DATA MINING - 5334

Last Name: Alagiya
First Name: Ravi
UTA ID: 1001452485 | rxa2485
Assignment: Assignment 6
Due Date: 11/20/2016 11:28 PM

Questions:

A) Construct a Linear Regression Classifier
- Plot it
- Test it

B) Construct Decision Tree Classifier
- Plot it
- Test it

Answer:

About Data set:

The Insurance Company Benchmark (COIL 2000) from http://kdd.ics.uci.edu/

This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was collected to answer the following question: Can you predict who would be interested in buying a caravan insurance policy?

Link of the raw data: http://kdd.ics.uci.edu/databases/tic/tic.html
For the regression I have used year wise housing price data.

Data Mining Task:

In this dataset, the data mining task is to create regression rules and Decision tree based on 85 attributes of the user. Such rule can help The Insurance Company to improve its prediction of sales and knowing their target customer.

This assignment shows how we can create regression and decision tree classifier in R.

Decision Tree Classifier

R Code:

Below is the R code for creating Decision Tree

```
library(party)

#Training Data
Training.Data<-read.table("/Users/Ravi/Documents/Data
Mining/Assignments/Assignment_6/trainingData.txt",header=T,sep="¥t")

#Test Data
Test.Data<-read.table("/Users/Ravi/Documents/Data
Mining/Assignments/Assignment_6/TestData_Insurance.txt",header=T,sep="¥t")

#creating tree formula based on important attributes
myFormula <- Number_of_mobile_home_policies ~ Customer_Subtype_see_L0__ +
  Number_of_houses + Avg_size_household + Avg_age_see_L1 +
  Customer_main_type_see + Roman_catholic_see_L3__ + Protestant +
  Other_religion____ + No_religion____ + Married_____ + Living_together____ +
  Other_relation____ + Singles_____ + Household_without_children___ +
  Household_with_children___ + High_level_education___ +
Medium_level_education___ +
  Lower_level_education___ + High_status____ + Entrepreneur_____ +
  Farmer_____ + Middle_management____ + Skilled_labourers____ +
  Unskilled_labourers____ + Social_class_A___ + Social_class_B1___
```

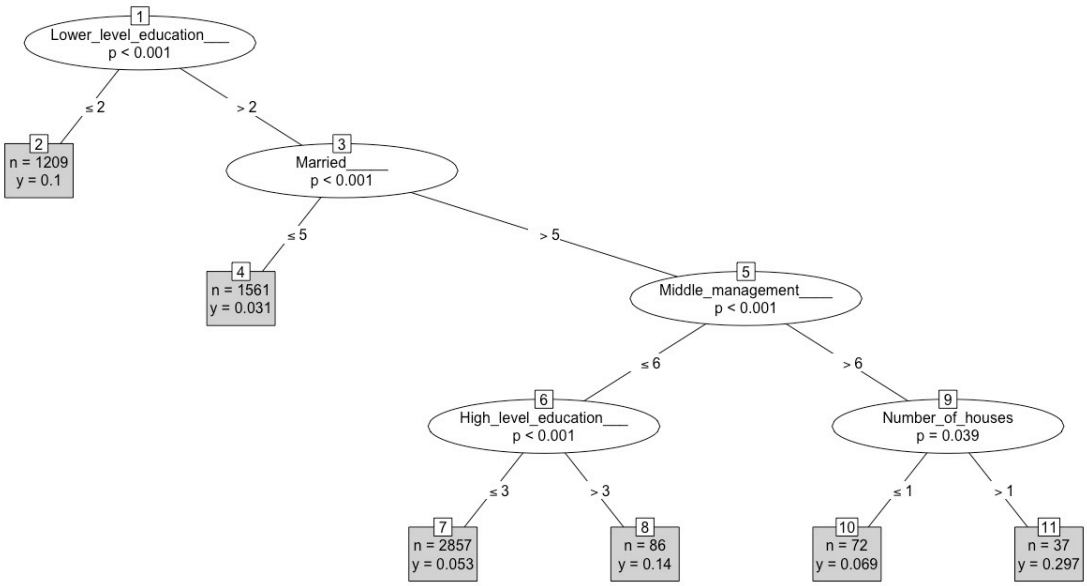
```
# creating Tree using ctree function
iris_ctree <- ctree(myFormula, data = Training.Data)

# check the prediction
Print(iris_ctree)

#plot the created Decision Tree
plot(iris_ctree, type = "simple")

#Testing data on created tree modal
predicted <- predict(iris_ctree, newdata = Test.Data)
#classify in 0 &1
Class<-ifelse(predicted < 0.05, 0, 1)
print(Class)
```

