FORMERLY Industries Helping Hands

THE iSCALE

Data Science | 30 Days of Machine Learning | Day - 16

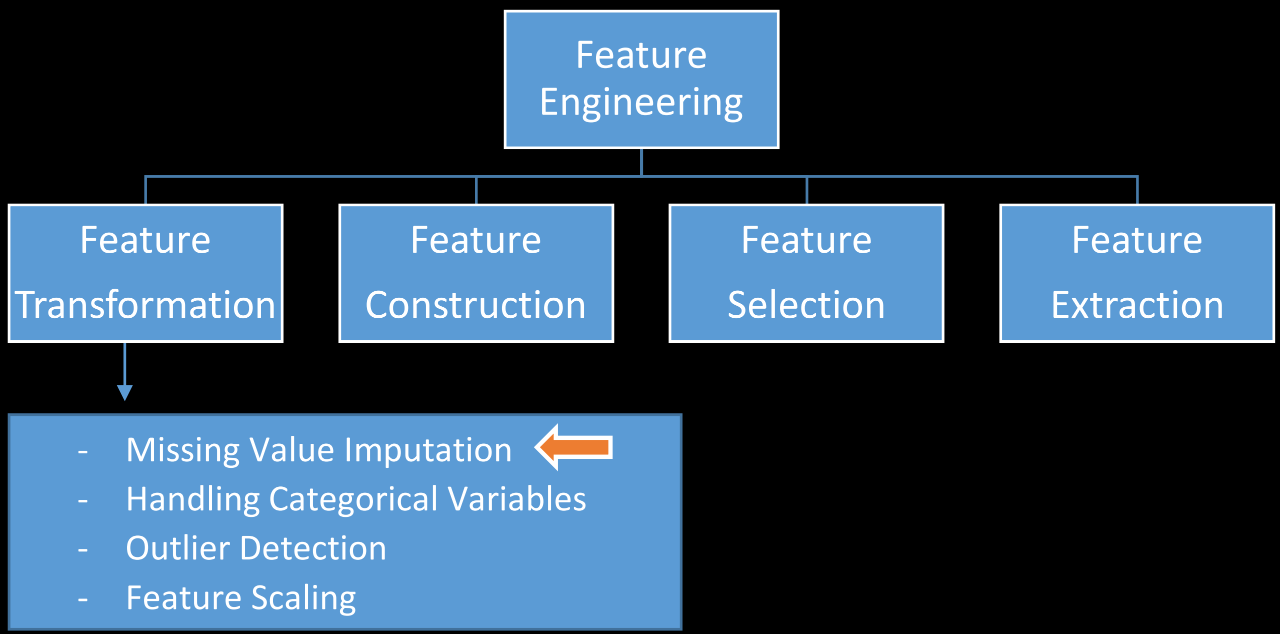
Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 16----

