**Data Mining For Business – IDS 572**

Assignment -10

- Gunasagar Kaaparthy Jayaram (UIN:675544801)

- Sudhir Suhas Bagalkotker (UIN: 660388617)

Problem 1

Data given for the problem:

None 22

sunroof (S) 0

manual (M) 5

audio (A) 3

S & M 8

S & A 4

M & A 14

S & M & A 18

a) P(S)= 0+8+4+18/0+8+5+3+8+4+14+18+22 = 30/74 = 0.4054 =40.5%

b) Confidence, Lift and Support of the rule sunroof ⇒ improved audio

Support = P(S ∩ A) = (4+18) /74=22/74 = 11/37 =0.29

Confidence = Support / P(S) = (11/37) / (30/74) =22/30 = .733

Lift = Confidence / P(A) = (22/30) / (39/74) = 1628/1170 = 1.3914

c) Confidence, Lift and Support of the rule (M and A) ⇒ S

Support = P((M and A) ∩ S ) =18/74 =0.243

Confidence = Support/ P((M and A)) = (18/74) / (32/74) = 18/32 = 0.5625

Lift = Confidence / P(S) = (18/32) /( 30/74) = 1332/960=1.3875

d) Confidence, Lift and Support of the rule M ⇒ S

Support = P(M ∩ S) = 26/74 = 0.351

Confidence = Support/ P( M) = (26/74) / (45/74) = 26/45 =0.577

Lift = Confidence / P(S) = (26/45)/(30/74)= 1924/1350 = 1.425

e) Confidence, Lift and Support of the rule (not M) ⇒ S

Support = P( (notM) ∩ S) = 4/74 =0.054

Confidence = Support/ P( (not M)) =(4/74)/(29/74) =4/29 =.137

Lift = Confidence / P(S) =(4/29)/ (30/74) =296/870 = 0.34

Problem 2

(a) Fraction of all your customers who bought both Apples and Bananas

This is the Support of the rule (Apples -> Bannanas) =0.4

(b) Of those customers who bought Apples, the fraction who did not buy Pears

P(Apples) =0.6 P(Apples and Pears) = Support (Apples->Pears) =0.3

Fraction of customers who brought apples but not pears =0.3/0.6=0.5

(c) Fraction of your customers who did not buy Bananas.

P(Bananas) =Confidence( Apples-> Bananas )/ Lift( Apples-> Bananas)

= 0.666/1.333=0.5

Fraction of people who don’t buy bananas=1-P(Bananas)=1-0.5=0.5

(d) Fraction of your customers who bought Apples.

P(Apples) = Support(Apples-> Pears) /Confidence(Apples-> Pears)

= 0.3/0.5=0.6

Problem 3:

Data for Problem 3:



Minimum support = 20% =20% of 800=160.

Generating the item set frequency

1-itemset:

|  |  |
| --- | --- |
| **Majors** |  |
| A | 193 |
| E | 420 |
| S | 187 |
| **Status** |  |
| G | 159 |
| U | 641 |
| **Age** |  |
| O | 188 |
| Y | 612 |
| **Nationality** |
| U | 579 | |
| I | 221 | |

From the one item frequency, we can exclude the graduates column because it doesn’t have a frequency that is greater than the minimum support.

Generating the two item dataset:

|  |  |
| --- | --- |
| A,U | 137 |
| A,O | 85 |
| A,Y | 108 |
| A,US | 175 |
| A,I | 18 |
| E,U | 342 |
| E,O | 78 |
| E,Y | 342 |
| E,US | 242 |
| E,I | 178 |
| S,U | 162 |
| S,O | 25 |
| S,Y | 162 |
| S,US | 162 |
| S,I | 25 |
| U,O | 29 |
| U,Y | 612 |
| U,I | 100 |
| U,US | 541 |
| O,US | 67 |
| O,I | 121 |
| Y,US | 512 |
| Y,I | 100 |

The rows with the values marked in the green don’t have the minimum support value, hence they can be excluded from further analysis

Three item dataset:

|  |  |
| --- | --- |
| E,U,Y | 342 |
| E,Y,US | 242 |
| S,Y,US | 162 |
| S,U,US | 162 |
| Y,U,US | 512 |
| E,U,US | 242 |
| S,U,Y | 162 |

Here all the item combinations can be considered, the combination with the highest support her has been highlighted.

Four item dataset:

|  |  |
| --- | --- |
| E,U,Y,US | 242 |
| S,U,Y,US | 162 |

Here both the combinations can be considered and the combination with the highest support has been highlighted.

b) From the Apriori algorithm that has been used to generate the frequency item datasets, the combinations with the highest support for the three and four item dataset have been highlighted and they are:

For three item dataset: Young,Undergraduate,US student with frequency of 512 and a support of 512/800 = 64%

For the four item dataset : Engineering,Undergraduate,Young,US student with frequency of 242 and a support of 242/800 =30.25%.

Now generating the rules for the above combinations.

For the rule selection there is a requirement of the support to be at least 20% and having a confidence value that is greater than 90%.

For the three item dataset,

|  |  |
| --- | --- |
| **Y,U,US** | **Confidence** |
| Y->U,US | 84% |
| U->Y,US | 80% |
| US->Y,U | 88% |
| Y,U->US | 84% |
| Y,US->U | 100% |
| U,US->Y | 95% |

From the given set of rules, only the ones that have been highlighted satisfy the requirement of having support greater than 20% and confidence greater than 90%.

For the four-item dataset,

|  |  |
| --- | --- |
| **E,U,Y,US** | **Confidence** |
| E->U,Y,US | 58% |
| U->E,Y,US | 38% |
| Y->E,Y,US | 40% |
| US->E,U,Y | 42% |
| E,U->Y,US | 71% |
| E,Y->U,US | 71% |
| E,US->U,Y | 100% |
| U,Y->E,US | 40% |
| U,US->E,Y | 45% |
| Y,US->E,U | 47% |
| E,U,Y->US | 71% |
| E,U,US->Y | 100% |
| E,Y,US->U | 100% |
| U,Y,US->E | 47% |

For the four-item dataset, the rules that have been highlighted are the one’s that satisfy the established requirement of support greater than 20% and confidence greater than 90%.

From the above two datasets, the rules that have been highlighted compose the association rules and these are the rules that will be used for the generation of actionable insights.