Samantha Misurda

Smisurda

HW4

Problem 1

a) 1.205741

b)

information gain: 0.08681003

relative information gain: 0.07199724

c)

(i) As noted in the tables below, as the size of feature sets increases, the information gain and relative information gain increases.

```
# Feature set n = 1
feature.set.1 <- features(my.data.no.predictions, 1)
feature.set.1
```

```
##    feature          IG         RIG
## 4   safety 0.261399361 0.216795620
## 2 persons 0.219662963 0.182180890
## 5   price 0.086810025 0.071997243
## 3   trunk 0.030008141 0.024887718
## 1   doors 0.004485717 0.003720299
```

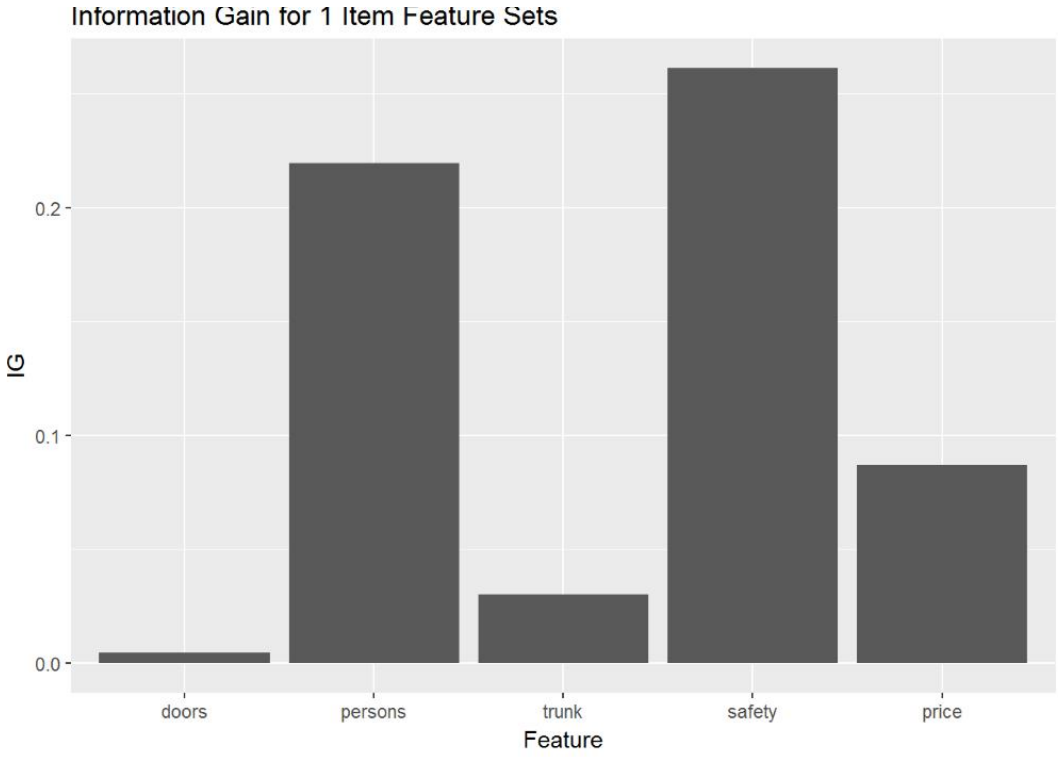
```
# Feature set n = 2
feature.set.2 <- features(my.data.no.predictions, 2)
feature.set.2
```

```
##           feature         IG        RIG
## 6  persons safety 0.52319987 0.43392394
## 10   safety price 0.36364237 0.30159245
## 8    trunk safety 0.32080048 0.26606086
## 7   persons price 0.31469535 0.26099747
## 3    doors safety 0.26882309 0.22295261
## 5   persons trunk 0.25467014 0.21121464
## 1   doors persons 0.22773086 0.18887213
## 9     trunk price 0.12199301 0.10117679
## 4     doors price 0.09174707 0.07609186
## 2     doors trunk 0.04248684 0.03523712
```

