Homework 2 Shapley Values - Piotr Sotniczuk

Dataset sklearn-diabetes

Description of dataset 442 samples, each has 10 body measures, s1-s6 were taken from blood sample. All variables have been mean centered and scaled by the standard deviation times `n_samples` . Target is to measure progression of diabetes after one year.

```
.. _diabetes_dataset:

Diabetes dataset
----------------

Ten baseline variables, age, sex, body mass index, average blood
pressure, and six blood serum measurements were obtained for each of n =
442 diabetes patients, as well as the response of interest, a
quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

  :Number of Instances: 442

  :Number of Attributes: First 10 columns are numeric predictive values

  :Target: Column 11 is a quantitative measure of disease progression one year after baseline

  :Attribute Information:
      - age     age in years
      - sex
      - bmi     body mass index
      - bp      average blood pressure
      - s1      tc, total serum cholesterol
      - s2      ldl, low-density lipoproteins
      - s3      hdl, high-density lipoproteins
      - s4      tch, total cholesterol / HDL
      - s5      ltg, possibly log of serum triglycerides level
      - s6      glu, blood sugar level

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times `n_samples
` (i.e. the sum of squares of each column totals 1).

Source URL:
https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

For more information see:
Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statisti
cs (with discussion), 407-499.
(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)
```

Quick look into the data features

```
<bound method NDFrame.head of           age       sex       bmi        bp        s1        s2        s3  \
0    0.038076  0.050680  0.061696  0.021872 -0.044223 -0.034821 -0.043401
1   -0.001882 -0.044642 -0.051474 -0.026328 -0.008449 -0.019163  0.074412
2    0.085299  0.050680  0.044451 -0.005671 -0.045599 -0.034194 -0.032356
3   -0.089063 -0.044642 -0.011595 -0.036656  0.012191  0.024991 -0.036038
4    0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142
..        ...       ...       ...       ...       ...       ...       ...
437  0.041708  0.050680  0.019662  0.059744 -0.005697 -0.002566 -0.028674
438 -0.005515  0.050680 -0.015906 -0.067642  0.049341  0.079165 -0.028674
439  0.041708  0.050680 -0.015906  0.017282 -0.037344 -0.013840 -0.024993
440 -0.045472 -0.044642  0.039062  0.001215  0.016318  0.015283 -0.028674
441 -0.045472 -0.044642 -0.073030 -0.081414  0.083740  0.027809  0.173816

           s4        s5        s6
0   -0.002592  0.019908 -0.017646
1   -0.039493 -0.068330 -0.092204
2   -0.002592  0.002864 -0.025930
3    0.034309  0.022692 -0.009362
4   -0.002592 -0.031991 -0.046641
..        ...       ...       ...
437 -0.002592  0.031193  0.007207
438  0.034309 -0.018118  0.044485
439 -0.011080 -0.046879  0.015491
440  0.026560  0.044528 -0.025930
441 -0.039493 -0.004220  0.003064

[442 rows x 10 columns]>
```

Target:

```
0      151.0
1       75.0
2      141.0
3      206.0
4      135.0
       ...
437    178.0
438    104.0
```

```
439    132.0
440    220.0
441     57.0
```

First Model

As the first model I've choosen Random Forest. Then I have splitted the dataset for train and test data, test data is about 10%.

