Home Work 1

DATA MINING II

Ravi Teja Sunkara
UBIT NAME: RSUNKARA | UBIT NUMBER: 50292191

Question 1:

User - User

Item-id ⇒ user-id ⇓	1	2	3	4	5	6	Mean Rating	Pearson (i,2) (user-user)
1	5	6	7	4	3	?	5	-1
2	4	?	3	?	5	4	4	1
3	?	3	4	1	1	?	2.25	-0.99
4	7	4	3	6	?	4	4.8	0.61
5	1	?	3	2	2	5	2.6	-0.24

$$\text{Pearson}(1,2) = \frac{(5-5)(4-4)+(7-5)(3-4)+(3-5)(5-4)}{\sqrt{2^2+2^2}\ \sqrt{1^2+1^2}}$$

$$= \frac{-2-2}{4} = -1$$

$$\text{Pearson}(3,2) = \frac{(1-2.25)(4-2.25)(3-4)+(1-2.25)(5-4)}{\sqrt{1.75^2+1.25^2}\ \sqrt{1^2+1^2}}$$

$$= \frac{-1.75-1.25}{3.04} = -0.99$$

$$\text{Pearson}(4,2) = \frac{(7-4.8)(0)+(3-4.8)(3-4)+(6-4.8)(0)}{\sqrt{2.2^2+1.8^2+0.8^2}\ \sqrt{1^2}}$$

$$= \frac{-1.8\times -1}{2.95} = 0.61$$

$$\text{Pearson}(5,2) = \frac{(-1.6)(0)+(0.4)(-1)+(-0.6)(1)+(2.4)(0)}{\sqrt{1.6^2+0.4^2+0.6^2+2.4^2}\ \sqrt{1^2+1^2}}$$

$$= \frac{-1}{4.2} = -0.24$$

User 4 is the closest to user 2. The mean centred ratings ~~of the~~ prediction is:

$$\hat{r}_{22} = 4 + \frac{-0.8 \times 0.61}{0.61} = 3.2$$

$$\hat{r}_{24} = 4 + \frac{1.2 \times 0.61}{0.61} = 5.2$$

Item - Item Based :

Item-id → / user-id ↓	1	2	3	4	5	6
1	0	1	2	-1	-2	?
2	0	?	-1	?	1	0
3	?	0.75	1.75	-1.25	-1.25	?
4	2.2	-0.8	-1.8	1.2	?	-0.8
5	-1.6	?	0.4	-0.6	-0.6	2.4
Cosine(2,j)	-0.62	1	0.99	-0.97	-0.996	1
Cosine(4,j)	0.78	-0.97	-0.97	1	0.94	-0.7

$$\text{AdjCosine}(2,1) = \frac{0\times1 + 2.2\times-0.8}{\sqrt{2.2^2}\ \sqrt{1^2+0.8^2}} = \frac{-1.76}{2.82} = -0.62$$

$$\text{AdjCosine}(2,3) = \frac{(1\times2) + (0.75)(1.75) + (0.8\times1.8)}{\sqrt{1^2+0.75^2+0.8^2}\ \sqrt{2^2+1.75^2+1.8^2}}$$

$$= \frac{4.75}{4.76} = 1$$

$$\text{AdjCosine}(2,4) = \frac{(1\times-1) + (0.75\times-1.25) + (-0.8\times1.2)}{\sqrt{1^2+0.75^2+0.8^2}\ \sqrt{1^2+1.25^2+1.2^2}} = -0.97$$

$$\text{AdjCosine}(2,5) = \frac{(1\times-2) + (0.75\times-1.25)}{\sqrt{1^2+0.75^2}\ \sqrt{2^2+1.25^2}} = -0.996$$

$$\text{AdjCosine}(2,6) = \frac{(-0.8\times-0.8)}{\sqrt{0.8^2}\ \sqrt{0.8^2}} = +1$$

$$\text{AdjCosine}(4,1) = \frac{(-1\times 0) + (1.2\times 2.2) + (-0.6\times -1.6)}{\sqrt{1^2+1.2^2+0.6^2}\ \sqrt{2.2^2+1.6^2}} = 0.78$$

$$\text{AdjCosine}(4,2) = \frac{(-1\times 1) + (-1.25\times 0.75) + (1.2\times -0.8)}{\sqrt{1^2+1.25^2+1.2^2}\ \sqrt{1^2+0.75^2+0.8^2}}$$