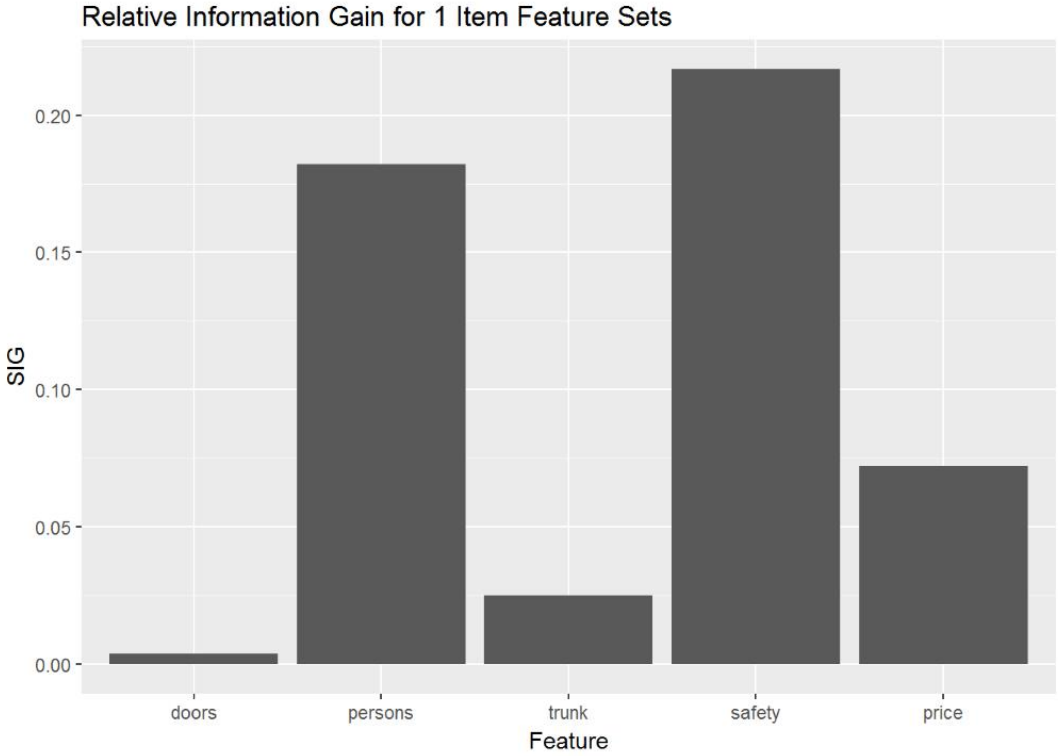
```
# Feature set n = 3
feature.set.3 <- features(my.data.no.predictions, 3)
feature.set.3
```

```
##                 feature         IG        RIG
## 9  persons safety price 0.6626263 0.5495594
## 7  persons trunk safety 0.5951881 0.4936285
## 2  doors persons safety 0.5371800 0.4455186
## 10   trunk safety price 0.4395222 0.3645246
## 6    doors safety price 0.3724507 0.3088978
## 8   persons trunk price 0.3552130 0.2946014
## 4    doors trunk safety 0.3487651 0.2892537
## 3   doors persons price 0.3237841 0.2685353
## 1   doors persons trunk 0.2819851 0.2338687
## 5     doors trunk price 0.1365830 0.1132772
```

(ii). See the bar charts below




```
# Plot Relative information gain
ggplot(data = feature.set.1, aes(x = feature, y = RIG)) +
    geom_bar(stat = "identity")+ ggtitle("Relative Information Gain for 1 Item Feature Sets")
  labs(x="Feature",y="SIG")
```



(iii.) As expected, customers seem most concerned about having a safe car that can accommodate a reasonable amount of passengers. Although price was important, I was surprised that it was not more so.

d.

If we used ID3, I would predict that the safety feature would be used. This is primarily due to the lecture notes specifying that the root node should be the best predictor of the output.

Problem 2

A.

- Number of rules found that meet the set criteria
  21

- The top five rules sorted by support
  [1] {class=unacc} => {price=high} 0.3958333 0.5652893 1.130579
  [2] {safety=high} => {price=high} 0.1684028 0.5069686 1.013937
  [3] {trunk=med} => {price=high} 0.1666667 0.5000000 1.000000
  [4] {trunk=small} => {price=high} 0.1666667 0.5000000 1.000000
  [5] {persons=4} => {price=high} 0.1666667 0.5000000 1.000000

- The top five rules sorted by the confidence score.

[1] {persons=4,class=unacc} => {price=high} 0.1145833 0.6346154 1.269231
[2] {safety=high,class=unacc} => {price=high} 0.1006944 0.6327273 1.265455
[3] {safety=med,class=unacc} => {price=high} 0.1302083 0.6181319 1.236264
[4] {persons=more,class=unacc} => {price=high} 0.1145833 0.6149068 1.229814
[5] {trunk=med,class=unacc} => {price=high} 0.1290509 0.5688776 1.137755

C.

I'm not entirely sure what it means to be conservative in this case. I am assuming it means the tightest confidence interval range, which in this case would be:

{persons=more,class=unacc} => {price=high} 0.1145833 0.6149068 1.229814 0.5919627 0.6378510

The rule with the most confidence however is:

{persons=4,class=unacc} => {price=high} 0.1145833 0.6346154 1.269231 0.6119107 0.6573200

Problem 3

A. [1] 0.6076389

B. AUC Score : [1] 0.6099537
The prices of cars in good and vgood class can be predicted with the highest certainty.

C. Looking at the complexity of our trees, we should be more concerned with overfitting the model based on our training set. In order to better tune the models to prevent overfitting, we could prune the tree after training. Additionally, cross validation techniques could be used.