Home Work-1 Report

Akshay Adlakha (akshayad) - 50317479

Problem 1:

1) Given,

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

a) for point A and B

Jaccard Index $(A,B) = \frac{4}{8} = \frac{1}{2} = 0.5$ Jaccard Distance $(A,B) = 1-0.5 = 0.5$

for point B and C

Jaccard Index $(B,C) = \frac{4}{8} = \frac{1}{2} = 0.5$ Jaccard Distance $(B,C) = 1-0.5 = 0.5$

for point A and C

Jaccard Index $(A,C) = \frac{4}{8} = \frac{1}{2} = 0.5$ Jaccard Distance $(A,C) = 1-0.5 = 0.5$

$$\text{Cos}(A,B) = \frac{5\times3 + 5\times3 + 1\times1 + 3\times1}{\sqrt{4^2+5^2+5^2+1^2+3^2+2^2} \cdot \sqrt{3^2+4^2+3^2+1^2+2^2+1^2}}$$

$$= \frac{15+15+1+3}{\sqrt{16+25+25+1+9+4} \cdot \sqrt{9+16+9+1+4+1}}$$

$$= \frac{34}{\sqrt{80} \cdot \sqrt{40}} = \frac{34}{8.944 \times 6.324}$$

$$= \frac{34}{56.56} = 0.6011$$

$$\text{Cos}(B,C) = \frac{4\times1 + 3\times3 + 2\times4 + 1\times5}{\sqrt{3^2+4^2+3^2+1^2+2^2+1^2} \cdot \sqrt{2^2+1^2+3^2+4^2+5^2+3^2}}$$

$$= \frac{26}{\sqrt{40} \cdot \sqrt{64}} = \frac{26}{6.3245 \times 8} = \frac{26}{50.596} = 0.513874 \simeq 0.514$$

$$\cos(A,C) = \frac{4\times2 + 5\times3 + 3\times5 + 2\times3}{\sqrt{4^2+3^2+5^2+1^2+3^2+2^2} \cdot \sqrt{2^2+1^2+3^2+4^2+5^2+3^2}}$$

$$= \frac{8+15+15+6}{\sqrt{80} \cdot \sqrt{64}} = \frac{44}{\sqrt{80} \cdot 8} = \frac{44}{71.55419}$$

$$= 0.614919$$

$$\simeq 0.615$$

b) Treating 3, 4, 5 as 1 and 1, 2 and blank as 0.

we have

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	1	0	1	1	1

Jaccard Index $(A,B) = \frac{2}{5} = 0.4$ Jaccard Distance $(A,B) = 1 - 0.4 = 0.6$

Jaccard Index $(B,C) = \frac{1}{6} = 0.167$ Jaccard Distance $(B,C) = 1 - 0.167 = 0.833$

Jaccard Index $(A,C) = \frac{2}{6} = 0.333$ Jaccard Distance $(A,C) = 1 - 0.333 = \cancel{0.667}$ 0.667

$$\cos(A,B) = \frac{1\times1 + 1\times1}{\sqrt{1^2+1^2+1^2+1^2} \cdot \sqrt{1^2+1^2+1^2}} = \frac{2}{\sqrt{4}\cdot\sqrt{3}} = \frac{\cancel{2}}{\cancel{2}\sqrt{3}}$$

$$= \frac{1}{\sqrt{3}} = 0.5780$$

$$\cos(B,C) = \frac{1\times1}{\sqrt{1^2+1^2+1^2} \cdot \sqrt{1^2+1^2+1^2+1^2}} = \frac{1}{\sqrt{3}\cdot\sqrt{4}} = \frac{1}{2\sqrt{3}}$$

$$= 0.28867$$

$$\simeq 0.289$$

$$\cos(A,C) = \frac{1\times1 + 1\times1}{\sqrt{1^2+1^2+1^2+1^2} \cdot \sqrt{1^2+1^2+1^2+1^2}} = \frac{2}{\sqrt{4}\cdot\sqrt{4}} = \frac{\cancel{2}}{\cancel{4}} = \frac{1}{2}$$

$$= 0.5$$

Jaccard Distance, we know, it is a measure of how dissimilar two sets are.

