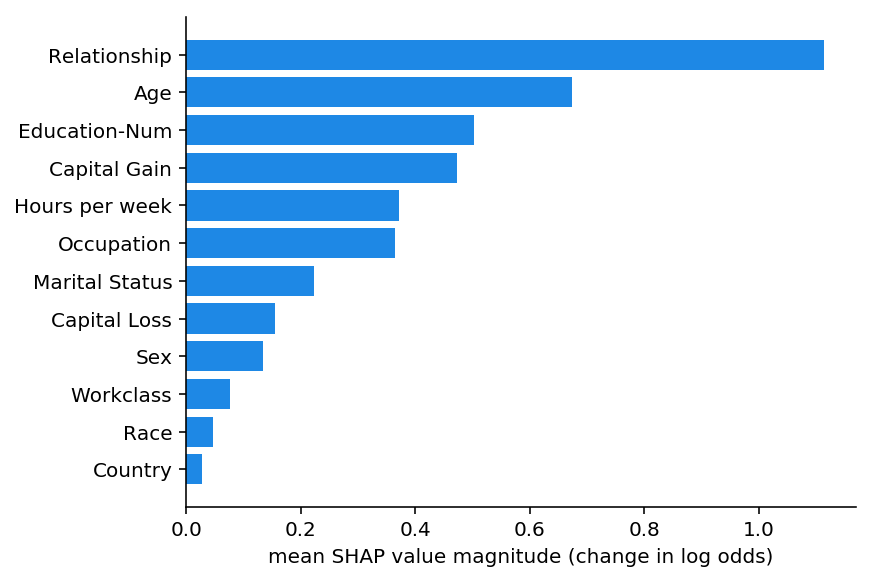
The documentation is taken from this link:

<https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernelexplainer-79de9464897a>

1. SHAP (Shapley Additive explanations) values which offer a high level of interpretability for a model. The SHAP values provide the *Global interpretability* — the SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable. This is like the variable importance plot, but it can show the positive or negative relationship for each variable with the target (see the summary plots below).
2. *Local interpretability* — each observation gets its own set of SHAP values (see the individual force plots below). This greatly increases its transparency. We can explain why a case receives its prediction and the contributions of the predictors. Traditional variable importance algorithms only show the results across the entire population but not on each individual case. The local interpretability enables us to pinpoint and contrast the impacts of the factors.

We get a graph with the SHAP values for all the features like the one shown below. The features which are the important ones comes at the top.



**How to interpret Shap results:**

We also get two more plots with SHAP to understand the positive effect and the negative effect on our target variable. The same is explained below.

Note: The example used in the above and below graphs are for different dataset.

1. The summary plot

This plot has loaded information. The biggest difference of this plot with the regular variable importance plot (Figure A) is that it shows the positive and negative relationships of the predictors with the target variable. It looks dotty because it is made of all the dots in the train data. Let me walk you through:

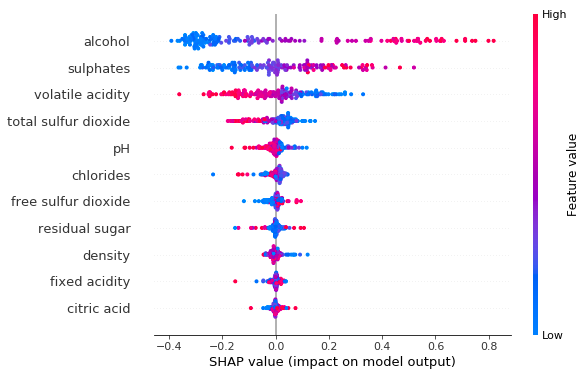
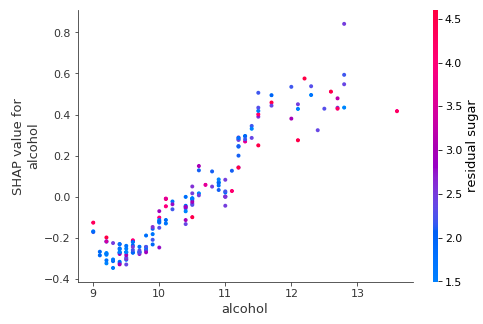


Figure A

* *Feature importance:* Variables are ranked in descending order.
* *Impact:* The horizontal location shows whether the effect of that value *is associated with a higher or lower prediction*.
* *Original value:* Colour shows whether that variable is high (in red) or low (in blue) for that observation.
* *Correlation:* A *high* level of the “alcohol” content has a high and *positive* impact on the quality rating. The “high” comes from the red colour, and the “positive” impact is shown on the X-axis. Similarly, we will say “volatile acidity” is negatively correlated with the target variable.

2. The dependence plot

It shows the marginal effect that one or two variables have on the predicted outcome. It tells whether the relationship between the target and the variable is linear, monotonic or more complex. Suppose we want to get the dependence plot of “alcohol”. The Python module SHAP includes automatically another variable that “alcohol” interacts most with. The following plot shows that there is an approximately linear and positive trend between “alcohol” and the target variable, and “alcohol” interacts with “residual sugar” frequently.



These two links also additionally explain the same concepts we can refer to:

<https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27>

<https://github.com/slundberg/shap>