R Report

1.Decision Tree

1.1 Business Understanding

Using decision tree way to analysis this data car evaluation.
The car data comes from UCI: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation. In total, collect consumers feedback, according six features(car purchase cost, maintenance fee, the door number, capacity of car, the trunk capacity of a vehicle, estimated safety of the car) to explore the relationship with car evaluation(acceptability).

1.2 Data Understand & Preparation

1. buying_price: buying car needs to pay (high, low, med, vhigh)
2. maint_price: maintain costing (high, low, med, vhigh)
3. doors: the trunk capacity of a vehicle (2, 3, 4 or 5more)
4. persons: capacity of car (2, 4, more)
5. lug_boot: size of luggage boot (big, med, small)
6. safety: estimated safety of the car (high, low, med)
7. acceptability: consumer acceptance of cars (acc, good, unacc, vgood)

```
> summary(datacar)
 buying_price maint_price   doors     persons     lug_boot    safety    acceptability
 high :432    high :432   2    :432   2    :576   big  :576   high:576   acc  : 384
 low  :432    low  :432   3    :432   4    :576   med  :576   low :576   good :  69
 med  :432    med  :432   4    :432   more:576   small:576   med :576   unacc:1210
 vhigh:432    vhigh:432   5more:432                                      vgood:  65
```

```
table(datacar$acceptability)

acc  good unacc vgood
384    69  1210    65
```

1-1

In this data, in total has 1728 data. The most of results belong to unacc, has 1210, it accounts for 70% of all results. So selecting the first 1200 data as training. 1201-1728 data as testing. Using runif() order the data, because checking at the data, you can see that the data is not evenly distributed. When finishing the reorder, it can set training and testing data.

1.3 Modeling

Using C5.0 to explore to deal with car data. Building a model to judge the car's acceptability,

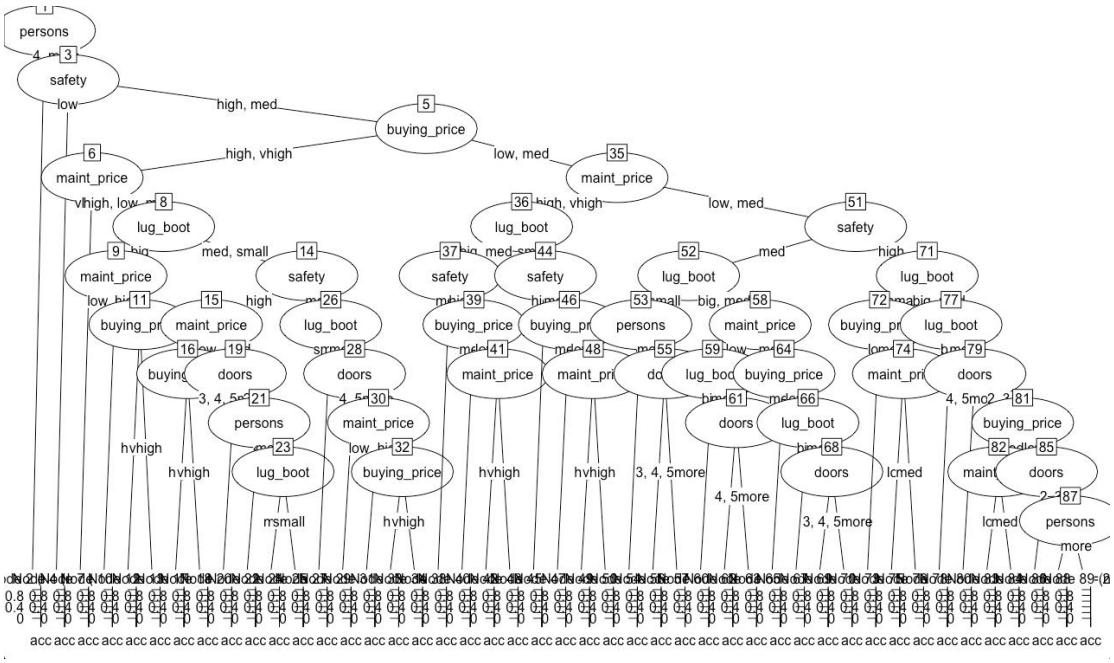
according to the buying_price, maint_price, doors, persons, lug_boot and safety.

```
> model=C5.0(acceptability~.,data=datacar)
> summary(model)

Evaluation on training data (1200 cases):

     Decision Tree
   ----------------
   Size      Errors

     44   17( 1.4%)   <<


    (a)   (b)   (c)   (d)    <-classified as
   ----  ----  ----  ----
   250     6           2    (a): class acc
          49           1    (b): class good
     5     3   839          (c): class unacc
                      45    (d): class vgood


  Attribute usage:

  100.00% safety
   65.75% persons
   43.58% buying_price
   43.58% maint_price
   35.75% lug_boot
    9.50% doors


Time: 0.0 secs
```

1-2

From the 3-2, we can see the summary(model)'s result. Training data(1200 cases) has 17 error. The acceptability has four class, acc (acceptability), good, unacc (unacceptability), vgood (very good). the acc has 250, good has 49, unacc has 839, vgood has 45.

Consumers care about the car's information. Attribute usage: the top is safety, has 100%. Persons has 65.75%, buying_price and maint_price are same, 43.58%. lug_boot and doors are 35.75% and 9.5%,respectively.



1-3

Choosing the features which have relationship with acceptability, it can generate decision tree. In this model I choose all features to build, because all features have relative with acceptability. The persons is the root node.

```
> CrossTable(predictions,datacar_test$acceptability,prop.chisq = FALSE,
+            prop.c=FALSE,prop.r = FALSE,
+            dnn=c('predicted','actual'))
```

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  528


             | actual
   predicted |       acc |      good |     unacc |     vgood | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|
         acc |       126 |         0 |         1 |         0 |       127 |
             |     0.239 |     0.000 |     0.002 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
        good |         2 |        17 |         1 |         0 |        20 |
             |     0.004 |     0.032 |     0.002 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
       unacc |         0 |         0 |       359 |         0 |       359 |
             |     0.000 |     0.000 |     0.680 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
       vgood |         0 |         0 |         0 |        22 |        22 |
             |     0.000 |     0.000 |     0.000 |     0.042 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
Column Total |       128 |        17 |       361 |        22 |       528 |
-------------|-----------|-----------|-----------|-----------|-----------|
```

1-4

Testing has 528 data, via build model and predict model. We can see the model has high accuracy. Comparing with the predicted and actual. About the acc, only one data predicts error.

Good has three data error. Unacc and vgood not have prediction error. The forecast agrees with the actual result. Predicted data: acc has 127, good has 20, unacc ahs 359 and vgood has 22. Actual data: acc has 128, good has 17, unacc has 361, vgood has 22.

1.4 Evaluation

In conclude, the model is very close to actual.Through CrossTable, we can see the comparison between the predicted and actual. Training data and testing data were selected appropriately, and training data accounted for 70% of the total.