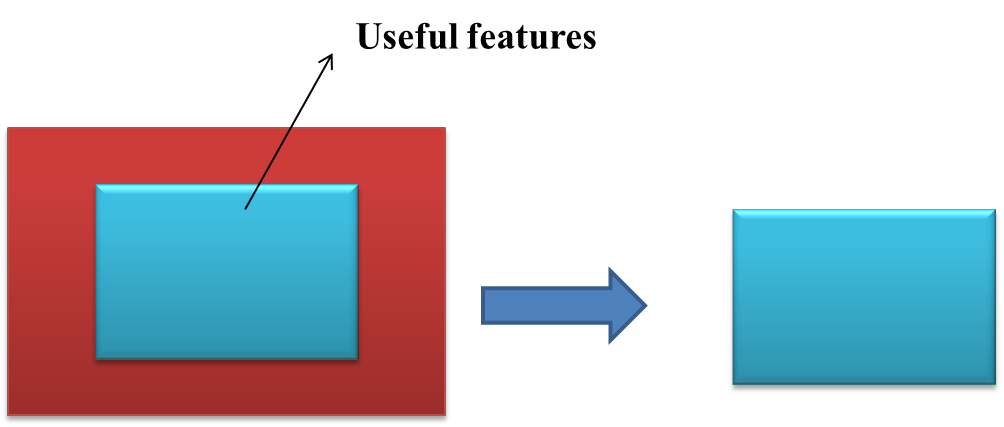
**What is Feature selection?**

*Feature selection is primarily the removal of non-informative or redundant features from the model.*

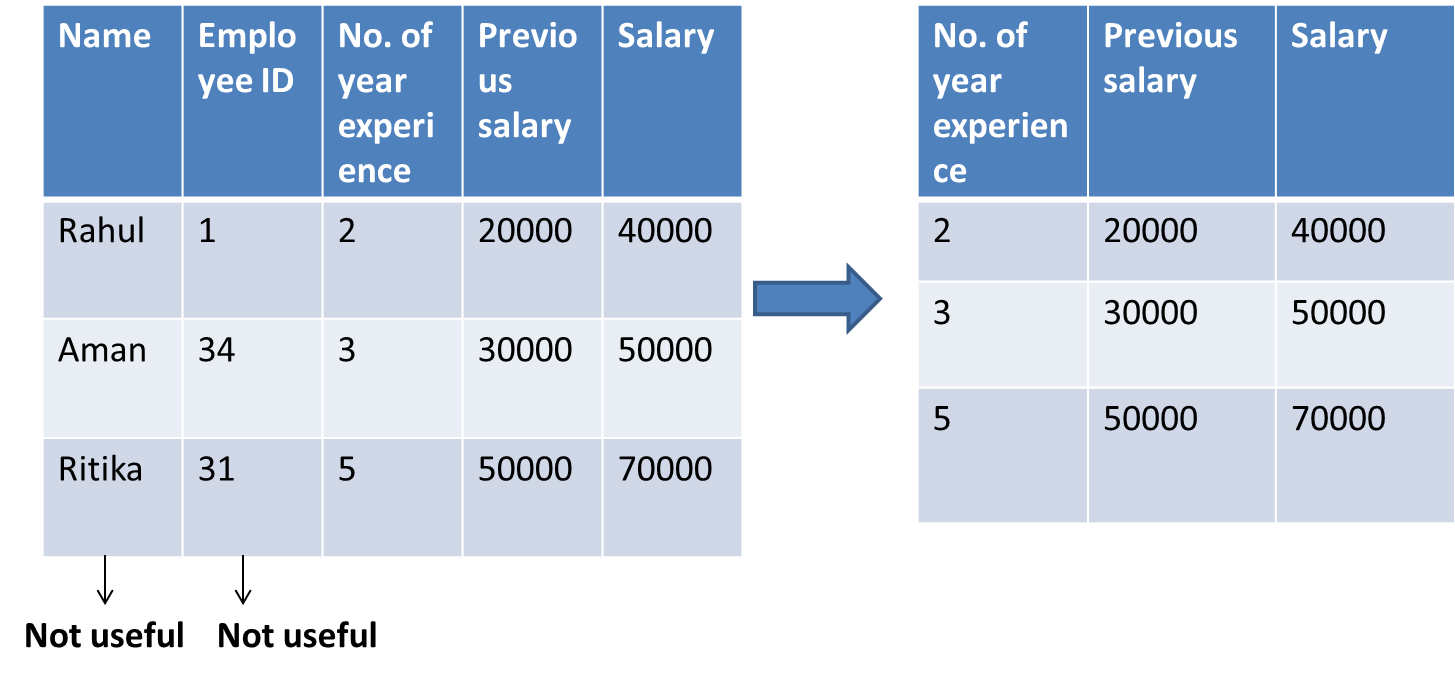
**Feature selection** is the automatic process of reducing the number of input from the set of features by choosing the relevant variables before feeding them to the machine learning model. The motive is to reduce the irrelevant data(noise) and choose relevant features. It is obvious that giving noise data to the model will not serve a good purpose. Now, what is noise data? It could be

