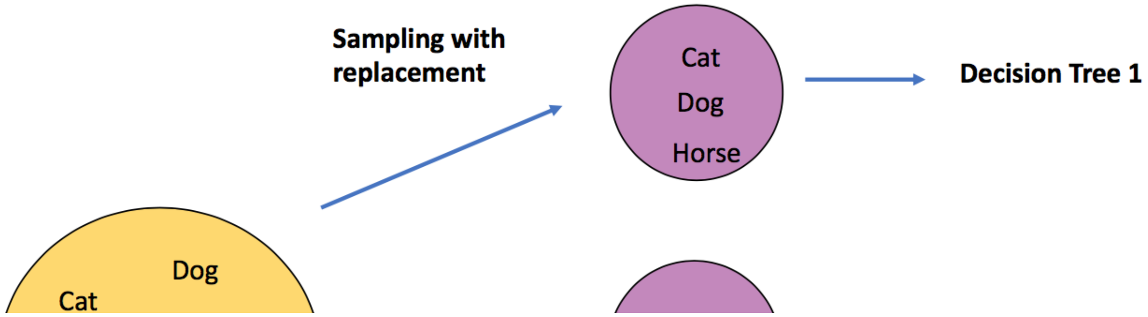
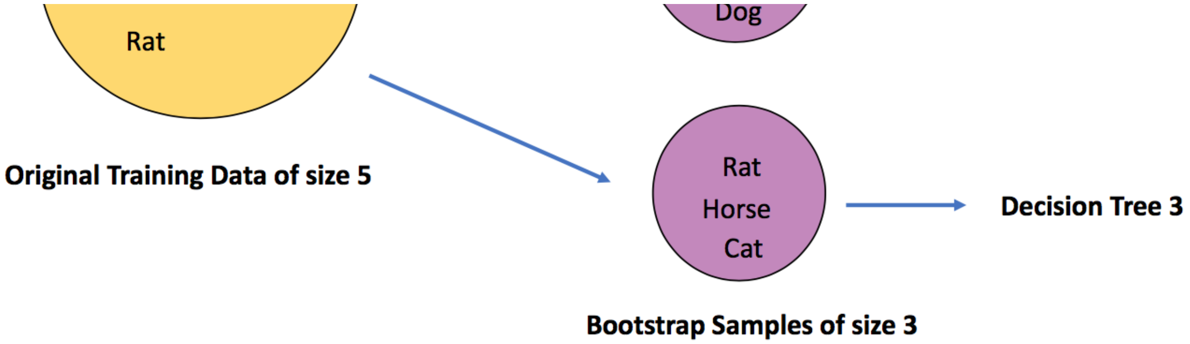
What is Out of Bag (OOB) score in Random Forest?

Navnina Bhatia Follow
Jun 26, 2019 · 5 min read ★

This blog attempts to explain the internal functioning of oob_score when it is set as true in the "RandomForestClassifier" in "Scikit learn" framework. This blog describes the intuition behind the Out of Bag (OOB) score in Random forest, how it is calculated and where it is useful.

In the applications that require good interpretability of the model, DTs work very well especially if they are of small depth. However, DTs with real-world datasets can have large depths. Higher depth DTs are more prone to overfitting and thus lead to higher variance in the model. This shortcoming of DT is explored by the Random Forest model. In the Random Forest model, the original training data is **randomly** sampled-with-replacement generating small subsets of data (see the image below). These subsets are also known as bootstrap samples. These bootstrap samples are then fed as training data to many DTs of large depths. Each of these DTs is trained separately on these bootstrap samples. This aggregation of DTs is called the Random Forest ensemble. The concluding result of the ensemble model is determined by counting a majority vote from all the DTs. This concept is known as Bagging or Bootstrap Aggregation. Since each DT takes a different set of training data as input, the deviations in the original training dataset do not impact the final result obtained from the aggregation of DTs. Therefore, bagging as a concept reduces variance without changing the bias of the complete ensemble.



