ECO 395 Homework 3:Chengkan Tao

chengkan_tao

```
devtools::install_github(c('rstudio/rmarkdown', 'yihui/tinytex'))
tinytex::install_tinytex()
```

URL:https://github.com/chengkan-Tao/eco395-hw3 ## Q1 We can easily find correlations between police and crime through data, but not causality. There may exist some unknown causalities. We need to find what cause what, not just get a messy result when running the regression of "Crime" on "Police".

They find an example where they have a lot of police unrelated to crime. Researchers use terrorism alert system, and they want to find out what happens to crime if unrelated police increases. They also use ridership to measure victims. In the figure, the first column shows that high alert(police) has a negative effect on crime, and R^2 =0.14shows that only 12% of variance of crime is explained by this model. In this model, The second column shows that high alert(police) has a negative effect on crime and ridership has a huge positive effect on crime. And 17% of variance of crime is explained by these two variable.

In the conversation, researchers assume ridership is highly correlated to victims and want to find the relationship between crime and alert through the change of ridership after there is a alert. And, they find ridership levels on the Metro system were not diminished on high-terror days, which means the number of victims remained same.

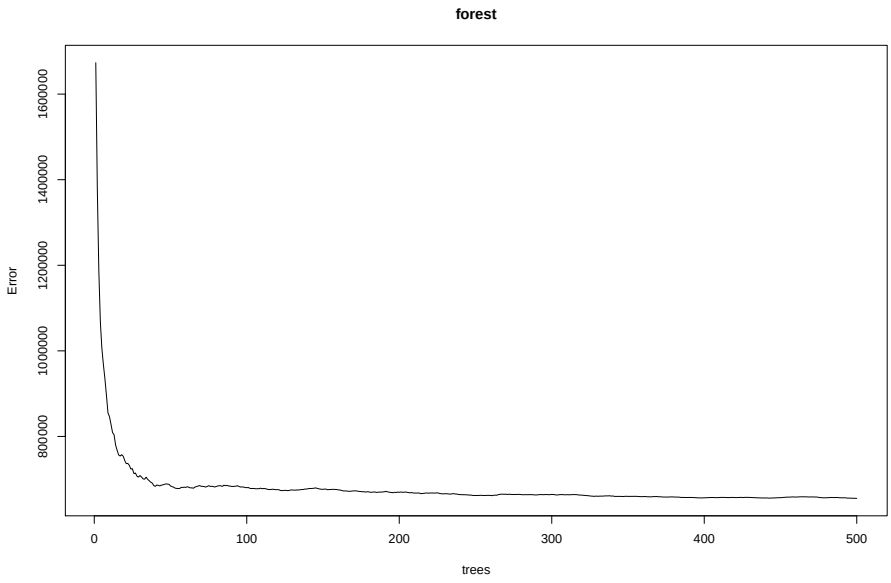
In the first column, if there's a high alert in District 1, there are about 2.621 units decreasing in daily total number of crime in D.C. holding all else fixed,

and the coefficient is statistically significant at the 1% level. And there's about 0.571 units decreasing in daily total number of crime in D.C. if a high alert happens in other districts, hold all else fixed. But the coefficient is not statistically significant. 1 percent increase in midday ridership will lead to 2.477units increase in daily total number of crime in D.C. so, ridership or tourists are positively correlated to crime. Hiring more cops in District 1 can let local crime go down.

#Q2

1) I try to use a predict model for revenue per square foot per calendar year and find DP of each features.