* Unwanted features that have no relevance in achieving accurate results
* Outliers
* Null values
* Corrupt data
* Unstructured data
* Data entry errors
* Missing values

But here feature selection deals with unwanted or extra features that if removed will either have **no effect** or will **increase the accuracy** of the model.



Let’s take an example to suppose we have a data set of employees



* **Independent features**-Name, Employee ID, Number of years experience, previous salary,
* **Dependent featur**e-Salary

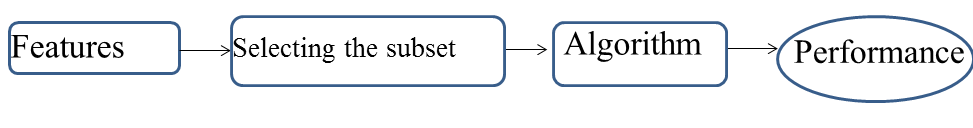
In this, we had two columns **Name**and **Employee ID** which have no relevance to the **Salary** feature. So these two features can be eliminated.

**Why Feature selection**

1. **Reduces Overfitting**: Feature selection removes redundant data also which ultimately reduces the chances of overfitting and high [bias](https://www.naukri.com/learning/articles/bias-and-variance/).
2. **Improves Accuracy**: Less [overfitting](https://www.naukri.com/learning/articles/overfitting-and-underfitting-with-a-real-life-example/) means the model will perform well on test data also and will give good accuracy.
3. **Reduces Training Time**: fewer data means less complexity. So less computation time to train the model.
4. Better generalized model.
5. Easy to explain and interpret features.

**Feature selection techniques**

1. **Filter Method**



This is a feature dropping method in which the features are dropped based on the score of features by some statistical test. This score indicates their relation to the output, or how much they are correlating to the target output(dependent feature). Each feature is evaluated with the outcome of a statistical test like

* Information gain
* Fisher’s score
* Chi-square
* Pearson correlation

And the score is generated. Later features are ranked based on their score and lower score features are removed. These methods are less computationally expensive and faster than wrapper methods. And are suitable to deal with high-dimensional data.

Let’s, discuss some of these techniques:

**Information Gain**

Information gain indicates the calculation of the reduction in entropy(disorder)from the transformation of a dataset. Information gain is defined as,

|  |
| --- |
| **IG(Y/X) = H(Y) – H(Y/X) where X and Y are the features.** |

Information gain is calculated for each feature(Independent feature)in the context of the target variable or dependent feature. More information gain indicates more entropy removal and also more information feature X carries about Y.

**Chi-square Test**

A chi-square test is used for categorical features to test the independence of two events. Chi-square value is calculated between the target feature and each feature and the desired number of features with the best Chi-square scores is selected. Now what to deduce from this chi-square score? The higher the Chi-Square value, the higher will be the dependence of the feature on the dependent feature will be and it can be now selected for model training. So on the basis of this value, the features are selected or rejected.

**Fisher’s Score**

On the basis of the fisher’s score, the algorithm we use will return the ranks of the variables in descending order. The Fisher score helps in selecting features with the larger scores.

**Correlation Coefficient**

Correlation is a measure of the relationship between two or more variables. Good features will be related to the dependent variable. That means the feature is relevant so now it can be selected. The idea is to select those features which have some relationship with the dependent feature. Let’s suppose we have a dataset in which we have a feature **Roll no**. And the target feature is **Marks**. Now**Roll no.**has no effect on **Marks** and is not related to **Marks**. So we can say that **Rollno** has less value for the correlation coefficient. We cannot predict marks with the help of **Roll no**.