Output:

```
Conditional inference tree with 6 terminal nodes

Response:  Number_of_mobile_home_policies
Inputs:  Customer_Subtype_see_L0__, Number_of_houses, Avg_size_household,
    Avg_age_see_L1, Customer_main_type_see, Roman_catholic_see_L3__, Protestant,
    Other_religion____, No_religion____, Married_____, Living_together____,
    Other_relation____, Singles_____, Household_without_children___,
    Household_with_children___, High_level_education___,
    Medium_level_education___, Lower_level_education___, High_status____,
    Entrepreneur_____, Farmer_____, Middle_management____, Skilled_labourers____,
    Unskilled_labourers____, Social_class_A___, Social_class_B1___
Number of observations:  5822

1) Lower_level_education___ <= 2; criterion = 1, statistic = 47.74
  2)*  weights = 1209
1) Lower_level_education___ > 2
  3) Married_____ <= 5; criterion = 1, statistic = 21.95
    4)*  weights = 1561
  3) Married_____ > 5
    5) Middle_management____ <= 6; criterion = 0.999, statistic = 17.079
      6) High_level_education___ <= 3; criterion = 0.999, statistic = 18.178
        7)*  weights = 2857
```

6) High_level_education___ > 3
8)* weights = 86
5) Middle_management____ > 6
9) Number_of_houses <= 1; criterion = 0.961, statistic = 10.038
10)* weights = 72
9) Number_of_houses > 1
11)* weights = 37

Plot of the Graph:



Classification of Test data:

[3002,]	0
[3003,]	1
[3004,]	0
[3005,]	1
[3006,]	0
[3007,]	1
[3008,]	1
[3009,]	0
[3010,]	1
[3011,]	1

[3012,]	1
[3013,]	1
[3014,]	1
[3015,]	1
[3016,]	1
[3017,]	1
[3018,]	1
[3019,]	1

In above data 0 represents “No” and 1 represents “Yes”. So 1 represents that the customer is likely to buy Car insurance.

Linear Regression Classifier

R Code:

```
#Training Data (house price in year with interest rate)
Training.Data<-read.table("/Users/Ravi/Documents/Data
Mining/Assignments/Assignment_6/newData12.txt",header=T,sep=" ")

#Test Data
Test.Data<-read.table("/Users/Ravi/Documents/Data
Mining/Assignments/Assignment_6/testData12.txt",header=T,sep=" ")

#creating Regression using lm funcation
lm.out=lm(Median_home_price~Year,data=Training.Data)

#summary
summary(lm.out);

#plot outcome of the regression modal
plot(lm.out);

#plotting price vs year
plot(Median_home_price~Year, data=Training.Data, main=" plot")

#drawing line in graph
abline(lm.out, col="red")

#Testing newly created modal using regression modal
next_year <- predict(lm.out, newdata = Test.Data)

#printing next three years predicted values based on created regression model
print(next_year)

data<-c(Test.Data,next_year)
```

Output:

```
Call:
lm(formula = Median_home_price ~ Year, data = Training.Data)

Residuals:
   Min     1Q Median     3Q    Max
-59429 -39859   6875  34983  66809

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25833898    3312555  -7.799 3.52e-07 ***
Year            13054       1658   7.872 3.08e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42760 on 18 degrees of freedom
Multiple R-squared:  0.7749,  Adjusted R-squared:  0.7624
F-statistic: 61.96 on 1 and 18 DF,  p-value: 3.084e-07
```