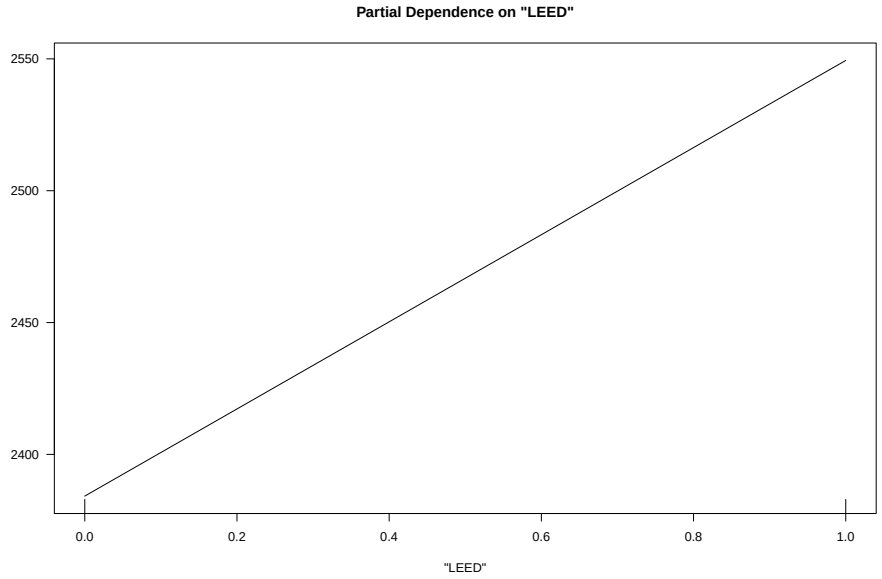
2) I use age, size, stories, renovated, amenities, cluster, Precipitation, LEED, Energystar, cd_total_07, hd_total07, Gas_Costs, Electricity_Costs, City_Market_Rent to build a random forest model for revenue.

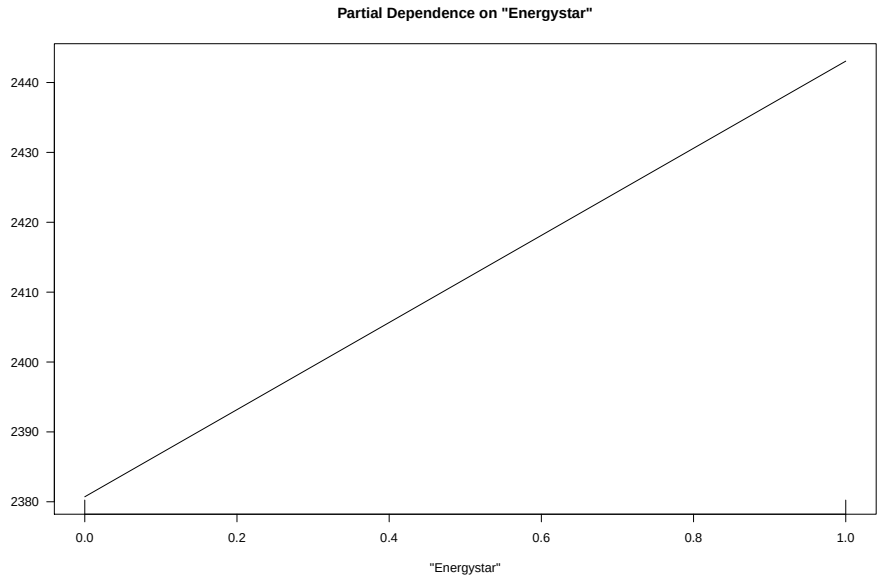


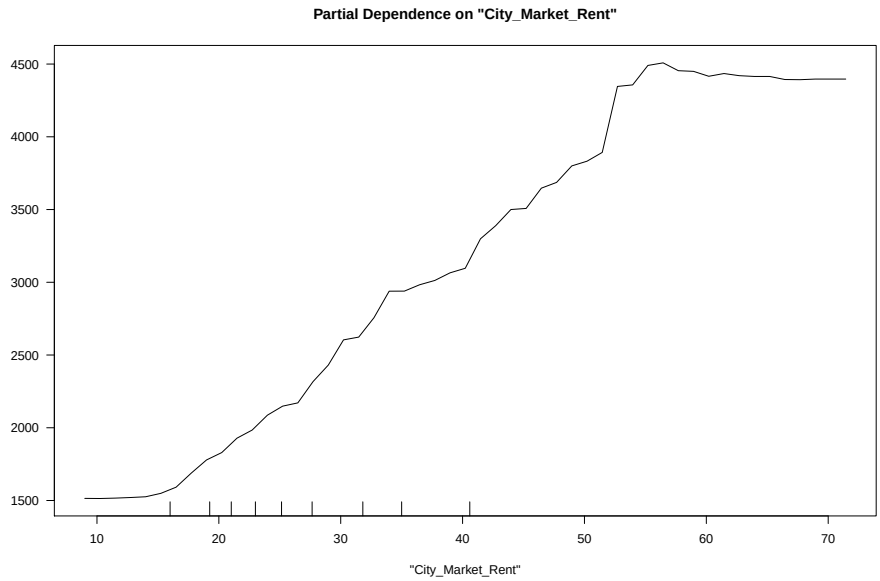
The plot is estimated using out-of-bag.We can find when there are about 300 trees, the mean squared error doesn't go down largely.

```
## [1] 767.3773
```