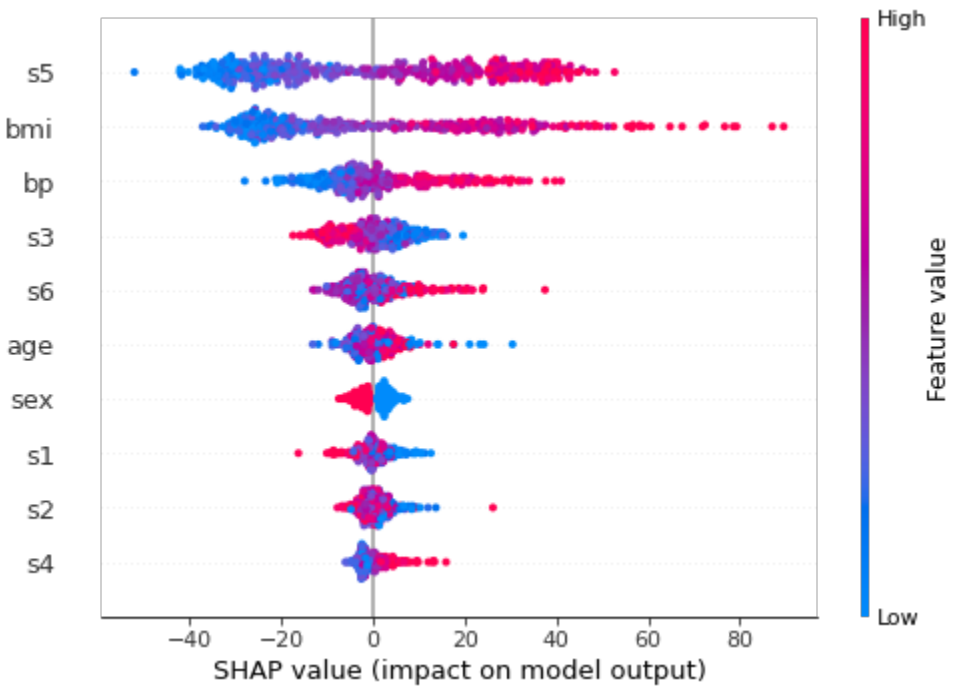
I looked into all predicitons on test data and tried to think of some interesting observations. I've choosen 3 observations, one that the prediction was very precise (284), one that prediction was too high (75) and one that prediction was to low (283). Results of those predictions are below.

```
index, value, prediction
284 156.0 162.91
75 42.0 120.38
283 181.0 86.36
```

Just to be sure I also outputed their features

```
          age       sex       bmi        bp        s1        s2        s3  \
284  0.041708  0.050680 -0.022373  0.028758 -0.066239 -0.045155 -0.061809
75  -0.009147  0.050680 -0.030996 -0.026328 -0.011201 -0.001001 -0.021311
283 -0.016412 -0.044642 -0.052552 -0.033214 -0.044223 -0.036387  0.019187

           s4        s5        s6
284 -0.002592  0.002864 -0.054925
75  -0.002592  0.006209  0.027917
283 -0.039493 -0.068330 -0.030072
```

Beeswarm plot of the Random Forest Regressor

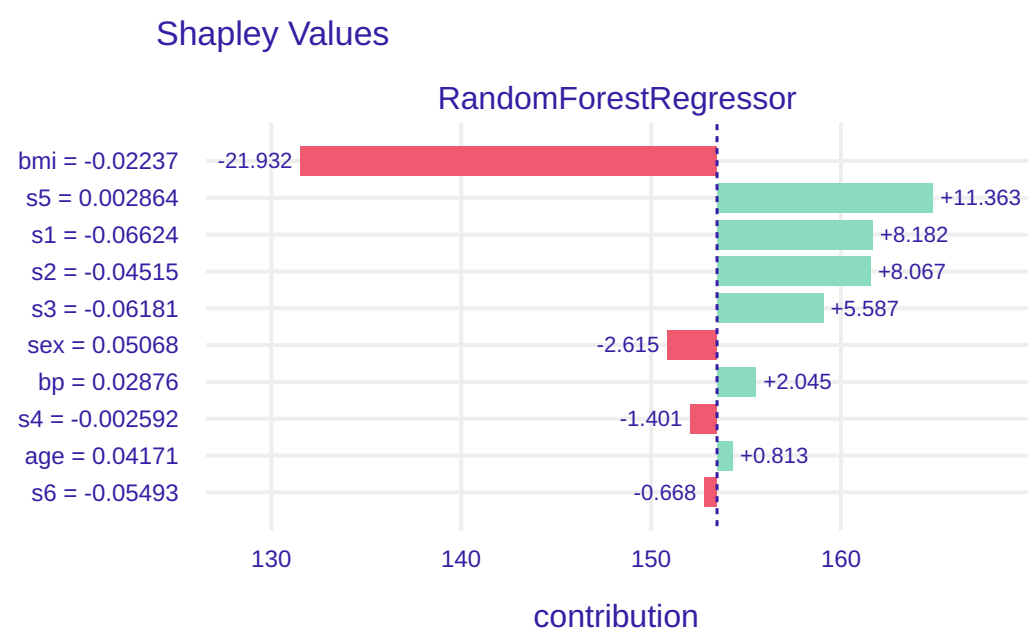


From this plot we can see that most significant features are s5 (triglycerides), bmi, bp (blood preasure). Triglycerides are basicily type of fat that is stored in human cells and then used for energy purpuses when needed. You can learn more about s5 here: https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186

