Challenge B

Cosma, Bianca & García Bouyssou, Clara

December 2nd 2017

Task 1B - Predicting house prices in Ames, Iowa (continued)

Step 1

Random forest algorithms are used for regression and classification models. Compare to other models it relies on fewer assumptions. The prediction is made on 2/3 of the training sample. This subsample is created by random selection of cases(observations) with replacement and it is used to grown the trees. Then a subsample of inputs (independent variables) are selected at random. Afterwards the remaining part of the sample it is used to calculate the Out-Of-the-Bag (OOB) error rate. Each tree gives a classifation on the OOB, then the classification having the most votes is choosen. In the case is tackled here, a regression model, the classification is the average of the outcome(dependent variable).

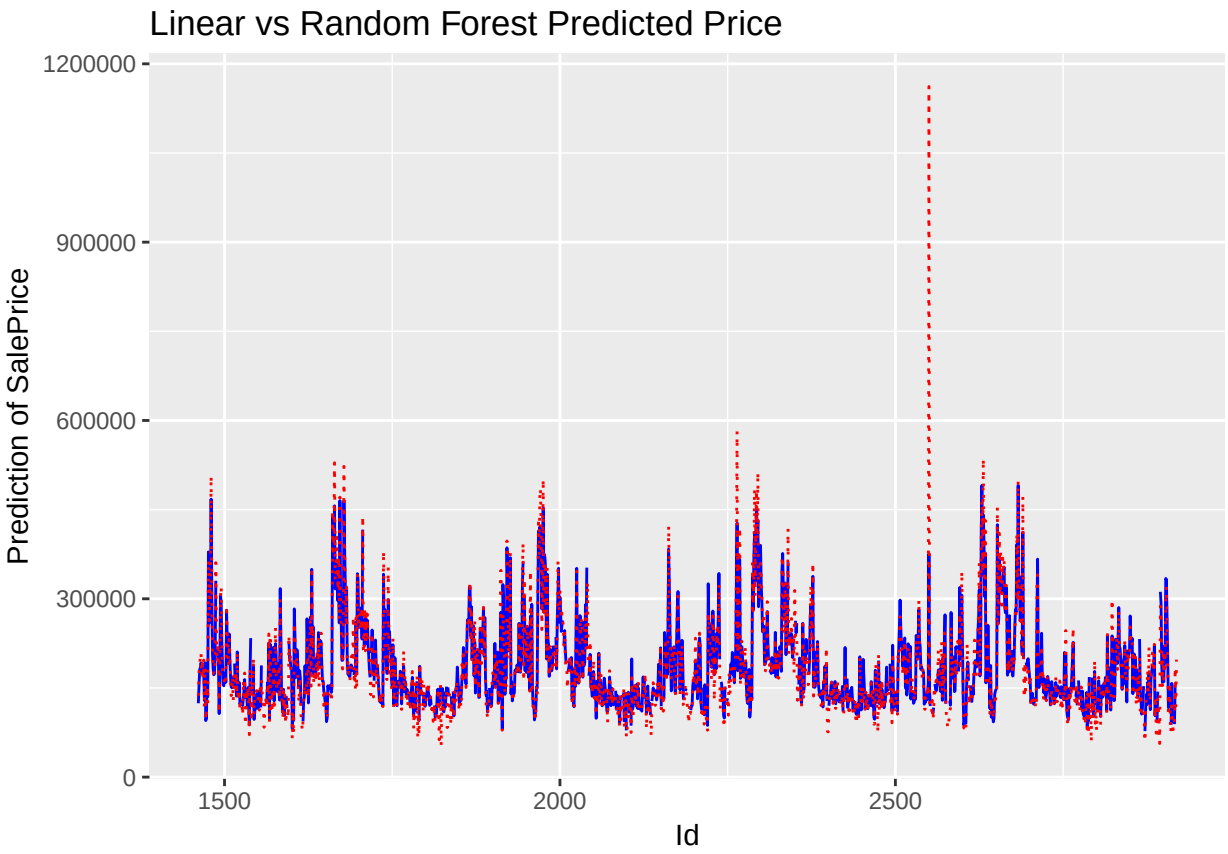
Step 2

A model is created with random forest and the results are presented below:

```
                Length Class   Mode
call               4   -none-  call
type               1   -none-  character
predicted       1338   -none-  numeric
mse              500   -none-  numeric
rsq              500   -none-  numeric
oob.times       1338   -none-  numeric
importance        73   -none-  numeric
importanceSD       0   -none-  NULL
localImportance    0   -none-  NULL
proximity          0   -none-  NULL
