Pre-processing Overview

* Refers to the additional, deletion or transformation of **training set data**.
* Can have a huge impact on the accuracy of models.
* Different models have different sensivities to the type of predictors in the model.
* How predictors enter the model is important.
* **Actually Use Them**: You are a step ahead if you are thinking about and using data transforms to prepare your data. It is an easy step to forget or skip over and often has a huge impact on the accuracy of your ﬁnal models.
* **Use a Variety**: Try a number of diﬀerent data transforms on your data with a suite of diﬀerent machine learning algorithms.
* **Review a Summary**: It is a good idea to summarize your data before and after applying a transform to understand the eﬀect it had. The summary() function can be very useful.
* **Visualize Data**: It is also a good idea to visualize the distribution of your data before and after to get a spatial intuition for the eﬀect of the transform.

## Data Pre-Processing Methods

It is hard to know which data pre-processing methods to use.

You can use rules of thumb such as:

* Instance-based methods are more eﬀective if the input attributes have the same scale.
* Regression methods can work better if the input attributes are standardized.

These are heuristics, not hard and fast laws of machine learning, because sometimes you can get better results if you ignore them. You should try a range of data transforms with a suite of diﬀerent machine learning algorithms. This will help you discover both good representations for your data and algorithms that are better at exploiting the structure that those representations expose.

It is a good idea to spot-check a number of transforms both in isolation as well as combinations of transforms. In the next section you will discover how you can apply data transforms in order to prepare your data in R using the caret package.

## Summary of Transform Methods

Below is a quick summary of all of the transform methods supported in the argument to the preProcess() function in caret.

* **BoxCox**: apply a Box-Cox transform, values must be non-zero and positive.
* **YeoJohnson**: apply a Yeo-Johnson transform, like a BoxCox, but values can be negative.
* **expoTrans**: apply a power transform like BoxCox and YeoJohnson.
* **zv**: remove attributes with a zero variance (all the same value).
* **nzv**: remove attributes with a near zero variance (close to the same value). + **center**: subtract mean from values.
* **scale**: divide values by standard deviation.
* **range**: normalize values.
* **pca**: transform data to the principal components.
* **ica**: transform data to the independent components.
* **spatialSign**: project data onto a unit circle.