IST707- Applied Machine Learning

Homework 3 - Association Rules

Hendi Kushta

Data Preprocessing

The first steps involve loading a list of clients into the work environment and examining the observations and attributes. To prepare the data for data mining algorithms to find significant relationships, the data must be transformed into discrete values and any unneeded information must be removed. The following steps were taken to accomplish this:

- The client **id** attribute was removed as it does not provide important information about any relationships.
- The values for the age attribute divided in 7 categories 'children', 'teens', 'twenties', "thirties",
- 'fourties','fifties','sixties', ‘elderly’
- The income was divided into 5 categories "very_low", "low", "medium", "high", "very_high" based on the minimum and maximum values. However, a better categorization could be established by the bank with more domain expertise.
- The children attribute was divided into 4 categories too "No Children", "1 Child", "2 Children", “3 or more Children”.
- The remaining preprocessing steps involved converting the data into specific data types and renaming the column names for easier value extraction.
- Finally, the dataset was converted into a transaction format where all attributes and their values become the columns of each transaction. The transaction list is represented as a matrix of true/false values, with a call having a true value if the attribute value occurs in a specific transaction (row).

Mine rules with the `Apriori` algorithm, experiments to obtain strong rules

1st experiment, support = 0,02, confidence = 0.8, sort by confidence. From 30 rules, I am showing the first. Support: 0.022 (2.2%) - This means that 2.2% of the transactions in the data set contain that the teenagers that have a saving account, have also a current active account.
Confidence: 1 – is 100%
Lift: 1.3 - This means that the likelihood that a teenager that have a saving account has also a current active account is 1.3 times higher than for the general population.

```
     lhs                                rhs                support confidence coverage lift count
[1]  {age=teens,
      save_act=YES}                  => {current_act=YES}   0.022          1    0.022  1.3    13
[2]  {income=very_high,
      children=3 or more Children}   => {save_act=YES}      0.022          1    0.022  1.4    13
[3]  {children=3 or more Children,
      save_act=NO}                   => {pep=NO}            0.037          1    0.037  1.8    22
[4]  {children=3 or more Children,
      pep=YES}                       => {save_act=YES}      0.022          1    0.022  1.4    13
[5]  {region=RURAL,
      income=very_high}              => {save_act=YES}      0.047          1    0.047  1.4    28
[6]  {age=sixties,
      income=very_high}              => {save_act=YES}      0.102          1    0.102  1.4    61
[7]  {income=very_high,
      children=1 Child}              => {save_act=YES}      0.038          1    0.038  1.4    23
[8]  {region=TOWN,
      income=very_high}              => {save_act=YES}      0.040          1    0.040  1.4    24
[9]  {age=forties,
      children=1 Child}              => {pep=YES}           0.060          1    0.060  2.2    36
[10] {region=SUBURBAN,
```

2nd experiment support = 0.005, confidence = 0.7, shows that the teenagers that are living in a suburban area, are less likely to have a car. Support is 5%, confidence is 100% and lift is 2.0
3rd experiment support = 0.008, confidence = 0.9, teenagers that have a car, probably have a current active account. Support = 1,3%, confidence = 100%, lift =1.8
4th experiment support = 0.002, confidence = 0.5, shows that the teenagers that are living in a suburban area, are less likely to have a car. Support is 5%, confidence is 100% and lift is 2.0
5th experiment support = 0.01, confidence = 0.7, shows that teenagers that have mortgage to pay, have a current active account. Support = 1.2%, confidence= 100% and lift = 1.3

Association Rule Mining Report

To generate a list of "rules," or relationships between attributes, several parameters must be established. The first parameter is support, which represents the fraction of transactions that contain an item set (a set of one or more items in a transaction). The higher the support, the more frequently the item set should appear in the transactions list. The second parameter is confidence, which is the proportion of transactions where the presence of item or item set x leads to the presence of item or item set y. In this case, the goal is to identify the item set in x that will result in Pep (Personal Equity Plan) being set to either "yes" or "no." The third parameter that could be considered is the minimum number of items in a set, which is set to two to obtain interesting associations.

To generate the rules, different values of support and confidence are tried, with the right-hand rule always being set to a specific value of Pep (either "yes" or "no").

This report details the process and results of an association rule mining process performed on a data set. The data set consisted of various demographic information, including age, income, region, number of children, and mortgage status, among others. The goal of the analysis was to identify patterns in the data that could indicate a high likelihood of purchasing a Personal Equity Plan (PEP).

The association rule mining process was performed using the Apriori algorithm, a popular algorithm for finding association rules in large data sets. The algorithm uses a set of parameters, such as support, confidence, and lift, to determine which rules to include in the final results.