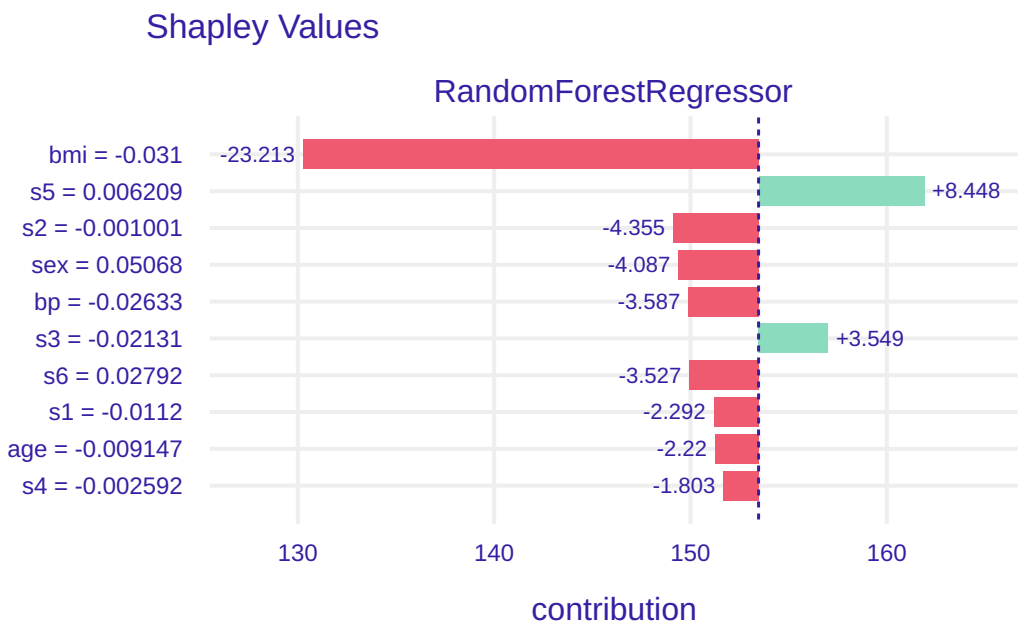
We can see from the plot that every feature has some impact on the model output. For features like s5, bmi, bp, s6, s4 high value of the feature impacts positively on the model. This fact is similar to my intuition because usually people with high bmi and blood preasure tend to worsen their condition.

