Data Mining & Machine Learning

Computer Exercise 9 - Random Forests

Steinarr Hrafn Höskuldsson

October 2022

Section 1.2

The confusion matrix was :

$$\text{CM} = \begin{bmatrix} 61 & 2 \\ 4 & 104 \end{bmatrix}$$

Accuracy was 96.5%, Precision was 98.1%, recall was 96.3% and cross validation accuracy was 91.5%.

Precision and recall gives a better picture than accuracy regarding false positives and false negatives. For an issue such as detecting cancer a false positive is preferable over a false negative. In such a case we want a model with high recall.

There is a noticeable discrepancy between accuracy and cross validation accuracy. By random chance the test/train split of the data set happened to result in a favourable set for the model, it's unlikely the model will get lucky 10 times in a row thus the cross validation score gives a much more accurate view of the accuracy.

If we wanted a confusion matrix from the cross validation we could implement our own cross validation method that averages up the confusion matrices into one. Precision and recall could then be calculated from the averaged confusion matrix.

Section 2.1

The confusion matrix was:

$$\text{CM}= \begin{bmatrix} 59 & 4 \\ 0 & 108 \end{bmatrix}$$

Accuracy was 97.66% , Precision was 96.43% , recall was 100.00% and cross validation accuracy was 96.14% .

The best combination of hyperparameters found were `n_estimators = 120` and `max_features = 4`.

Section 2.2

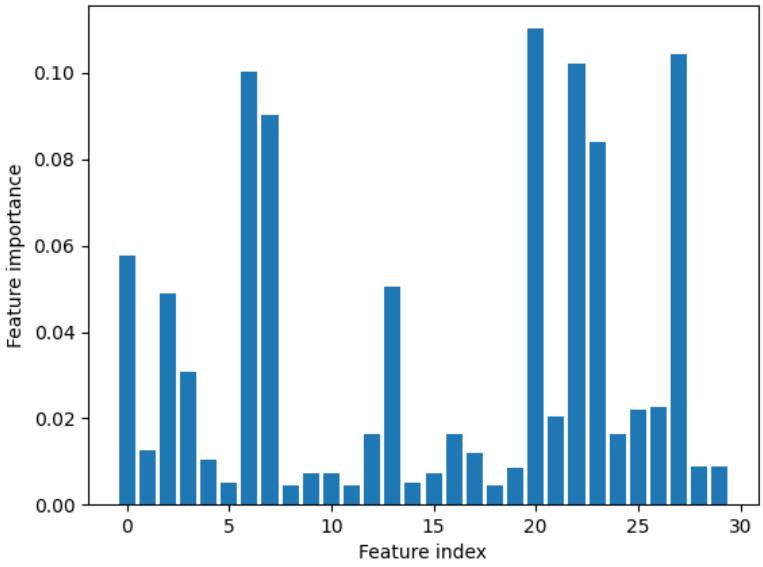

Figure 1:

Section 2.3

The most important feature was the 21rd feature, number 20, named `radius_worst`. The least important feature was the nineteenth feature, number 18, named `symmetry_se`

Section 2.4

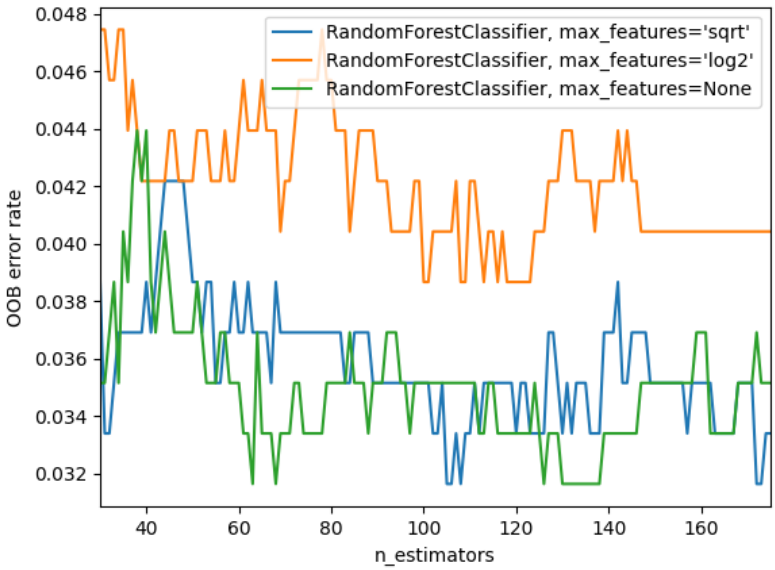

Figure 2: The Out Of Bag error for different settings of parameters. The plots have been smoothed slightly using scipy's uniformfilter1d function, using n=5.

