

## **SMOTE - Oversampling**

```
def smote( data, y, k = 5, samp_method = "balance", drop_na_col = True,  
drop_na_row = True, rel_thres = 0.5, rel_method = "auto", rel_xtrm_type = "both",  
rel_coef = 1.5, rel_ctrl_pts_rg = None):
```

SMOTE represents Synthetic Minority Over-sampling TEchnique and oversamples the minority class.

Function designed to help solve the problem of imbalanced data for regression.

### **Parameters:**

#### **main arguments / inputs:**

data: pandas dataframe, the training set.

y: string, response variable y by name. It should be a header name found in the dataframe data.

k: int (positive), default = 5, number of neighbors for over-sampling.

samp\_method: {'balance', 'extreme'}, default = 'balance', specified method to determine over sampling percentage.

drop\_na\_col: bool, default = 'True', if 'True', auto drop columns with NaN's.

drop\_na\_row: bool, default = 'True', if 'True', auto drop rows with NaN's.

#### **phi relevance function arguments / inputs:**

rel\_thres: float, positive real number, default = 0.5, define the relevance threshold considered rare in phi relevance function.

rel\_method: {'auto', 'manual'}, default = 'auto', the relevance method in phi relevance function.

rel\_xtrm\_type: {'low', 'high', 'both'}, default = 'both', distribution focus on high, low or both.

rel\_coef: float, positive real number, default = 1.5, coefficient for box plot in phi relevance function to consider rare.

rel\_ctrl\_pts\_rg: 2d array, default = None, when rel\_method = 'manual', it inputs for "manual" rel method.

## References:

N. V. Chawla, K. W. Bowyer, L. O'Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, 321-357, 2002.

<https://arxiv.org/pdf/1106.1813.pdf>.

Branco, P., Torgo, L., Ribeiro, R. (2017). SMOGN: A Pre-Processing Approach for Imbalanced Regression. Proceedings of Machine Learning Research, 74:36-50.

<http://proceedings.mlr.press/v74/branco17a/branco17a.pdf>.

Kunz, N. (2019). SMOGN: Synthetic Minority Over-Sampling for Regression with Gaussian Noise (Version 0.1.0). Python Package

Index. <https://pypi.org/project/smogn>.