uday sai

4 Followers About

Feature Engineering : Imputation Techniques

uday sai Apr 19, 2020 · 6 min read

The real-world datasets often contains null values/ missing values, outliers and high-cardinality categorical variables. Please refer the below link to read my article on “EDA: Detecting Anomalies”. In this article I will be discussing about few imputation techniques besides the KNNImputer and IterativeImputer present in sklearn package.

How to perform EDA which complements our Data Science story: Detecting Anomalies

“Data Anomalies are not always noise in our data, they might be inbuilt patterns which needs critical examination and...

medium.com

I am just delineating the definitions of MAR, MCAR, MNAR. Refer the above link for elucidation.

MCAR: This case occurs when the occurrence of missing values is equally likely for all the instances of data

MAR: This context occurs when the probability of missing values is dependent on other features or a specific cause to some extent.

MNAR: This case occurs when the reason for missing values is firmly known

above and based on the exploration use one of the following techniques:

> *Complete-Case Analysis (CCA)*
>
> *Simple Mean/Median Imputation*
>
> *Random Sample Imputation*
>
> *Capturing Missing Values with Additional Variables*
>
> *End of Distribution Imputation*
>
> *Imputing with Arbitrary Values*

Complete-Case Analysis (CCA)

This technique is used when the missing values belong to the MCAR category. This is because when the missing values are completely at random we can not figure out what is the cause of these missing values.If we can conclude that these instances does not add much information and add noise to our dataset, we can delete the rows which contain missing values. As the data is not manipulated we call this case as Complete-Case-Analysis technically.

Advantages:

This method is easy to implement.

The distribution of the variables is not distorted as we are not imputing any of the variables.

Disadvantages:

The deletion of the rows might cause a potential information loss.

The sample is biased if the cause of the missing values is due to MAR or MNAR.

Simple Mean/Median/Mode Imputation

the distribution of the variable in choosing which statistical measure (mean, median, mode) to impute. This technique is also followed in the context of MCAR. It is a convention to use mean/median for numerical variables and mode for categorical variable.

Advantages:

This is fast and easy to implement as we have pre-built libraries

Disadvantages:

The distribution of the original value is distorted and potentially the correlation and covariance might also gets impacted.

The addition of values which are already frequent in our dataset might underweigh the other values/labels of a variable.

Random Sample Imputation

This is one of the advancement of the simple mean/median/mode imputation. In this method we will impute the missing values by creating a random sample. The purpose of this imputation is to preserve the distribution of the variable. The main benefit of this technique is besides using the mean/median imputation some other uncommon values will also be used for substitution. The rationale is to replace the population of missing values with a population of values with the same distribution of the variable. In doing so the mean, variance and standard deviation is preserved. This technique is also followed in the context of MCAR.

Note: This technique is also used for categorical variables.

Advantages:

This technique preserves the distribution and variance of the variable

Disadvantages:

This technique in a sense imparts some randomness in to our dataset.

missing. Let us say we have missing values in a column 'X' then we introduce a column 'X_missing' in which we have 1 if the value is present in 'X' column or else 0 if the value is missing. This technique is to emphasize the importance of missing values, besides imputing the missing values with a simple mean/median in the actual column. The downside of this technique is it increases the feature space and should be avoided in case where already more features are present. This technique is used in the case of MAR.

Note: In case of categorical values besides introducing a different variable we can also use a different label to replace these missing values and treat it as another label for the respective categorical variable.

Advantages:

This technique captures the importance of missing. This can be materialistic in some of the datasets.

Disadvantages:

This technique adds a great value when there are tens of variables are present. However, if several tens or hundreds of variables are present then the feature space will be increased. This will be the case when multiple columns have missing values which has a negative effect on model performance.

End of Distribution Imputation

From the above technique we emphasize the importance of capturing the missing values. However it may not be effective in the cases where we have several hundreds of variables. So in order to capture the importance of missing and impute missing values we use a naive technique. We choose the values which are present in the tail end of the distribution to impute the missing values. This technique is also used in the case of MAR.

The rationale is that if the value is missing, it has to be for a reason, therefore, we would not like to replace missing values with the mean and make that observation look like the majority of our observations. Instead, we want to flag that observation as different, and therefore we assign a value that is at the tail of the distribution, where observations are rarely represented in the population.

This technique captures the importance of missing. This can be materialistic in some of the datasets.

Disadvantages:

This technique will mask the true outliers if there are more missing values present in our dataset. Conversely if the missing values are small then these substituted values will be treated as outliers in subsequent steps

This will distorts the distribution of the variable.

If the missing values does not have any importance then this will mask the predictive power of the variable.

Imputing with Arbitrary Values

This technique is based up on the domain expertise. In this method we will choose an arbitrary value based on the domain expertise and replace the missing values in the variable. This arbitrary value is different from the mean/median value. However, which value to choose is often demanding and requires vigorous research. This technique is also used in the case of MAR.

Advantages:

This technique captures the importance of missing.

Disadvantages:

Distorts the original distribution of the variable

Hard to decide which value to use If the value is outside the distribution it may mask or create outliers

Custom Technique to impute missing values due to MNAR :

This is the case where missing values of a variable are due to the other variables. To illustrate this let us consider a dataset in which we have missing values for a variable named ‘TotalCharges’ in a telecom customer churn. After careful inspection of the

Image: Missing Values because of MNAR

Conclusion

These are some of the common techniques used in data science projects at industry level and some data science challenges. We must categorize the case of causation of missing values as either of MCAR, MAR and MNAR and use one of the above imputation techniques to replace missing values based on the dataset taking the disadvantages of respective methods in to consideration.

Please refer my github for the code to implement the above techniques.

References

udaysai50/Full-Stack-Data-Science

Permalink Dismiss GitHub is home to over 40 million developers working together to host and review code, manage...

github.com

krishnaik06 - Overview

Dismiss Sign up for your own profile on GitHub, the best place to host code, manage projects, and build software...

https://raw.githubusercontent.com/srivatsan88/YouTubeLI/master/dataset/churn_data_st.csv

Data Science Machine Learning Feature Engineering Imputation Python