In part A)

The Jaccard distance between each users A, B and C is 0.5. They all are 50% similar.

In part B)

when we replaced ratings 3,4,5 as 1 and ratings 1,2 and blank as 0. And, computed the Jaccard distance,

The dissimilarity b/w A and B increases. ~~the~~ Same is the case between {A ~~and~~ C} and B ~~and~~ C.

The Jaccard distance increases between each of the users.

And, the Cosine Similarity is a measure how similar they are irrespective of the size. It measures the Cosine of the angle which tells how closely they are oriented together.

The cosine similarity ~~dese~~ decreases b/w each user when we replace the ratings by 0 and 1.

Smaller the angle, higher the similarity.

c) Average value for A $= \frac{4+5+5+1+3+2}{6}$

$= \frac{20}{6} = 3.33$

Average value ~~A~~ for B $= \frac{3+4+3+1+2+1}{6}$

$= 2.33$

Average value for C $= \frac{2+1+3+4+5+3}{6}$

$= \frac{18}{6} = 3.$

Normalizing the matrix by subtracting from each non-blank entry the average value. we get.

	a	b	c	d	e	f	g	h
A	0.67	1.67		1.67	-2.33		-0.33	-1.33
B		0.67	1.67	0.67	-1.33	-0.33	-1.33	
C	-1		-2	0		1	2	0

$$\cos(A,B) = \frac{1.67\times0.67 + 1.67\times0.67 + (-2.33)\times(-1.33) + (-0.33)\times(-1.33)}{\sqrt{(0.67)^2+(1.67)^2+(1.67)^2+(-2.33)^2+(-0.33)^2+(-1.33)^2} \times \sqrt{(0.67)^2+(1.67)^2+(0.67)^2+(-1.33)^2+(-0.33)^2+(-1.33)^2}}$$

$$= \frac{1.1189 + 1.1189 + 3.0989 + 0.4522}{\sqrt{0.4489+2.7889+2.7889+5.4289+0.1089+1.7689} \times \sqrt{0.4489+2.7889+0.4489+1.7689+0.1089+1.7689}}$$

$$= \frac{5.7889}{\sqrt{13.3334}\times\sqrt{7.334}} = \frac{5.7889}{9.8883} = 0.5854$$

Scanned by CamScanner

$$\cos(B, C) = \frac{-2\times1.67 + (-0.33)\times1 + (-1.33)\times2}{\sqrt{(0.67)^2 + (1.67)^2 + (0.67)^2 + (-1.33)^2 + (-0.33)^2 + (-1.33)^2} * \sqrt{(-1)^2 + (-2)^2 + 1^2 + (2)^2}}$$

