Challenge B

Cosma, Bianca & García Bouyssou, Clara

December 5nd 2017

The code is available online in our correspondent github profiles: https://github.com/ClaraGBouyssou https://github.com/bianca-16?tab=repositories

The SIRC dataset could not be uploaded, and we renamed the file SIRC.

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 in the Appendix.

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 in the Appendix.

Step 2

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

Step 3

The first scatterplot in the Appendix 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 second scatterplot in the Appendix 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. The summary is displayed in the Appendix.

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.

Step 3

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

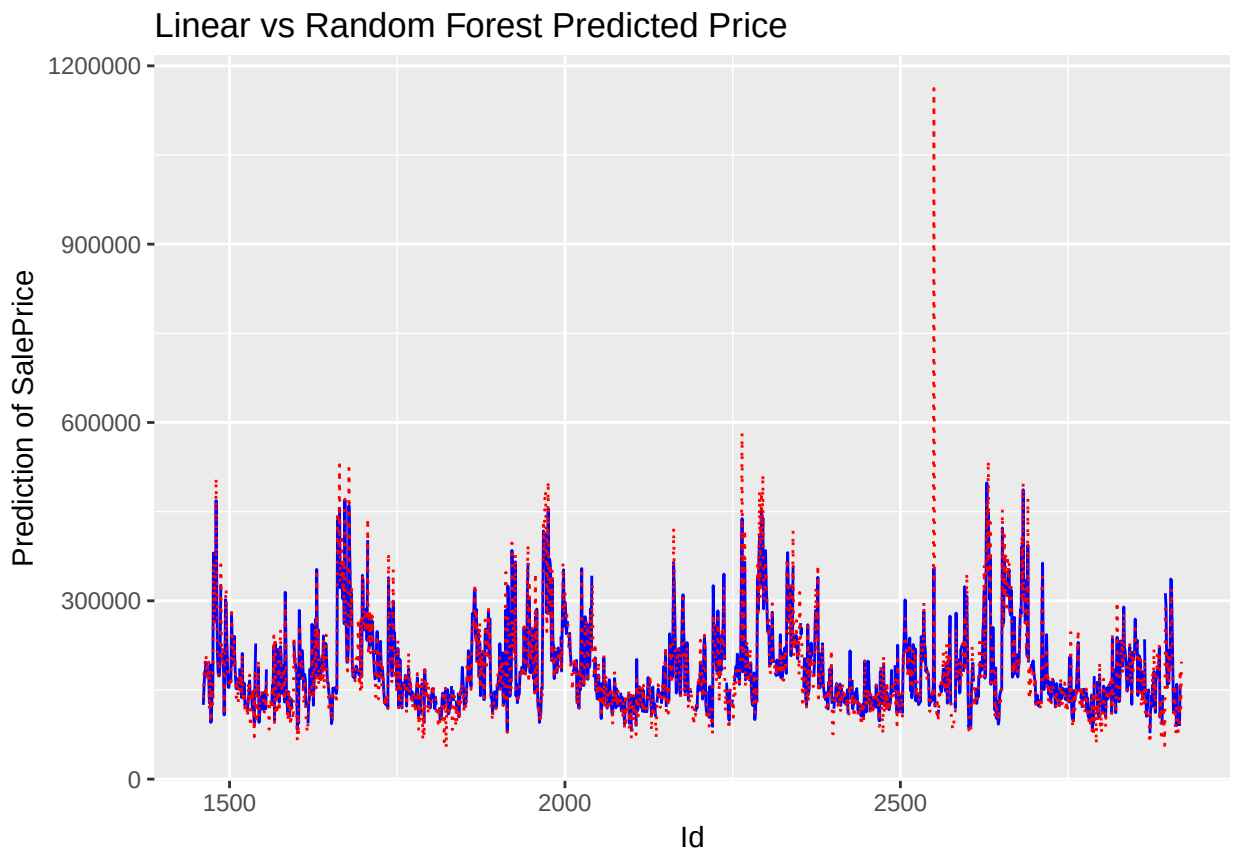
Step 4

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.

