BOSTON HOUSING PRICE

RANDOM FOREST MODEL FOR EXAMPLE

109258026 經研碩二 盧禹叡
108753122 資科碩二 曾偉綱
110753165 資科碩一 張修誠
110753110 資科碩一 邱顯安

OUT LINE

- Goal
- Data
- Preprocess
- Model Construct
- Result & Comparison
- Reference

MAIN GOAL

- To predict the housing value in Boston.
- Improve the model predicting power by data preprocessing.
- Select the most appropriate model for predict the housing value.
- Test our prediction on Kaggle competition.
- Evaluation by RMSE

QUICK LOOK FOR DATA

- Variables
  - Target variable : MDEV (median value of owner-occupied homes in \$1000s)
  - Features :

crim
per capita crime rate by town.

zn
proportion of residential land zoned for lots over 25,000 sq.ft.

indus
proportion of non-retail business acres per town.

chas
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox
nitrogen oxides concentration (parts per 10 million).

rm
average number of rooms per dwelling.

age
proportion of owner-occupied units built prior to 1940.

dis
weighted mean of distances to five Boston employment centres.

rad
index of accessibility to radial highways.

tax
full-value property-tax rate per \$10,000.

ptratio
pupil-teacher ratio by town.

black
1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat
lower status of the population (percent).

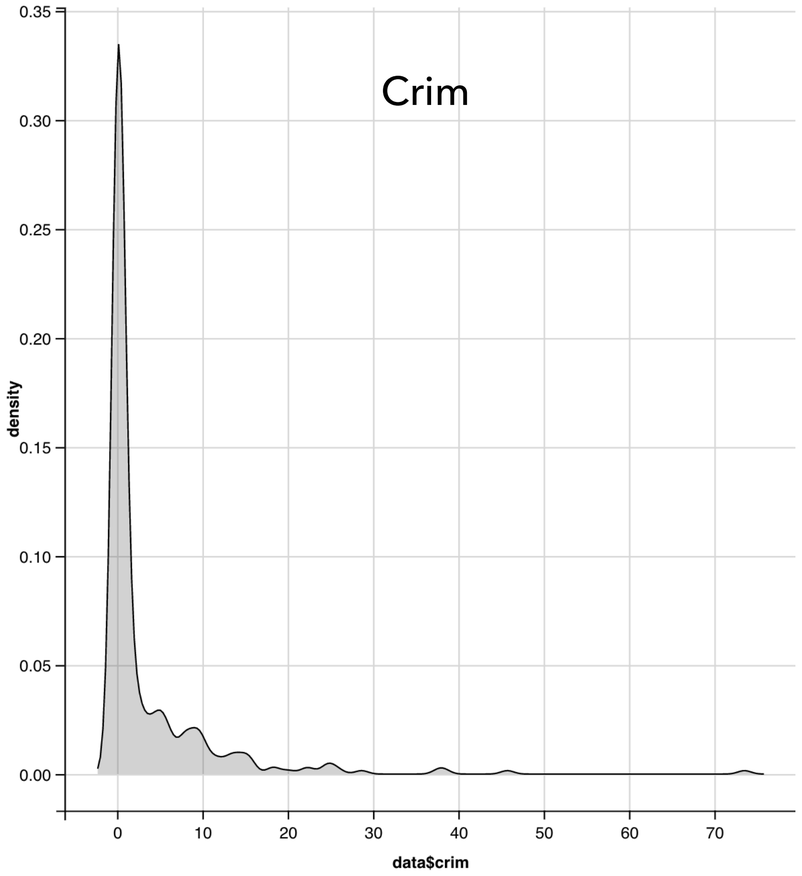
QUICK LOOK FOR DATA

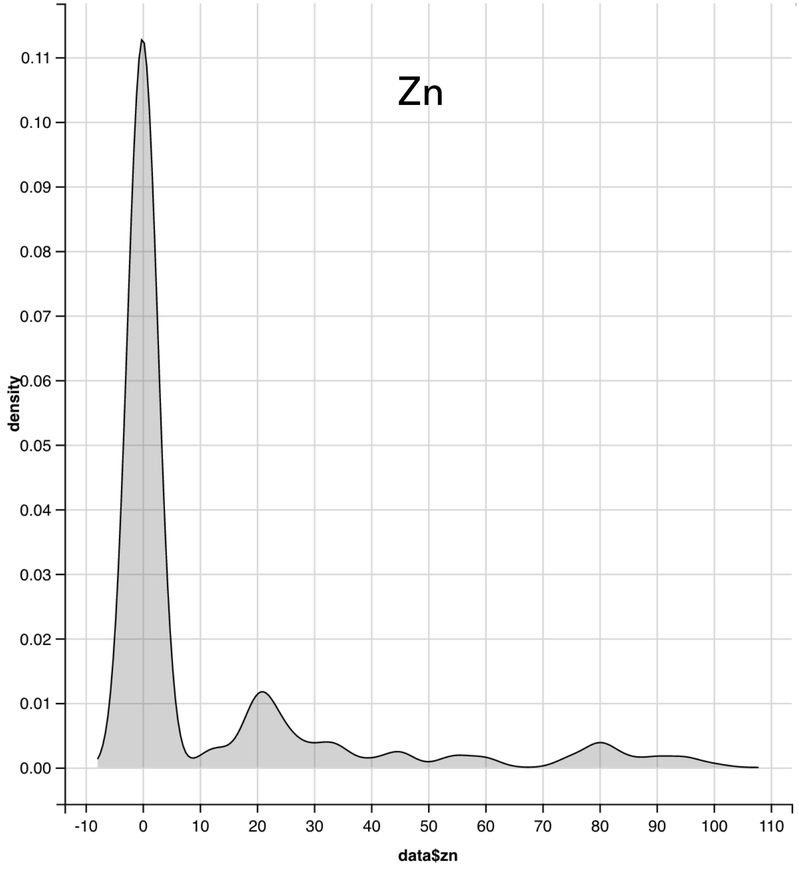
- Missing value
  - ```
    colSums(is.na(data))
      ID    crim      zn   indus    chas     nox      rm     age     dis     rad     tax ptratio   black   lstat    medv
       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
    ```

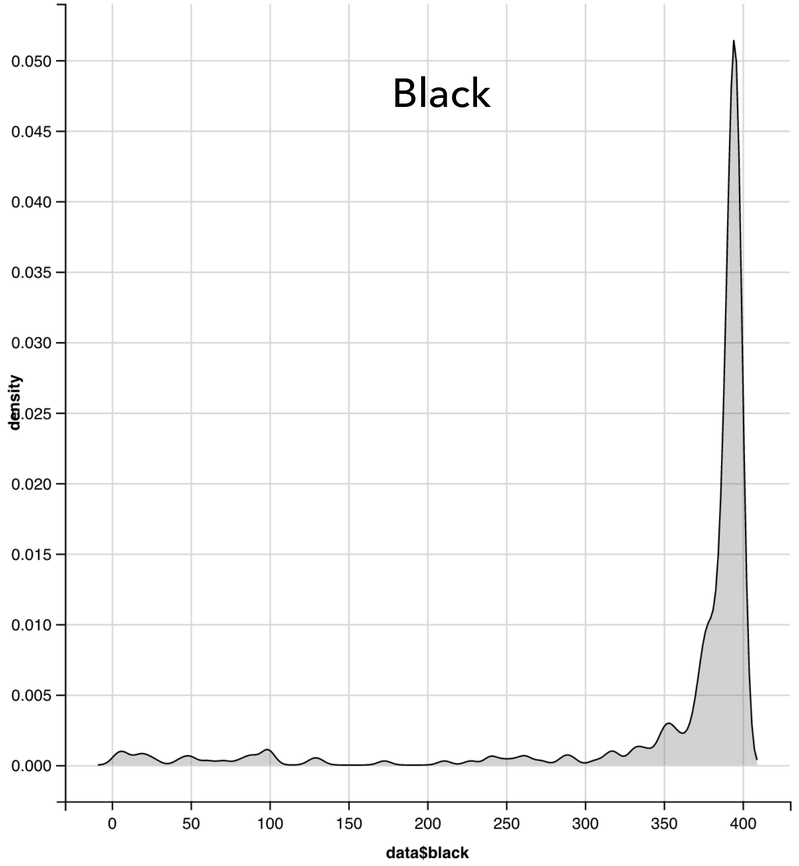
- Summary
  - ```
    > summary(data[,2:15])
          crim                zn              indus            chas              nox               rm             age       
     Min.   : 0.00632   Min.   :  0.00   Min.   : 0.74   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  6.00  
     1st Qu.: 0.07896   1st Qu.:  0.00   1st Qu.: 5.13   1st Qu.:0.00000   1st Qu.:0.4530   1st Qu.:5.884   1st Qu.: 45.40  
     Median : 0.26169   Median :  0.00   Median : 9.90   Median :0.00000   Median :0.5380   Median :6.202   Median : 76.70  
     Mean   : 3.36034   Mean   : 10.69   Mean   :11.29   Mean   :0.06006   Mean   :0.5571   Mean   :6.266   Mean   : 68.23  
     3rd Qu.: 3.67822   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6310   3rd Qu.:6.595   3rd Qu.: 93.80  
     Max.   :73.53410   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.725   Max.   :100.00  
          dis              rad              tax            ptratio          black            lstat            medv      
     Min.   : 1.130   Min.   : 1.000   Min.   :188.0   Min.   :12.60   Min.   :  3.5   Min.   : 1.73   Min.   : 5.00  
     1st Qu.: 2.122   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:376.7   1st Qu.: 7.18   1st Qu.:17.40  
     Median : 3.092   Median : 5.000   Median :330.0   Median :19.00   Median :392.1   Median :10.97   Median :21.60  
     Mean   : 3.710   Mean   : 9.634   Mean   :409.3   Mean   :18.45   Mean   :359.5   Mean   :12.52   Mean   :22.77  
     3rd Qu.: 5.117   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.2   3rd Qu.:16.42   3rd Qu.:25.00  
     Max.   :10.710   Max.   :24.000   Max.   :711.0   Max.   :21.20   Max.   :396.9   Max.   :37.97   Max.   :50.00  
    ```