$$= -0.97$$

$$\text{AdjCosine}(4,3) = \frac{(-1\times 2) + (-1.25\times 1.75) + (1.2\times -1.8) + (-0.6\times 0.4)}{\sqrt{1^2+1.25^2+1.2^2+0.6^2}\ \sqrt{2^2+1.75^2+1.8^2+0.4^2}}$$

$$= -0.97$$

$$\text{AdjCosine}(4,5) = \frac{(-1\times -2) + (-1.25\times -1.25) + (-0.6\times -0.6)}{\sqrt{1^2+1.25^2+0.6^2}\ \sqrt{2^2+1.25^2+0.6^2}}$$

$$= 0.94$$

$$\text{Adj Cosine}(4,6) = \frac{(1.2\times -0.8) + (-0.6\times 2.4)}{\sqrt{1.2^2+0.6^2}\ \sqrt{0.8^2+2.4^2}} = -0.7$$

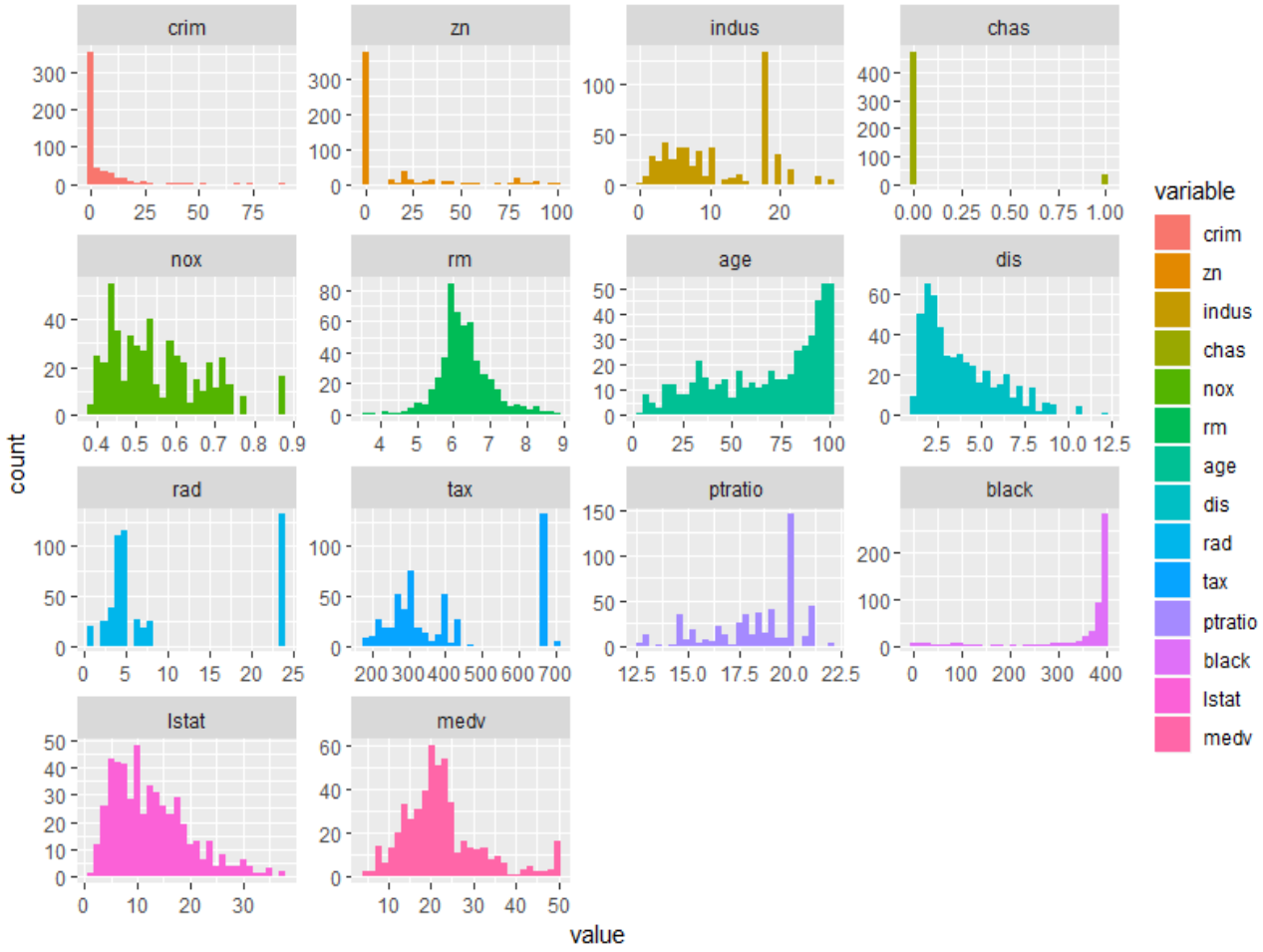
It is evident that items 3 & 6 are most similar to 2.
And items 1 & 5 are most similar to 4.

