate total value | Moderate to higher total values, influenced by recent remodeling |
| **Lot Size** | Limited land area | Larger land areas | Varied lot sizes | Diverse lot sizes, influenced by recent remodeling and potential expansion |
| **Year Built** | Very old constructions | Newer constructions | Mix of construction years | Varied construction years, focus on recent remodeling |
| **Kitchens** | Scarcity of kitchens | Sufficient kitchens | Varied kitchen counts | Diverse kitchen features, influenced by recent remodeling |
| **Fireplaces** | No fireplaces | Fireplaces present | Varied fireplace presence | Mix of properties with or without fireplaces, influenced by recent remodeling |
| **Remodeling** | None | Mix of remodeling types | Varied remodeling statuses | Focus on recent remodeling, attracting buyers seeking modern and updated living spaces |

**(Overall good)**

**Total 58/60 points**