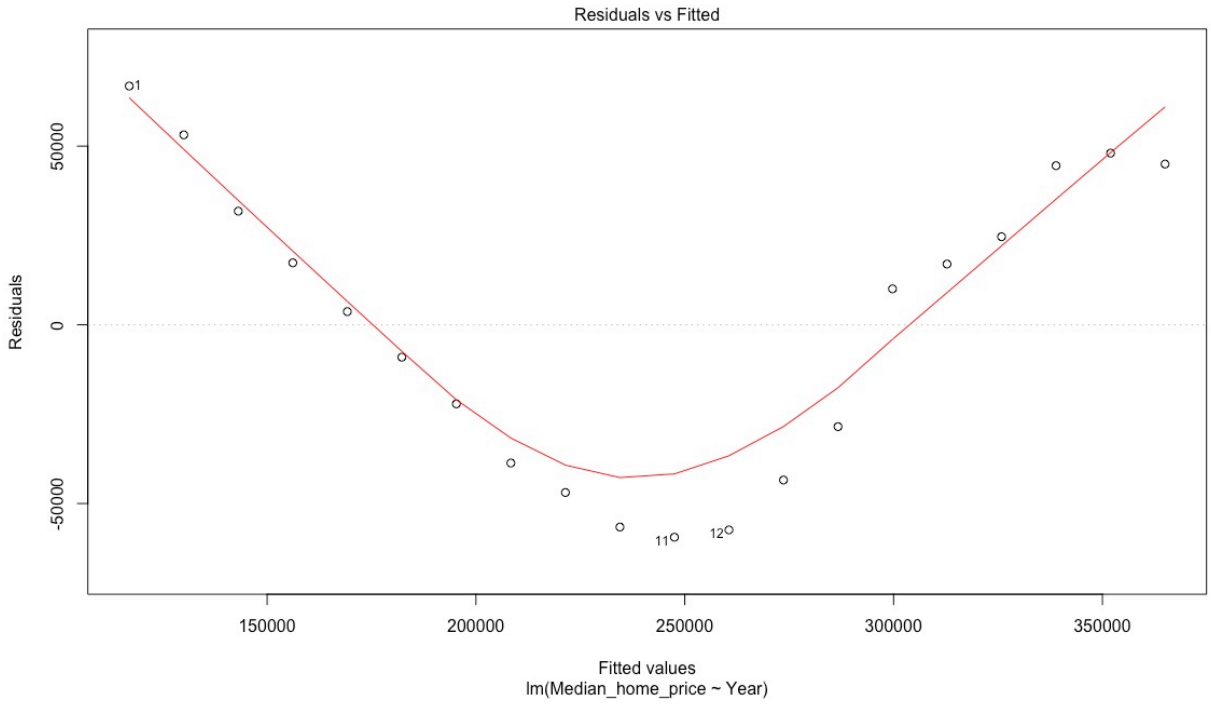
Below is classified value for the year 2008 2009 2010 using regression classifier :