Looking into choosen observations

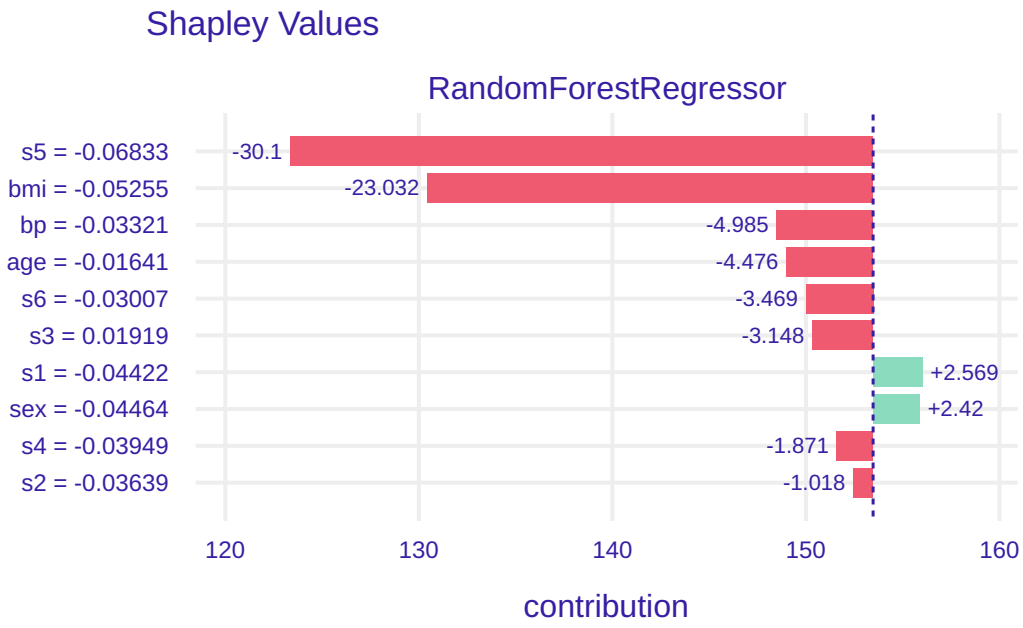
284



75



283



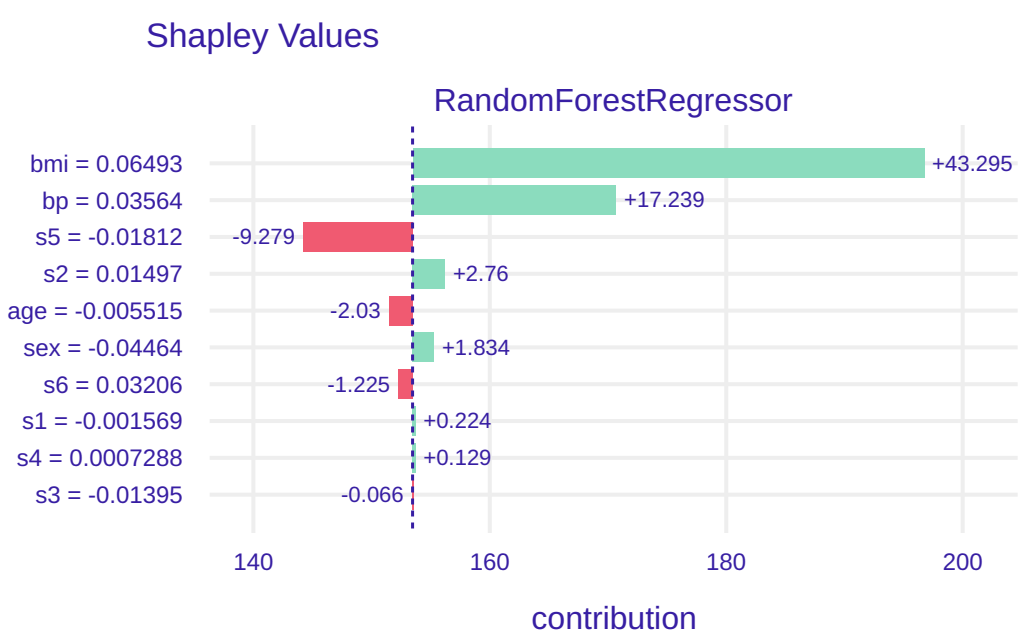
All of observations have s5 and bmi as 2 most significant features. This complies with the beeswarm plot above. We can see that feature 's5' has positive effect on observations 284 and 75, but negative effect in observation 283 whatsmore in this observation 's5' is the most significant feature.

Looking for observations with diffrent features of highest importance

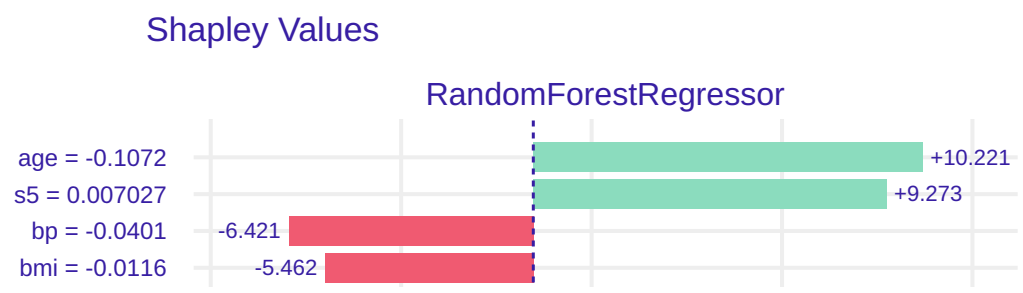
As 's5', 'bmi' and 'bp' tend to be of most importance it is hard to find observations outside this pattern. I looked into all observations from test data and managed to find one or two examples.

