Using the Apriori Algorithm for Employee Retention Analysis

A Case Study on Employee Satisfaction and Retention

By: Luis Reynoso-Perez

6/26/2024

```python
# Apriori Algorithm Implementation
def apriori(transactions, min_support):
    def get_frequent_itemsets(transactions, itemsets, min_support):
        itemset_counts = {itemset: 0 for itemset in itemsets}
        for transaction in transactions:
            for itemset in itemsets:
                if all(item in transaction for item in itemset):
                    itemset_counts[itemset] += 1
        return {itemset: count for itemset, count in itemset_counts.items() if count / len(transactions) >= min_support}
```

Introduction & Data Description

Overview:

- Analyzing employee data to discover patterns and rules influencing retention using the Apriori algorithm.

Dataset Overview:

- **Features:**
  - Job Satisfaction, Training Opportunities, Years of Service, Work-Life Balance, Age, Department, Commute Time, Promotion History, Performance Score, etc.

Initial DataFrame:

	JobSatisfaction	TrainingOpportunities	YearsOfService	WorkLifeBalance	PerformanceScore	CommuteTime	PromotionHistory	Department	Age	Left
0	Medium	Few	6-10	Poor	Low	60-90min	Never	Sales	57	No
1	High	Moderate	<1	Average	High	30-60min	Never	Sales	47	Yes
2	Medium	Few	1-2	Average	Medium	30-60min	Never	HR	48	No
3	Medium	Moderate	6-10	Average	High	60-90min	Never	Engineering	26	No
4	Low	Many	6-10	Good	Medium	60-90min	Never	Engineering	41	Yes

Data Preprocessing & Applying Apriori

- **Data Preprocessing:**
  - Handling Missing Values: Removed or imputed missing values.
  - Encoding Categorical Variables: Used One-Hot Encoding for categorical features.
- **Applying Apriori:**
  - Transformed data into a transactional format.
  - Set minimum support, confidence, and lift thresholds.
  - Ran the Apriori algorithm to find frequent itemsets and generate rules.

Transformed DataFrame (One-Hot Encoded):

	Age	JobSatisfaction_High	JobSatisfaction_Low	JobSatisfaction_Medium	\
0	57	False	False	True	
1	47	True	False	False	
2	48	False	False	True	
3	26	False	False	True	
4	41	False	True	False	

	TrainingOpportunities_Few	TrainingOpportunities_Many	\
0	True	False	
1	False	False	
2	True	False	
3	False	False	
4	False	True	

	TrainingOpportunities_Moderate	YearsOfService_1-2	YearsOfService_10+	\
0	False	False	False	
1	True	False	False	
2	False	True	False	
3	True	False	False	
4	False	False	False	

	YearsOfService_3-5	...	PromotionHistory_Once	PromotionHistory_Thrice+	\
0	False	...	False	False	
1	False	...	False	False	
2	False	...	False	False	
3	False	...	False	False	
4	False	...	False	False	

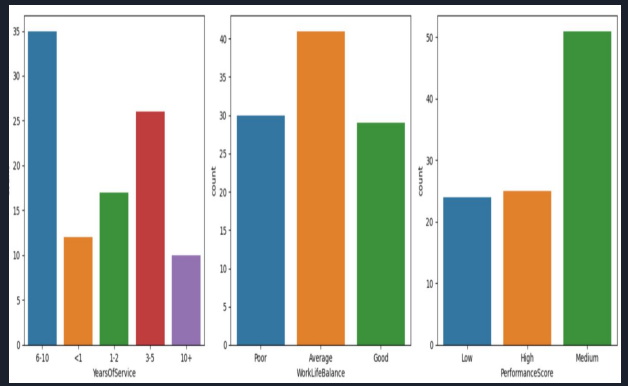
	PromotionHistory_Twice	Department_Engineering	Department_Finance	\
0	False	False	False	
1	False	False	False	
2	False	False	False	
3	False	True	False	
4	False	True	False	

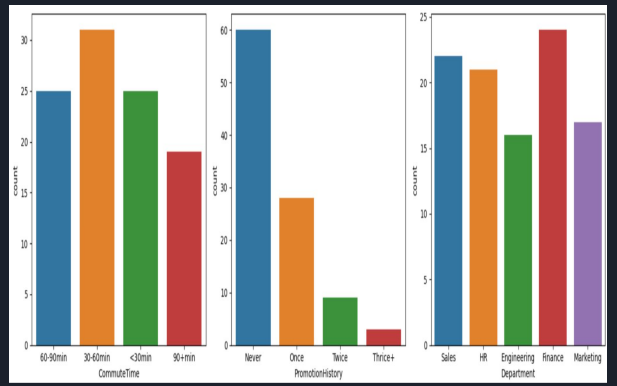
	Department_HR	Department_Marketing	Department_Sales	Left_No	Left_Yes
0	False	False	True	True	False
1	False	False	True	False	True
2	True	False	False	True	False
3	False	False	False	True	False
4	False	False	False	False	True

[5 rows x 33 columns]

Results & Complexity Analysis

- **Frequent Itemsets & Rules:**
  - Examples:
    - {'JobSatisfaction_Low'}: Support = [value]
    - {'YearsOfService_1-2', 'Left_Yes'}: Support = [value]
    - Rule: {'JobSatisfaction_Low'} -> {'Left_Yes'}: Confidence = [value], Lift = [value]
- **Time and Space Complexity:**
  - **Time Complexity:**
    - Exponential in the worst case due to the generation of candidate itemsets.
    - O(2^n) for generating all possible itemsets, where n is the number of unique items.
  - **Space Complexity:**
    - Also exponential in the worst case due to storage of itemsets and candidate sets.
    - O(k * 2^n), where k is the average length of itemsets.
- **Visualization:**
  - Employee Data Distributions: Bar charts and histograms for various features.
  - Network graphs of frequent itemsets and association rules.





Insights, Filtered Rules & Key Findings

Filtered Retention Rules:

- The sections "Filtered Retention Rules" and "Retention Rules DataFrame" are empty due to:
  - Low support or confidence thresholds not being met by any rule.
  - The filtering criteria being too strict for the dataset.

Key Findings:

- Low job satisfaction and few training opportunities are linked to higher turnover.
- Major factors affecting retention: job satisfaction, training opportunities, and years of service.