Handling Missing Values using KNN Imputation

KNN (K-Nearest Neighbors) Imputation is a technique used to handle the missing values in a dataset. It identifies the k nearest neighbors (most similar data points) from the given missing value record and use the value of nearest neighbors to impute the missing observations.

How it works:

It identifies the k nearest neighbors based on the distance metric from the observation which is called nan-euclidean distance

Calculating nan_euclidean distance

dist(x,y) = sqrt(weight * sq. distance from present coordinates) where, weight = Total # of coordinates / # of present coordinates

For example, the distance between [3, na, na, 6] and [1, na, 4, 5] is:

$$sqrt(\ 4/3((3-1)^2 + (6-5)^2)))$$

Example taken from sklearn documentation

Missing Value Imputation

Each sample's missing values are imputed using the value from *n-neighbors* nearest neighbors found in the training set either by giving equal weights to the distances or inverse distance weights.

Key Parameters:

1) Number of neighbouring samples to use for imputation.
2) Weight function tells us how the calculated distance is treated, two possible values – uniform and distance
3) In unform weights all the distance are given equal weightages
4) In the distance parameter – it gives the weight points by the inverse of their distances i.e. close neighbours will have higher impact on the imputed value

How it impacts the data distribution? Mean Imputation Vs KNN Imputations

1) In Mean Imputation, all missing values of the column are replaced with mean. This can lead to a spike at the mean value, distorting the natural distribution of the feature
2) In KNN Imputation missing values are filled with similar data points nearer to the missing values so it avoids the spike in the mean area and distribution is better preserved (relatively smooth)
3) Mean Imputation ignore outliers and relationship with other feature which gets accounted in KNN based imputation
4) Variance gets reduced with the mean imputation, since all the missing values are replaced by a single mean value. There is a comparatively lesser decrease in variance in KNN Imputation compared to mean imputation so spread is likely to be preserved

I have used the Titanic dataset to Analyze how the distribution of the "Age" feature is impacted when applying Mean Imputation versus KNN Imputation. Additionally, I explored how varying parameters in KNN Imputation influence the accuracy score of the predictive model.

Disadvantage:

In production we need to keep all the training dataset which consumes high space and increases processing time with the huge dataset. https://tinyurl.com/knnimputer