The choosen observations are 310 and 344.

310



344



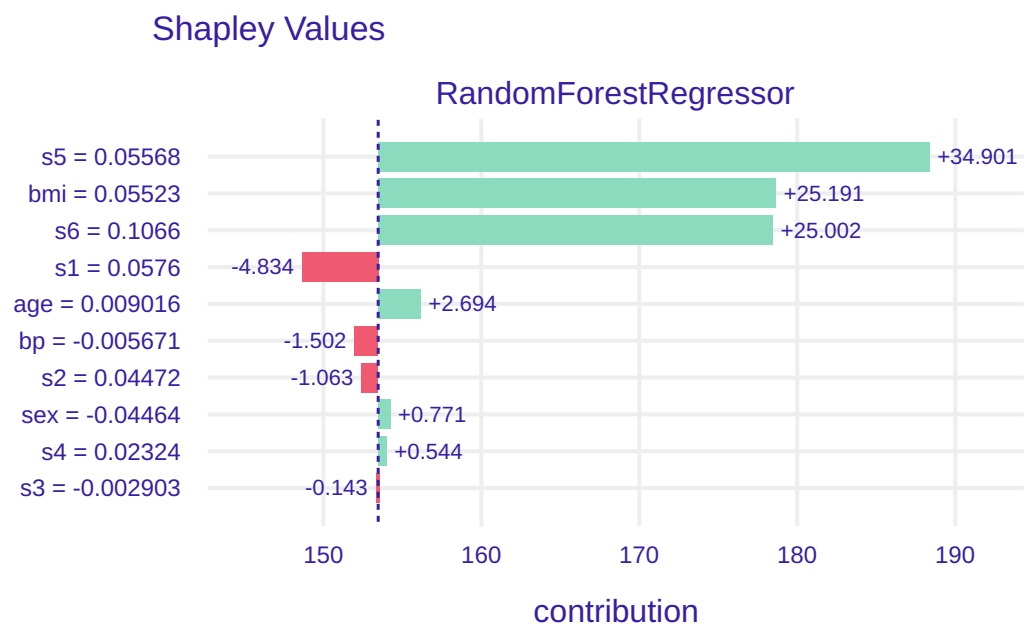
Let's look into their features maybe there is something interesting inside.

```
          age       sex       bmi        bp        s1        s2        s3  \
310 -0.005515 -0.044642  0.064930  0.035644 -0.001569  0.014970 -0.013948
344 -0.107226 -0.044642 -0.011595 -0.040099  0.049341  0.064447 -0.013948

           s4        s5        s6
310  0.000729 -0.018118  0.032059
344  0.034309  0.007027 -0.030072

index, value, prediction
310 109.0 206.35
344 200.0 158.39
```

