

Using the Apriori Algorithm for Employee Retention Analysis

A Case Study on Employee Satisfaction and Retention

By: Luis Reynoso-Perez

6/26/2024

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

- $\{\text{JobSatisfaction_Low}\}$: Support = [value]
 - $\{\text{YearsOfService_1-2}, \text{'Left_Yes'}\}$: Support = [value]
 - Rule: $\{\text{JobSatisfaction_Low}\} \rightarrow \{\text{'Left_Yes'}\}$: Confidence = [value], Lift = [value]

- **Time and Space Complexity:**

- **Time Complexity:**

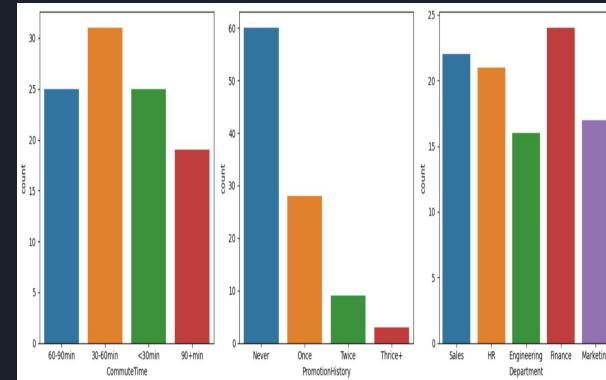
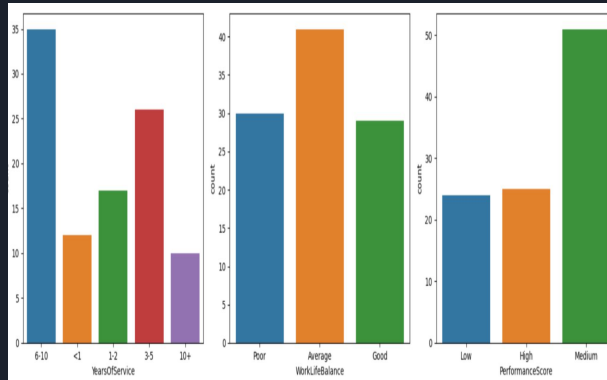
- Exponential in the worst case due to the generation of candidate itemsets.
 - $O(2^n)$ for generating all possible itemsets, where n is the number of unique items.

- **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.

Frequent Itemsets:

```
{{'TrainingOpportunities_Few',): 44, ('YearsOfService_6-10',): 35, ('PromotionHistory_Never',): 60, ('Department_Sales',): 22, ('PerformanceScore_Low',): 24, ('CommuteTime_60-90min',): 25, ('WorkLifeBalance_Poor',): 30, ('Left_No',): 58, ('JobSatisfaction_Medium',): 48, ('TrainingOpportunities_Few', 'YearsOfService_6-10',): 16, ('TrainingOpportunities_Few', 'PromotionHistory_Never',): 25, ('TrainingOpportunities_Few', 'CommuteTime_60-90min',): 15, ('TrainingOpportunities_Few', 'WorkLifeBalance_Poor',): 13, ('TrainingOpportunities_Few', 'Left_No',): 25, ('TrainingOpportunities_Few', 'JobSatisfaction_Medium',): 20, ('YearsOfService_6-10', 'PromotionHistory_Never',): 23, ('YearsOfService_6-10', 'Department_Sales',): 10, ('YearsOfService_6-10', 'CommuteTime_60-90min',): 10, ('YearsOfService_6-10', 'WorkLifeBalance_Poor',): 14, ('YearsOfService_6-10', 'Left_No',): 18, ('YearsOfService_6-10', 'JobSatisfaction_Medium',): 15, ('PromotionHistory_Never', 'Department_Sales',): 16, ('PromotionHistory_Never', 'PerformanceScore_Low',): 15, ('PromotionHistory_Never', 'CommuteTime_60-90min',): 15, ('PromotionHistory_Never', 'WorkLifeBalance_Poor',): 17, ('PromotionHistory_Never', 'Left_No',): 33, ('PromotionHistory_Never', 'JobSatisfaction_Medium',): 27, ('Department_Sales', 'Left_No',): 11, ('Department_Sales', 'JobSatisfaction_Medium',): 13, ('PerformanceScore_Low', 'CommuteTime_60-90min',): 11, ('PerformanceScore_Low', 'WorkLifeBalance_Poor',): 11, ('PerformanceScore_Low', 'Left_No',): 14, ('PerformanceScore_Low', 'JobSatisfaction_Medium',): 14, ('CommuteTime_60-90min', 'Left_No',): 16, ('CommuteTime_60-90min', 'JobSatisfaction_Medium',): 15, ('WorkLifeBalance_Poor', 'Left_No',): 18, ('WorkLifeBalance_Poor', 'JobSatisfaction_Medium',): 12, ('Left_No', 'JobSatisfaction_Medium',): 26, ('TrainingOpportunities_Few', 'YearsOfService_6-10', 'PromotionHistory_Never',): 11, ('TrainingOpportunities_Few', 'PromotionHistory_Never', 'JobSatisfaction_Medium',): 11, ('TrainingOpportunities_Few', 'Left_No', 'JobSatisfaction_Medium',): 11, ('YearsOfService_6-10', 'PromotionHistory_Never', 'Left_No',): 11, ('YearsOfService_6-10', 'Pr
```

```
{'antecedent': ('YearsOfService_6-10', 'JobSatisfaction_Medium'), 'consequent': ('PromotionHistory_Never',), 'support': 0.11, 'confidence': 0.7333333333333333}  
{'antecedent': ('WorkLifeBalance_Poor', 'Left_No'), 'consequent': ('PromotionHistory_Never',), 'support': 0.1, 'confidence': 0.5555555555555556}  
{'antecedent': ('PromotionHistory_Never', 'WorkLifeBalance_Poor'), 'consequent': ('Left_No',), 'support': 0.1, 'confidence': 0.5882352941176471}  
{'antecedent': ('Left_No', 'JobSatisfaction_Medium'), 'consequent': ('PromotionHistory_Never',), 'support': 0.14, 'confidence': 0.5384615384615384}  
{'antecedent': ('PromotionHistory_Never', 'JobSatisfaction_Medium'), 'consequent': ('Left_No',), 'support': 0.14, 'confidence': 0.5185185185185185}
```

Filtered Retention Rules:

```
[]
```

Retention Rules DataFrame:

```
Empty DataFrame
```

```
Columns: []
```

```
Index: []
```

```
No rules to visualize.
```