Forward Feature Selection

Forward Feature Selection is a feature selection technique that iteratively builds a model by adding one feature at a time, selecting the feature that maximizes model performance. It starts with an empty set of features and adds the most predictive feature in each iteration until a stopping criterion is met. This method is particularly useful when dealing with a large number of features, as it incrementally builds the model based on the most informative features. This process involves assessing new features, evaluating combinations of features, and selecting the optimal subset of features that best contribute to model accuracy.

Backward Feature Selection: Backward feature selection is closely related, and as you may have guessed starts with the entire set of features and works backward from there, removing features to find the optimal subset of a predefined size.

We'll use the same example of fitness level prediction based on the three independent variables:

ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

So the first step in Forward Feature Selection is to train n models using each feature individually and checking the performance. So if you have three independent variables, we will train three models using each of these three features individually. Let's say we trained the model using the **Calories_Burnt** feature and the target variable, **Fitness_Level** and we've got an accuracy of **87%**:

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 87%

Next, we'll train the model using the **Gender** feature, and we get an accuracy of **80%:**

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 80%

And similarly, the **Plays_sport** variable gives us an accuracy of **85%**:

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 85%

Now we will choose the variable, which gives us the best performance. When you look at this table:

Variable used	Accuracy
Calories_burnt	87.00%
Gender	80.00%
Plays_Sport?	85.00%

As you can see Calories_Burnt alone gives an accuracy of 87% and Gender give 80% and the Plays_Sport variable gives 85%. When we compare these values, of course, Calories_Burnt produced the best result. And hence, we will select this variable.

Next, we will repeat this process and add one variable at a time. So of course we'll keep the **Calories_Burnt** variable and keep adding one variable. So let's take **Gender** here and using this we get an accuracy of **88%**:

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 88%

\

When you take **Plays_Sport** along with **Calories_Burnt**, we get an accuracy of **91%.** A variable that produces the highest improvement will be retained. That intuitively makes sense. As you can see, Plays_Sport gives us a better accuracy when we combined it with the Calories_Burnt. Hence we will retain that and select it in our model. We will repeat the entire process until there is no significant improvement in the model's performance.

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 91%

Summary:

Steps to Perform Forward Feature Selection

1. Train n model using feature (n) individually and check the performance.
2. Choose the variable which gives the best performance.
3. Repeat the process and add one variable at a time.
4. Variable Producing the highest improvement is retained.
5. Repeat the entire process until there is no significant improvement in the model's performance.