**FEATURE SELECTION**

Machine learning works on a simple rule - if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data. This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. I have myself witnessed feature subsets giving better results than complete set of feature for the same algorithm.

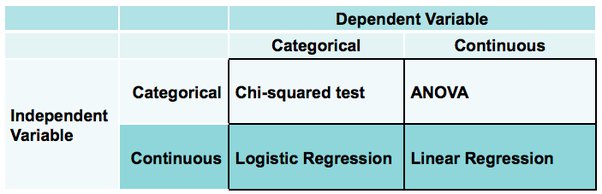
Top reasons to use feature selection are:

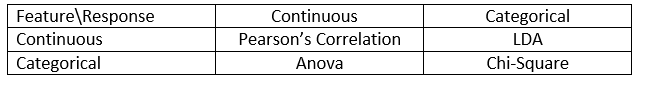
It enables the machine learning algorithm to train faster.

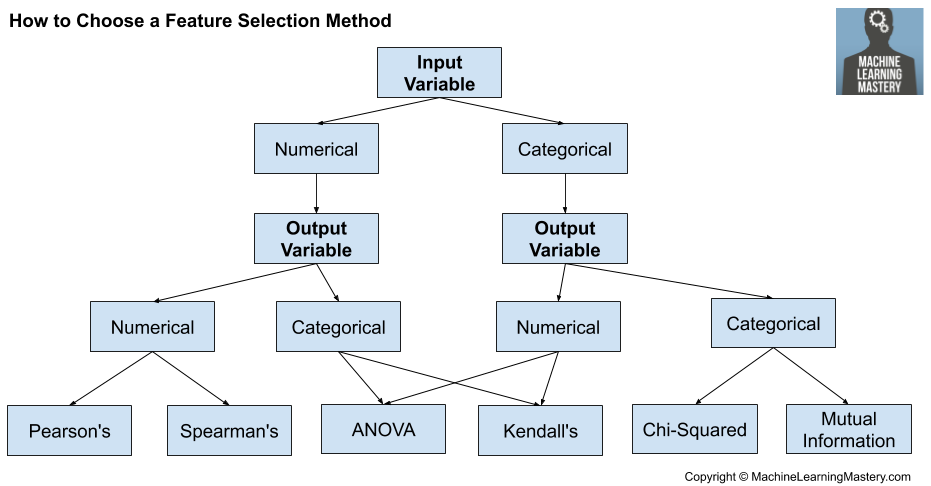
It improves the accuracy of a model

It reduces overfitting

**FILTER METHODS**

****





Numerical Input, Numerical Output

Pearson’s Correlation - It is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. The Pearson correlation coefficient (named for Karl Pearson) can be used to summarize the strength of the linear relationship between two data samples. The Pearson’s correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score.

Pearson's correlation coefficient = covariance (X, Y) / (stdv(X) \* stdv(Y))

where Covariance is one of the statistical measurement to know the relationship of the variance between the two variables. Cov (x, y) = SUM [(xi - xm) \* (yi - ym)] / (n - 1)

A value of 0 means no correlation. The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation

Spearman’s Rank Correlation - Two variables may be related by a nonlinear relationship, such that the relationship is stronger or weaker across the distribution of the variables. Further, the two variables being considered may have a non-Gaussian distribution.

In this case, the Spearman’s correlation coefficient (named for Charles Spearman) can be used to summarize the strength between the two data samples. This test of relationship can also be used if there is a linear relationship between the variables, but will have slightly less power (e.g. may result in lower coefficient scores).

Instead of calculating the coefficient using covariance and standard deviations on the samples themselves, these statistics are calculated from the relative rank of values on each sample. This is a common approach used in non-parametric statistics, e.g. statistical methods where we do not assume a distribution of the data such as Gaussian.