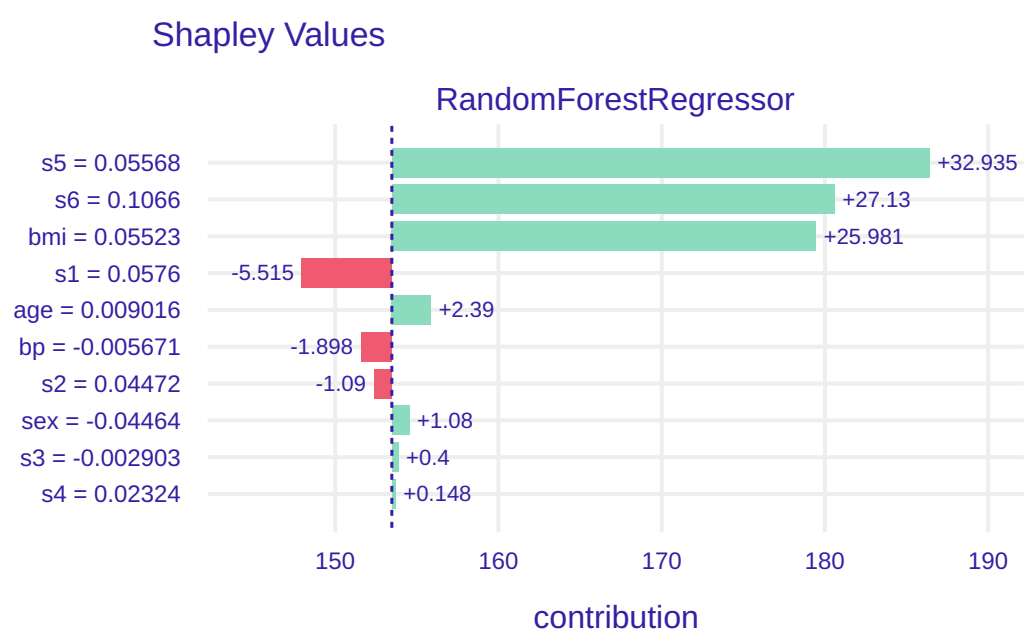
Strange output

As I was looking into different outputs of Shapley values I noticed something strange. I plotted the same object two times and diffrent plots appeared. The plots for (432) are similar, but there is a significant difference. The second and third most important features switch places. I thought that the alghorythm for counting shapley values is deterministic and now I am a bit confused.

432



432

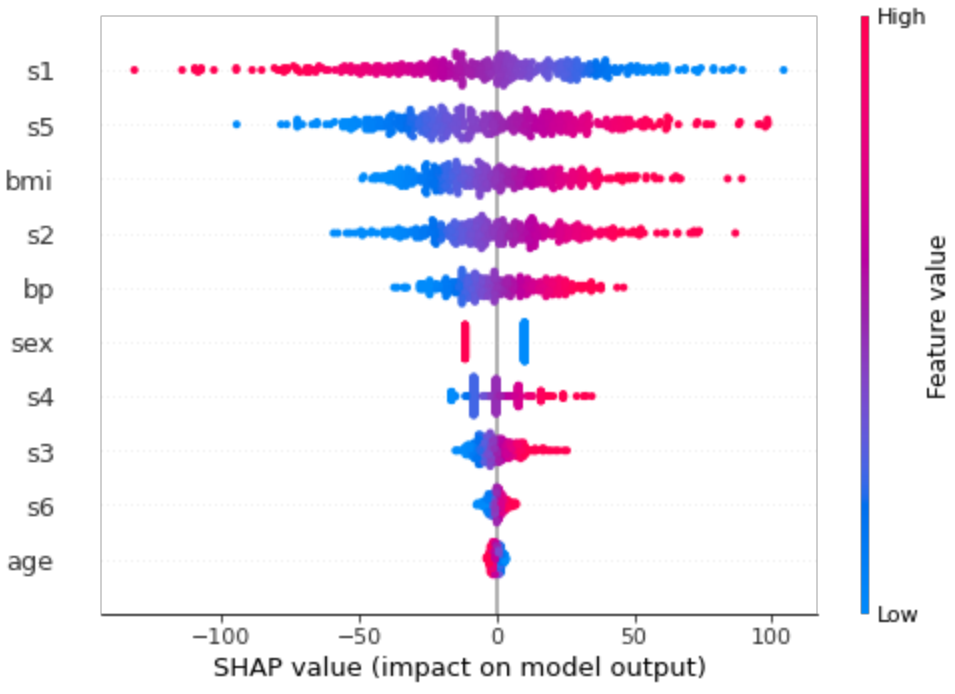


Second model

For the second model I picked LinearRegression. I trained it on the same data as the model above.

Beeswarm plot

Firstly lets see how the new summary plot looks like.