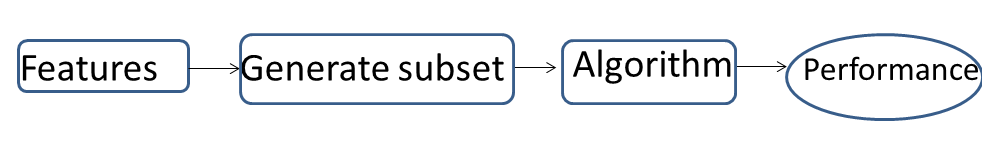
**NOTE:** Variables should be correlated with the target but should not be correlated among themselves.

An absolute value, say 0.5 is set as the threshold for selecting the variables. We can drop the variable if the feature has a lower correlation coefficient value with the target variable or dependent feature and are correlated among themselves. In case more than two variables are correlated to each other by computing multiple correlation coefficients. This phenomenon is known as multicollinearity.

We can use Pearson Correlation here.

2. **Wrapper method**

It is a feature selection method where you take a group of features and decide what is working best for your model. First different unset of features are created. Now with this set of features, the machine learning algorithm is trained on sample data, and model performance is evaluated. The set of features that gives the best performance is considered as selected features.



1. **Forward Selection**: Forward selection is an iterative method in which we start with the best performing feature with respect to the target feature. And then we keep on adding the features and checking the accuracy in each iteration and will check which combination of features is improving the model. We will perform this until the addition of a new variable does not improve the performance of the model. So in short we take a different combination of features and select the best combination.

2. **Backward Elimination**: In backward elimination is the opposite of forward selection, we take all features in starting and remove the least significant feature in each iteration until no improvement is observed.

3. **Recursive Feature elimination**: It is a greedy optimization algorithm that expects to observe the best performing feature subset. On every iteration, it creates models and keeps aside the worst or the best performing feature. Features are then ranked based on the order of their elimination.

Blogs that may interest you:

3. **Embedded methods**

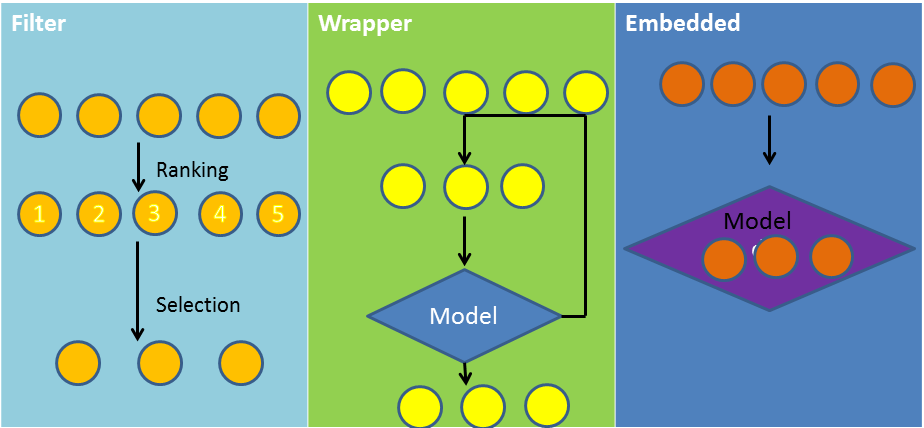
Embedded methods use machine learning models to automatically choose your features. This technique will select the best features during the training itself. Here you train the model and then you see which features are good enough for you.

**L1 Regularization**

L1 (or LASSO) regression is for generalized linear models that can add a penalty equivalent to the absolute value of the magnitude of coefficients in order to reduce the overfitting or variance of a model.

**Ridge regression**

It performs L2 regularization by adding a penalty equivalent to the square of the magnitude of coefficients.



**Difference between Filter and Wrapper methods**

|  |  |
| --- | --- |
| **Filter method** | **Wrapper method** |
| Measure the relevance of features with the dependent variable | Measure the usefulness of a subset of feature |
| This method is fast and is computationally less expensive | This method is slow and is computationally more expensive |
| Useful for large datasets | Useful for small datasets |
| Might fail to find the best subset of features | Always provide the best subset of features |
| Avoid overfitting | Prone to overfitting |

**NOTE:** The filter method is useful in case of more number of features. The wrapper method performs better when the embedded method lies in between the other two methods.