Feature Selection

While training a machine learning model, higher the number of features passed to the model, greater would be its complexity and it would require more data to be trained. So we should always try to remove the unnecessary features. This process of selecting only the essential features is referred to as Feature Selection.

Formally, feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion. The aim of feature selection is to maximize relevance and minimize redundancy

How do we identify which features are relevant for modeling our target variable?
There are many methods to perform feature selection. Let's start with the easiest one, that is assessing the correlation between the variables. If a variable is highly correlated with our target variable, then it is relevant. To measure correlation, first we must determine the type of variables we have.

1. If both are continuous variables:
    a. Pearson Correlation Coefficient
2. If both are categorical variables:
    a. If they have more than two categories:
        i. Cramer's V
    b. If they have only two categories:
        i. Phi coefficient
        ii. Tetrachoric correlation
3. If atleast one of them is ordinal:
    a. Spearman's Rho
4. If one variable is dichotomous and the other is continuous:
    a. Point Biserial Correlation

We can discuss how to calculate these correlations in some other article. Now let's look at another approach for feature selection.

1. The First approach is known as **Filter methods**. In this method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step. The advantage of using filter methods is that it needs low computational time and does not overfit the data.
   Some techniques under this approach are:
    a. **Information Gain**: It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Entropy is the measure of uncertainty or unpredictability. Information gain calculates the decrease in entropy. If a feature doesn't reduce the entropy of modeling our target dataset, then it isn't a useful feature and can be dropped.
    b. **Chi-square Test**: Chi-square test is a technique to determine the relationship between the categorical variables. The chi-square value is calculated between each feature and the target variable, and the desired number of features with the best chi-square value is selected.

c. **Fisher's score** is one of the popular supervised techniques of feature selection. It returns the rank of the variable on the fisher's criteria in descending order. Then we can select the variables with a large fisher's score.

2. The next approach is known as **Wrapper methods.** They are also referred to as greedy algorithms. Here the algorithm is trained by using a subset of features in an iterative manner. Based on the conclusions made from training prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved. The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive. Some techniques under this approach are:

a. **Forward selection** – This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
b. **Backward elimination** – This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
c. **Exhaustive selection** – This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model's performance is best.

3. Lastly, there is also an approach called Embedded methods. Here, the feature selection algorithm is blended as part of the learning algorithm, thus having its own built-in feature selection methods. Embedded methods encounter the drawbacks of filter and wrapper methods and merge their advantages. These methods are faster like those of filter methods and more accurate than the filter methods and take into consideration a combination of features as well. Some techniques under this approach are:

a. **Regularization** – This method adds a penalty to different parameters of the machine learning model to avoid overfitting of the model. This approach of feature selection uses Lasso (L1 regularization), Ridge (L2 regularization) and Elastic nets (L1 and L2 regularization). The penalty is applied over the coefficients, thus bringing down some coefficients to zero. The features having zero coefficient can be removed from the dataset. In L1 regularization, we create the penalty term by taking the sum of absolute values of the coefficients. In L2 regularization, the penalty term is created by taking a weighted sum of squares of the coefficients.
b. **Random Forest Importance** - Different tree-based methods of feature selection help us with feature importance to provide a way of selecting features. Here, feature importance specifies which feature has more importance in model building or has a great impact on the target variable. Random Forest is such a tree-based method, which is a type of bagging algorithm that aggregates a different number of decision trees. It automatically ranks the nodes by their performance or decrease in the impurity (Gini impurity) over all

the trees. Nodes are arranged as per the impurity values, and thus it allows pruning of trees below a specific node. The remaining nodes create a subset of the most important features.

So that’s a high level overview of Feature selection techniques in Machine Learning.