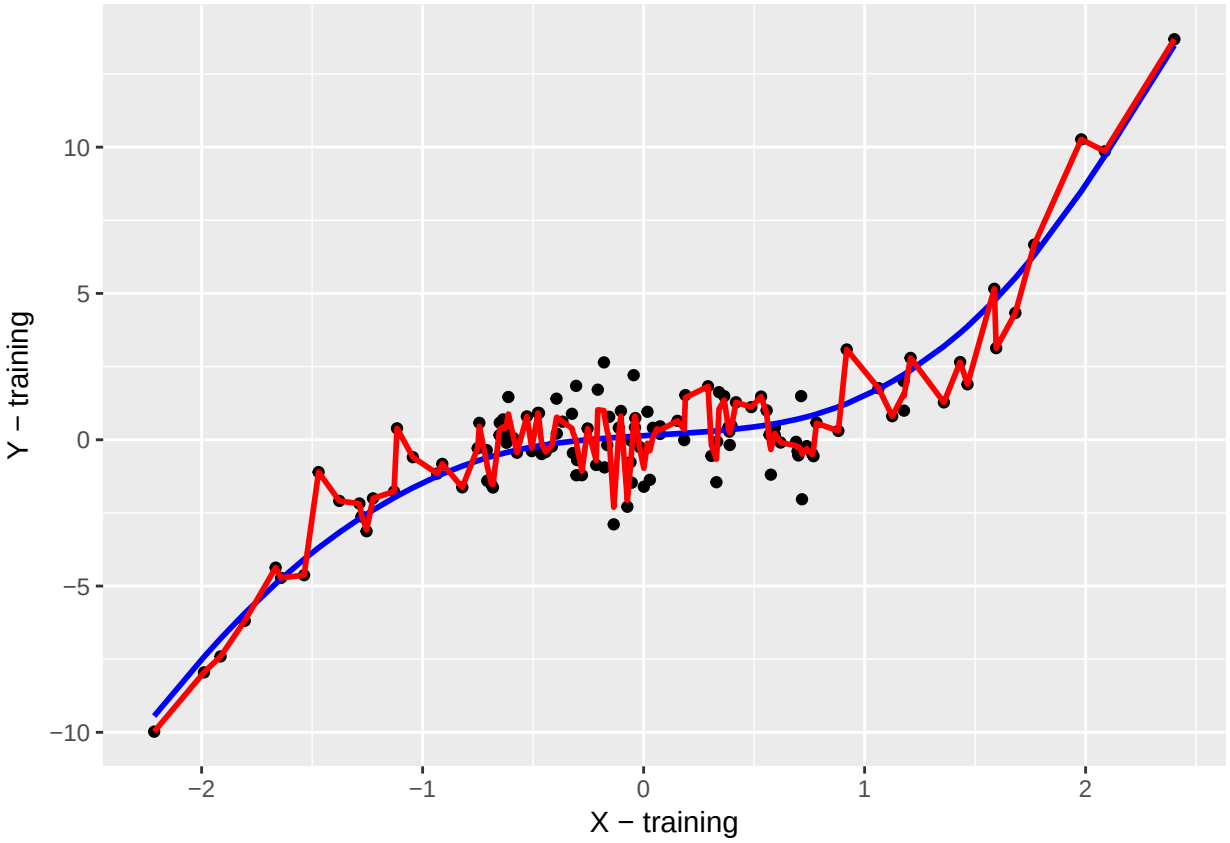
Appendix

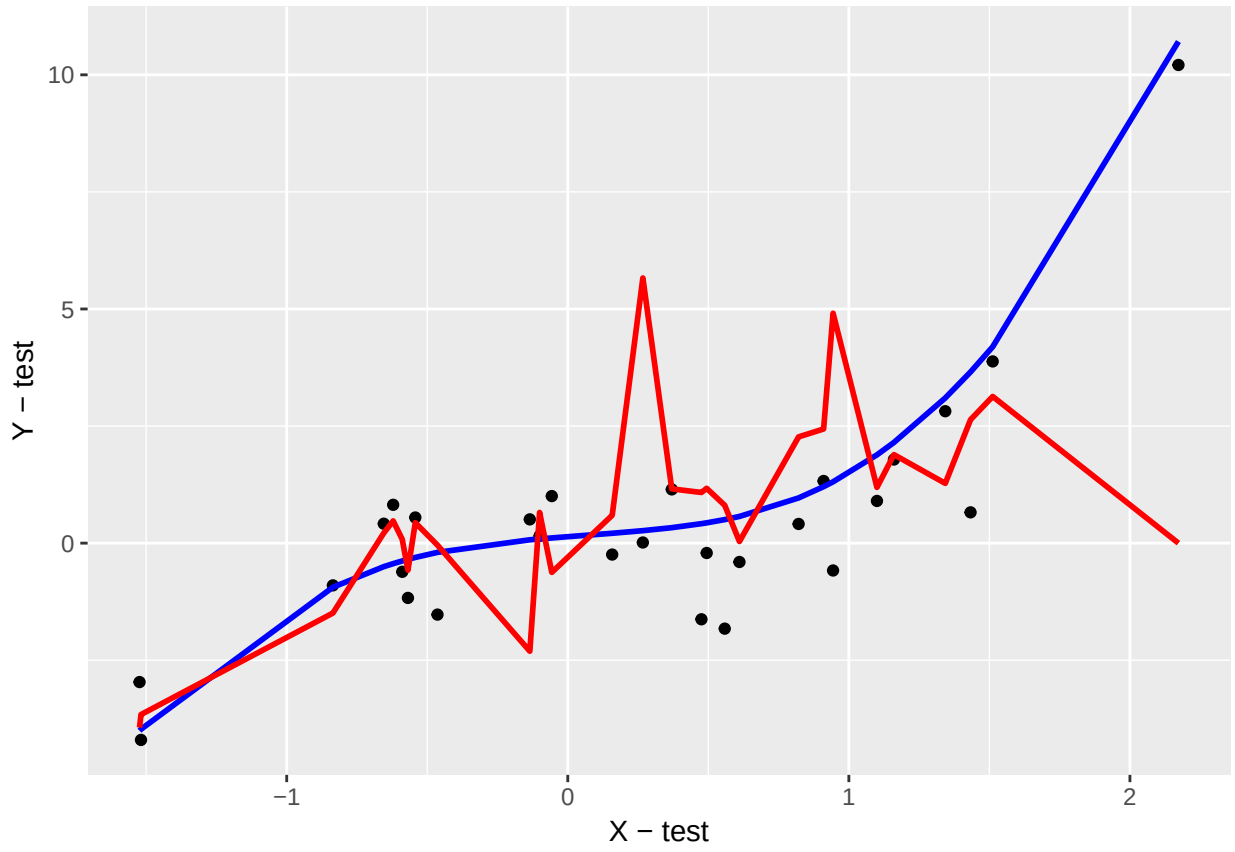
```
                Length Class   Mode
call               4   -none-  call
type               1   -none-  character
predicted       1338   -none-  numeric
mse              500   -none-  numeric
rsq              500   -none-  numeric
oob.times       1338   -none-  numeric
importance        73   -none-  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
```

