**Why feature selection techniques are important.**

When building a machine learning model in real-life, it’s almost rare that all the variables in the dataset are useful to build a model. Adding redundant variables reduces the generalization capability of the model and may also reduce the overall accuracy of a classifier. Furthermore adding more and more variables to a model increases the overall complexity of the model.[1]

Increasing variables(features) may build a complex model which overfit the data.

**Goal**

The goal of feature selection in machine learning is to find the best set of features that allows one to build useful models of studied phenomena.

The techniques for feature selection in machine learning can be broadly classified into the following categories:

**Supervised Techniques:** These techniques can be used for labeled data, and are used to identify the relevant features for increasing the efficiency of supervised models like classification and regression.

Correlation coefficient

Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves.

If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information. We will use the Pearson Correlation here.

Feature importance and weight determination  
<https://machinelearningmastery.com/calculate-feature-importance-with-python/>

Detecting Multicollinearity with VIF – Python

* Last Updated : 29 Aug, 2020

Multicollinearity occurs when there are two or more independent variables in a multiple regression model, which have a high correlation among themselves. When some features are highly correlated, we might have difficulty in distinguishing between their individual effects on the dependent variable. Multicollinearity can be detected using various techniques, one such technique being the **Variance Inflation Factor**(**VIF**).

In VIF method, we pick each feature and regress it against all of the other features. For each regression, the factor is calculated as :

Where, [R-squared](https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/) is the coefficient of determination in linear regression. Its value lies between 0 and 1. [2]

A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

[https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis#:~:text=If%20the%20VIF%20is%20equal,predictors%20may%20be%20moderately%20correlated.&text=And%20if%20the%20VIF%20goes,poorly%20estimated%20due%20to%20multicollinearity.]

Feature selection using lasso regularization[3]\

We don’t have class imbalanced problem, because dataset includes 4897 phishing websites and 6158 legitimate websites. Almost 50% for each class

**mean absolute difference (MAD)**

‘The mean absolute difference (MAD) computes the absolute difference from the mean value. The main difference between the variance and MAD measures is the absence of the square in the latter. The MAD, like the variance, is also a scale variant. This means that higher the MAD, higher the discriminatory power.

<https://www.sciencedirect.com/science/article/abs/pii/S0167865512001870>

ROC curves and precision-recall curves

* ROC curves are appropriate when the observations are balanced between each class, whereas precision-recall curves are appropriate for imbalanced datasets.

<https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>

[pandas](https://pandas.pydata.org/)— Allows easy manipulation of data structures.

[numpy](http://www.numpy.org/)— Allows scientific computing.

[sklearn](https://scikit-learn.org/stable/index.html)— A machine learning library for python.

**Future works**

To integrate all the feature extraction codes and build a website for facilitating the phishing detection

**References**

[1] <https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>

[2] <https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/>

[3] <https://towardsdatascience.com/feature-selection-using-regularisation-a3678b71e499>