Association rules Interview Questions:

1.     What is lift and why is it important in Association rules?

2.     What is support and Confidence. How do you calculate them?

3.     What are some limitations or challenges of Association rules mining?

**1. What is lift and why is it important in Association rules?**

Lift is a measure of the strength of the association between two items in a dataset. It quantifies how much more likely two items are to occur together compared to their individual probabilities.

**Importance of Lift:**

* **Identifying Strong Associations:** High lift values indicate strong associations between items, helping to uncover meaningful patterns.
* **Prioritizing Rules:** Rules with higher lift values are more likely to be actionable and valuable for decision-making.
* **Evaluating Rule Strength:** Lift provides a standardized way to compare the strength of different association rules.

**2. What is support and Confidence. How do you calculate them?**

**Support:**

* Measures the frequency of occurrence of an itemset (a set of items) in the dataset.
* Calculated as the number of transactions containing the itemset divided by the total number of transactions.

**Confidence:**

* Measures the conditional probability of one item occurring given that another item has occurred.
* Calculated as the support of the itemset divided by the support of the antecedent itemset.

**Example:**

Consider a dataset of grocery store transactions:

|  |  |
| --- | --- |
| **Transaction ID** | **Items** |
| 1 | Bread, Milk, Eggs |
| 2 | Milk, Diapers, Beer |
| 3 | Bread, Milk, Diapers, Beer |
| 4 | Bread, Milk, Diapers |

* **Support(Bread, Milk):** 3/4 = 0.75
* **Confidence(Milk -> Bread):** Support(Bread, Milk) / Support(Milk) = 0.75 / 0.75 = 1

**3. What are some limitations or challenges of Association rules mining?**

* **Scalability:** As the dataset size and dimensionality increase, the computational cost of mining association rules can become significant.
* **Interpretability:** A large number of rules can be generated, making it difficult to identify the most meaningful ones.
* **Sparse Data:** In many real-world datasets, items may occur infrequently, leading to sparse data and potentially weak associations.
* **Noise and Outliers:** Noisy data or outliers can affect the quality of the discovered rules.
* **Domain Knowledge:** Understanding the underlying domain knowledge is crucial for interpreting and applying the discovered rules effectively.
* **Contextual Relevance:** Association rules may not be relevant to all contexts or scenarios.