$$\hat{r}_{22} = \frac{3\times 0.99 + 4\times 1}{0.99+1} = 3.5$$

$$\hat{r}_{24} = \frac{4\times 0.78 + 5\times 0.94}{0.78+0.94} = 4.55$$

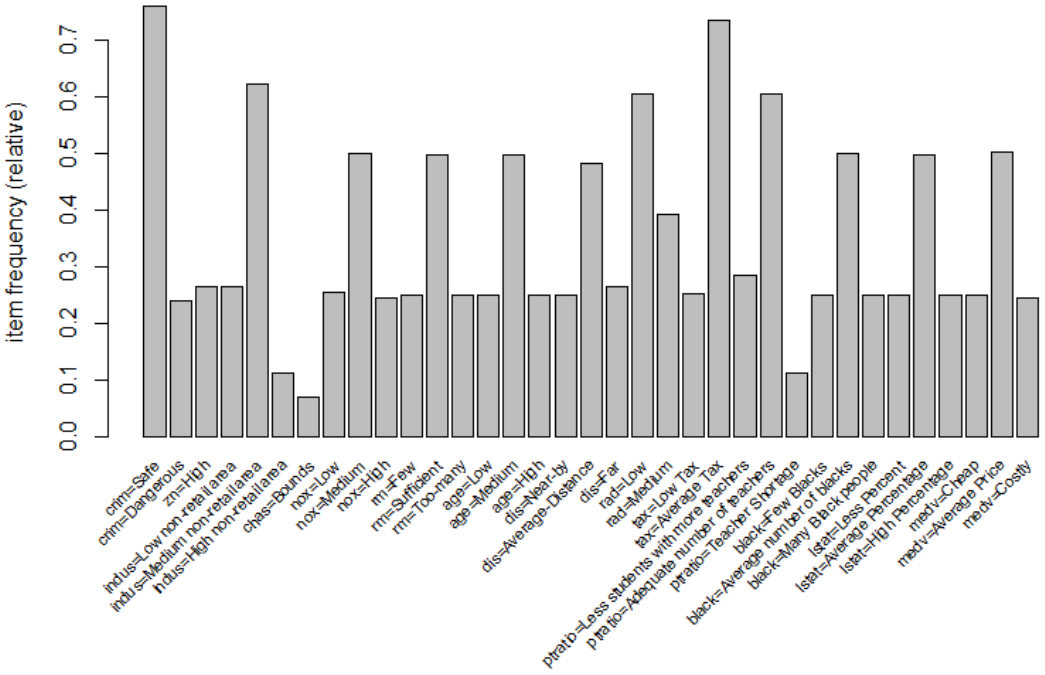
Question 2:

a) ***Histograms***



The grouping into categories has been done based on average value(mean) or, 1st quartile and 3rd quartile values.

b) *Item Frequency Plot*



c) *Student is interested in low crime area as close to the city as possible*

Feature	Value
proportion of non-retail business acres per town.	High non-retail area
Nitric oxide concentration	High
Charles river	Bounded (Yes)
pupil-teacher ratio by town	Teacher Shortage
% Lower status of the population	High
average number of rooms per dwelling	High

d) *For Schools with low-pupil teacher ratio*

Feature	Value
proportion of non-retail business acres per town.	High non-retail area
Charles river	Bounds (Yes)
Median Value of owner-occupied homes	Costly
nitric oxides concentration	High
% lower status of the population	Less
full-value property-tax rate	Low
accessibility to radial highways	Low

e) Regression model results

The results are comparable with NOX (nitric oxide concentration), RAD (accessibility to radial highways), Medv (median value of owner-occupied homes, Indus (proportion of non-retail business acres per town) being common indicator variables among both the methods.

