“Null Drop Threshold” Module Transcript

Chapter 1

Introduction to the Module

Hi. I'm Hayley, and I'm happy to welcome you back to the Global Settings Module Series. In this mini-module, we will talk about the null drop threshold, which is particularly important for ensuring data quality by removing input features with excessive missing values.

Chapter 2

Null Drop Threshold Overview

The null drop threshold determines if input features in the model dataset should be dropped based on the proportion of missing values they contain. For One AI, this threshold specifies the percentage of null data an input feature column can have before it's automatically dropped and excluded from the model during preprocessing. Zeros are not considered nulls. Nulls are missing or blank values.

Automatically dropping columns with a high percentage of null rows lends to stronger performance and enhanced model interpretability. This is because these columns contribute little-to-no useful information and unnecessarily increase dataset dimensionality.

Removing them reduces the complexity of the dataset and can lead to more efficient model training and improve performance. Additionally, including high-null columns can bias statistical estimates and model parameters resulting in inaccurate predictions. By removing these columns, the model focuses on meaningful representative data, reducing the risk of biased estimates.

And finally, dropping columns with many null values minimizes noise and variability, enhancing the model's robustness and generalizability.

This approach allows analysts to concentrate on the most relevant features, making models easier to understand and leverage strategically. The default null drop threshold in One AI is 0.05, meaning columns with 5% or more null values will automatically be dropped and excluded from the model during preprocessing.

Users may want to adjust this threshold if certain columns are intentionally partially null, but still relevant for the model. For instance, not everyone receives a performance review, so that column might naturally have more null values, but could be very predictive in a promotion model.

Alternatively, users can intervene on a per-column basis to prevent specific columns from being dropped due to nullness, allowing the null drop threshold to still apply to all other columns. Additionally, users can use null fill strategies, which replace missing values with estimated or calculated ones available in the Per Column Intervention section. Check out that module for more information.

Chapter 3

Adjusting the Null Drop Threshold in One Model

You can adjust this threshold by toggling the override slider to "On" and entering a new value between 0 and 1 in the designated field. For example, setting it to 0.2 means that columns can have up to 20% null values before One AI will automatically drop them from the model dataset during preprocessing. Don't forget to scroll to the bottom and save your threshold adjustments before rerunning your model.

Chapter 4

Conclusion & What's Next

Thanks for joining me to take a look at the null drop threshold setting. In the next mini-module, we will examine random state settings. Happy modeling!