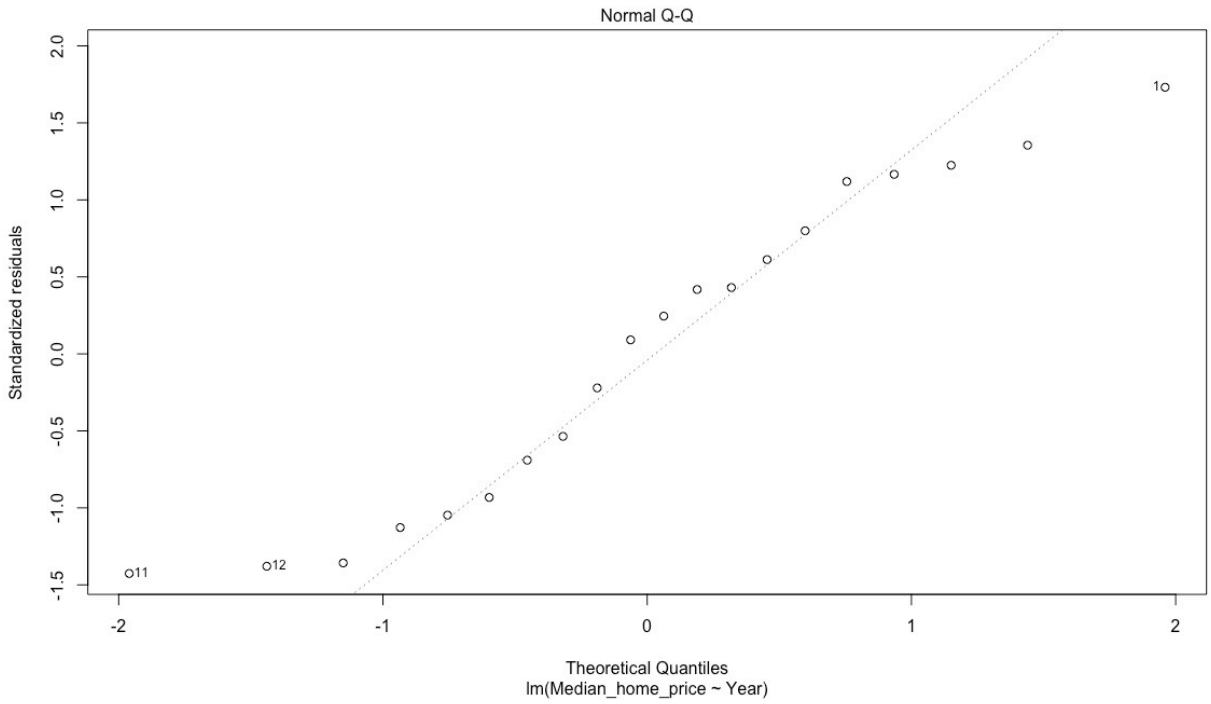
```
$Year
[1] 2008 2009 2010

$interest_rate
[1] 10.3 10.3 10.1

$`1`
[1] 378066.3

$`2`
[1] 391120

$`3`
[1] 404173.8
```

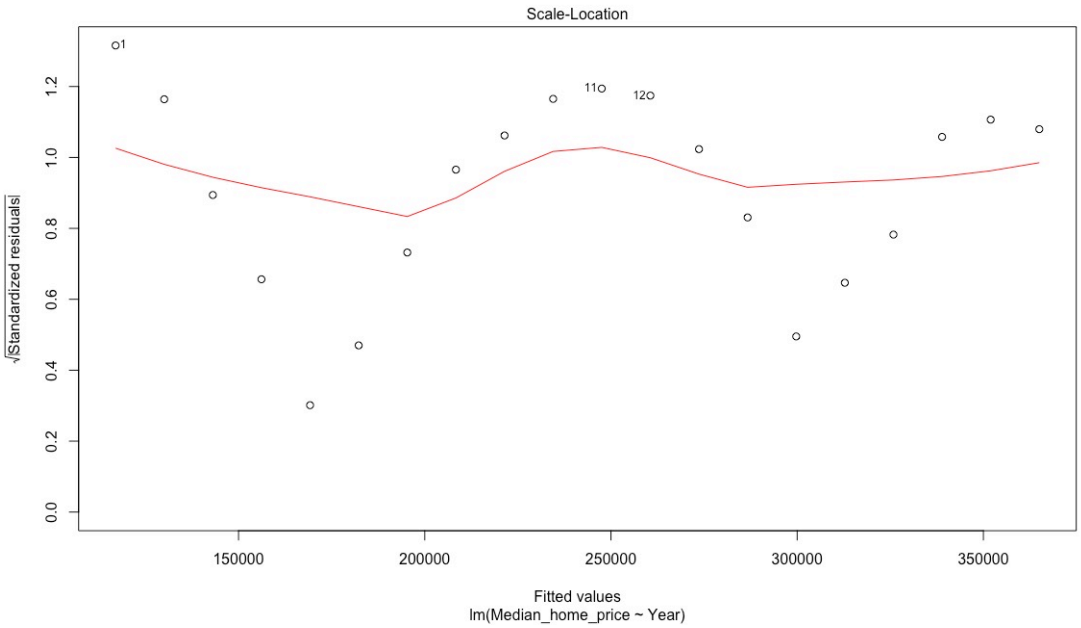
Formula derived using regression:

Home Price= 13054 * (Year) - 25833898

Graph: Residuals vs Fitted







Plot of Year vs Price of houses:

On X axis year, and on Y axis median home price is drawn.
The Line represents modal derived from regression classifiers.