$$= \frac{-3.34 - 0.33 - 2.66}{\sqrt{7\cancel{3}.334} * \sqrt{10}}$$

$$= \cancel{\frac{-6.33}{3.6515 * 3.1622}} = \cancel{\frac{-6.33}{11.5467}}$$

$$= \frac{-6.33}{2.7081\times3.1622} = \frac{-6.33}{8.5637}$$

$$= -0.739166$$

$$\simeq -0.74$$

$$\cos(A, C) = \frac{0.67 * -1 + (-0.33 * 2)}{\sqrt{13.3334} * \sqrt{10}}$$

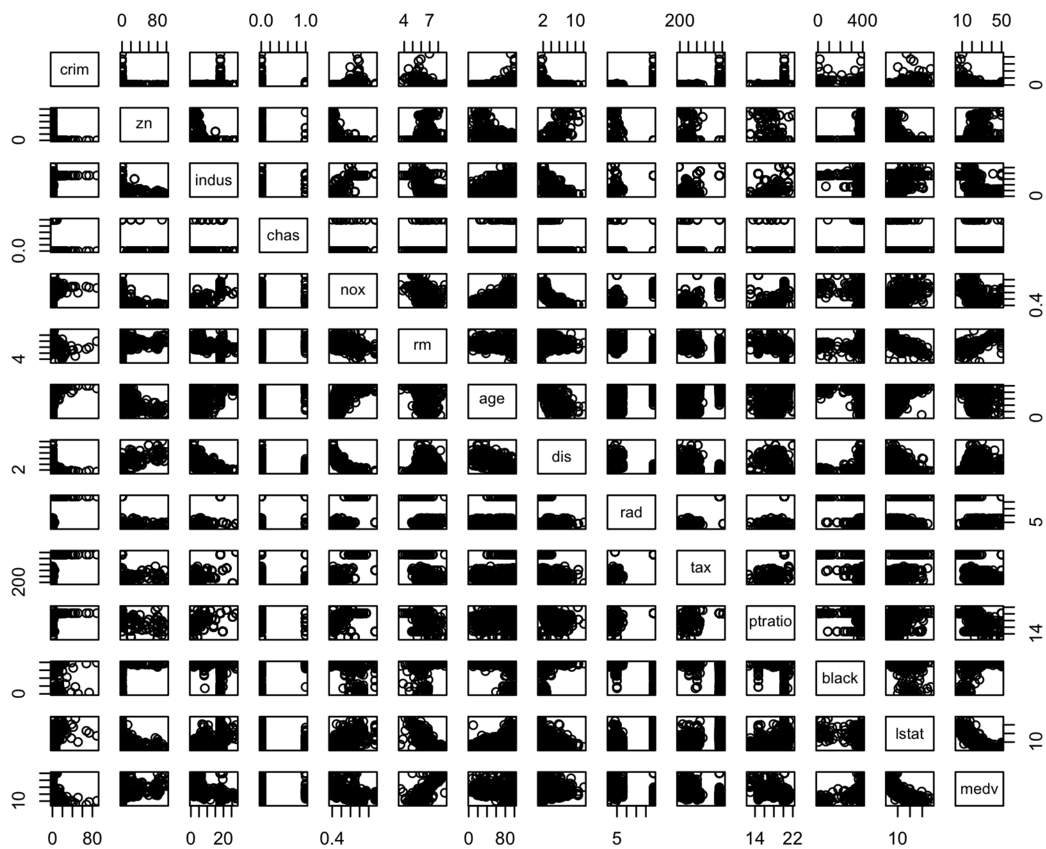
$$= \frac{-1.33}{3.6514 * 3.1622} = \frac{-1.33}{11.5464}$$