SKWENESS

- Density plot



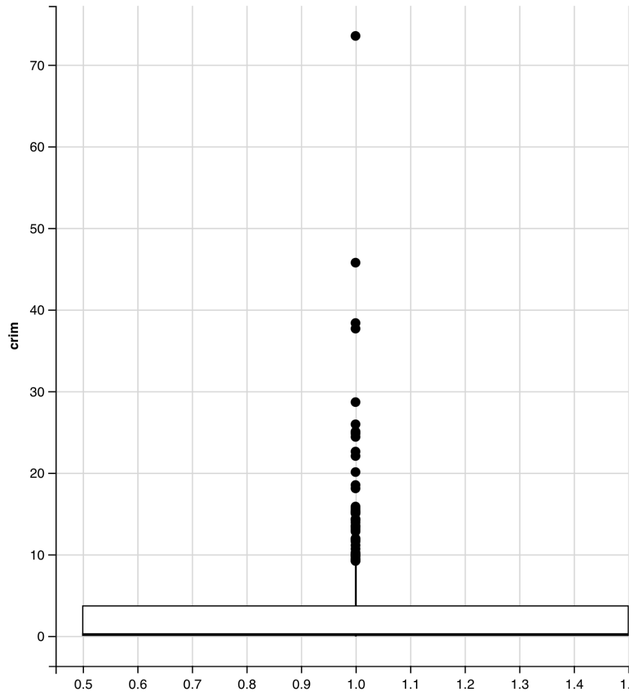




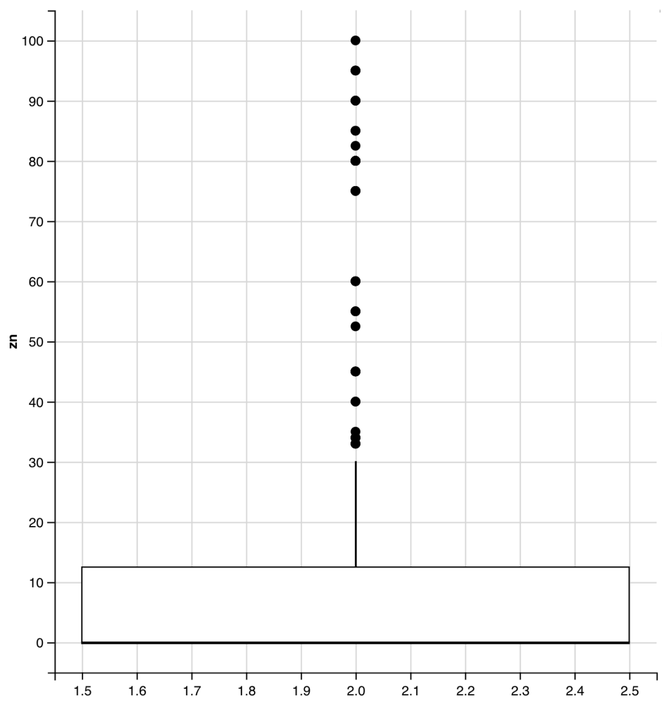
OUTLIER

- Box plot

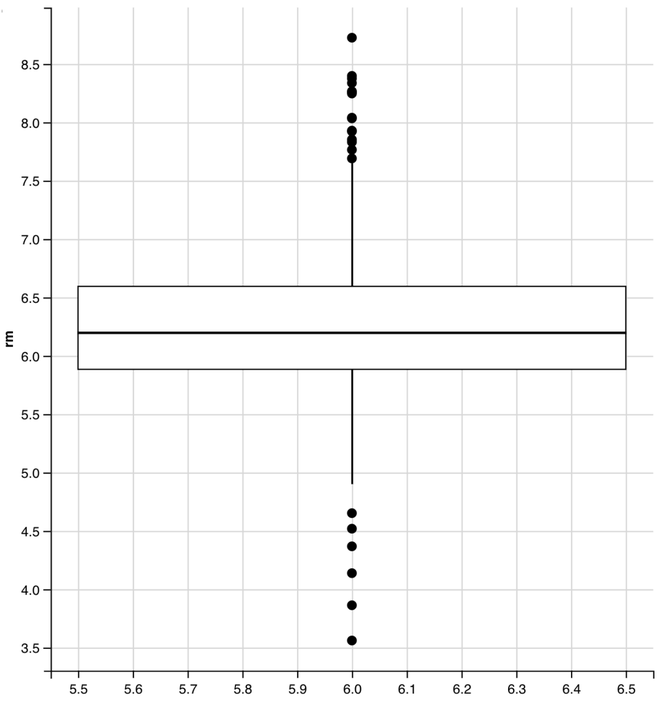
crim



zn

