PREDICTION COMPETITION 4

The MSE from the training data, I got : 0.6249354

The MSE from the test data, I got : 0.4407304

Loading data, creating factors and omitting data :

```
library(readr)


Auto <- read.csv("Desktop/test_data_predcomp4.csv", header = T, na.strings = "?", stringsAsFactors = T)
Autodata = Auto
Autodata = na.omit(Auto) ####  we want to omit instead of predicitng the NAS

Autodata$id <- as.factor(Autodata$id)
Autodata$description_credit <- as.factor(ifelse(Autodata$description_credit == 1, "yes", "no"))
Autodata$description_owner <- as.factor(ifelse(Autodata$description_owner == 1, "yes", "no"))
Autodata$description_badcredit <- as.factor(ifelse(Autodata$description_badcredit == 1, "yes", "no"))
Autodata$description_length <- as.factor(Autodata$description_length)
Autodata$cylinders<- as.factor(Autodata$cylinders)
Autodata$transmission<-as.factor(Autodata$transmission)
Autodata$drive<- as.factor(Autodata$drive)
Autodata$type<-as.factor(Autodata$type)
Autodata$paint_color <-as.factor(Autodata$paint_color)
Autodata$size<- as.factor(Autodata$size)
Autodata$title_status<- as.factor(Autodata$title_status)
Autodata$condition<- as.factor(Autodata$condition)
Autodata$fuel<- as.factor(Autodata$fuel)
Autodata$state<- as.factor(Autodata$state)
Autodata$manufacturer<- as.factor(Autodata$manufacturer)
cars_data<- Autodata[, c("year", "odometer", "cylinders", "transmission","drive", "type", "paint_color", "size",


Autodata$year <- as.numeric(Autodata$year)
Autodata$odometer <- as.numeric(Autodata$odometer)

```

Due to the huge number of categories and levels in ID and description_length, I made a random forest of all the data except the two.

```
library(rando sample(x, size, replace = FALSE, prob = NULL)

train <- sample(1:nrow(Autodata),500)
train2<- sample(1:nrow(Autodata),250)
Autodata$price <- log(Autodata$price) ### taking log of price
## top 10 feautres in our random forest
## removing ID and description_length due to the number of levels
rf.autodata <-randomForest(price~ .-id-description_length, data = Autodata, subset =train)
importance(rf.autodata)

## Reducing the number of levels in a variable by taking the most important ones:
Autodata$type<- factor(Autodata$type,levels = c("SUV","sedan","pickup"))
Autodata$state<- factor(Autodata$state,levels = c("tx","cl","fl"))
Autodata$manufacturer<- factor(Autodata$manufacturer,levels = c("toyota","chevrolet","ford"))
##
library(tree)
tree.auto<- tree(price~ .-id-description_length, data = Autodata, subset =train)

```

After creating a random forest of the data, we can see the importance of the various variables:

```
> importance(rf.autodata)
                     IncNodePurity
year                    76.8285693
odometer                56.1153577
cylinders               12.1842713
transmission             3.0365754
drive                   21.5163262
type                    20.9172622
paint_color              8.6341027
size                     3.4116880
condition                4.8260829
fuel                     2.6847047
state                   38.5110115
manufacturer            29.4649344
title_status             0.3327056
description_credit       5.8786700
description_owner        0.7530969
description_badcredit    0.4802507
```

Reducing the number of levels in these following 3 variables in order to make a tree as we cannot use more than 32 levels. For this I used the summary method in order to see the levels which were the most popular.
And also creating a tree using all the variables as in the random forest(except ID and description_length):

```
library(tree)
tree.auto<- tree(price~ .-id-description_length, data = Autodata, subset =train)


yhat<- predict(tree.auto, newdata = Autodata[-train, ])
auto.test<- Autodata[-train,"price"]
plot(yhat, auto.test)
abline(0,1)
mean((yhat-auto.test)^2)
library(gbm)
boost.auto = gbm(price~.-id-description_length, data= Autodata[train2,], distribution = "gaussian",n.trees = 100
summary(boost.auto)
yhat.boost <- predict(boost.auto, newdata = Autodata[train,], n.trees = 5000)
mean((yhat.boost- auto.test)^2)
```

The results I got :
The MSE I got from the data :

```
> yhat<- predict(tree.auto, newdata = Autodata[-train, ])
> auto.test<- Autodata[-train,"price"]
> plot(yhat, auto.test)
> abline(0,1)
> mean((yhat-auto.test)^2)
[1] 0.4407304
```

Using boosting, I got the following data:

```
> summary(boost.auto)
                                          var    rel.inf
year                                     year 37.0886088
odometer                             odometer 17.1400004
paint_color                       paint_color 13.9447090
cylinders                           cylinders 12.4029563
drive                                   drive  6.5093695
condition                           condition  3.0399995
size                                     size  3.0348900
type                                     type  2.0534339
manufacturer                     manufacturer  1.6819550
fuel                                     fuel  1.3918098
description_credit         description_credit  0.9913428
description_owner           description_owner  0.3662220
transmission                     transmission  0.2244456
description_badcredit description_badcredit  0.1302575
state                                   state  0.0000000
title_status                     title_status  0.0000000
> yhat.boost <- predict(boost.auto, newdata = Autodata[train,], n.trees = 5000)
> mean((yhat.boost- auto.test)^2)
[1] 1.075836
```

Q3 (bonus) :
I think there is a huge difference between the craigslist data and the data in the paper given as the data goes into deep into the biases that may even exist in a given variable. For example, in the paper, there is a difference between a fleet/lease vs a dealer car and also conducted an analysis separately for the two sellers.
In the robustness checks and alternative explanations, the author even goes into the published price information which might even affect the result such as the kelly blue book value and edmunds.com.
The biggest difference which leads to the difference of results in the paper and the result I got is the amount of biases considered, for example, odometer tampering.
This pattern will be very different from the craigslist data as the amount of variables aren't just different, but also certified from leasing or car-renting companies. The variables mentioned by the author were : “make, model, model year, body style of car, and auction year” and the analysis was done separately for different types of buyers which is different from the craigslist data as the sellers aren't certified and the sellers happen to be thousands of individuals. The number of biases in the craigslist data will be way larger when taken into account that the data from the paper which happens to be certified.