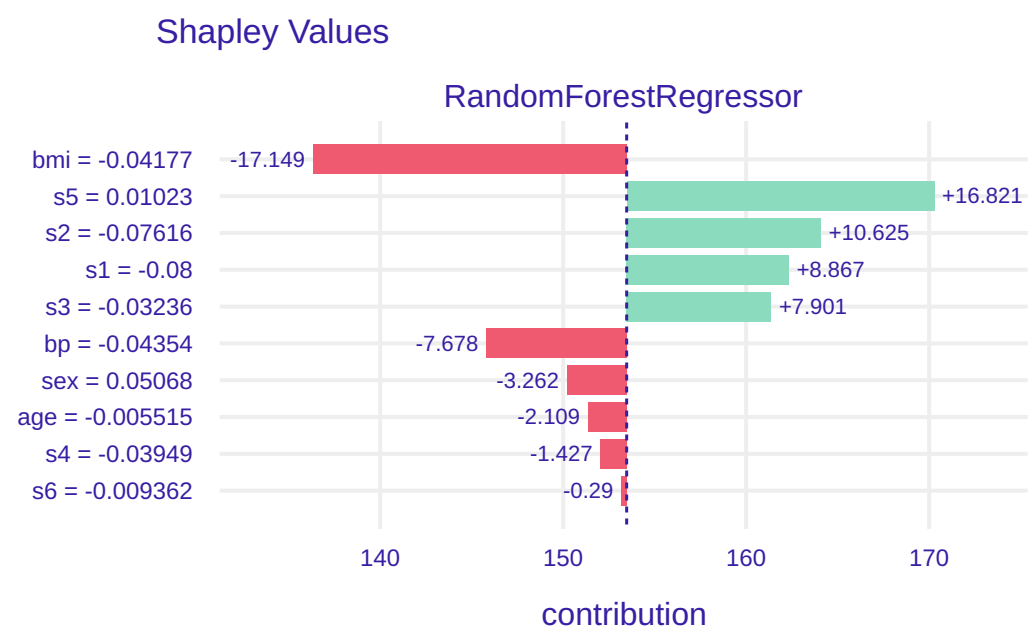


This time feature 's1' (total serum cholesterol) seems to be much more important. Still features like 's5' and 'bmi' are in top 3 features. 'S1' has negative impact on model, so high values lowers predictions. This doesn't meet my expectations. High cholesterol should be bad for the health of the patient.

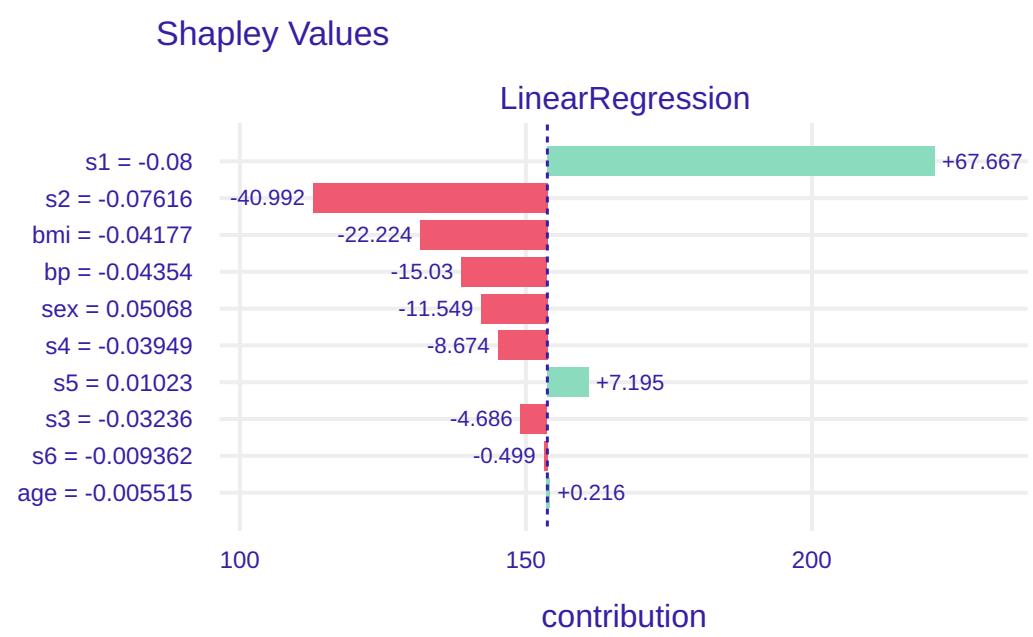
Comparing models

I looked in few observations from Linear Regression and it was rather easy to find an observation that is different in Random Forest. I choosed observation 191.

191



191



Features 's1' and 's2' (low-density lipoproteins) are most significant in those observations. It was hardly possible to find any observation like this in Random Forest.

Thoughts for the future

It would be nice to have better inforamtion background on the features. I don't really understand how features 's1'-'s6' work but they are all connected to fat/cholesterol. The only knowledge I have is that high cholesterol is bad but as the second beeswarm plot showed I may be wrong. Those features might also be dependent from each other.