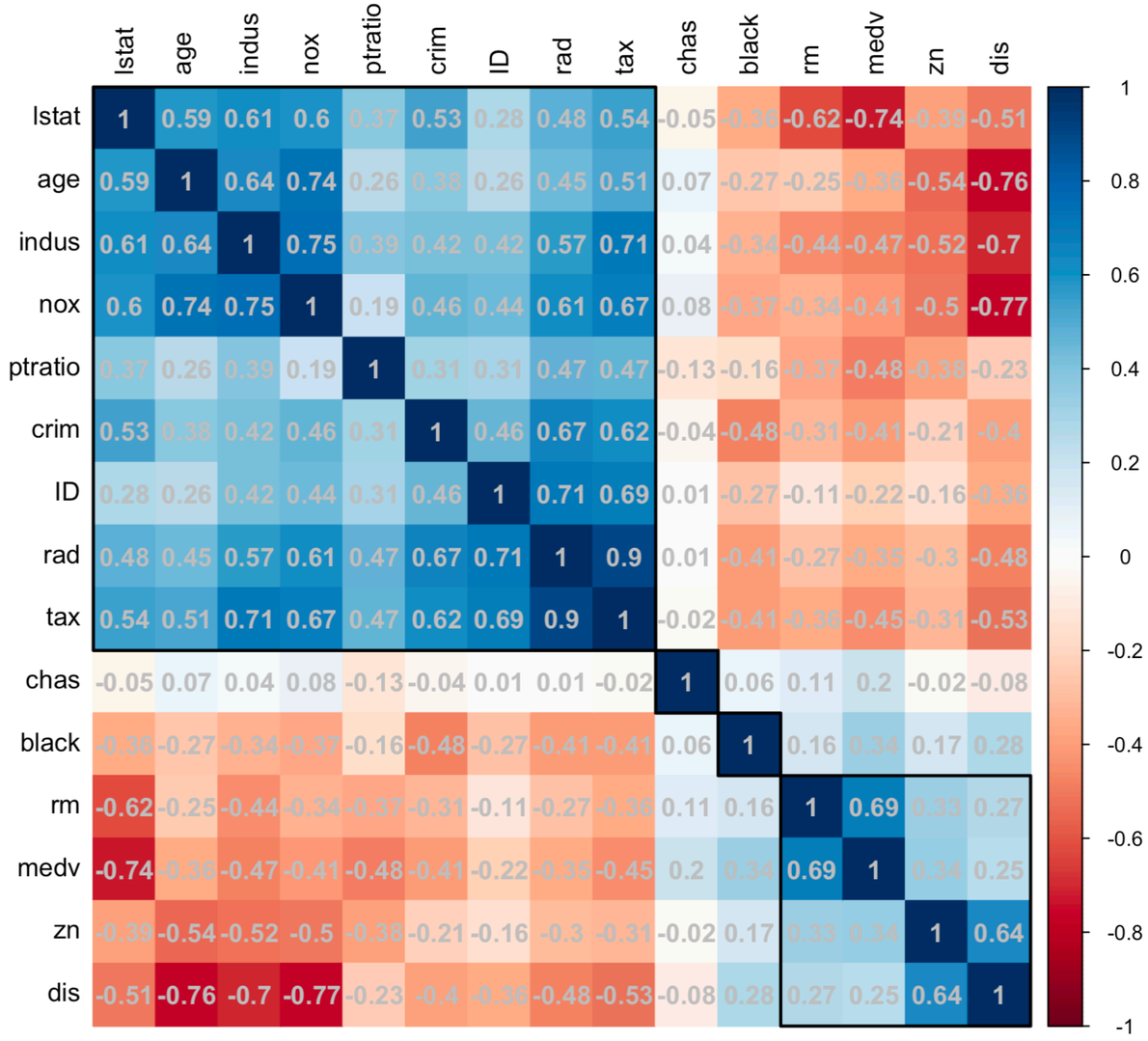


rm



CORRELATION

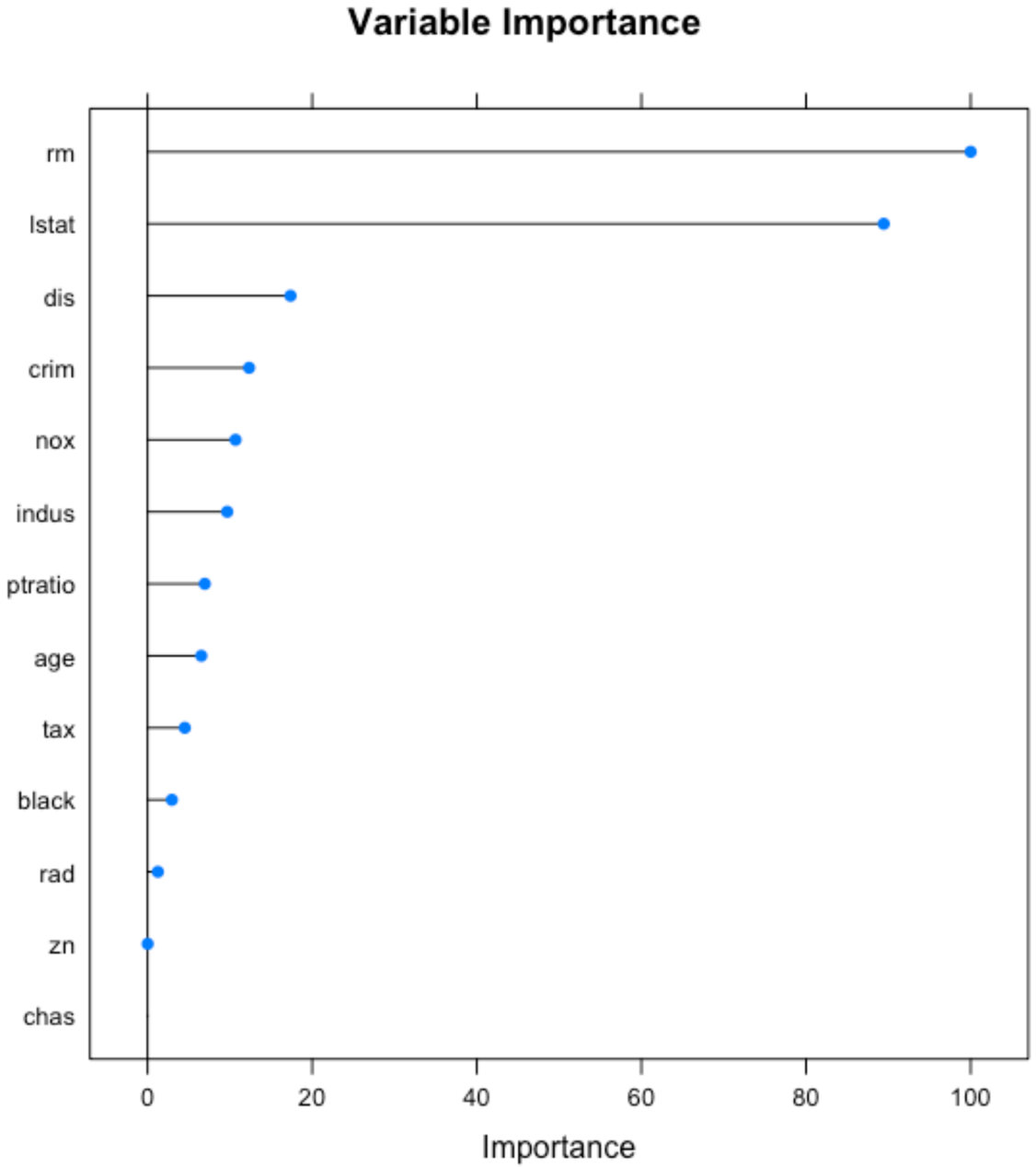
- More higher correlate to MDEV
  - ptratio
  - lstat
  - indus
  - tax
  - rm
  - nox
  - rad
  - crim



