Feature Transformation

Inspecting the distributions of the features is a very important step particularly when we are using parametric statistical tests such as t-test, ANOVA or Linear Regression. If the features are not normally distributed, have a high Skewness then using the data as is may produce misleading results.

When the distribution of the continuous data is non-normal, transformations of data are applied to make the data as "normal" as possible and, thus, increase the validity of the associated statistical analyses. Transformation is the replacement of a variable by a function of that variable: for example, replacing a variable x by the logarithm of x (log-transform) or square root of x. The core idea being, if the predictor variable doesn't show linear relationship with the target variable, or doesn't have a normal distribution then make appropriate transformations so the a linear relationship is achieved, and the distribution becomes near normal.

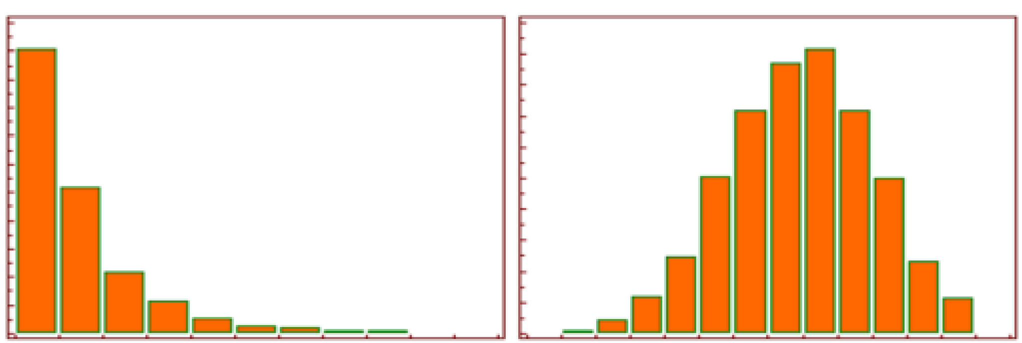
Like most decisions in Machine Learning, the decision to choose an appropriate transformation also involves a "trial and error" approach. Let us now look at the most commonly used transformations.

Types of Feature Transformations-

1. The Logarithmic Transformation –

   The log transformation is the most commonly used method which involves taking the natural algorithm (or a different base) of each data point.

   If the original data follows a log-normal distribution or approximately so, or has a heavy right-skewness, then the log-transformed data follows a normal or near normal distribution. In this case, the log-transformation does remove or reduce skewness. The log transformation is particularly relevant when the data vary a lot on the relative scale.



a) Original Data Distribution b) Post Log Normal Transformation

From the above graph, we see that taking log has two major effects -

I. Small values that were close together are spread further out
II. Large values that were spread out are brought closer together

The combined effect of these is that the distribution is transformed to be a near-normal distribution.

Note - If you have zeros or negative numbers, you can't take the log; you should add a constant to each number to make them positive and non-zero. (Because natural logarithm function ln(x) is defined only for x>0)

2. Cube Root Transformation -

   The cube root, x to x^(1/3). This is a fairly strong transformation with a substantial effect on distribution shape. It is also used for reducing right skewness, and has the advantage that it can be applied to zero and negative values. Rest of the properties are very similar to the Log-transformation.

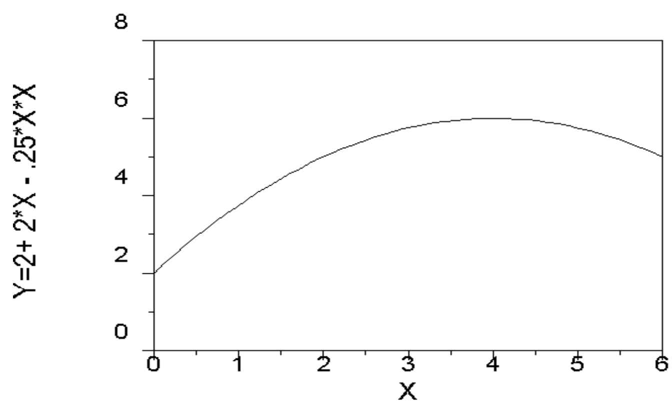
3. Square Root Transformation –
   The square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape: it is weaker than the logarithm and the cube root (the explanation is beyond the scope of this document). It is also used for reducing right skewness, and has the advantage that it can be applied to zero values. Rest of the properties are very similar to the Log-transformation.

4. Square and Cube Transformations –
   Unlike the previous transformations, the Square and Cube transformations are used to reduce the left skewness of the data. It involves taking the square of x, or the cube of x respectively.
   Squaring usually makes sense only if the variable concerned is zero or positive, given that (-x)^2 and x^2 are identical.

Choosing the appropriate transformation –

Following are some tips that can help in arriving at best possible transformation for your case study.

1. Inspect the skewness of the feature, apply Log, Square root or Cube root for removing Right-skewness. Square, Cube or exponential to remove the left skewness.
2. Be wary of the range of the distribution, certain transformations have restrictions as to the range that they work on.
3. Inspect the scatter plot of feature with the target variable, and use the transformation that closely describes the relationship between them. For example, a curvilinear relationship can be converted to a linear relationship by using Square transformation.



Example of Curvilinear relationship between x & y (Quadratic relationship)

4. Prefer the transformations that lead to simpler or interpretable results, for example choose square root transformations when dealing with Area feature, cube root with Volume feature.
5. Refer the literature/research papers to understand transformations performed on similar features and their effect on the results.

Inverse Transformation –

Often after the modelling, we need to perform inverse transformation to achieve interpretability.

Following are the set of transformation pairs, which can be used to transform and then inverse-transform the features -

Log base 10	10 to the power

t = log_10 x	x = 10^t
Log base e t = log_e x = ln x	**e to the power** x = exp(t)
Cube root t = x^(1/3)	**Cube** x = t^3
Square root t = x^(1/2)	**Square** x = t^2

Python Codes –

Unlike feature scaling (z-score , min-max etc.) transformation usually doesn't require the machine to learn parameters of the distribution, thus it can be applied both before and after the sampling (Train test split).

1. Logarithmic: data['log_ref_age'] = np.log(data['referral_age'])
2. Exponential: data['exp_ref_age'] = np.exp(data['referral_age'])
3. Square root: data['sqrt_ref_age'] = np.sqrt(data['referral_age'])
4. Cube root: data['cbrt_ref_age'] = np.cbrt(data['referral_age'])
5. Square: data['sqr_ref_age'] = np.square(data['referral_age'])
6. Cube: data['cube_ref_age'] = np.power(data['referral_age'],3)

References –

1. Penn State study material
2. Data Preprocessing in Data Mining By - Salvador García , Julián Luengo , Francisco Herrera