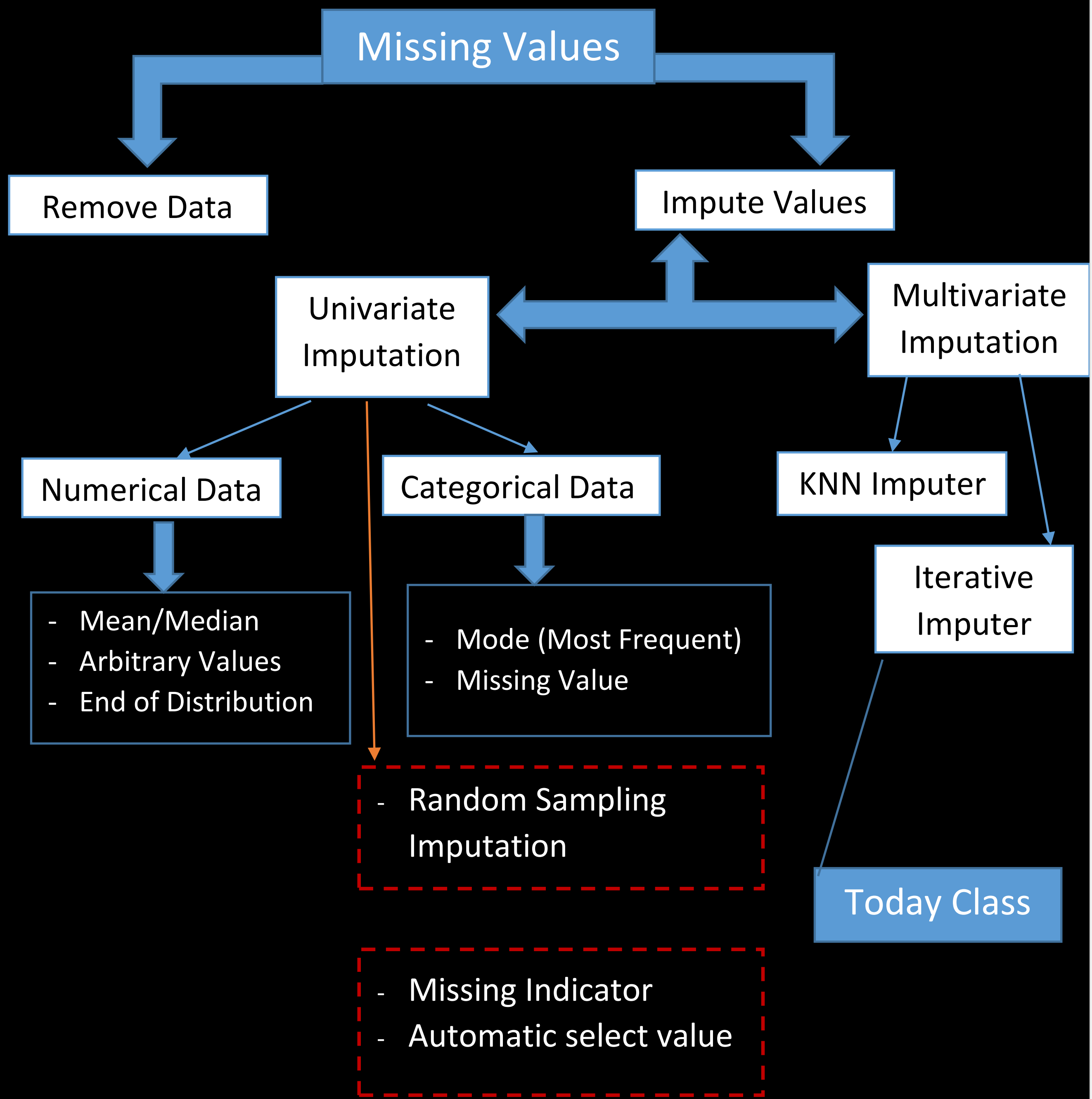
Feature Engineering (Missing Value Imputation)

- Iterative Imputer
- MICE- Multiple Imputation by Chained Equations
- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)
- Find Predictive Value for Iterative Imputer Technique

Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



Today's Topics:



- **Multivariate Imputation:** Multiple imputations can be used in cases where the data are MCAR, MAR, and even when the data are MNAR. Multiple imputation methods are known as multivariate imputation.

- **Iterative Imputer:** Iterative Imputer is a multivariate imputing strategy that models a column with the missing values (target variable) as a function.

========================== ==============================

MICE stands for "Multivariate Imputation by Chained Equations":

Having a better understanding of the reasoning for the missing ness of your data can help you determine what type of imputation method you can use. In general, there are three types of missing data.

1. Missing Completely at Random (MCAR):
2. Missing at Random (MAR):
3. Missing Not at Random (MNAR):

Missing Completely at Random (MCAR): The values in your dataset are missing completely at random. This is when there is no clear reasoning as to why a certain value in your dataset is missing. (There is no reason why your data was not collected.)

Missing at Random (MAR): The values in your dataset are missing at random. This is when we can determine some correlation to why the data value may be missing. An example of this is if a certain question in a survey is blank for multiple surveys of the same gender. A way in which we can handle this situation is *by using other features* to do a grouped mean/median replacement— the data missing is still recoverable.

Missing Not at Random (MNAR): The values in your dataset are not missing at random. This is when we can see a clear pattern to the missing values. An example of this is if a certain question category in a survey is left blank by surveys because of the question itself, as it may be a sensitive question to the surveyed (the missing ness depends on the missing data) — the data missing will be hard to recover unless further research is done. Unlike the other two types of missing data, MNAR is non ignorable.

