Feature Engineering – Imputation Techniques

Missing Data:

• Missing data, or missing values, occur when no data is stored for a certain observation in a variable.

• In many organizations, information is collected into a form by a person talking with a client on the phone, or, by customers filling forms online. Often, the person entering the data does not complete all the fields in the form. Many of the fields are not compulsory, which may lead to missing values.

• The reasons for omitting the information can vary: perhaps the person does not want to disclose some information, for example, income, or they do not know the answer, or the answer is not applicable for a certain circumstance, or on the contrary, the person in the organization wants to spare the customer some time and therefore omits to ask questions they think are not so relevant.

• There are other cases where the value for a certain variable does not exist. For example, in the variable 'total debt as a percentage of total income' (very common in financial data), if the person has no income, then the total percentage of 0 does not exist, and therefore it will be a missing value.

Missing Data: Causes

Lost:

A value is missing because it was forgotten, lost, or not stored properly.

Don't Exist:

A variable is created from the division of 2 variables and the denominator takes 0.

Not Found:

when matching data against postcode, or date of birth, to enrich with more variables, and the postcode or DOB are wrong or don't exist, the new variables will take NA.

Missing Data: Impacts

- Incompatible with Scikitlearn
- Missing data imputation may distort the variable distribution.
- Affects all Machine Learning Models.

Missing Data: Mechanisms:

Understanding the missing data mechanisms may help us choose the right missing data imputation technique. 3 mechanisms lead to missing data, 2 of them involve missing data randomly or almost-randomly, and the third one involves a systematic loss of data.

1. Missing Data Completely at Random (MCAR):

- The probability of being missing is the same for all the observations
- There is no relationship between the data missing and any other values, observed or missing, within the dataset
- A variable is missing completely at random (MCAR) if the probability of being missing is the same for all the observations. When data is MCAR, there is no relationship between the data missing and any other values, observed or missing, within the dataset. In other words, those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others. If values for observations are missing completely at random, then disregarding those cases would not bias the inferences made.

2. Missing Data at Random (MAR):

- In missing at Random the data is missing at a certain rate but that rate depends on some other variable in the data.
- example, if men are more likely to disclose their weight than women, weight is MAR. The weight information will be missing at random for those men and women who do not disclose their weight, but as men are more prone to disclose it, there will be more missing values for women than for men

3. Missing Data not at Random (MNAR):

There is a relationship between the propensity of a value to be missing and its values. In other words, data are missing not at random when the missing values of a variable are related to the values of that variable itself, even after controlling for other variables. An example would be a survey about drug usage. Individuals being surveyed could potentially leave fields blank if they used drugs that are currently illegal out of fear of being prosecuted. So, the fields aren't blank out of randomness but are left null on purpose.

Missing Data Imputation:

• Imputation is the act of replacing missing data with statistical estimates of the missing values.

• The goal of any imputation technique is to produce a complete dataset that can be used to train machine learning models.

Numerical Variables

1. Mean / Median Imputation
2. Arbitrary value imputation
3. End of tail imputation

Categorical Variables

1. Frequent category imputation or mode Imputation
2. Adding a “missing” category

Both

1. Complete Case Analysis
2. Adding a “Missing” indicator
3. Random sample imputation

1. Complete Case Analysis(CCA)

This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete

directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing.

Assumptions:

- Data is Missing At Random(MAR).
- Missing data is completely removed from the table.

Advantages:

- Easy to implement.
- No Data manipulation required.

Limitations:

- Deleted data can be informative.
- Can lead to the deletion of a large part of the data.
- Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.
- The production model will not know what to do with Missing data.

When to Use:

- Data is MAR(Missing At Random).
- Good for Mixed, Numerical, and Categorical data.
- Missing data is not more than 5% – 6% of the dataset.

2. Arbitrary Value Imputation

Imputation as it can handle both the Numerical and Categorical variables.

This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column.

Mostly we use values like 99999999 or -9999999 or “Missing” or “Not defined” for numerical & categorical variables.

Assumptions:

- Data is not Missing At Random.
- The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.

Advantages:

- Easy to implement.
- We can use it in production.
- It retains the importance of “missing values” if it exists.

Disadvantages:

- Can distort original variable distribution.
- Arbitrary values can create outliers.
- Extra caution required in selecting the Arbitrary value.

When to Use:

- When data is not MAR(Missing At Random).
- Suitable for All.

3. Frequent Category Imputation

This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as Mode Imputation.

Assumptions:

- Data is missing at random.
- There is a high probability that the missing data looks like the majority of the data.

Advantages:

- Implementation is easy.
- We can obtain a complete dataset in very little time.
- We can use this technique in the production model.

Disadvantages:

- The higher the percentage of missing values, the higher will be the distortion.
- May lead to over-representation of a particular category.
- Can distort original variable distribution.

When to Use:

- Data is Missing at Random(MAR)
- Missing data is not more than 5% – 6% of the dataset.

4. The missing values can be replaced using the following techniques:

1. Mean value
2. Median value
3. Mode (most frequent)
4. Constant value

The goal is to find out which is a better measure of the central tendency of data and use that value for replacing missing values appropriately.

Plots such as box plots and distribution plots come very handily in deciding which techniques to use. You can use the following code to print different plots such as box and distribution plots.

You may note that the data is skewed. There are several or large numbers of data points that act as outliers

Outlier's data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values.

For symmetric data distribution, one can use the mean value for imputing missing values.

Impute / Replace Missing Values with Mean

The missing values are replaced with the mean value of the entire feature column

If the data points are skewed not recommended.

Note that imputing missing data with mean values can only be done with numerical data.

Impute / Replace Missing Values with Median

The missing values are replaced with the median value of the entire feature column.

When the data is skewed, it is good to consider using the median value for replacing the missing values

Note that imputing missing data with median value can only be done with numerical data.

Impute / Replace Missing Values with Mode

The missing values are replaced with the mode value or most frequent value of the entire feature column.

When the data is skewed, it is good to consider using mode values for replacing the missing values.

Note that imputing missing data with mode values can be done with numerical and categorical data.

5. End Tail Imputation / End of Distribution Imputation:

Missing Value is not at random (MNAR) then the information is important, we want to replace missing data with values that are at the tail of the distribution of the variable.

Note:

- when we do this, the outliers are covered by when we do this imputation.
- End of the distribution means the data points from 3rd deviation.

6. Categorical Imputation:

Which is the technique replacing the missing value in a categorical variable with the string “Missing” or “Most Frequent Category”

It works only with categorical variables

A list of variables can be indicated or imputer will automatically select all categorical variables in the dataset.

7. KNN Imputation

The **KNN Imputer** class provides imputation for filling in missing values using the k-Nearest Neighbors approach.

By default, a Euclidean distance metric that supports missing values, **nan_euclidean_distances**, is used to find the nearest neighbors.

Each missing feature is imputed using values from **n_neighbors** nearest neighbors that have a value for the feature.

The feature of the neighbors is averaged uniformly or weighted by distance to each neighbor.

If a sample has more than one feature missing, then the neighbors for that sample can be different depending on the particular feature being imputed.

When the number of available neighbors is less than **n_neighbors** and there are no defined distances to the training set, the training set average for that feature is used during imputation.

If there is at least one neighbor with a defined distance, the weighted or unweighted average of the remaining neighbors will be used during imputation. If a feature is always missing in training, it is removed during **transform**.

8. Regression Imputation

With regression imputation the information of other variables is used to predict the missing values in a variable by using a regression model.

Commonly, first the regression model is estimated in the observed data and subsequently using the regression weights