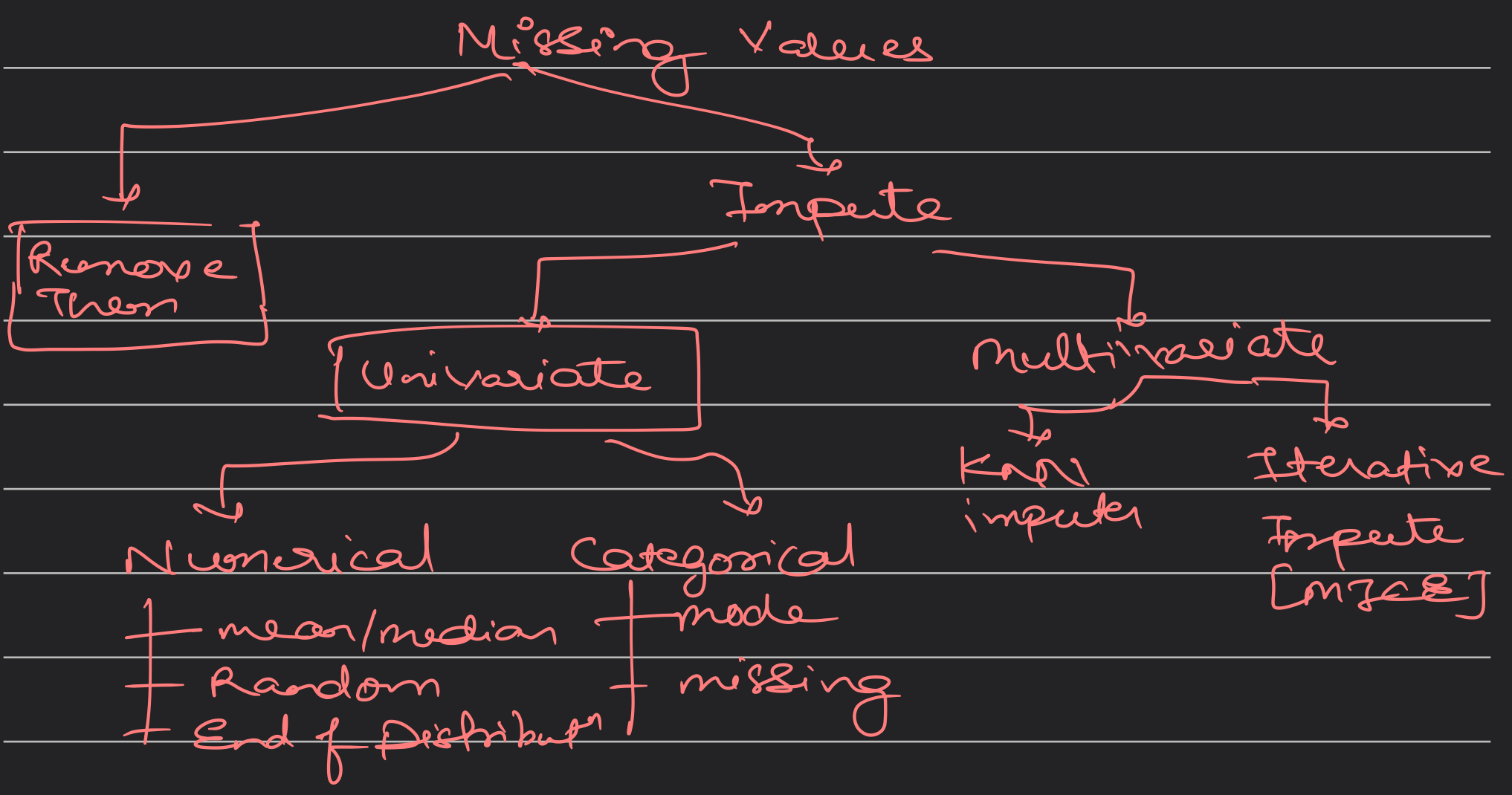
Handling missing Data - Imputation to filling missing data.



Removing the Data (Complete Case Analysis)

Approach

Also called List-wise deletion of cases consist in discarding observations where values in any of the variables are missing.

Column Row

Complete case analysis literally means analysing only those rows for which information is present in each column.

No missing value in any of the column for rows.

Assumption for CCA

- Data should be missing completely at Random.

for Ex 1000 values 50 are empty but these 50 values should be randomly missing. No starting, middle or end, randomly only.

Advantages:

① Easy to implement, no data manipulation is required

② Data before and after removal is same

Disadvantages

① Can exclude large part of observation

② Excluded values could be useful for analysis

③ When using models in production, model will not know how to handle missing data.

- When to use CCA?

We use CCA when missing data is less than 5% and second data should be missing at Random.

what if 95% of missing data?

② We drop those columns where we have missing 95% data in columns.

- Thing to remeber when dropping this nf values?

histogram

well the distribution before & after should some what match.
This tell us that MCRA existed and eel CCA is successful for that variable.

For categorical data the ratio of different categories in variable should be maintained before & after CCA.

Handling Missing Numerical Data.

Numerical

Univariate Imputation

Multivariate Imputation

Lets say we have

1	2	3	4

and column 2 has missing values, and we use some statistical method (mean, max min etc) to fill the values in column "2" using values present in colum 2 we are doing "univariate Imputation"

Similarly if we use multiple columns to fill missing value in a particular column its called multivariate imputation

- Techniques for univariate Imputation

a.) Mean / Median Imputation

When to use mean or median

If Data is normally distributed use mean or median any of them.