Section 2.5

Looking at Figure 2 there does not seem to be a correlation between `n_estimators` and the Error rate.

This seems to be true for all three types of ensambles.

1 Section 3.1

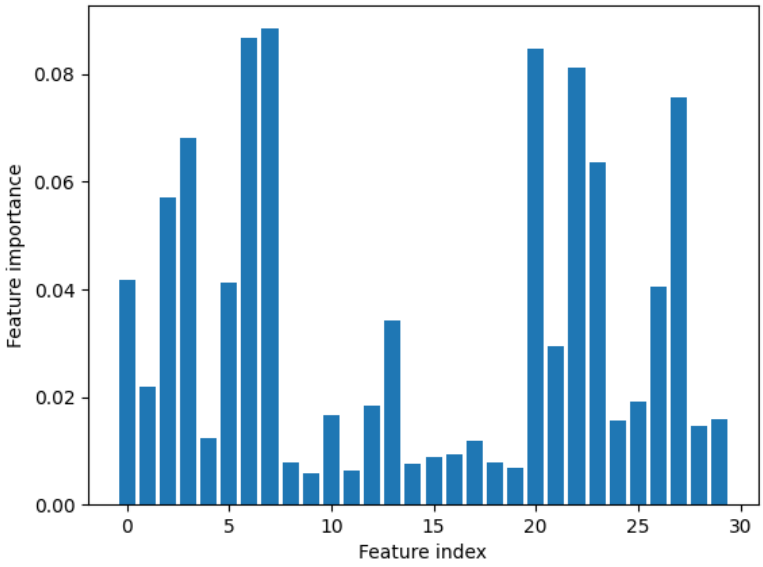

Figure 3: the Feature importance for Extremely Random Forest

The confusion matrix was:

$$\text{CM}= \begin{bmatrix} 58 & 50 & 108 \end{bmatrix}$$

Accuracy was 97.08% , Precision was 95.58% , recall was 100.00% and cross validation accuracy was 97.01%

The most important feature was the eighth feature, number 7, named `concave_points_mean`. The least important feature was the tenth feature, number 9, named `fractal_dimension_mean`

2 Section 3.2

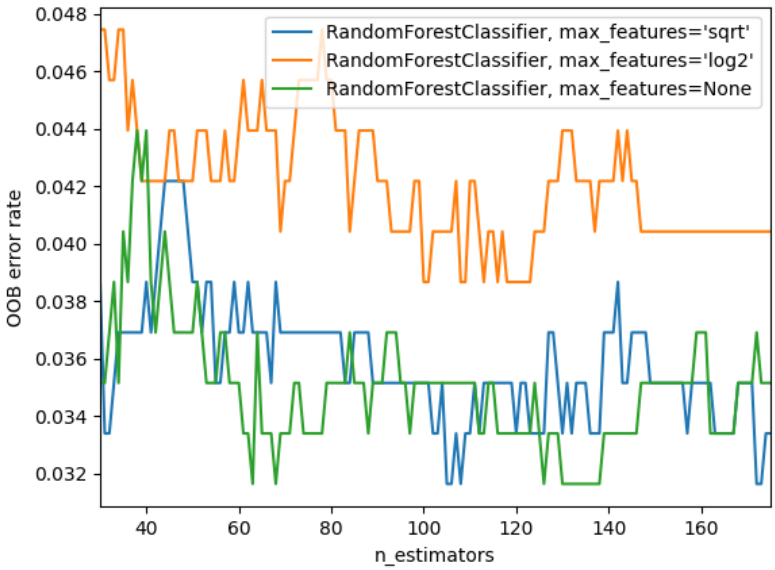

Figure 4: The Out Of Bag error for different settings of parameters of Extremely Random Forest Classifier. The plots have been smoothed slightly using scipy's uniformfilter1d function, using n=5.

Independent

Not attempted