The root mean squared error is about 800, which uses the test subset. And I think it is comparably small.







we can find partial dependence of two dummy variables increase when variables goes from 0 to 1, and partial dependence of city_market_rent also goes up when the feature increases.

#Q3

In this question, I try to build a model to predict medianhousevalue. And I will use forward selection to build a model including three order polynomial expansion and interaction. The model contains all the variables, and there a log transformation of totalrooms.

```
##                                (Intercept)
##                              -2.782541e+06
##                               medianIncome
##                               4.424876e+05
##                          housingMedianAge
##                              -1.646158e+05
##                             totalBedrooms
##                               3.940800e+02
##                                 population
##                               3.392241e+01
```

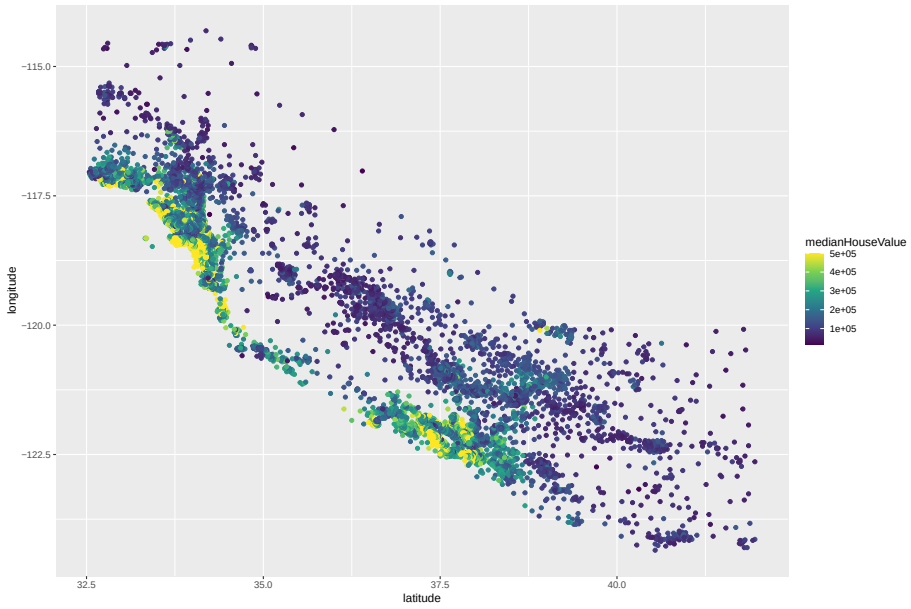
```
##                                latitude
##                            1.222383e+05
##                               longitude
##                           -2.356741e+04
##                          log_totalrooms
##                           -6.219989e+04
##          housingMedianAge:totalBedrooms
##                            2.725511e+01
##             housingMedianAge:population
##                           -1.910838e+01
##              housingMedianAge:latitude
##                            1.558136e+03
##             housingMedianAge:longitude
##                           -1.597477e+03
##           medianIncome:totalBedrooms
##                           -6.162785e+00
##              medianIncome:population
##                            2.488739e+01
##            totalBedrooms:log_totalrooms
##                           -5.654792e+01
##               totalBedrooms:population
##                           -1.098880e-02
##            medianIncome:log_totalrooms
##                            1.021727e+04
##          housingMedianAge:log_totalrooms
##                            8.098778e+02
##                population:log_totalrooms
##                            3.149031e+01
##           medianIncome:housingMedianAge
##                            2.711066e+03
##                       latitude:longitude
##                            9.226535e+02
##                  medianIncome:latitude
```

```
##                                  -4.075445e+04
##                         medianIncome:longitude
##                                   1.567118e+03
##                         totalBedrooms:latitude
##                                   4.295885e+00
##                           population:longitude
##                                   2.350074e+00
##    housingMedianAge:population:log_totalrooms
##                                  -7.512244e-01
##   medianIncome:housingMedianAge:totalBedrooms
##                                   1.658538e+00
##  medianIncome:housingMedianAge:log_totalrooms
##                                  -1.541576e+02
##         medianIncome:totalBedrooms:population
##                                   2.252617e-03
##     housingMedianAge:totalBedrooms:population
##                                   4.018803e-04
##        medianIncome:population:log_totalrooms
##                                  -3.383022e+00
##      medianIncome:housingMedianAge:population
##                                  -3.642597e-01
## housingMedianAge:totalBedrooms:log_totalrooms
##                                  -1.194296e+00
##         housingMedianAge:population:longitude
##                                  -2.041691e-01
##       housingMedianAge:totalBedrooms:latitude
##                                  -4.684540e-01
##               medianIncome:latitude:longitude
##                                  -2.722400e+02
##       medianIncome:housingMedianAge:longitude
##                                   1.410242e+01
##           housingMedianAge:latitude:longitude
##                                   2.060182e+01
```