If data is somewhat skewed we use median data for Imputation

Its very simple to use, but not used in production.

Not reliable if missing data > 5%

1) Changes the shape of distribution.
2) we also get extra outliers
3) Changes or creates changes in Co-variance or co-relation btw variables.

when to use

1) Data is MCAR
2) Less than 5% data is missing.

- Arbitrary value Imputation

Missing values gets replaced with word "Missing", this help m2 model to differentiate between observation is that had data and the ones that didn't.

In numerical data we use a "data" of "value" that is not present in our variable

The whole idea is create difference observation that had data with others that didn't have it.

PDF graph gets distort.

Variance gets effected

Co-variance & correlation get affected.

- End of Destribution Imputation

An extension of last technique, we just use any "end value" either from starting or ending to fill the value

for a normally distributed data

value = mean + 3σ

mean - 3σ

for skewed data

Value = Q1 - 1.5 IQR → (Q1 - 25th)
(Q3 - Q1)

or Value = Q3 + 1.5 IQR

↳ 75th percentile

Handling missing categorical data.

Most frequent value

Missing Category Imputation

Most Frequent values: → we use mode of the data categories, here we have assumptions

1) Missing Category at Random
2) Mode should have higher frequency compared to other data in category.

Changes the distribution of the data.

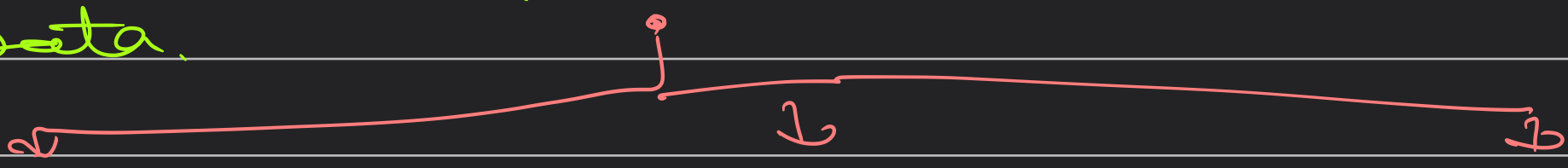
Missing Category Imputation

Here we create another variable called "Missing" and put the missing categories here for observations

Done when data missing >10% else at data is "not" missing at random.

Just a randomness gets created in data.

Technique used for Both Univariate Imputation of Numerical and Categorical Data.



Random Value Imputation | Missing Indicator | Automatic selection of Parameter.

Random value Imputation

Filling missing value through random numbers, we randomly select from the numbers present in our variable, randomly.

distribution

Helps preserve the variance of variable. We do this task using pandas.

*

Memory heavy for deployment as we need data for dataset to set and exact value & replace NA in coming observation for prediction

Good for linear models as it preserve the distribution.

Not good for Decision ~~Based~~ tree based algorithms.

other
Co-variance with data gets a bit distributed.

- **Missing Indicator**

For any column that has a missing data, we create a new column for each such column and keep value True or False for every observation

Age	Fare	Age - Missing
37	32	False
21	35	false
NA	46	True
40	100	false

arbitrary imputing
Rely
One the most used technique, but we have estimate that the ML Algo learns to different data with missing observations & non-missing observations

- **Auto Select value**

We use something called Grid Search CV where sci-kit learn used auto matically selects different imputat technique and select the best one.

Multivariate Imputation

KNN Imputer — Iterative Imputer.

- KNN Imputer.

works on the KNN Algorithm, you are like your neighbours, your behaviour is in accordance with your neighbor.

we fill a missing value in row by taking values from row which is most similar to our missing value row.

This similarity is calculated using Euclidean distance.

Each row can be used as a coordinate where distance of other rows is from missing row is calculated & data from closest row is used to fill the missing values.

Here K represent number of neighbour
ex K = 3 nearest one

$$\text{Missing Value} = \frac{\text{value row1} + \text{value row2} + \text{value row3}}{3}$$

we sklearn algo to get distances.

Adv & Dis Adv

1) More Accurate
2) more no. of calculation
3) Deploying on production is heavy

as missing value gets calculated from remote X-train data.

v) Good for smaller & medium size dataset.

concept of Uniform & Distance

we calculate weight either by taking mean K=2

Uniform = $\frac{70 + 50}{2}$ = 60

Distance = multiply by reciprocal of distance with value of data

→ Every one has same weightage

→ Closest neighbour has more value given, farthest one has less.

- Iterative Impute (MICE)

This technique use something called M.I.C.E algorithm

Multivariate Imputation By Chained Equation (MICE)

We use MICE undersome assumption

Few categories of missing data.

1) Missing completely at random
2) " at Random
3) Missing Not at Random

For few data couldn't get collected

→ In source people didn't filled data

Data removed conciously.

Assumption

we use MICE in Missing at Random

This technique is quicte Accurate, but is slow as we M2 Algo to predict missing value

we need to put training data on server. Problematic for bigger data.

How mice works?

works on Input variables

Step 1 → Replace NAN values with mean values of each columns respectively

Step 2 → Move Left to Right and whereve we had missing values we replace again with NAN.

Step 3 → we use ML Algo to predict

this missing value for that column using other columns present in the data.

Step 4 → same from Left to Right we fill missing values with NAN and repeat step 3

Step 5 → After getting done for each row. we calculate difference for each iteration Before & After. [mean - prediction]

Step 6 → we repeat this iteration until our difference becomes 0. At each iteration, the previous iteration becomes base,

when difference becomes "0" the ML Algo cannot improve more and we got most accurate value to replace this missing value.