The resulting five interesting rules are listed below, along with explanations and recommendations.

```
     lhs                          rhs       support confidence coverage lift count
[1]  {age=forties,
      children=1 Child}       => {pep=YES}   0.060          1    0.060  2.2    36
[2]  {age=forties,
      children=1 Child,
      save_act=YES}           => {pep=YES}   0.048          1    0.048  2.2    29
[3]  {age=forties,
      children=1 Child,
      current_act=YES}        => {pep=YES}   0.047          1    0.047  2.2    28
[4]  {age=forties,
      children=1 Child,
      save_act=YES,
      current_act=YES}        => {pep=YES}   0.040          1    0.040  2.2    24
[5]  {age=forties,
      married=YES,
```

The association rule mining process was performed on a customer demographic data set. The data set contained information about customer demographics such as age, income, region, mortgage, children, save_act, current_act, and whether or not the customer has bought a PEP (Personal Equity Plan) product. The process involved identifying the frequent patterns in the data and finding the associations between different demographic variables and the purchase of PEP products.

Results:
From the analysis, 5 interesting rules were discovered with the following characteristics:

Rule 1: {age=forties, children=1 Child} => {pep=YES}
Support: 0.060 (6%) - This means that 6% of the transactions in the data set contain the items in the antecedent and the consequent.
Confidence: 1 - This means that 100% of the transactions that contain the items in the antecedent also contain the items in the consequent.
Lift: 2.2 - This means that the likelihood of buying a PEP product for customers with the antecedent characteristics is 2.2 times higher than for the general population.
Recommendations: This rule suggests that customers in their forties with one child are more likely to purchase a PEP product. Marketing campaigns targeted towards this demographic could have higher success rates.

Rule 2: {age=forties, children=1 Child, save_act=YES} => {pep=YES}
Support: 0.048 (4.8%) - This means that 4.8% of the transactions in the data set contain the items in the antecedent and the consequent.
Confidence: 1 - This means that 100% of the transactions that contain the items in the antecedent also contain the items in the consequent.
Lift: 2.2 - This means that the likelihood of buying a PEP product for customers with the antecedent characteristics is 2.2 times higher than for the general population.
Recommendations: This rule suggests that customers in their forties with one child who have a savings account are more likely to purchase a PEP product. Offering tailored financial products for this demographic could have higher success rates.

Rule 3: {age=forties, children=1 Child, current_act=YES} => {pep=YES}
Support: 0.047 (4.7%) - This means that 4.7% of the transactions in the data set contain the items in the antecedent and the consequent.

Confidence: 1 - This means that 100% of the transactions that contain the items in the antecedent also contain the items in the consequent.
Lift: 2.2 - This means that the likelihood of buying a PEP product for customers with the antecedent characteristics is 2.2 times higher than for the general population.
Recommendations: This rule suggests that customers in their forties with one child who have a current account are more likely to purchase a PEP product. Offering tailored financial products for this demographic could have higher success rates.

Rule 4: {age=forties, children=1 Child, save_act=YES, current_act=YES} => {pep=YES}
Support: 0.040 (4%) - This means that 4% of the transactions in the data set contain the items in the antecedent and the consequent.
Confidence: 1 - This means that 100% of the transactions that contain the items in the antecedent also contain the item in the consequent.
Coverage: 0.040 (4%) - This means that 4% of the total transactions in the data set contain the antecedent and consequent.
Lift: 2.2 - This means that the presence of the antecedent items increases the likelihood of the consequent item being present by a factor of 2.2.
Explanation: This rule suggests that people in their forties with one child and who have both savings and current accounts are likely to buy a personal equity plan (PEP).
Recommendations: For financial institutions, this is a useful rule as it indicates that individuals in this demographic with both savings and current accounts are likely to invest in PEPs. Thus, they can focus their marketing efforts on this demographic to increase PEP sales. Additionally, financial institutions can offer special incentives for individuals in this demographic who have both savings and current accounts, to encourage them to invest in PEPs.

Rule 5: {age=forties, married=YES, children=1 Child} => {pep=YES}
Support: 0.038 (3.8%) - This means that 3.8% of the transactions in the data set contain the items in the antecedent and the consequent.
Confidence: 1 - This means that 100% of the transactions that contain the items in the antecedent also contain the item in the consequent.
Coverage: 0.038 (3.8%) - This is the percentage of transactions in the data set that contain the items in the antecedent.
Lift: 2.2 - This is the ratio of the observed support to that expected if the antecedent and consequent were independent. A lift value greater than 1 indicates that the items in the antecedent and consequent are more likely to occur together than expected by chance.
Explanation:
Based on this association rule, it can be concluded that people in their forties who are married and have one child are likely to be a customer of the personal equity plan (PEP). This rule has 100% confidence, indicating that all transactions with the antecedent items also have the consequent item. This pattern has a lift value of 2.2, which suggests that it is a strong association.
Recommendations:
Marketing campaigns aimed at people in their forties who are married and have one child could be more effective if they focus on promoting the Personal Equity Plan (PEP). This group is a strong target market for PEPs, so these campaigns should be designed to emphasize the benefits of PEPs and how they can help meet their financial goals. Additionally, financial advisors could use this information to target customers in this demographic with information about the PEPs, helping to increase awareness and sales of PEPs.