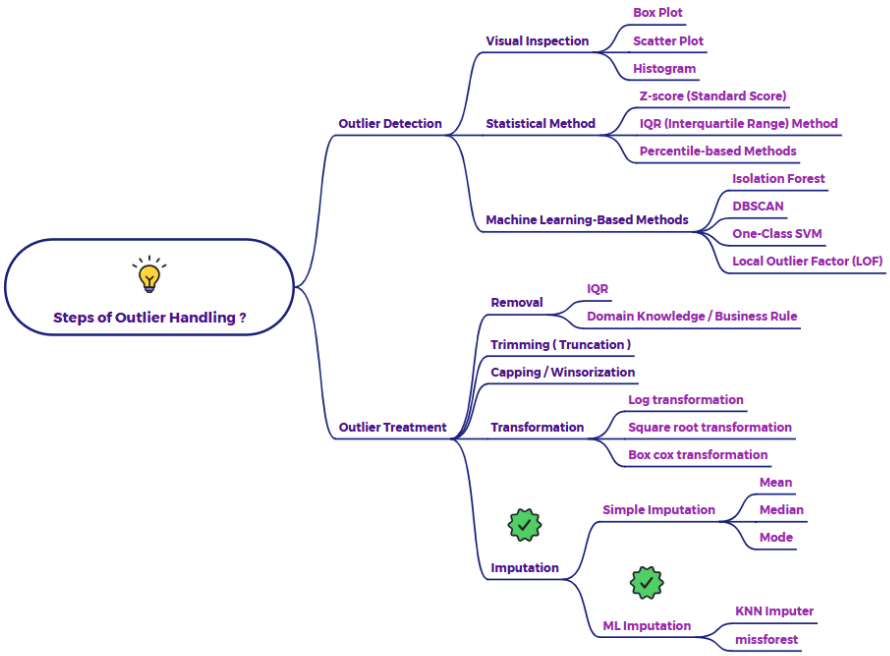
Explain Outlier treatment with ML based Imputation



Outlier Treatment: ML-Based Imputation

Machine learning-based imputation methods use algorithms to predict and replace outlier values with more accurate estimates, considering the relationships between variables. These methods are generally more sophisticated than simple imputation.

1. KNN Imputer

- **Concept:** The KNN Imputer algorithm imputes missing values (which we'll treat our outliers as) by finding the k-nearest neighbors of the data point with the missing value and using the values from those neighbors to estimate the missing value.
- **How it Works:**
    1. Identify a data point with a missing value (outlier).
    2. Find the 'k' nearest data points (neighbors) to this data point, based on a distance metric (e.g., Euclidean distance).
    3. Impute the missing value by taking the average (for numerical variables) or the mode (for categorical variables) of the corresponding values from the k-nearest neighbors.

- **Example:**
    - Dataset:

Age	Income
25	50000
30	60000
35	75000
40	90000
45	100000
50	110000
55	120000
60	130000
65	140000
20	500000

- Let's impute the outlier value 500000 for Income using KNN Imputer with k=3.
- The 3 nearest neighbors to the outlier data point (20, 500000) are (25, 50000), (30, 60000), and (35, 75000).
- Impute the Income value as the average of the Income values of the neighbors: (50000 + 60000 + 75000) / 3 = 61666.67
- The imputed dataset becomes:

Age	Income
25	50000
30	60000
35	75000
40	90000
45	100000
50	110000
55	120000
60	130000
65	140000
20	61666.67

2. MissForest

- **Concept:** MissForest is a non-parametric imputation method that uses a Random Forest algorithm to impute missing values. It iteratively predicts the missing values based on the other variables in the dataset.
- **How it Works:**
  1. Initialize the missing values (outliers) with some initial estimates (e.g., mean, median).
  2. Treat the variable with the missing values as the target variable and the other variables as predictor variables.

3. Train a Random Forest model to predict the missing values of that variable.
4. Replace the initial estimates with the predictions from the Random Forest model.
5. Repeat steps 2-4 for each variable with missing values, iterating until the imputation converges (i.e., the imputed values don't change significantly between iterations).

Example:

- Dataset:

Age	Income	Education
25	50000	Bachelor's
30	60000	Master's
35	75000	PhD
40	90000	Bachelor's
45	100000	Master's
50	110000	PhD
55	120000	Bachelor's
60	130000	Master's
65	140000	PhD
20	**500000**	High School

- Let's impute the outlier value 500000 for Income using MissForest.
- MissForest will use Age and Education to predict a new value for Income. It will iterate, refining its predictions. A Random Forest model will be trained to predict income based on age and education for all non-outlier rows. Then that model will predict the income for the outlier row.
- The imputed dataset might become (the exact value depends on the MissForest model's predictions):

Age	Income	Education
25	50000	Bachelor's
30	60000	Master's
35	75000	PhD
40	90000	Bachelor's
45	100000	Master's
50	110000	PhD
55	120000	Bachelor's
60	130000	Master's
65	140000	PhD
20	**70000**	High School

When to Use ML-Based Imputation for Outliers:

- When you suspect that outliers are not random errors and are related to other variables in your dataset.
- When you want to leverage the relationships between variables to obtain more accurate estimates for the outlier values.
- When you are working with a dataset that has a moderate amount of missing data in addition to the outliers.

Benefits:

- More accurate imputation than simple methods, as it considers relationships between variables.
- Can handle both numerical and categorical data.
- MissForest is robust to non-linear relationships and mixed data types.

Cautions:

- Computationally more expensive than simple imputation.
- Can be more complex to implement and tune.
- May still distort the original distribution of the data, although less so than simple imputation.