Spearman's correlation coefficient = covariance(rank(X), rank(Y)) / (stdv(rank(X)) \* stdv(rank(Y)))

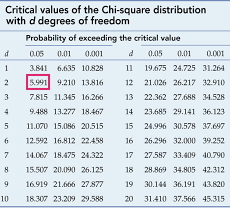
A linear relationship between the variables is not assumed, although a monotonic relationship is assumed. This is a mathematical name for an increasing or decreasing relationship between the two variables.

If you are unsure of the distribution and possible relationships between two variables, Spearman correlation coefficient is a good tool to use.

Categorical Input, Categorical Output

Pearson’s Chi-Square - The Pearson’s Chi-Square statistical hypothesis is a test for independence between categorical variables. The Chi-Squared test is a [statistical hypothesis test](https://machinelearningmastery.com/statistical-hypothesis-tests/) that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same. The test returns test statistic value (chi-square value), p value, degree of freedom, table of expected frequencies. This is calculated using a contingency table which is a table (also called crosstab) that used in statistics to summarise the relationship between several categorical variables.

After calculating chi square value, we need to find the critical value of chi-square and alpha (significance level). The alpha is chosen by user. We can obtain critical value of chi square from a table using Critical values of the Chi square distribution with d degrees of freedom table which takes the degree of freedom and gives the critical value of chi square. The degree of freedom is (no. of rows – 1) \* (no. of columns – 1). Now, let us look at the table and find the value corresponding to calculated degree of freedom and 0.05 significance factor



Here is a look on how test statistics value and calculated degree of freedom is calculated by contingency table.

Here is out dataset for the example.

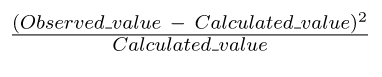
|  | dog | cat | bird | total |
| --- | --- | --- | --- | --- |
| men | 207 | 282 | 241 | 730 |
| women | 234 | 242 | 232 | 708 |
| total | 441 | 524 | 473 | 1438 |

Next, we prepare a similar table of expected values. To do this we need to calculate each item in the new table as



|  | Dog | Cat | bird | total |
| --- | --- | --- | --- | --- |
| men | 223.87343533 | 266.00834492 | 240.11821975 | 730 |
| women | 217.12656467 | 257.99165508 | 232.88178025 | 708 |
| total | 441 | 524 | 473 | 1438 |

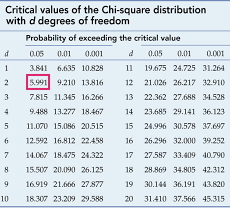
Next, we prepare a similar chi square table. To do this we need to calculate each item in the new table as



|  | observed (o) | calculated (c) | (o-c)^2 / c |
| --- | --- | --- | --- |
|  | 207 | 223.87343533 | 1.2717579435607573 |
|  | 282 | 266.00834492 | 0.9613722161954465 |
|  | 241 | 240.11821975 | 0.003238139990850831 |
|  | 234 | 217.12656467 | 1.3112758457617977 |
|  | 242 | 257.99165508 | 0.991245364156322 |
|  | 232 | 232.88178025 | 0.0033387601600580606 |
| Total |  |  | 4.542228269825232 |

From this table, the calculated value of chi-square is 4.542228269825232.

After we have got chi-square value, now we need to find the critical value of chi-square. The degree of freedom is (no. of rows – 1) \* (no. of columns – 1). Now, let us look at the table and find the value corresponding to degree of freedom 2 and 0.05 significance factor



For example, the tabular or critical value of chi-square for degrees of freedom 2 is 5.991.

