Univariate Imputation- Random Value (Num + Cat Data)

09 July 2023 17:26

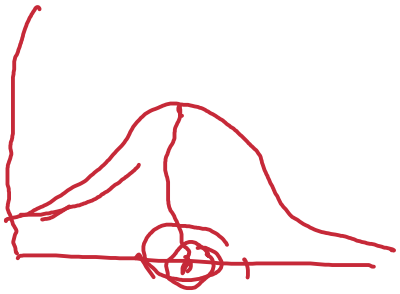
This is a **very simple technique** which includes **filling each missing value by randomly selecting value** from the existing non-missing values within the **same feature**.

Here, the process of selecting a random value is done for filling each missing value.

It is a **simple** imputation method that helps **preserve the statistical properties** of the original dataset.

This technique is used a lot and can be applied to **both Numerical and Categorical Data**.

ID	Gender	Age
emp1	M	**? = 36**
emp2	**? = F**	22
emp3	F	27
emp4	M	34
emp5	M	52
emp6	F	19
emp7	**? = M**	36
emp8	F	**? = 52**
emp9	M	47
emp10	F	43
emp11	**? = F**	23
emp12	M	**? = 19**
emp13	F	31
emp14	M	24
emp15	M	**? = 19 – Again Possible**



This technique is **good for Linear Algorithms** as the distribution does not change as such.
But this is **not good** in case of **Tree Based Algorithms** like **Decision Tree, Random Forest** etc. because of additional Randomness and **Increased Similarity among Observations** which in turn affects the unbiased splitting of data based on Original Information during model training.

There is **no class available in SkLearn** for this technique, we have to carry out this using **Pandas.**

Advantages:

a. **Easy** to implement.
b. **Preserves the Variance** because the randomly selected values are those with higher probability of getting selected

(like Values around the mean which are not significantly contributing in the Variance). So these value doesn't change the Variance.

Disadvantages:

a. This **changes the Covariance**.
b. This is **memory heavy** because we have to store the training data on the server to select random value while model is in production.