Note:
MICE "Multivariate Imputation by Chained Equations" generally used in Missing at Random (MAR):

- MICE used in Input column

Today Class we use Dataset 50 Start-up:
https://www.kaggle.com/datasets/abhishek14398/50startups/data

1. Import – Original Data / Screenshot Image

50

	R&D Spend	Administration	Marketing Spend	Profit
21	8.0	15.0	30.0	11.0
37	4.0	5.0	20.0	9.0
2	15.0	10.0	41.0	19.0
14	12.0	16.0	26.0	13.0
44	2.0	15.0	3.0	7.0

5 Sample

2. Remove the Target Column

Profit x

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	4.0	5.0	20.0
2	15.0	10.0	41.0
14	12.0	16.0	26.0
44	2.0	15.0	3.0

3. NaN value import (Manipulate the Data)

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	NaN	5.0	20.0
2	15.0	10.0	41.0
14	12.0	NaN	26.0
44	2.0	15.0	NaN

Step 1 - Impute all missing values with mean

df

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	9.25	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.25	26.00
44	2.00	15.00	29.25

Step 2 - Remove the column 1 imputed value (Left to Right)

L → R

	R&D Spend	Administration	Marketing Spend
21	8.0	15.00	30.00
37	NaN	5.00	20.00
2	15.0	10.00	41.00
14	12.0	11.25	26.00
44	2.0	15.00	29.25

I/P Data

Cor O/P Data

Training Data in X (Training Input)

```
In [19]: X = df1.iloc[[0,2,3,4],1:3]
         X
```

Out[19]:

	Administration	Marketing Spend
21	15.00	30.00
2	10.00	41.00
14	11.25	26.00
44	15.00	29.25

Training Data in Y (Corresponding Output)

```
In [18]: y = df1.iloc[[0,2,3,4],0]
         y
```

```
Out[18]: 21     8.0
         2     15.0
         14    12.0
         44     2.0
         Name: R&D Spend, dtype: float64
```

Step 3 - Predict missing value of column 1

dF1

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.25	26.00
44	2.00	15.00	29.25

Step 4 - Remove the column 2 imputed value (Left to Right)

	R&D Spend	Administration	Marketing Spend
21	8.00	15.0	30.00
37	23.14	5.0	20.00
2	15.00	10.0	41.00
14	12.00	NaN	26.00
44	2.00	15.0	29.25

Training Data in X (Training Input) | Column 2

```
X = df1.iloc[[0,1,2,4],[0,2]]
X
```

	R&D Spend	Marketing Spend
21	8.00	30.00
37	23.14	20.00
2	15.00	41.00
44	2.00	29.25

Training Data in Y (Corresponding Output) | Column 2

```
y = df1.iloc[[0,1,2,4],1]
y
```

```
21    15.0
37     5.0
2     10.0
44    15.0
Name: Administration, dtype: float64
```

Step 5 - Predict missing value of column 2

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	$\boxed{11.06}$	26.00
44	2.00	15.00	29.25

Step 6 - Remove the column 3 imputed value (Left to Right)

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.0
37	23.14	5.00	20.0
2	15.00	10.00	41.0
14	12.00	11.06	26.0
44	2.00	15.00	NaN

Training Data in X (Training Input) | Column 3

```
X = df1.iloc[0:4,0:2]
X
```

	R&D Spend	Administration
21	8.00	15.00
37	23.14	5.00
2	15.00	10.00
14	12.00	11.06

Training Data in Y (Corresponding Output) | Column 3

```
y = df1.iloc[0:4,-1]
y
```

```
21    30.0
37    20.0
2     41.0
14    26.0
Name: Marketing Spend, dtype: float64
```

Step 7 - Predict missing value of column 3

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.06	26.00
44	2.00	15.00	31.56

Step 8 - Subtract 0th (df0) iteration from 1st (df1) iteration

```
df1 - df0
```

	R&D Spend	Administration	Marketing Spend
21	0.00	0.00	0.00
37	13.89	0.00	0.00
2	0.00	0.00	0.00
14	0.00	-0.19	0.00
44	0.00	0.00	2.31

Again Iteration Process

What is the iterative process?

The iterative process is the practice of building, refining, and improving a project, product, or initiative. Teams that use the iterative development process create, test, and revise until they're satisfied with the end result.

Day 16: Curious Data Minds

<Suggest Next Class Topic>