“Category Size Threshold” Module Transcript

Chapter 1

Introduction to the Module

Hey. 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 category size threshold, which is particularly important to understand for categorical features.

Chapter 2

Category Size Threshold Overview

In machine learning, the category size threshold determines how categorical variables are handled, particularly those with a large number of unique categorical groupings or levels.

In One AI models, this threshold specifies the minimum size a categorical grouping must have before being grouped into an "Other" category through one hot encoding.

The default threshold is 0.05, meaning any categorical grouping representing less than 5% of the total will be placed in this "Other" grouping.

Let's go through an example to make this a bit more concrete. Suppose "Work City" is an input feature in your model dataset, and your workforce is distributed across 4 cities with 50% in L.A., 46% in New York, 2% in Dallas and 2% in Chicago. With the default threshold of 5%, L.A. and New York City would be treated as separate categorical groupings while Dallas and Chicago would be grouped together as "Other" because they do not make up 5% of the total. If you lowered the threshold to 0.01, then Dallas and Chicago would also be treated as separate categorical groupings.

Adjust the threshold if you believe there are significant differences between categorical groupings that don't meet the 5%threshold, but should be treated separately rather than being combined into other. Alternatively, you can raise the threshold if you want only larger category groupings to be considered individually.

Keep in mind that when category groupings are too small, they become less meaningful, making comparisons more difficult or even impossible.

Chapter 3

Impact on Model Performance

Setting a reasonable category size threshold is important for model performance, interpretability, and efficiency. High cardinality categorical variables can lead to overfitting, making it difficult for the model to generalize to unseen data because the groupings may be quite different as time passes. It also dilutes the significance of individual categories, obscuring meaningful patterns, reducing the number of categories or using appropriate encoding techniques helps the model focus on the most informative features, enhancing prediction accuracy.

Finally, algorithms like decision trees and random forests may become computationally inefficient when handling many unique categories.

Chapter 4

Adjusting the Category Size Threshold in One Model

The category size threshold can be adjusted by toggling the override to "On" and entering a new value ranging from 0 to 1 in the designated field. For example, inputting "0.1", like so would mean that categorical groupings must make up at least 10% of the total, or there'll be one hot encoded into the "Other" column.

Don't forget to save your threshold changes before rerunning your model.

Chapter 5

Conclusion & What's Next

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