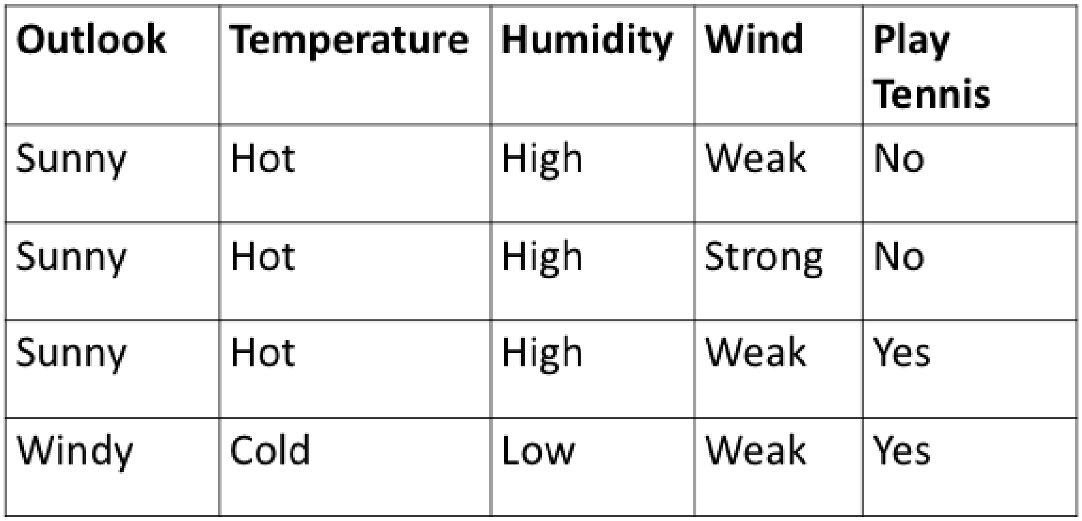


Generation of bootstrap samples with replacement. "Sampling-with-replacement" here means that if a data point is chosen in the first random draw it still remains in the original sample for choosing in another random draw that may follow with an equal probability. This can be seen in the image above as "Dog" is chosen twice in the second bootstrap sample.

What is the Out of Bag score in Random Forests?

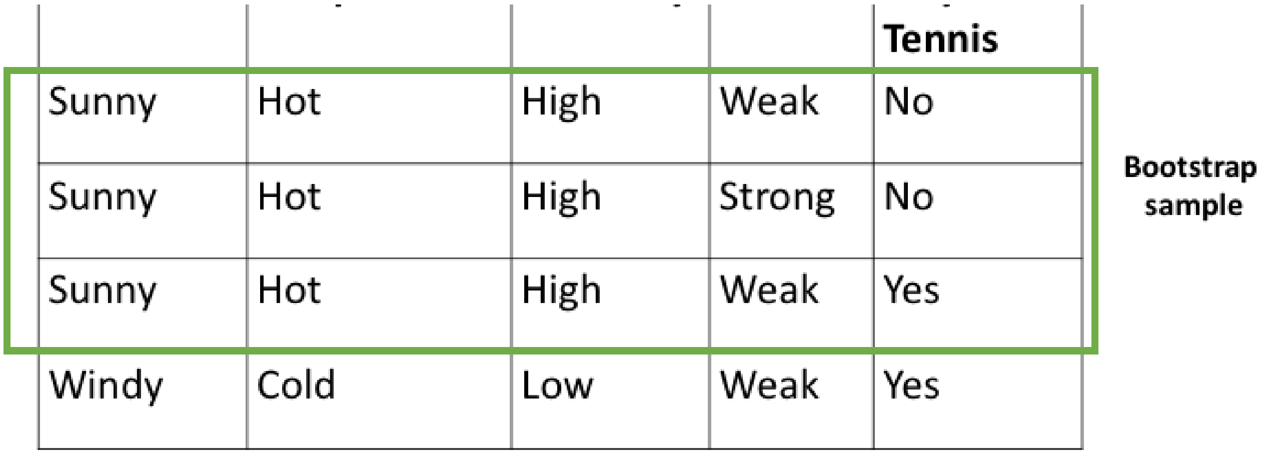
Out of bag (OOB) score is a way of validating the Random forest model. Below is a simple intuition of how is it calculated followed by a description of how it is different from validation score and where it is advantageous.

For the description of OOB score calculation, let's assume there are five DTs in the random forest ensemble labeled from 1 to 5. For simplicity, suppose we have a simple original training data set as below.



Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

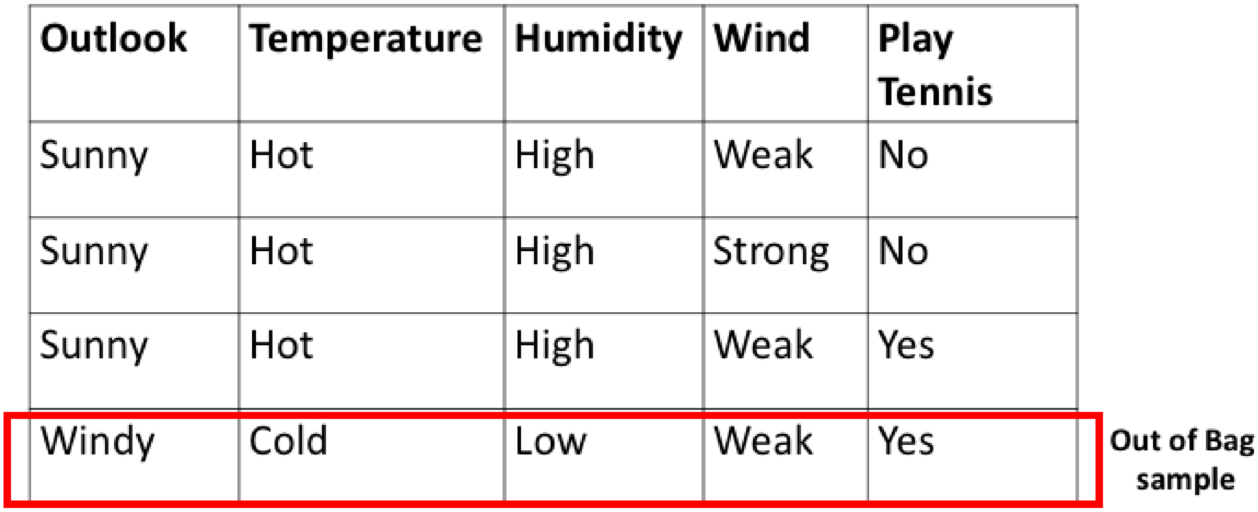
Let the first bootstrap sample is made of the first three rows of this data set as shown in the green box below. This bootstrap sample will be used as the training data for the DT "1".



				Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

Bootstrap sample

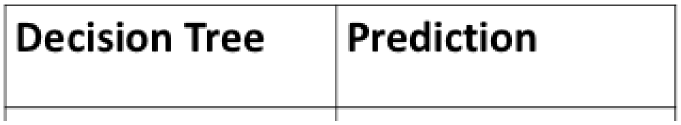
Then the last row that is “left out” in the original data (see the red box in the image below) is known as Out of Bag sample. This row will not be used as the training data for DT 1. Please note that in reality there will be several such rows which are left out as Out of Bag, here for simplicity only one is shown.



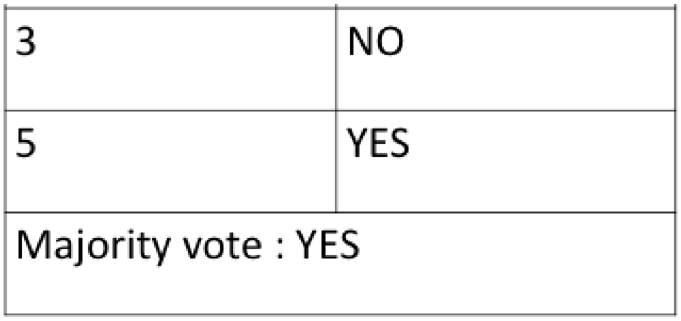
Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

Out of Bag sample

After the DTs models have been trained, this leftover row or the OOB sample will be given as unseen data to the DT 1. The DT 1 will predict the outcome of this row. Let DT 1 predicts this row correctly as “YES”. Similarly, this row will be passed through all the DTs that did not contain this row in their bootstrap training data. Let’s assume that apart from DT 1, DT 3 and DT 5 also did not have this row in their bootstrap training data. The predictions of this row by DT 1, 3, 5 are summarized in the table below.



Decision Tree	Prediction



3	NO
5	YES
Majority vote : YES	

We see that by a majority vote of 2 “YES” vs 1 “NO” the prediction of this row is “YES”. It is noted that the final prediction of this row by majority vote is a **correct prediction** since originally in the “Play Tennis” column of this row is also a “YES”.

Similarly, each of the OOB sample rows is passed through every DT that did not contain the OOB sample row in its bootstrap training data and a majority prediction is noted for each row.

And lastly, the OOB score is computed as **the number of correctly predicted rows from the out of bag sample.**

What is the difference between OOB score and validation score?

Since we have understood how OOB score is estimated let’s try to comprehend how it differs from the validation score.

As compared to the validation score OOB score is computed on data that was not necessarily used in the analysis of the model. Whereas for calculation validation score, a part of the original training dataset is actually set aside before training the models. Additionally, the OOB score is calculated using only a subset of DTs not containing the OOB sample in their bootstrap training dataset. While the validation score is calculated using all the DTs of the ensemble.

Where can OOB score be useful?

As noted above, only a subset of DTs is used for determining the OOB score. This leads to reducing the overall aggregation effect in bagging. Thus in general, validation on a full ensemble of DTs is better than a subset of DT for estimating the score. However, occasionally the dataset is not big enough and hence set aside a part of it for validation is unaffordable. Consequently, in cases where we do not have a large dataset and want to

Nonetheless, it should be noted that validation score and OOB score are unalike, computed in a different manner and should not be thus compared.

In an ideal case, about 36.8 % of the total training data forms the OOB sample. This can be shown as follows.

If there are N rows in the training data set. Then, the probability of not picking a row in a random draw is

$$\frac{N-1}{N}$$

Using sampling-with-replacement the probability of not picking N rows in random draws is

$$\left(\frac{N-1}{N}\right)^{N}$$

which in the limit of large N becomes equal to

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{N}=e^{-1}=0.368$$

Therefore, about 36.8 % of total training data are available as OOB sample for each DT and hence it can be used for evaluating or validating the random forest model.

Machine Learning Random Forest Data Science AI Towards Data Science