End of Distribution Imputation

Subject: **Data Mining**
Lecturers: **Phauk Sokkhey** & **Chan Sophal**
Presenter: **Choeng Veyseng**

Table of Content

Definition

Which variable is fit for this method?

How to use it?

Code Implementation

Advantages & Disadvantages

Definition

End of Distribution Imputation

is equivalent to arbitrary value imputation, but it automatically selecting arbitrary values at the end of the variable distribution known as outlier.

If the variable is normally distributed, can use the mean plus or minus 3 times the standard deviation(SD)

If the variable is skewed, can use the IQR(Inter-Quantile Range) proximity rule.

Which variable is fit for this method?

Suitable numerical variables

How to use it?

Skew Distributions

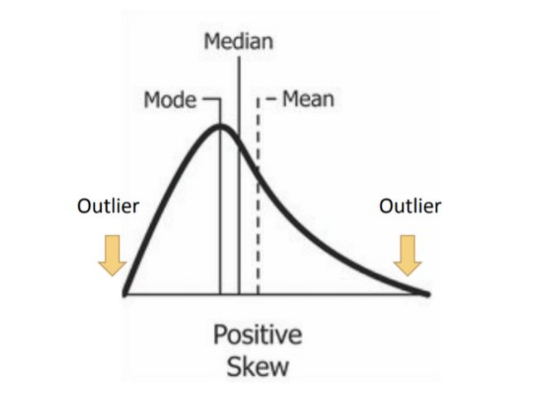
The general approach is to calculate the quantile, and then the inter-quantile range(IQR)

IQR = 75th Quantile - 25th Quantile

Upper limit = 75th Quantile + IQR x 1.5

lower limit = 25th Quantile - IQR x 1.5

Extreme outliers will time 3 instead of 1.5



Code Implementation

Advantages & Disadvantages

Advantages

Easy to implement
Rapid way of obtaining complete dataset
Can be integrated into production
Capture the importance of “Missingness” if there is one

Disadvantages

Distortion of the original variable distribution
Distortion of the original variance
Distortion of the covariance with the remaining variables of dataset
This technique may mask true outliers in the distribution

References:

https://medium.com/analytics-vidhya/feature-engineering-part-1-end-of-tail-imputation-c5069a41869a

https://www.kaggle.com/code/rushikeshlavate/end-of-distribution-imputation/notebook