ntree              1   -none-  numeric
mtry               1   -none-  numeric
forest            11   -none-  list
coefs              0   -none-  NULL
y               1338   -none-  numeric
test               0   -none-  NULL
inbag              0   -none-  NULL
terms              3   terms   call
na.action        122   exclude numeric
```

Step 3

Once the predictions using random forest model are computed, they are plotted together with the predictions from the linear model obtained in Challenge A. It is possible to observe a diference between the two predictions. Moreover, in the linear model we obtained an important outlier that does not appear in this machine learning technique.



Task 2B - Overfitting in Machine Learning (continued)

Step 1

By employing the training dataset a low flexibility local linear model is constructed. Summary is displayed below:

```
Regression Data: 122 training points, in 1 variable(s)
                x
Bandwidth(s): 0.5

Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed
Residual standard error: 1.085438
R-squared: 0.8540956

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Step 2

With the same procedure a high flexibility local linear model is created using the training dataset:

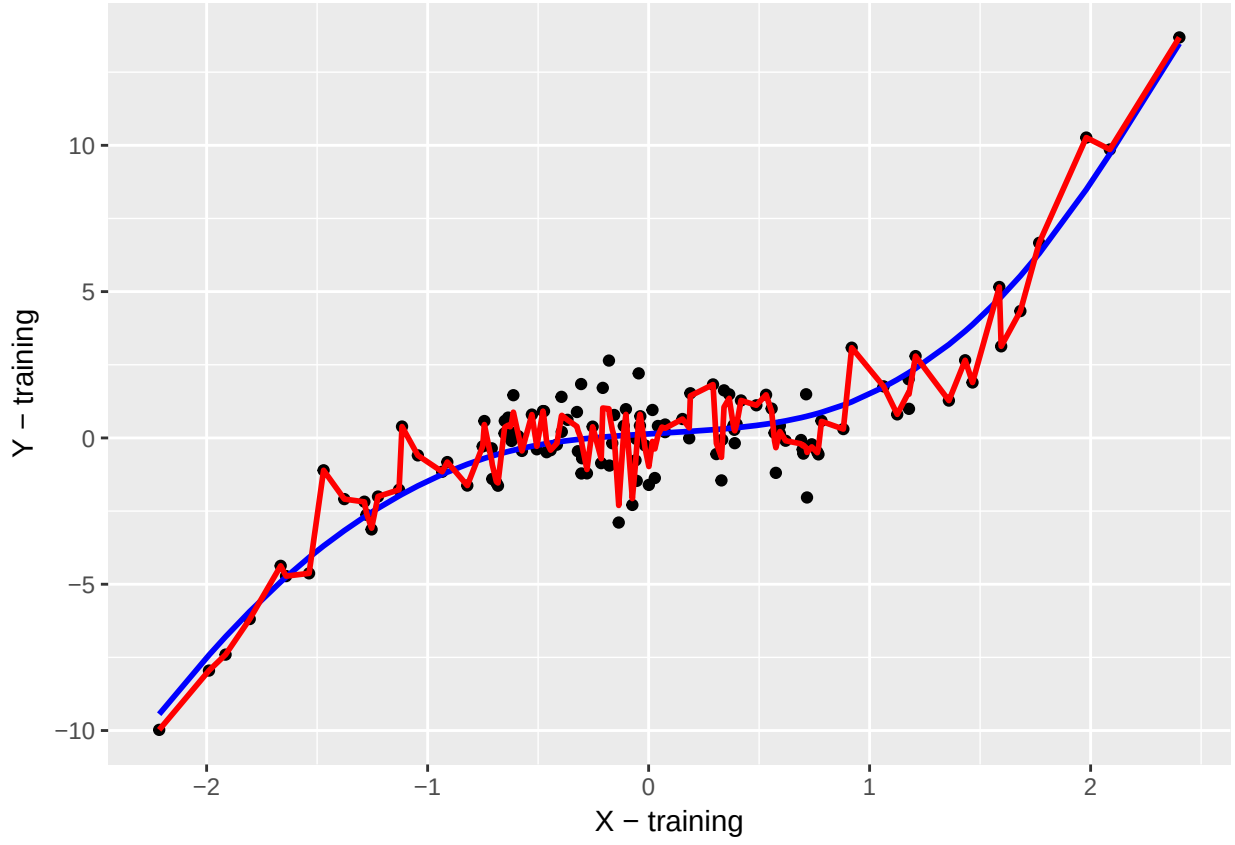
```
Regression Data: 122 training points, in 1 variable(s)
                 x
Bandwidth(s): 0.01

Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed
Residual standard error: 0.5070779
R-squared: 0.9680171

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Step 3

The following scatterplot reports the values of x and y and the prediction of the low (blue) and high (red) flexibility models in the training dataset.

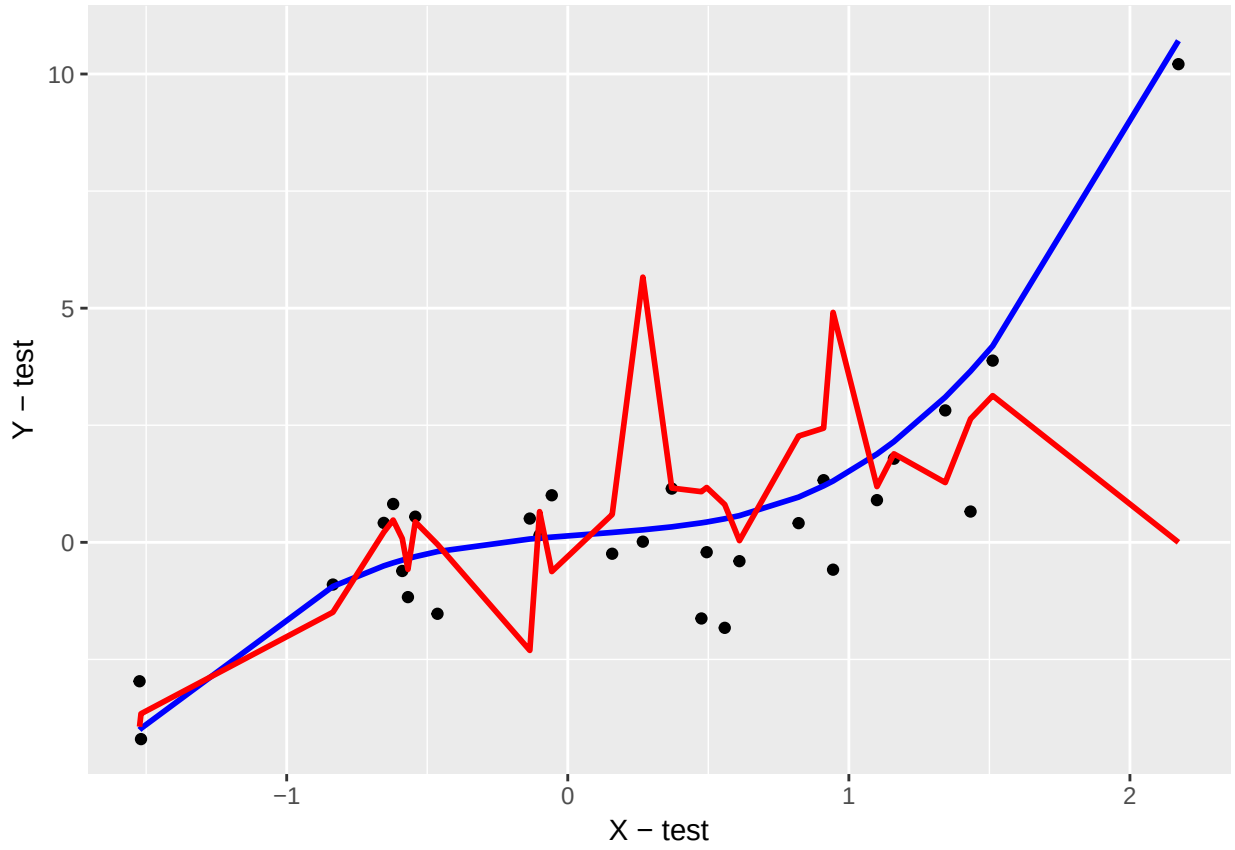


Step 4

It is possible to observe that the high flexibility model leads to more variance in the predictions. On the other hand, the low flexibility model increases the bias i.e. the distance between the prediction and the value. Therefore, is necesssary to find the right balance between overfitting and underfitting respectively.

Step 5

The following scatterplot presents the values of y and x along with the predictions of the high and low flexibility models in the test dataset.



The high flexibility model has still the highest variance among predictions. Besides, being this the least biased model it can be seen that the fitted model does not addapt properly to the new sample. In fact, the original variance of the model does not reflect the variance of the new sample.

Step 6

First of all, a vector with different bandwidths is created in order to construct models with different degrees of flexibility.

Step 7

A function was employed to creat a model for each value in the bandwidth interval in the training dataset.

Step 8

After computing the predictions in the training dataset for each model the mean squared error is computed.