The test can be interpreted as follows

Null Hypothesis: The features are independent (which means they are not associated).  
Alternate Hypothesis: The features are not independent (which means they are associated)

In terms of statistics and critical values

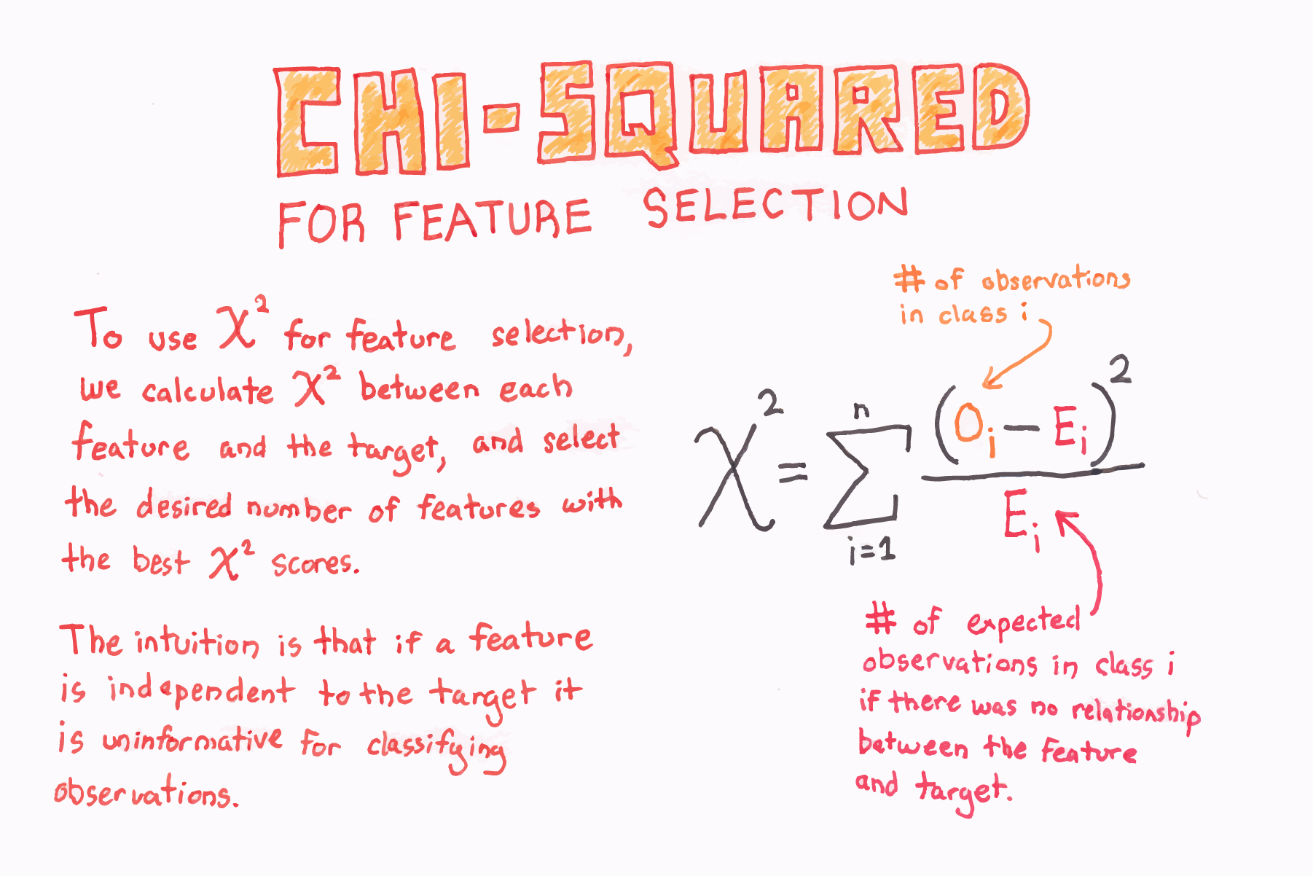
If test statistic value >= critical value of chi square: significant result, dependent, reject null hypothesis (H0)

If test statistic value < critical value of chi square: not significant result, independent, fail to reject null hypothesis (H0)

