interpreting features

Selecting the best features is essential to the optimal performance of machine learning models. Only features that contribute to measurably improving model performance should be used.

Using extraneous features can contribute noise to training and predictions from machine learning models. This behavior can prevent machine learning models from generalizing from training data to data received in production.

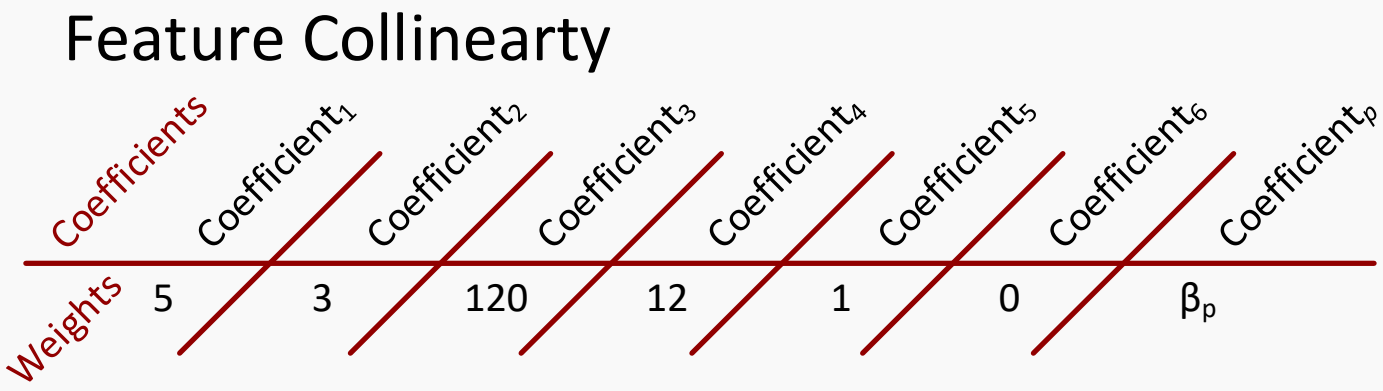
Collinear Features are features in which are highly correlated amongst themselves; they are essentially the same and thus will have comparable predictive power.

> **Collinear Features** are often the product of data scientists engineering new features to learn more information for a given topic or property of a dataset or domain.

However, **Collinear Features** often distort the interpretation of model parameterization (coefficients).

It is important to note that coefficients are not relevant to measuring their importance in a model.

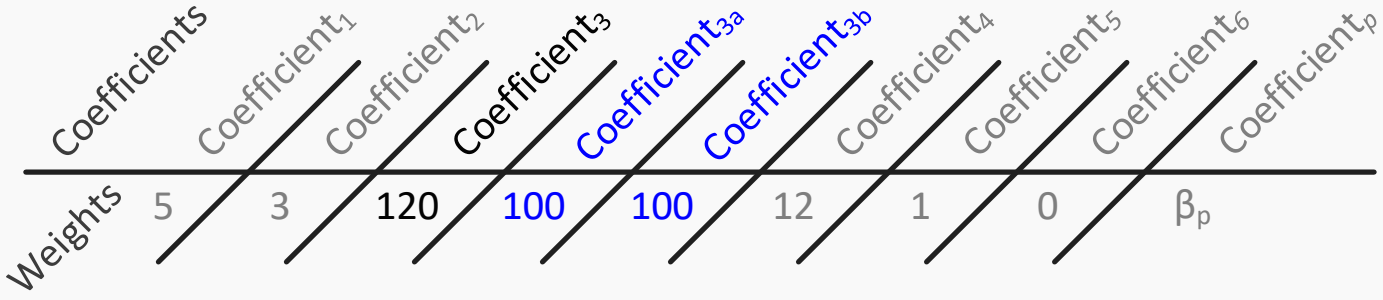
The following hypothetical illustrates **Colinearity** and Feature Selection:



Given the above algorithm parameters $\text{Coefficient}_1, \ldots, \text{Coefficient}_p$ and either **Regularization Term**:

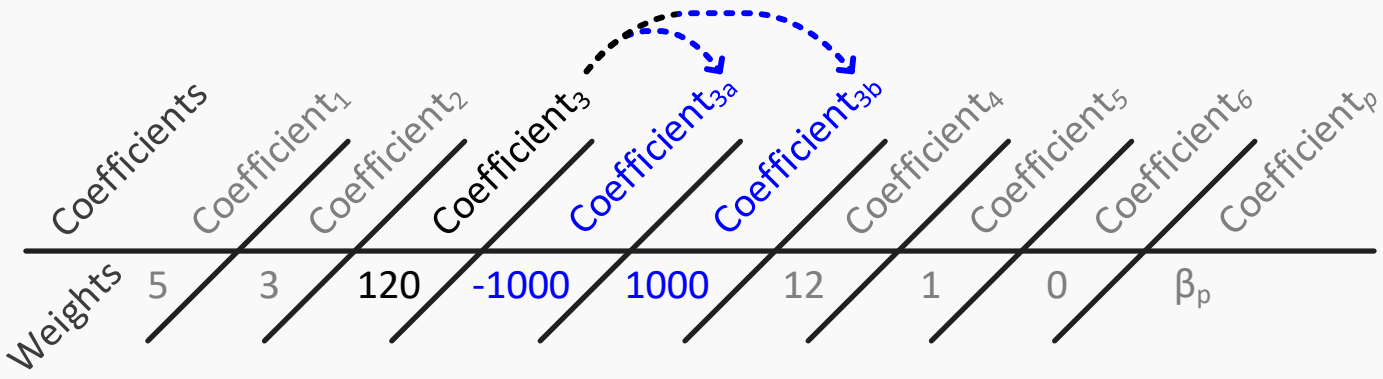
$$\text{Regularization}(f) = \beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots + \beta_p^2 = \|\beta\|_2^2 \text{ (referred to as } \ell_2)$$

$$\text{Regularization}(f) = |\beta_1^2| + |\beta_2^2| + |\beta_3^2| + \cdots + |\beta_p^2| = \|\beta\|_1 \text{ (referred to as } \ell_1)$$



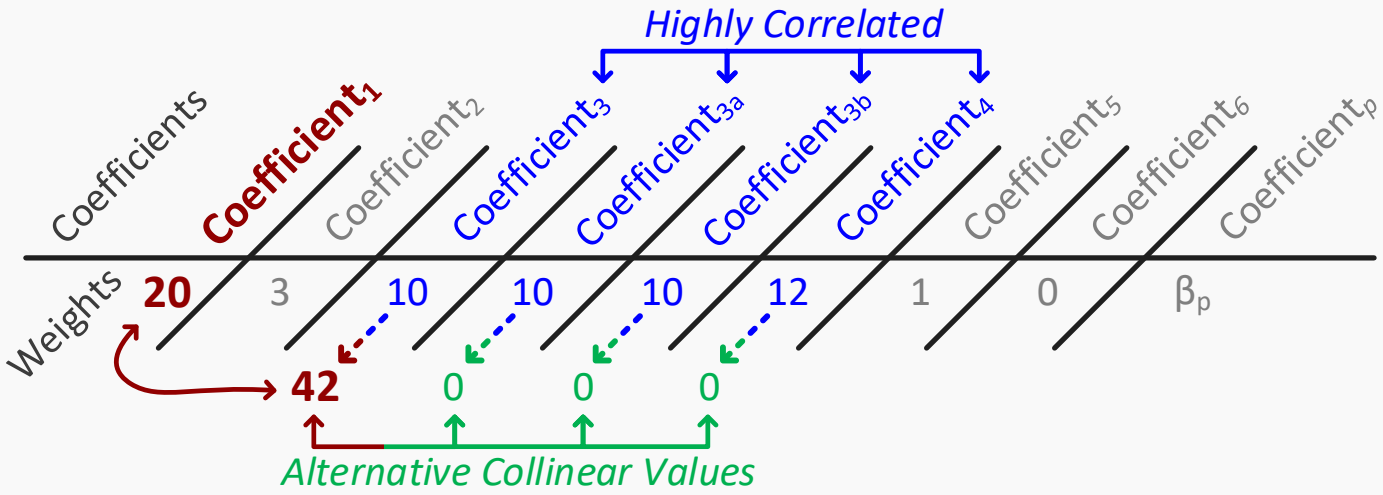
Assuming $\textbf{Coefficient}_3$ is determined to be valuable, a data scientist might derive multiple features (feature engineering) from the original as seen above Coefficient_{3a} and Coefficient_{3b}. The goal is to

achieve a higher rate of predictive power. However, the derived features are **highly correlated** amongst themselves and might lack independent predictive power on the model when evaluated.



Assuming Coefficient_{3a} and Coefficient_{3b} are assigned weights of -1000 and 1000 respectively. It might appear that the engineered features are highly valuable to the model. However this is a **learning fallacy** considering the two features effectively net to 0 and cancel each other out.

Again, **Coefficients** are **NOT** related to **feature importance** in a predictive model.



Assuming the above **Coefficient Weights**, it appears that the **four blue features** are **highly correlated**. Therefore, the most valuable feature with the given weights appears to be $\mathbf{Coefficient_1}$ with a given weight of **20**. However, because the **four blue features** are **highly correlated** with each other, they could have equally been weighted with the given **Alternative Values** instead. In the latter scenario, $\mathbf{Coefficient_3}$ appears to hold the highest value in predicting new labels from a dataset.