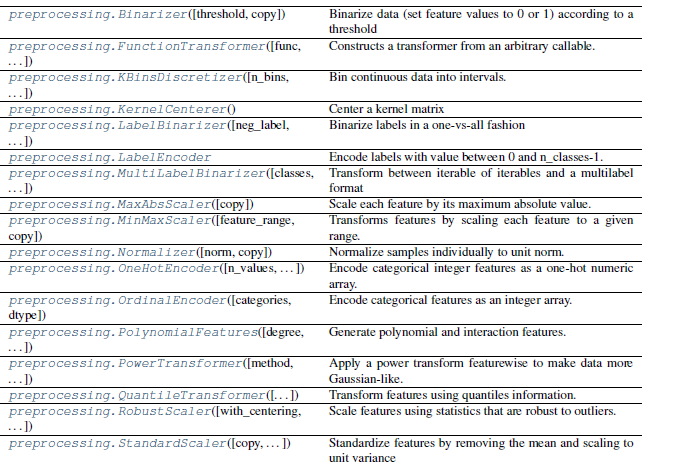
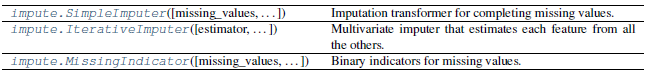
****

****

**Variables**

1. Numerical
2. Categorical
   1. Ordinal - Categorical variable in which categories can be meaningfully ordered are called ordinal.
   2. Nominal - There isn't an intrinsic order of the labels.
3. Date and Time

**Types of Categorical Encoding**

1. **Nominal Encoding –** No need of arranging the categories.

Eg: Gender(M,F)

1. **Ordinal Encoding –** Categories can be rearranged according to their rank. Eg:Degrees(PHD,M.Tech,Schooling,B.Tech)

**Nominal Encoding**

1. One Hot Encoding
2. One Hot Encoding with many categorical
3. Mean Encoding
4. Count Or Frequency Encoding
5. Probability Ratio Encoding

**Ordinal Recording**

1. Label Encoder / OrdinalEncoder
2. Target guided Ordinal Encoding

OrdinalEncoder for 2D data; shape (n\_samples, n\_features), LabelEncoder is for 1D data: for shape (n\_samples,)

**One Hot Encoding**

* It is applicable to nominal data values.
* It will have number of columns = number of unique categories – 1(to avoid dummy variable trap subtracting one column)
* Disadvantages:

If number of unique categories is more, then it will have more number of columns and it will lead to curse of dimensionality.

**One Hot Encoding with many categorical**

From all the unique categories we need to find how many top categories are having max. no. of frequency. And apply one hot encoding to all those top occurring categories..

**Target guided Ordinal Encoding**

Ordering the labels according to the target means assigning a number to the label, but this numbering, this ordering, is informed by the mean of the target within the label.

Briefly, we calculate the mean of the target for each label/category, then we order the labels according to these mean from smallest to biggest, and we number them accordingly.

**Mean Encoding**

Replacing labels by the risk factor means essentially replacing the label by the mean of the target for that label.

I have only seen this procedure applied in classifications scenarios, where the target can take just the values of 1 or 0. However, in principle, I don't see why this shouldn't be possible as well when the target is continuous. Just be mindful of over-fitting.

## Count or frequency encoding

## Another way to refer to variables that have a multitude of categories, is to call them variables with ****high cardinality****.

One approach, is to **replace each label of the categorical variable by the count, this is the amount of times each label appears in the dataset**. Or the frequency, this is the percentage of observations within that category. The 2 are equivalent.

There is not any rationale behind this transformation, other than its simplicity.

### Advantages

* Simple
* Does not expand the feature space

Disadvantages

* If 2 labels appear the same amount of times in the dataset, that is, contain the same number of observations, they will be merged: may loose valuable information
* Adds somewhat arbitrary numbers, and therefore weights to the different labels, that may not be related to their predictive power

**Probability ratio encoding**

For each label, we calculate the mean of target=1, that is the probability of being 1 ( P(1) ), and also the probability of the target=0 ( P(0) ). And then, we calculate the ratio P(1)/P(0), and replace the labels by that ratio.

The calculation of the probability ratios to replace the labels should be done considering ONLY on the training set, and then expanded it to the test set.

The methods discussed in this can be also used on numerical variables, after discretisation. This creates a monotonic relationship between the numerical variable and the target, and therefore improves the performance of linear models

### Advantages

* Capture information within the label, therefore rendering more predictive features
* Creates a monotonic relationship between the variable and the target
* Does not expand the feature space

### Disadvantage

* Prone to cause over-fitting

## Weight of evidence

Weight of Evidence (WoE) was developed primarily for the credit and financial industries to help build more predictive models to evaluate the risk of loan default. That is, to predict how likely the money lent to a person or institution is to be lost. Thus, Weight of Evidence is a measure of the "strength” of a grouping technique to separate good and bad risk (default).

It is computed from the basic odds ratio: ln( (Proportion of Good Credit Outcomes) / (Proportion of Bad Credit Outcomes))

WoE will be 0 if the P(Goods) / P(Bads) = 1. That is, if the outcome is random for that group. If P(Bads) > P(Goods) the odds ratio will be < 1 and the WoE will be < 0; if, on the other hand, P(Goods) > P(Bads) in a group, then WoE > 0.

WoE is well suited for Logistic Regression, because the Logit transformation is simply the log of the odds, i.e., ln(P(Goods)/P(Bads)). Therefore, by using WoE-coded predictors in logistic regression, the predictors are all prepared and coded to the same scale, and the parameters in the linear logistic regression equation can be directly compared.

The WoE transformation has three advantages:

* It establishes a monotonic relationship to the dependent variable.
* It orders the categories on a "logistic" scale which is natural for logistic regression
* The transformed variables, can then be compared because they are on the same scale. Therefore, it is possible to determine which one is more predictive.

The WoE also has three drawbacks:

* May incur in loss of information (variation) due to binning to few categories (we will discuss this further in the discretisation section)
* It does not take into account correlation between independent variables
* Prone to cause over-fitting