“Continuous Strategy & Null Indicators” Module Transcript

Chapter 1

Introduction to the Module

Hey. I'm Hayley, and I'm excited to welcome you back to the Global Settings Module Series. In this first mini-module, we'll examine continuous strategy and null indicators.

These two topics are grouped together because the null indicator setting is only applicable if the continuous strategy has been changed from scaling to binning. This will all make sense shortly. Both of these settings allow for greater flexibility in tailoring and handling data, particularly in the preprocessing stage.

Chapter 2

Continuous Strategy Overview

The continuous strategy determines how One AI handles numeric and date input features, which are known as continuous features in the model dataset. When you expand the global section, you'll find continuous strategy as the first setting. Toggling the override slider to "On" will reveal the configuration options available.

By default, One AI performs standard linear scaling on numeric and date variables. This ensures that multiple features are on the same scale, preventing them from being incorrectly weighted by the algorithm. This method preserves the original information in the data, allowing the model to retain the nuances and details present in the data without bias. Preserving this information allows the model to capture subtle patterns and relationships contributing to model flexibility.

The default scale type - 'Standard' - rescales the dataset to have a mean of 0 and a standard deviation of 1. This can be changed by using the dropdown menu to 'MinMax', which rescales the dataset so that each value falls between 0 and 1 or 'Robust', which scales values to have a median of 0 and an interquartile range of 1, making it more resistant to outliers.

Scaling assumes a normal data distribution for the model dataset, but it may be less effective if the distribution is significantly non-normal. In these cases, consider a binning continuous strategy, which involves grouping continuous data into discrete intervals or bins. This is done with one hot encoding because it's creating categories for your

numerical data. For example, salaries can be binned into salary ranges with $50,000 intervals.

The number of bins or intervals is determined by the bin type. The default is auto, which uses algorithmic methods to determine the number of bins, but you can also manually specify the number of bins by inputting a single value in the designated field.

By default, when using a binning strategy, null rows are set aside. Bins are only created for non-null rows of each continuous feature unless a null indicator override is applied, which we will discuss shortly. The binning strategy simplifies complex relationships between the target variable and continuous predictor by focusing on broader trends within each bin rather than modeling intricate patterns.

It's also effective for handling extreme outliers as it places them all in the same bin, preventing them from disproportionately influencing the model. However, this can mean that extreme values are not adequately captured, potentially impacting model performance. Additionally, binning can lead to a loss of information and reduce the ability to detect subtle patterns or trends in data.

If you configure continuous strategy, remember to scroll to the bottom of the screen to save before rerunning your model.

Chapter 3

Null Indicator Overview

Now, let's move to the null indicator section, which is only applicable if the continuous strategy was changed from scaling to binning. When using a binning strategy, the numerical values of continuous features are one hot encoded, creating category groupings or intervals.

By default, binning sets aside null rows and doesn't create bins for nulls to prevent algorithmic bias. If you want to bin the nulls, the null indicator field accepts a free text value as a label for the null bin indicating scalars that were originally null, but had to be filled to apply bin edges.

Basically, you are creating a category for nulls to be placed into when a binning strategy is used. For example, without a null indicator override, null salaries would be set aside as null as shown in this image. If you specified "my special null override" in the null indicator designated field, your null salaries would then be binned into that category as illustrated in the second image.

Null indicators help manage null data and preserve information, ensuring these values are not overlooked. This flexibility is beneficial as each model has unique requirements in which different strategies may be helpful.

Chapter 4

Conclusion & What's Next

Thanks for joining me to discuss the continuous strategy and null indicator settings. In the next mini-module, we will examine correlation type and general correlation threshold settings. Happy modeling!