FEATURE IMPORTANCE

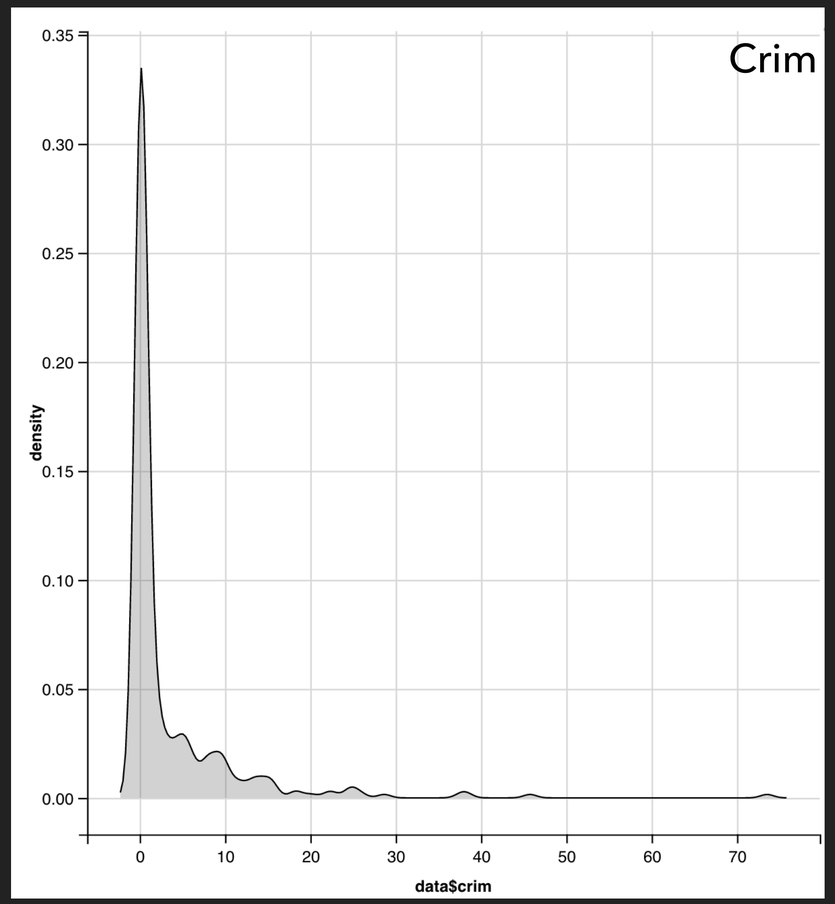
- Importance

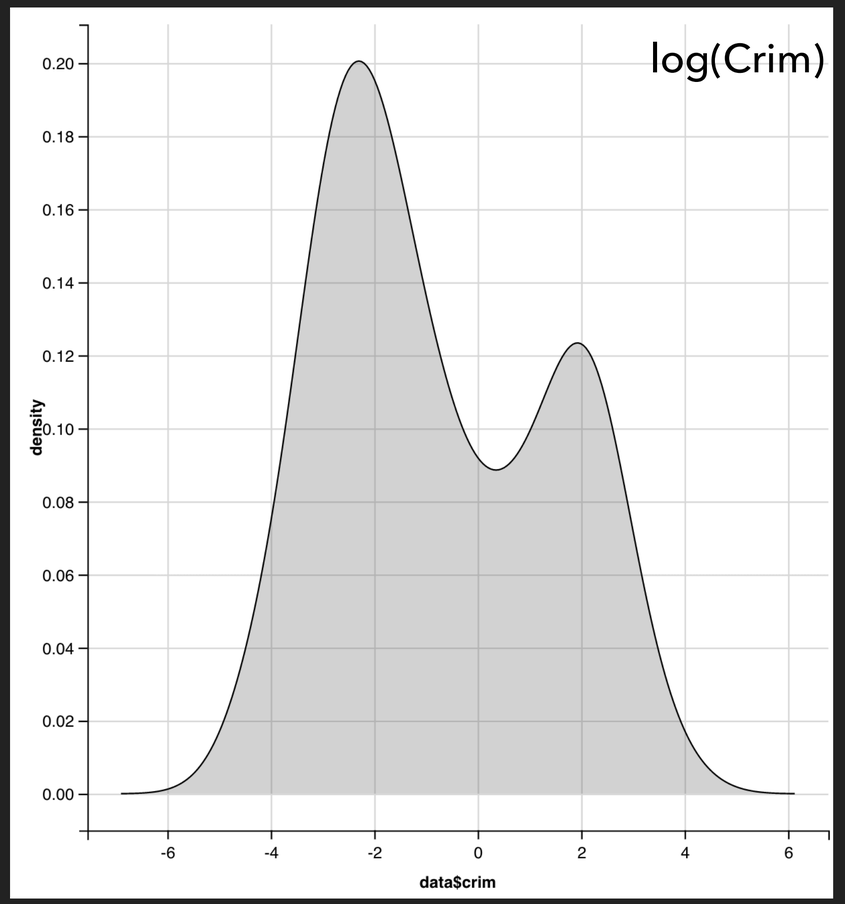
```
rf variable importance

          Overall
rm       100.0000
lstat     89.4408
dis       17.3649
crim      12.3327
nox       10.6838
indus      9.6736
ptratio    6.9504
age        6.5271
tax        4.5244
black      2.9530
rad        1.2500
zn         0.0127
chas       0.0000
```



SKWENESS

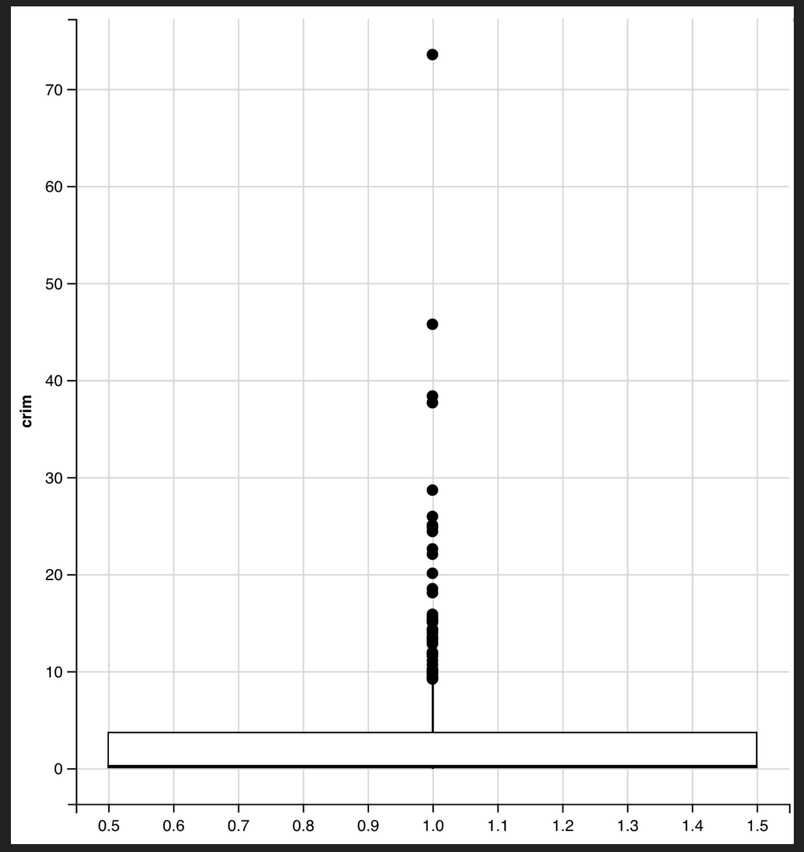
- Log transform ‘Crim’