The interpretation of linear regression model is not good when compared to Arules because in the regression model we just get co-efficient values of each variables in terms of positive and negative which have to interpreted by us and is confusing at times. Where as in Association Rules, due to the prior categorization of the variables we get a clear understanding of what level the variable should take to get the desired outcome.

Question 3:

Created a random data that resembles the original data and column target with class as 1. Named the class a 0 and randomly permuted the features in the dataset and then built a decision tree model.

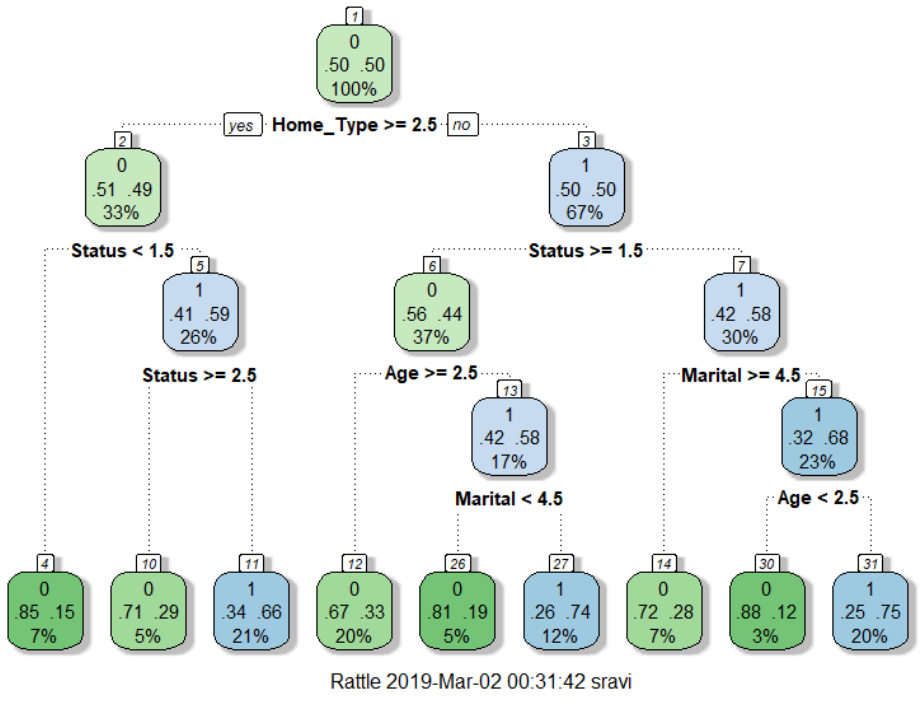
```
> summary(fit_combined)
Call:
rpart(formula = class ~ ., data = combined_data, method = "class",
    control = model.control)
  n= 17986

          CP nsplit rel error    xerror        xstd
1 0.06356796      0 1.0000000 1.0108974 0.007456017
2 0.06215946      3 0.8092961 0.9349494 0.007440667
3 0.05771155      4 0.7471367 0.8032914 0.007310776
4 0.04080952      6 0.6317136 0.6628489 0.007019889
5 0.03869676      7 0.5909040 0.6294896 0.006925771
6 0.02000000      8 0.5522073 0.5867897 0.006790118

Variable importance
    Marital      Status         Age Dual_Income      Income         Edu  Occupation
         33          33          24           4           2           2           1

Node number 1: 17986 observations,    complexity param=0.06356796
  predicted class=0  expected loss=0.5  P(node) =1
    class counts:  8993  8993
   probabilities: 0.500 0.500
  left son=2 (5929 obs) right son=3 (12057 obs)
```

Decision tree:




```
> predicted = predict(fit_combined, combined_data[, -c(15)])
> predicted
          0         1
1 0.2500000 0.7500000
2 0.2500000 0.7500000
3 0.3420421 0.6579579
4 0.2597222 0.7402778
5 0.2597222 0.7402778
6 0.2500000 0.7500000
7 0.3420421 0.6579579
8 0.3420421 0.6579579
9 0.3420421 0.6579579
10 0.3420421 0.6579579
11 0.3420421 0.6579579
```

- From the above model, we can observe that features have predictive power.
- To cross verify this, we can predict the model on the training set itself and observe the probabilities.
- The terminal node has a percentage of 20% with a class 1 probability of 0.75