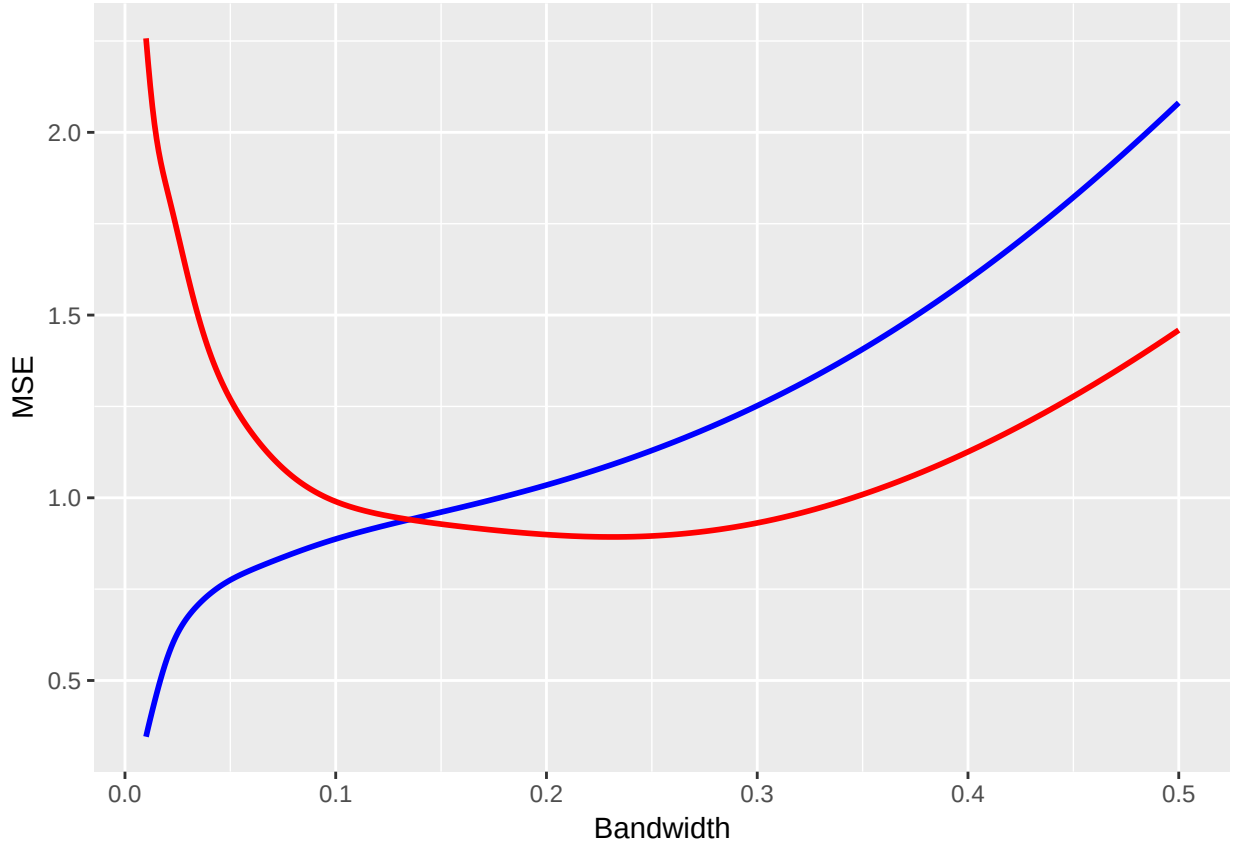
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3461  0.9369  1.1404  1.2262  1.5067  2.0810
```

Step 9

The previous step is repeated on the test dataset.

Step 10

The following graph summarises how de MSE changes with respect to the bandwidth in the two samples i.e. blue for the training dataset and red for the test. It can be concluded that the MSE in the test dataset reaches a minimum for intermediate values of the bandwidht i.e. 0.2, instead in the training dataset the mean squared error is almost proportionally increasing with the bandwidth. Therefore it is preferable to use a value around the median of the bandwidth.



Task 3B - Privacy regulation compliance in France

Step 1

For the next steps the dataset CNIL will be employed.

Step 2

The following table sumarizes how many organizations in France have adhered to the CNIL per department i.e. they have a CIL.

	Department	Number of organizations with CIL
1:	01	1247156
2:	02	1042223
3:	03	696910
4:	04	436588
5:	05	406750
6:	06	2758356
7:	07	490274
8:	08	501518
9:	09	198296
10:	10	861278
11:	11	932791
12:	12	472694
13:	13	4976823
14:	14	2342945
15:	15	307214
16:	16	924127
17:	17	1482943
18:	18	760128
19:	19	592323
20:	20	821015
21:	21	1330379
22:	22	1228357
23:	23	303156
24:	24	820373
25:	25	1113925
26:	26	1135166
27:	27	1104939
28:	28	867000
29:	29	1777066
30:	30	1282962
31:	31	3140816
32:	32	777053
33:	33	3692845
34:	34	2632928
35:	35	2427073
36:	36	523643
37:	37	1389506
38:	38	3420784
39:	39	479395
40:	40	1777811
41:	41	853743
42:	42	1841759
43:	43	541104
44:	44	3186444
45:	45	1786777
46:	46	517875
47:	47	841253
48:	48	120273
49:	49	1637051
50:	50	1392722
51:	51	1444169
52:	52	374540
53:	53	915805