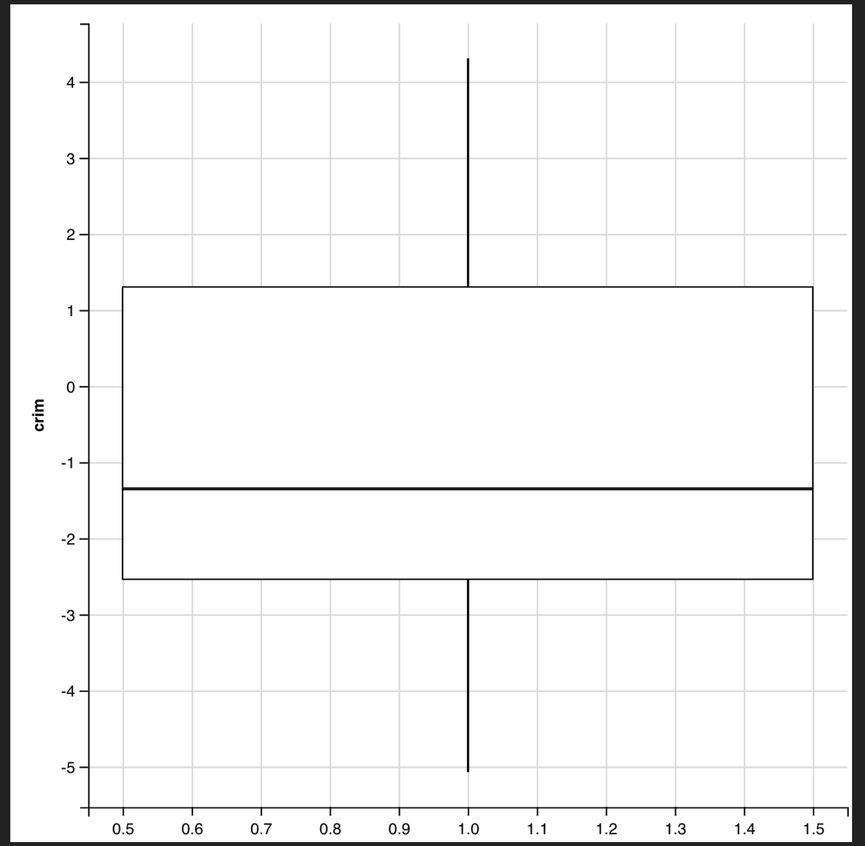




OUTLIER

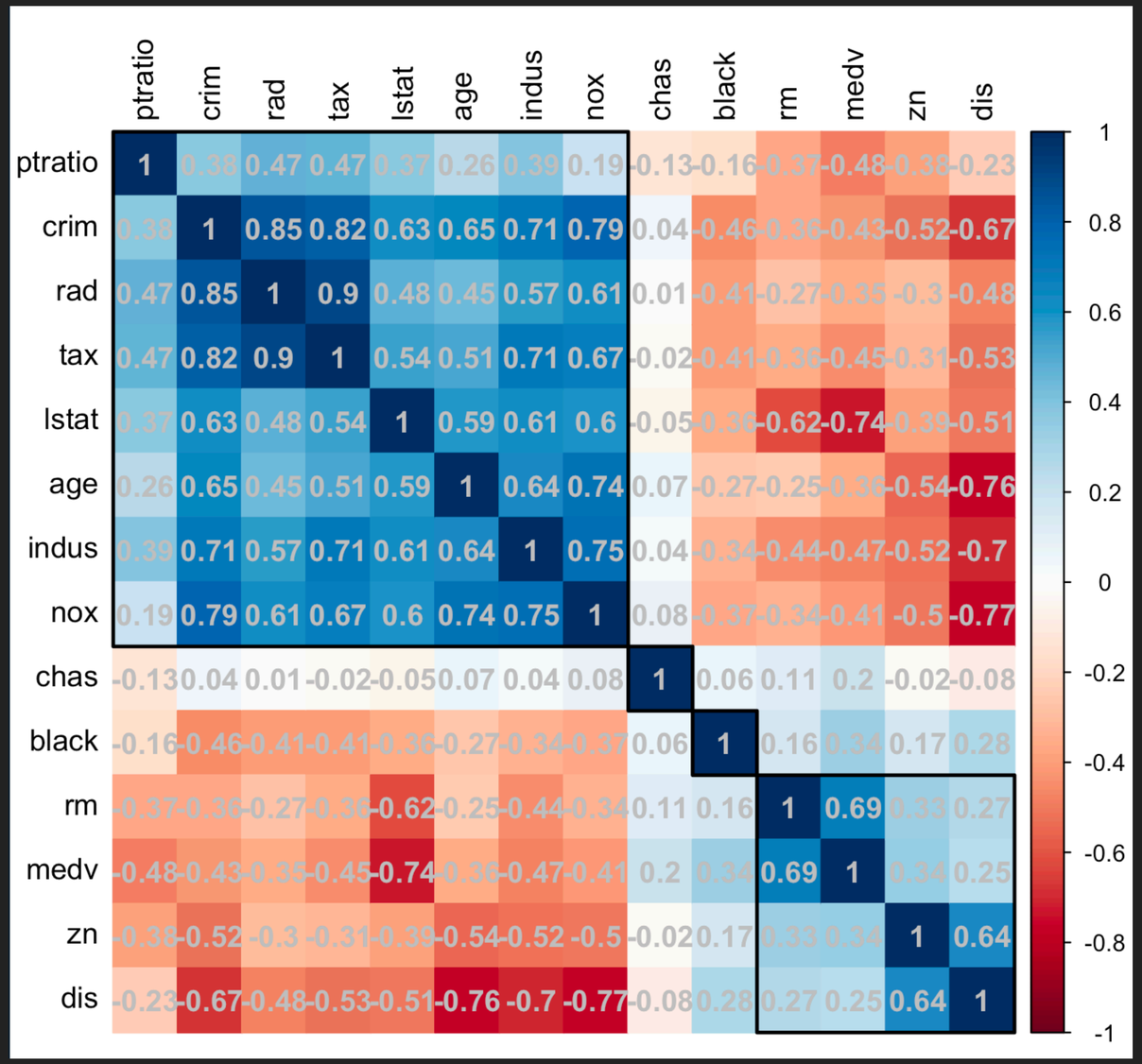
- Log transform ‘Crim’





CORRELATION

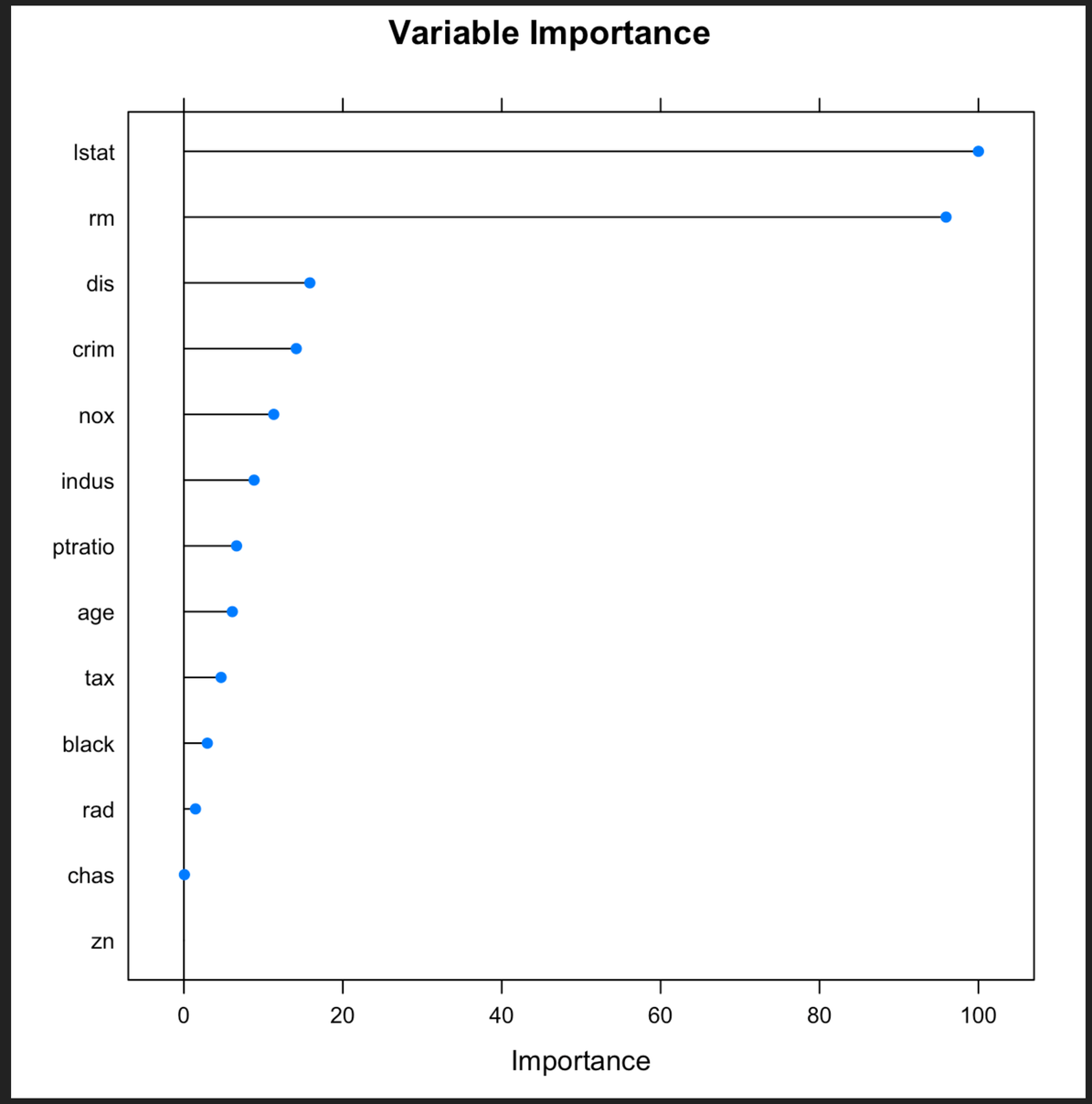
- More higher correlate to MDEV
  - lstat
  - rm
  - ptratio
  - indus
  - tax
  - crim
  - nox



FEATURE IMPORTANCE

- Importance

```
rf variable importance

          Overall
lstat   100.00000
rm       95.92526
dis      15.85576
crim     14.14329
nox      11.30949
indus     8.83528
ptratio   6.63035
age       6.09576
tax       4.68007
black     2.94926
rad       1.46006
chas      0.06912
zn        0.00000
```