$$= -0.115187418$$

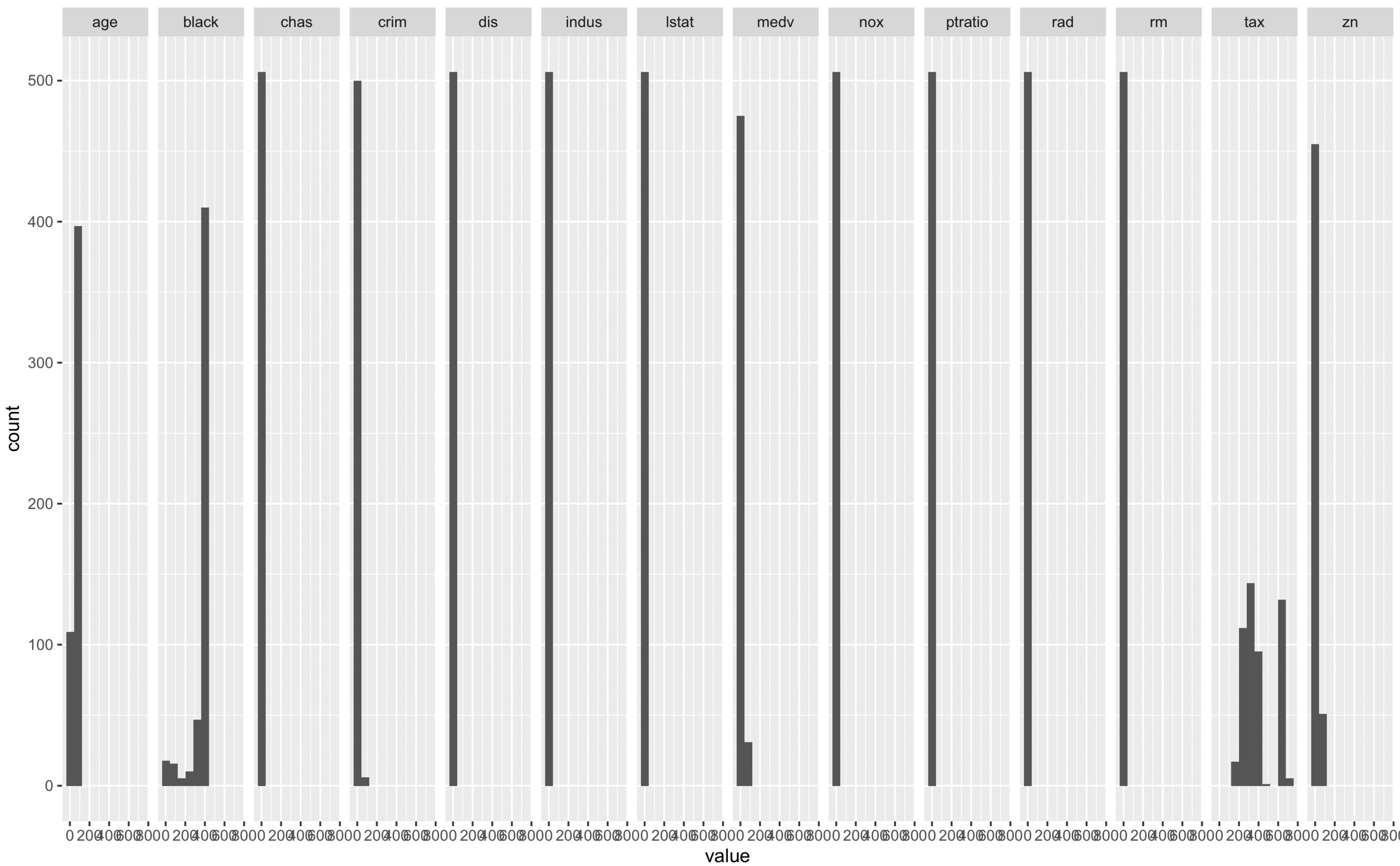
$$\simeq -0.12$$

Problem 2: In this problem, I was given Boston housing data from MASS package.

After loading the data, did a sanity by checking NA values. Then, I plotted the data to visualize it.