```
54:         54                          1876004
55:         55                           492392
56:         56                          1421324
57:         57                          2440196
58:         58                           377635
59:         59                          5342534
60:         60                          2095044
61:         61                           536246
62:         62                          2247181
63:         63                          1167134
64:         64                          1438362
65:         65                           456154
66:         66                          1033592
67:         67                          2741503
68:         68                          1695674
69:         69                          5740825
70:         70                           574975
71:         71                          1115832
72:         72                          1159826
73:         73                          1159507
74:         74                          1808049
75:         75                         20680495
76:         76                          3038177
77:         77                          2427699
78:         78                          3208572
79:         79                          1115939
80:         80                          1565386
81:         81                           680175
82:         82                           597479
83:         83                          2071381
84:         84                          1325238
85:         85                          1384561
86:         86                          1203310
87:         87                          1150170
88:         88                           986807
89:         89                           781574
90:         90                           275901
91:         91                          2327495
92:         92                          9467573
93:         93                          3235117
94:         94                          3059537
95:         95                          1889832
96:         97                          2541888
97:         98                           354983
    Department Number of organizations with CIL
```

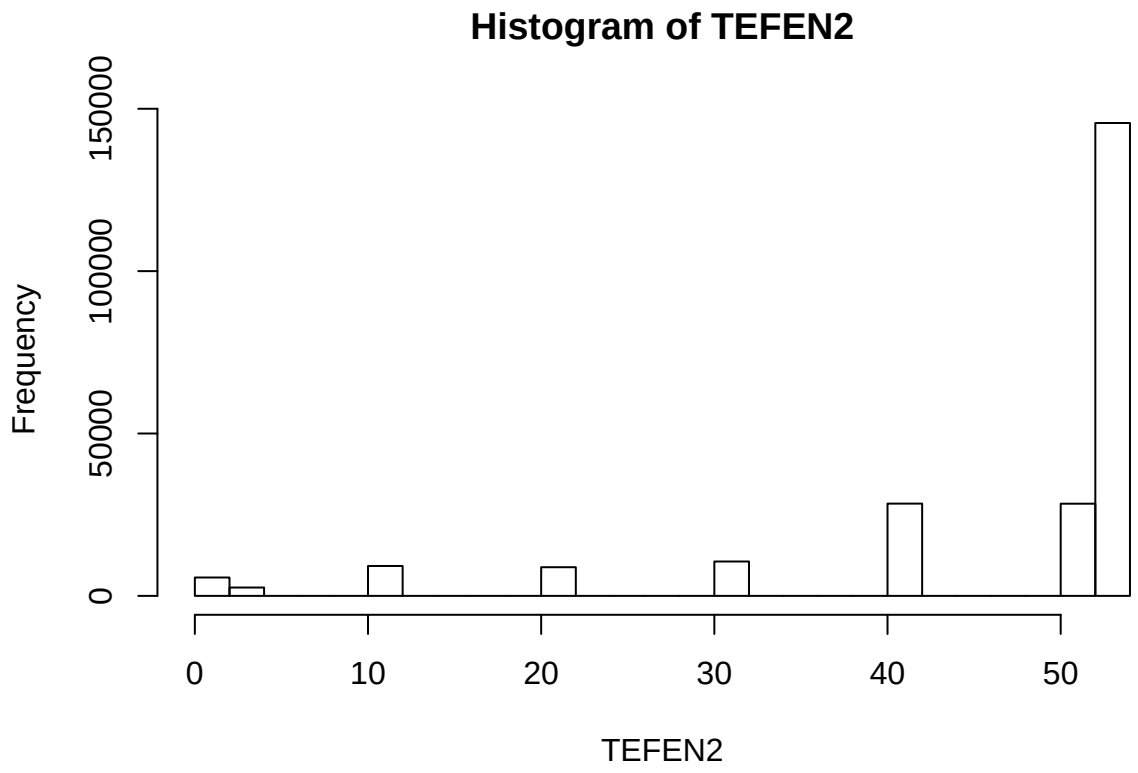
Step 3

First of all the dataset SIRC was imported. Then, it was merge whith the dataset cnil by the variable SIREN.

Step 4

The dataset presents information on the employment size of the various companies. This variable is categorical with the numeric values that represent the size. In the histogram it is seen that companies that have employment size greater than 2000 employees (categories greater than 51 [1]) have the highes frequency.

```
Warning: NAs introduced by coercion
```




```
$breaks
 [1]  0  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
[24] 46 48 50 52 54

$counts
 [1]   5654   2568      0      0      0   9208      0      0      0      0
[11]   8824      0      0      0      0  10587      0      0      0      0
[21]  28410      0      0      0      0  28390 145623

$density
 [1] 0.011815401 0.005366457 0.000000000 0.000000000 0.000000000
 [6] 0.019242343 0.000000000 0.000000000 0.000000000 0.000000000
[11] 0.018439882 0.000000000 0.000000000 0.000000000 0.000000000
[16] 0.022124097 0.000000000 0.000000000 0.000000000 0.000000000
[21] 0.059369567 0.000000000 0.000000000 0.000000000 0.000000000
[26] 0.059327772 0.304314481

$mids
```

[1] https://www.sirene.fr/sirene/public/variable/tefen

```
 [1]  1  3  5  7  9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
[24] 47 49 51 53

$xname
[1] "TEFEN2"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```