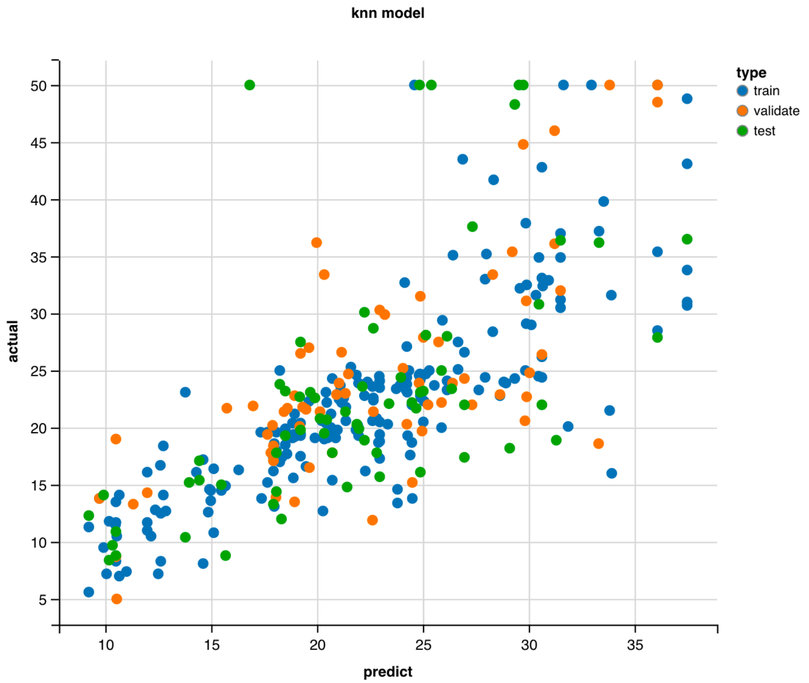


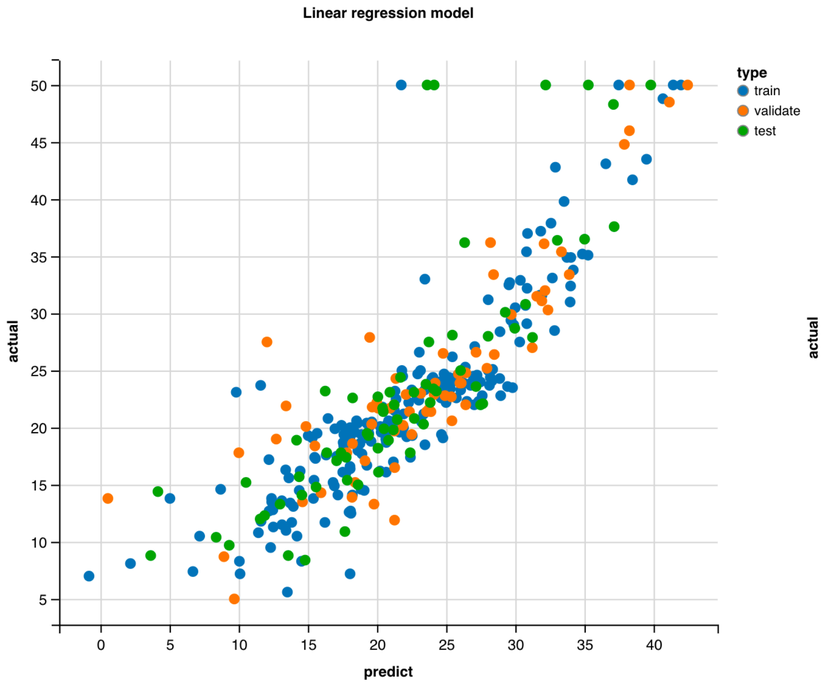
RECAP

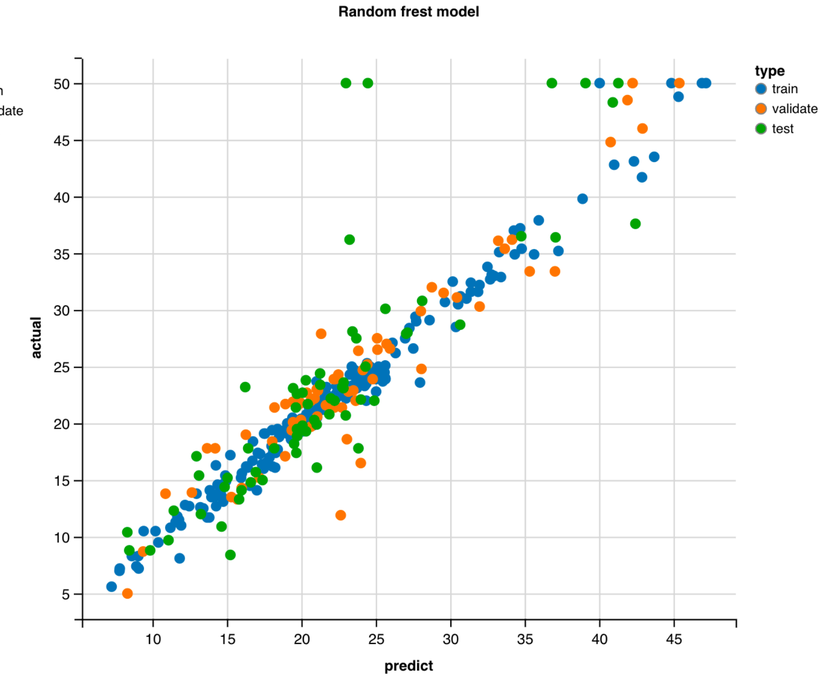
- Preprocess
  - Log transform the feature ‘Crim’
- Feature select
  - We select seven and eleven features according to the Heat map and Random forest Importance respectively
    - Correlation : lstat, rm , ptratio, indus, tax, crim, nox
    - Importance : lstat, rm, dis, crim, nox, indus, ptratio, age, tax, black, rad

MODEL CHOOSING

- We compare three model with raw data with 5-Fold
  - Apparently, Random forest model is more appropriate than the other two.