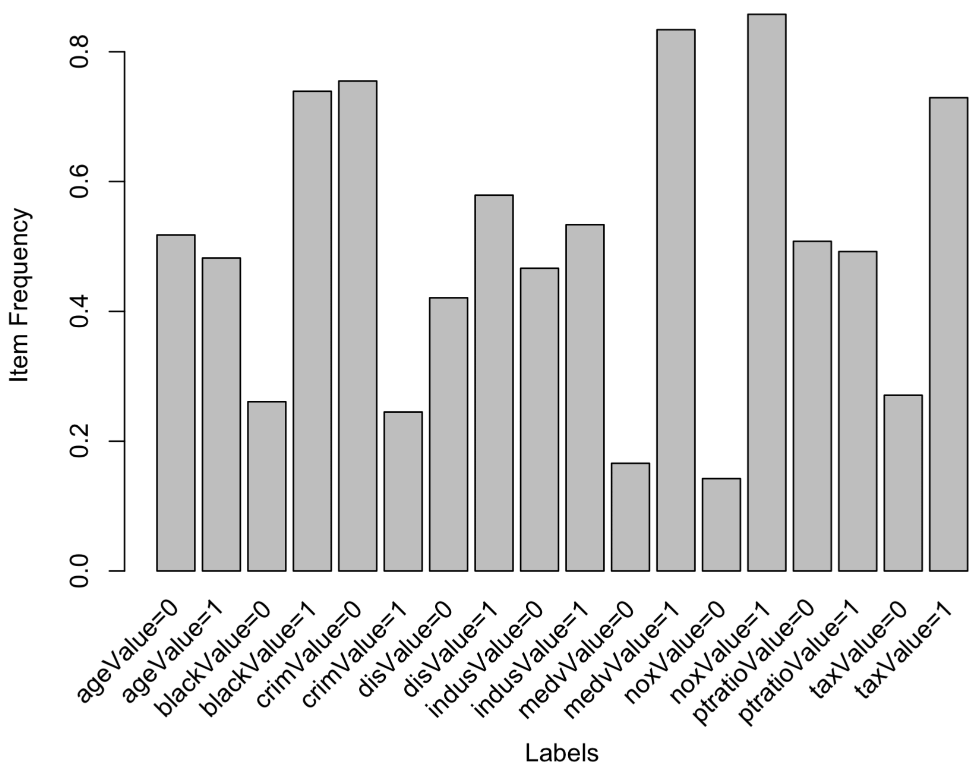


Then, used the histogram to understand the dataset. Here is the result:



Then, transformed data into binary form. For example, took the value of black > 396, tax > 500 and medv > 30 to split the data by looking out the summary of each feature in the dataset.

Plotted the frequency plot using itemFrequencyPlot. Here is the result:



From the above result, we can observe that the frequency for medv =1 and nox =1 is high. On taking the support =0.05 and the confidence of 0.4, apriori gives us 12976 rules and with confidence of 0.6 with same support gives us 10546 rules.

```
set of 10546 rules

rule length distribution (lhs + rhs):sizes
   1    2    3    4    5    6    7    8    9 
   5  117  739 2129 3243 2725 1254  302   32 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   4.000   5.000   5.217   6.000   9.000 

summary of quality measures:
    support          confidence          lift            count       
 Min.   :0.05138   Min.   :0.6000   Min.   :0.6995   Min.   : 26.00  
 1st Qu.:0.07115   1st Qu.:0.7561   1st Qu.:1.1480   1st Qu.: 36.00  
 Median :0.10277   Median :0.8857   Median :1.3458   Median : 52.00  
 Mean   :0.12838   Mean   :0.8652   Mean   :1.4489   Mean   : 64.96  
 3rd Qu.:0.15810   3rd Qu.:1.0000   3rd Qu.:1.7261   3rd Qu.: 80.00  
 Max.   :0.85771   Max.   :1.0000   Max.   :3.6934   Max.   :434.00  

mining info:
            data ntransactions support confidence
 transactionData           506    0.05        0.6
```

Result for low crime rules:

```
     lhs                 rhs             support confidence     lift count
[1] {blackValue=1,
     disValue=0,
     medvValue=0}     => {crimValue=1} 0.05533597        0.8 3.264516    28
[2] {ageValue=1,
     blackValue=1,
     disValue=0,
     medvValue=0}     => {crimValue=1} 0.05533597        0.8 3.264516    28
[3] {blackValue=1,
     disValue=0,
     medvValue=0,
     ptratioValue=1} => {crimValue=1} 0.05533597        0.8 3.264516    28
[4] {blackValue=1,
     disValue=0,
     indusValue=1,
     medvValue=0}     => {crimValue=1} 0.05533597        0.8 3.264516    28
[5] {blackValue=1,
     disValue=0,
     medvValue=0,
     taxValue=1}      => {crimValue=1} 0.05533597        0.8 3.264516    28
[6] {blackValue=1,
     disValue=0,
     medvValue=0,
     noxValue=1}      => {crimValue=1} 0.05533597        0.8 3.264516    28
```

From the above results, we can see that we are not having such rules that match both low crime data and to be as close to the city as possible.