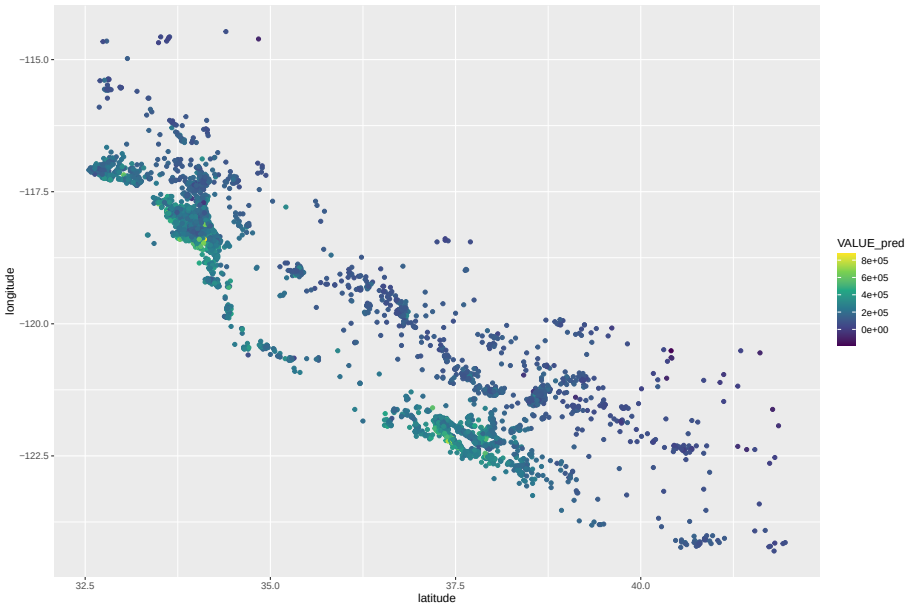
this is the model by forward selection

and we use bootstrape to get a root mean square error showing out of sample accuracy

```
##   result
## 64747.79
```



we can find some interactions. When latitude is equal to 40, the colour doesn't change a lot and when latitude is 33 or 37.5, the colour changes a lot.



The predict value is much lower because it is obvious there are more blue points. The interaction still exists and the relationship between value and latitude and relationship between longitude and value remain same