**Multi Collinearity**

This is an issue that is sometime ignored within machine learning, but needs to be considered, as it can have unintended consequences.

What is it?

* Multi collinearity is when you have multiple variables that are correlated to each other

How can you spot it?

* Inspecting the data with a simple heatmap (seaborn)
* Use the VIF (variance Inflation Factors) measures for each predictor
  + This tells us how much the value of a coefficient is inflated because of linear dependence with other predictors
  + Eg. “a VIF of 1.8 tells us that the variance (the square of the standard error) of a particular coefficient is 80% larger than it would be if that predictor was completely uncorrelated with all the other predictors.”
  + NB. VIF values start at 1 and have no upper bound. Some folks are cautious about VIFs over 5, some people over 10.

Why is it a problem?

* The aim of a regression is to isolate the the relationship between each IV and the DV
* The regression coefficients are supposed to represent the mean change in the dependent variable for each one-unit change in an independent variable, holding the other variables constant.
* But, if the independent variables are highly correlated, this isn’t possible.
* The model can’t estimate each independent variable’s effect individually.
* This causes the coefficient estimates to be inconsistent and highly sensitive to small changes in the model.
* If we cannot trust our coefficient estimates, causal inference is also no longer possible if the collinear variables are the ones we’re actually interested in, as we can’t see which of the collinear predictors is responsible for the prediction and the causal effect of interest.

Why is it sometimes ignored in ML?

* While multicollinearity affects coefficients and p-values, it does not influence the quality of predictions.
* In ML predictive performance is typically king, so in such cases where interpreting the coefficients is not important, neither is multicollinearity.
* Also, many ML algorithms overcome the problem of multicollinearity on their own naturally.
  + For example, decision trees will simply randomly choose one only of the collinear features to use. Regularization can also help overcome this problem.

Interpretable ML

* A prediction alone is not sufficient for creating actionable analyses in many cases.
* Multicollinearity can make interpreting the features challenging, even when using models that are robust for prediction.

What are the ways to deal with multi collinearity?

* Feature engineer a new feature that combines the separate collinear variables into one
* Use PCA
* Simply drop the variable that is less theoretically important, or less strongly correlated with the target variable