Result for low distance rules:

```
    lhs                             rhs          support   confidence lift     count
[1] {noxValue=0}                 => {disValue=1} 0.1422925 1          1.726962 72
[2] {noxValue=0,taxValue=0}      => {disValue=1} 0.1106719 1          1.726962 56
[3] {indusValue=0,noxValue=0}    => {disValue=1} 0.1422925 1          1.726962 72
[4] {noxValue=0,ptratioValue=0} => {disValue=1} 0.1106719 1          1.726962 56
[5] {ageValue=0,noxValue=0}      => {disValue=1} 0.1422925 1          1.726962 72
[6] {blackValue=1,noxValue=0}    => {disValue=1} 0.1146245 1          1.726962 58
```

Here we can see that the Nox values are so high for low distance.

Result for low putratio rules:

```
    lhs                                      rhs              support    confidence
[1] {blackValue=1,crimValue=1,medvValue=0}  => {ptratioValue=1} 0.05928854 1
[2] {blackValue=1,disValue=0,medvValue=0}   => {ptratioValue=1} 0.06916996 1
[3] {ageValue=1,blackValue=1,medvValue=0}   => {ptratioValue=1} 0.08300395 1
[4] {disValue=1,indusValue=1,medvValue=0}   => {ptratioValue=1} 0.05928854 1
[5] {blackValue=1,indusValue=1,medvValue=0} => {ptratioValue=1} 0.11462451 1
[6] {disValue=1,medvValue=0,taxValue=1}     => {ptratioValue=1} 0.06916996 1
    lift     count
[1] 2.032129 30
[2] 2.032129 35
[3] 2.032129 42
[4] 2.032129 30
[5] 2.032129 58
[6] 2.032129 35
```

From this, we can infer that the pupil teacher ratio is low where the proportion of black is low, distance is low, proportion of non-retail business is low and the median value of owner-occupied homes is relatively high.

Then build a regression model. Here is the result:

```
Call:
lm(formula = ptratioValue ~ ., data = lmdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97694 -0.07909  0.07719  0.32456  0.63415 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.30758    0.09197   3.344 0.000887 ***
ageValue     0.12954    0.04666   2.776 0.005706 ** 
blackValue   0.05416    0.03792   1.428 0.153800    
crimValue    0.12589    0.04585   2.746 0.006255 ** 
disValue     0.07302    0.04677   1.561 0.119119    
indusValue  -0.12126    0.05040  -2.406 0.016505 *  
medvValue   -0.35566    0.04769  -7.457 3.96e-13 ***
noxValue    -0.15629    0.05520  -2.831 0.004823 ** 
taxValue     0.69147    0.05135  13.466  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3696 on 497 degrees of freedom
Multiple R-squared:  0.4631,	Adjusted R-squared:  0.4544 
F-statistic: 53.58 on 8 and 497 DF,  p-value: < 2.2e-16
```

The regression model has somewhat the same results. As we can see that the crime rate low, distance is low and median value of owner occupied homes is more. And, the property tax is also high.

The regression is usually preferred when we have the features in numeric values. In this case, this becomes easier to interpret using association rules as we have the binary features.

Problem 3: In this problem, I was given a dataset of marketing from ElemStatLearn package. I had to cluster demographic data using a classification tree.

After loading the data, did a sanity check by checking NA values. It had 2694 missing values. I replaced the missing values by median value.

Then, took the target column as class 1 and randomly permuted the values within each feature and took the target column as class 0 in reference sample. Then, combined both the data.

Then, build a decision tree model. Here is the result of the model.

```
Call:
rpart(formula = target ~ ., data = dataset)
  n= 17986

    CP nsplit rel error xerror xstd
1 0.01      0         1      0    0

Node number 1: 17986 observations
  predicted class=0  expected loss=0.5  P(node) =1
    class counts:  8993  8993
   probabilities: 0.500 0.500
```

From this result, we can observe that it has 0.5 probabilities for each class and has one single root only which indicates that features in the dataset do not the predictive power to do a classification.

Then, predicted the model on the training data and got to see the probability for every row is 0.5.

Thus, conclude that it has no predictive power to do a classification.