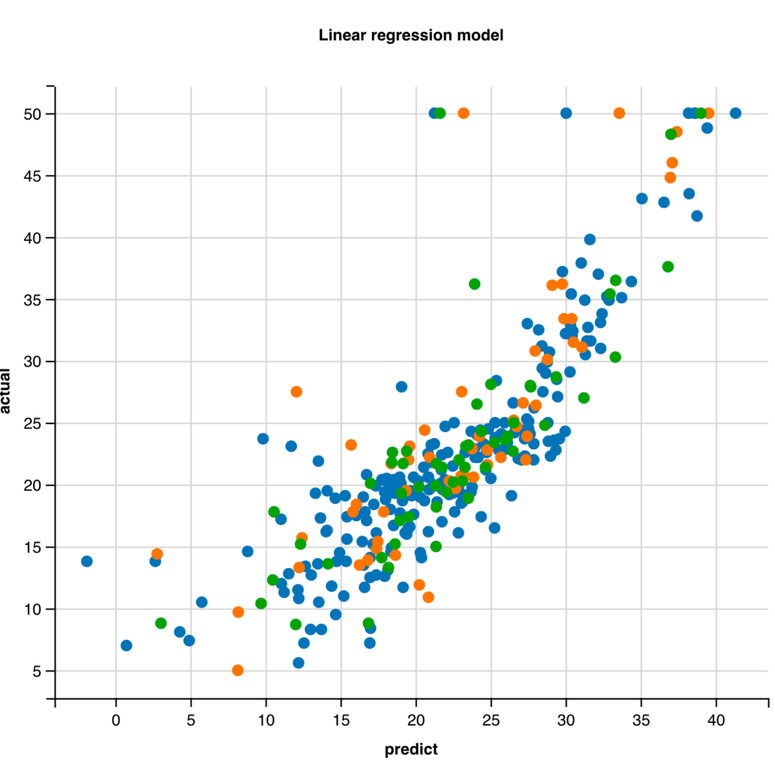


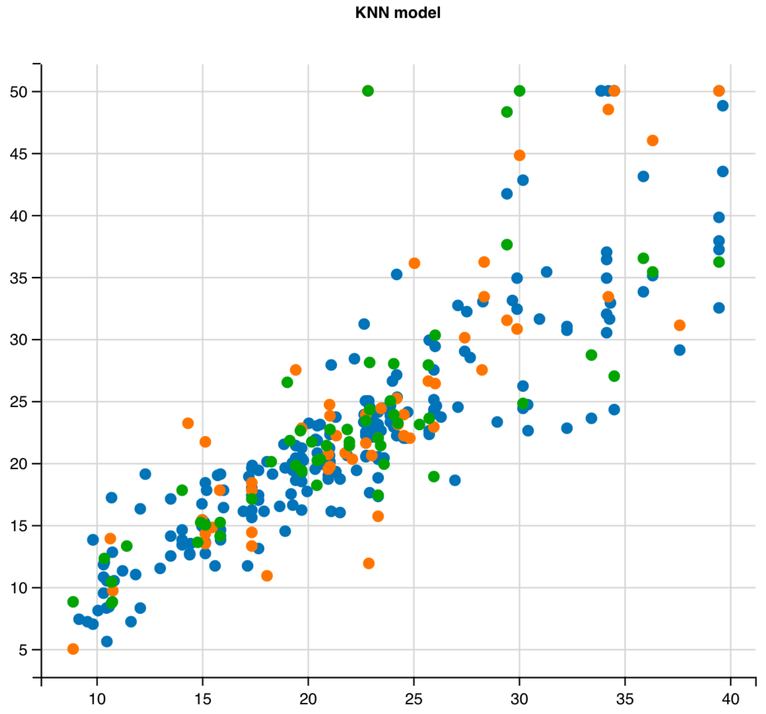


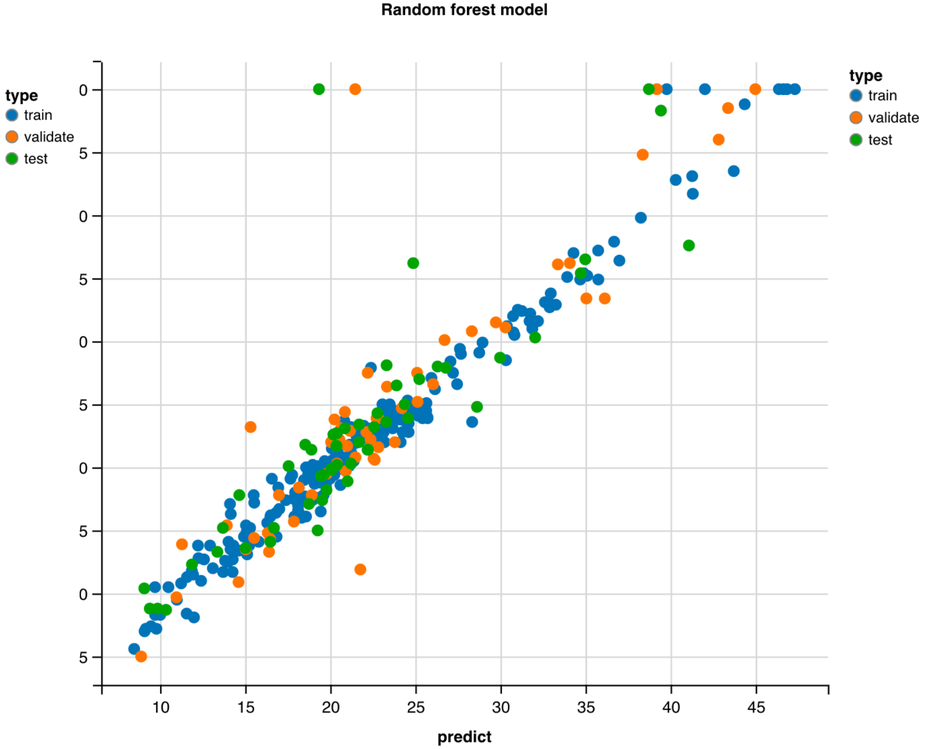


MODEL CHOOSING

- We compare three model with the feature chose from correlation with 5-Fold
  - Apparently, Random forest model is more appropriate than the other two.

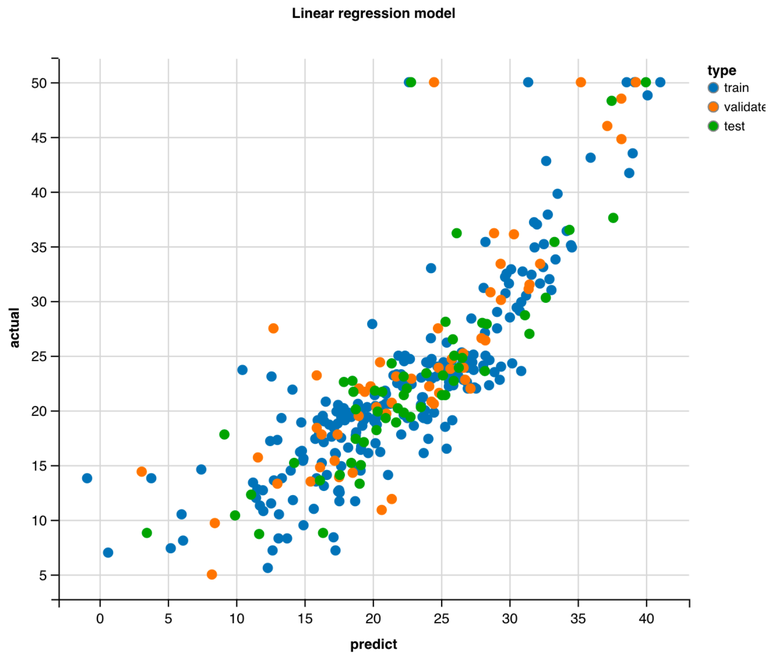


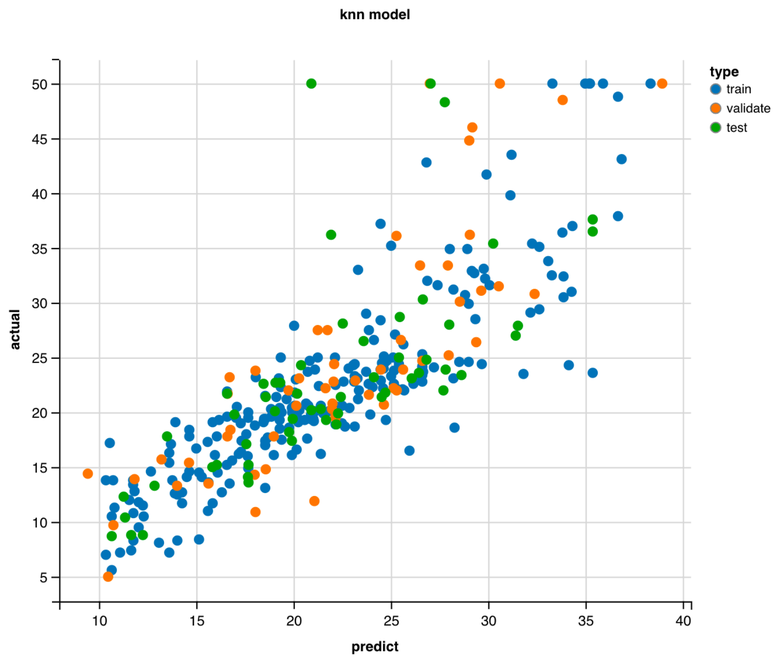


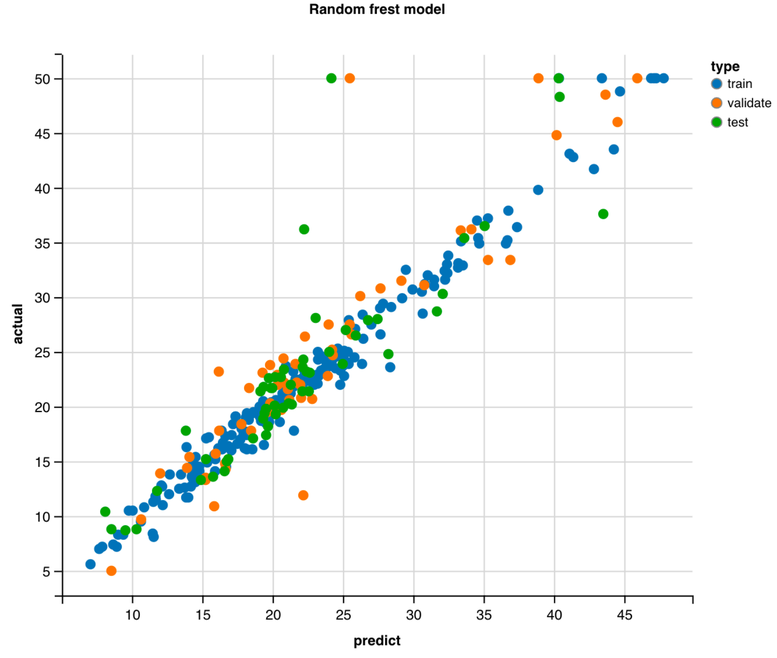


MODEL CHOOSING

- We compare three model with the feature chose from feature importance with 5-Fold
  - Apparently, Random forest model is more appropriate than the other two.