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

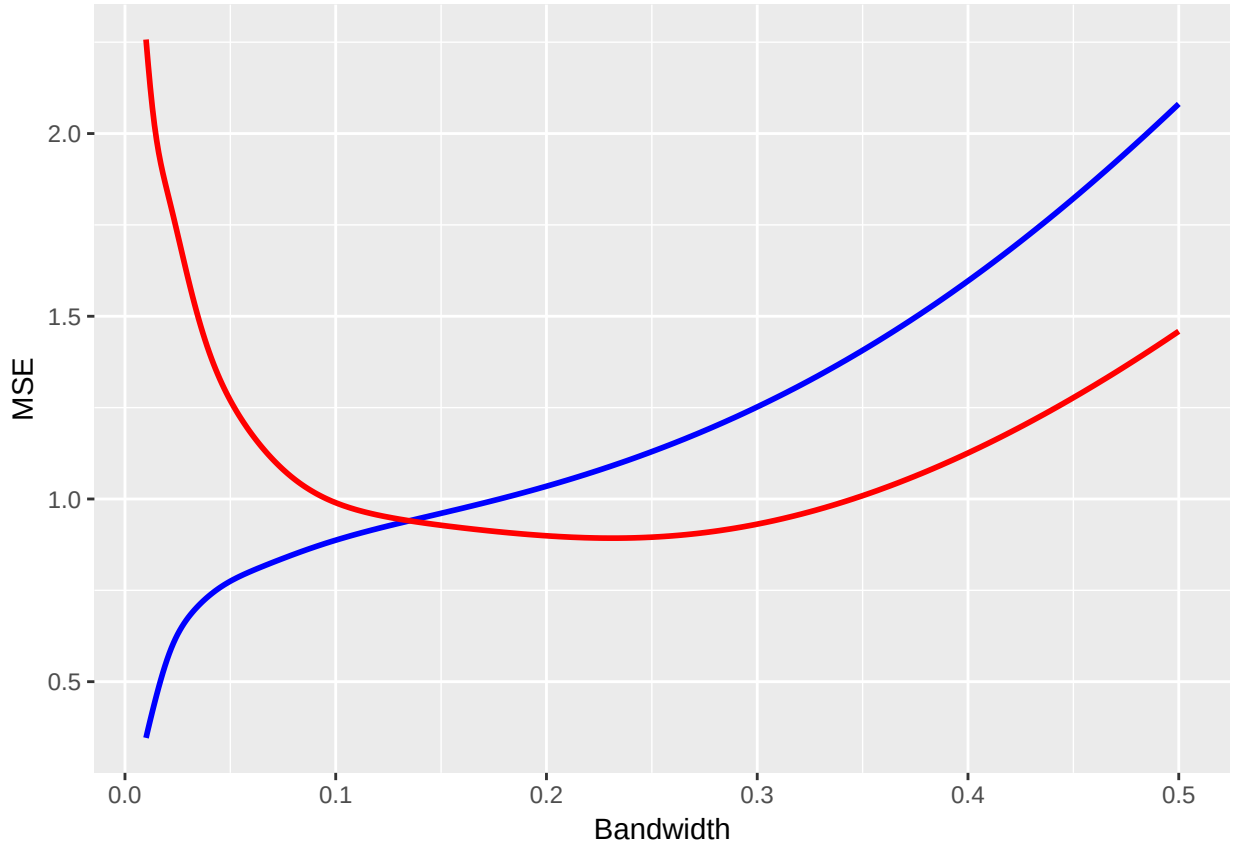



```
Regression Data: 122 training points, in 1 variable(s)
                x
Bandwidth(s): 0.5

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


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
```


10
5
0
−5
−10
Y − training
−2
−1
0
1
2
X − training




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




```
Time difference of 0.01562595 secs
Time difference of 7.726253 mins
Time difference of 0.09375906 secs
```