Charafeddine Mouzouni
@HeyCharafeddine

How to deal with missing data in 2023?

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

1- "Let's go fast" solution:

▪ How? You delete data. You impute data. Data imputation can be as simple as a median/mean or more complex like averaging values of the k-nearest neighbors, SMOTE with an auto-encoder (why not), etc.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

1- "Let's go fast" solution:

▪ Why? It's simple, fast, and easy. It works well if your imputation doesn't alter the distribution of points.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

1- "Let's go fast" solution:

▪ Why not? You might have a 'relatively' significant amount of missing data, so deleting rows/columns is not an option. Imputation might introduce an undesirable bias/variance—especially if missing data is not random but with a specific pattern.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

2- Use Gradient boosting trees:

▪ How? XGBoost and other Gradient Boosting trees handle missing data for you. You have nothing to do - just watch them do the work.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

2- Use Gradient boosting trees:

▪ Why? XGBoost uses a "sparsity-based split search" algorithm to learn how best to split the missing value in the tree. This is one reason why XGBoost is so widely used by Data Scientists when they have missing or sparse data. I love XGBoost 👌

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

2- Use Gradient boosting trees:

▪ Why not? Tree-based models are terrible at extrapolation - in regression problems. They will never generalize to values they haven't seen.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

3- Encode missing value as a new feature.

▪ How? You can add a binary column where you say this was a missing value and this wasn't.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

3- Encode missing value as a new feature.

▪ Why? The model has the same information as you; he knows that some data is missing, so he can differentiate the case where information is missing.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

3- Encode missing value as a new feature.

▪ Why not? It would help if you had quite a lot of data to learn something actionable.

7:32 AM · Dec 31, 2022 · undefined

Charafeddine Mouzouni
@HeyCharafeddine

How do you deal with missing data?

Like this post?

Please let me know what you think in the comments below. Also, follow me for more insights.

7:32 AM · Dec 31, 2022 · undefined