EVALUATION

- We compare three model with the feature chose from correlation, and run the k-Fold CV , k= 5~10.
  - Apparently, 10 - fold model is more appropriate than the other four.

set	training	validation	testing
fold 1	1.4	3.07	6.43
fold 2	1.77	3.07	2.75
fold 3	1.72	3.51	2.68
fold 4	1.68	4.62	2.68
fold 5	1.73	4.67	5.3
ave.	1.66	3.79	3.97

set	training	validation	testing
fold 1	1.64	5.06	5.16
fold 2	1.72	3.3	2.51
fold 3	1.86	2.49	3.78
fold 4	1.85	2.49	5.23
fold 5	1.54	3.63	6.18
fold 6	1.76	2.87	6.22
ave.	1.73	3.31	4.85

set	training	validation	testing
fold 1	1.6	5.55	5.35
fold 2	1.72	3.2	2.93
fold 3	1.91	4.09	2.43
fold 4	1.63	3.56	4.18
fold 5	1.89	3.34	4.46
fold 6	1.73	5.26	5.12
fold 7	1.56	3.57	3.23
ave.	1.72	4.08	3.96

set	training	validation	testing
fold 1	1.63	7.2	2.61
fold 2	1.65	2.5	2.84
fold 3	1.86	4.35	2.18
fold 4	1.74	3.08	5.56
fold 5	1.68	2.54	2.77
fold 6	1.7	5.24	5.61
fold 7	1.6	3.99	2.5
fold 8	1.94	2.8	4.45
ave.	1.73	3.96	3.56

set	training	validation	testing
fold 1	1.62	6.25	5.23
fold 2	1.63	2.7	2.99
fold 3	1.58	3.22	2.64
fold 4	1.72	5.92	2.9
fold 5	1.69	2.5	2.98
fold 6	1.7	5.18	5.66
fold 7	1.56	3.25	4.11
fold 8	1.7	2.63	2.49
fold 9	1.67	3.3	3.17
ave.	1.65	3.88	3.57

set	training	validation	testing
fold 1	1.64	6.22	5.5
fold 2	1.62	2.68	3.01
fold 3	1.61	3.26	2.59
fold 4	1.73	2.76	6.12
fold 5	1.66	2.47	2.96
fold 6	1.71	2.1	7.72
fold 7	1.55	4.35	2.58
fold 8	1.95	2.17	2.63
fold 9	1.87	3.14	3.19
fold 10	1.77	5.84	3.19
ave.	1.71	3.5	3.95

EVALUATION

- We compare three model with the feature chose from feature importance, and run the k-Fold CV , k = 5~10
  - Apparently, 5 - fold model is more appropriate than the other two.

set	training	validation	testing
fold 1	1.38	3.04	6.11
fold 2	1.55	2.97	2.69
fold 3	1.57	3.17	2.61
fold 4	1.52	4.22	2.6
fold 5	1.55	4.11	4.76
ave.	1.51	3.5	3.75

set	training	validation	testing
fold 1	1.65	5.03	5.17
fold 2	1.51	3.23	2.48
fold 3	1.57	2.5	3.24
fold 4	1.55	2.25	4.5
fold 5	1.48	3.55	5.48
fold 6	1.6	2.62	5.56
ave.	1.56	3.2	4.41

set	training	validation	testing
fold 1	1.57	5.48	4.97
fold 2	1.45	3.28	2.69
fold 3	1.63	3.21	2.56
fold 4	1.45	3.39	3.69
fold 5	1.63	2.81	3.8
fold 6	1.67	5.2	4.76
fold 7	1.4	3.45	3.37
ave.	1.54	3.83	3.69

set	training	validation	testing
fold 1	1.47	6.47	2.24
fold 2	1.45	2.49	2.55
fold 3	1.55	3.49	2.38
fold 4	1.48	3.14	4.56
fold 5	1.52	2.44	2.53
fold 6	1.63	4.94	5.45
fold 7	1.42	3.98	2.42
fold 8	1.57	2.79	3.19
ave.	1.51	3.72	3.17

set	training	validation	testing
fold 1	1.58	6.34	4.7
fold 2	1.48	2.68	2.98
fold 3	1.47	2.82	2.66
fold 4	1.48	4.88	2.84
fold 5	1.5	2.41	2.94
fold 6	1.66	4.68	5.51
fold 7	1.37	3.37	3.9
fold 8	1.52	2.63	2.43
fold 9	1.48	3.2	3.26
ave.	1.5	3.67	3.47

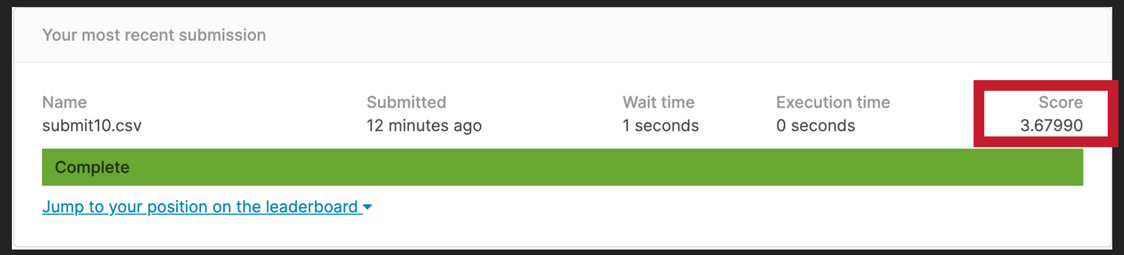
set	training	validation	testing
fold 1	1.49	5.56	4.73
fold 2	1.48	2.77	2.62
fold 3	1.44	2.97	2.41
fold 4	1.39	2.56	5.06
fold 5	1.54	2.28	2.93
fold 6	1.47	2.23	6.68
fold 7	1.3	4.28	2.61
fold 8	1.53	1.83	2.3
fold 9	1.44	3.1	3.18
fold 10	1.47	4.88	2.68
ave.	1.46	3.25	3.52

RECAP

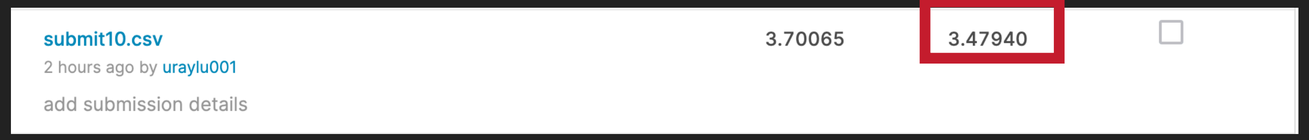
- Model
  - Random Forest
- Feature select
  - By random forest importance
- K - fold CV
  - Chose 5 fold model

KAGGLE SUBMISSION

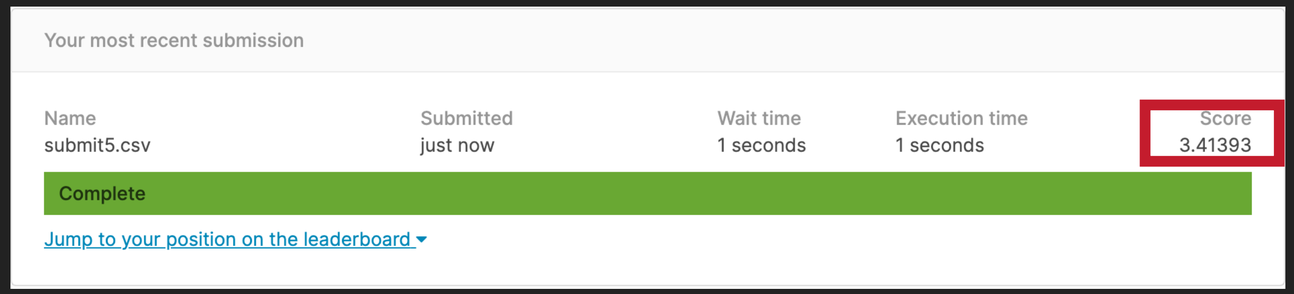
- Raw data model



- Correlation



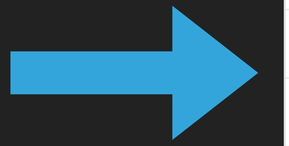
- Random forest Importance



KAGGLE SUBMISSION

- Ranking

We are here !



#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	amyaramine			0.00000	3	6Y
2	—	VjKR			0.00000	10	5Y
3	—	MayankSatnalika			1.33055	1	5Y
4	▲ 8	AbhijeetUpadhyay			2.90742	9	5Y
5	▲ 2	Valentino1992			3.20117	12	5Y
6	▲ 3	SivaTeja			3.35823	7	5Y
7	▲ 28	Thomas Tong			3.48979	3	5Y
8	▲ 7	PankajMishra			3.49795	6	5Y
9	▲ 4	Jaya Sureya			3.49923	7	5Y

RESULT

- Use Random Forest model with 5 - fold CV to train the model will have the best prediction.
- Check the visualization in our shiny app : https://brianchiu.shinyapps.io/finalproject/
- Git hub : https://github.com/1101-datascience/finalproject-finalproject_group11

THANKS YOU !

ANY QUESTION ?