In terms of a p value and a chosen significance level (alpha), we define a significance factor to determine whether the relation between the variables is of considerable significance. Generally, a significance factor or alpha value of 0.05 is chosen. This alpha value denotes the probability of erroneously rejecting H0 when it is true. A lower alpha value is chosen in cases where we expect more precision. If the p-value for the test comes out to be strictly greater than the alpha value, then H0 holds true

If p value <= alpha: significant result, dependent, reject null hypothesis (H0)

If p value > alpha: not significant result, independent, fail to reject null hypothesis (H0)



Numerical Input, Categorical Output

ANOVA - ANOVA stands for Analysis of variance. It provides a statistical test of whether the means of several groups are equal or not. An Analysis of Variance Test, or ANOVA, can be thought of as a generalization of the t-tests for more than 2 groups. The independent t-test is used to compare the means of a condition between two groups. ANOVA is used when we want to compare the means of a condition between more than two groups.

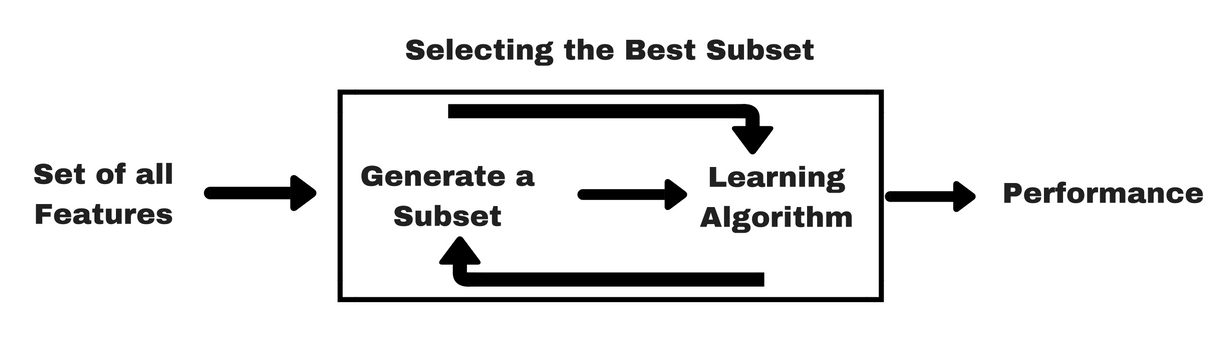
An [F-statistic](https://en.wikipedia.org/wiki/F-test), or F-test, is a class of statistical tests that calculate the ratio between variances values, such as the variance from two different samples or the explained and unexplained variance by a statistical test, like ANOVA. The ANOVA method is a type of F-statistic referred to here as an ANOVA f-test.

For checking of correlation between continuous input and categorical output we can use logistics regression as well.

Categorical Input, Numerical Output

ANOVA

**WRAPPER METHODS**



F\_regression: Used only for numeric targets and based on linear regression performance

Chi2: Performs the chi-square statistic for categorical targets which is less sensible to the nonlinear relationship between the predictive variable and its target.

F\_classify: Used only for categorical targets and based on the Analysis of Variance (ANOVA) statistical test.

Mutual exclusion: Used when target variable is categorical. Mutual information from the field of information theory is the application of information gain to feature selection. Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. Mutual information is straightforward when considering the distribution of two discrete (categorical or ordinal) variables such as categorical input and categorical output data. Nevertheless, it can be adapted for use with numerical input and categorical output

Forward Selection: Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

Backward Elimination: In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

Recursive Feature elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination. RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains

Boruta Package: Boruta by default uses random forest although it works with other algorithms like LightGBM, XGBoost etc. It selects the features that actually have relationship with outcome variable. The algorithm reshuffles the data to create shadow features. It eliminates the features that have significantly worst importance than shadow ones. There are few limitations in the approach. It is required to remove multicollinearity and missing values from data before running Boruta. It can be used for both regression and classification problems

**INTRINSIC METHODS**

Embedded methods use algorithms that have built-in feature selection methods. For example: Decision Trees and RF have their own feature selection methods.