“Correlation Type & General Correlation Threshold” 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 mini-module, we will explore correlation type and general correlation threshold configurations. These topics are grouped together because both require a basic understanding of correlations.

Chapter 2

Correlation Type Overview

We will begin with Correlation Type, but first let me give you a little background information to get us all on the same page. During a One AI machine learning run, a correlation check is performed to determine how correlated each input feature column is with the target column, which is what we're predicting. This is important because if any input variable columns are too highly correlated with the target column, One AI will automatically drop them alongside a "Leakage" drop label in the EDA report, since this usually indicates data leakage. One AI also checks the correlation between each input variable and every other input variable, which I will discuss in more detail in the next section.

When performing these tests, the default behavior is to use a Cramér correlation type test for categorical variables and a Pearson correlation type test for continuous variables.

The Cramér correlation check calculates Cramér's V to assess the strength and significance of the association between a pair of categorical variables. Cramér's V, a correlation coefficient, measures the effect size for the chi-square test of independence. It assesses the strength and significance of the relationship between two nominal or ordinal variables, ranging from 0 to 1. Values closer to 0 indicate low correlation, while values closer to 1 indicate a high correlation, meaning the variables are highly associated with each other.

The Pearson correlation check uses Pearson's correlation coefficient to measure the linear relationship between each independent feature (which are the input features) and the target variable to assess the degree of linear correlation.

The Pearson correlation coefficient, often denoted as "r", quantifies the strength and direction of a linear relationship between two continuous variables.

It ranges from -1 to 1 where 1 indicates a perfect positive linear correlation, which means as one variable increases, the other also increases. -1 indicates a perfect negative linear correlation, which means as one variable increases, the other decreases. And 0 indicates no linear correlation between the variables.

Chapter 3

Adjusting Correlation Type in One Model

To check out these settings, toggle the Override slider to 'On' next to "Correlation Type". This will reveal a dropdown menu. You can select 'None' if you do not wish to perform a correlation check. You will also see that you can select 'Cramér's' or 'Pearson' for the type of correlation check from the dropdown menu. However, it is not recommended to change these settings because Cramér's V should only be used for categorical variables, and Pearson's correlation should only be used for continuous variables.

Most model datasets contain both types of variables. So if you select either Cramér's or Pearson for all input features, it will try to apply these tests to variables it shouldn't and do a bad job. If you do configure correlation type, remember to scroll to the bottom of the screen to save before rerunning your model.

Chapter 4

General Correlation Threshold Overview

As I mentioned earlier, One AI runs a correlation test to check the correlation between input variables using Cramér's V for categorical variables and Pearson's correlation for continuous variables. The general correlation threshold determines how correlated two or more input variables must be for the less performant variable(s) to be automatically dropped by One AI.

The purpose of this test is to detect if two or more input variables are too correlated, meaning they're too similar, to both be included in the model dataset. For example, date

of birth and age are both often available to One AI models as input variables, but shouldn't be selected in the same model because they are effectively the same thing and highly correlated.

Multicollinearity occurs when models contain highly correlated input variables. This negatively impacts model training and performance because models struggle with similar input columns. This is why we want to avoid multicollinearity.

In One AI, the default threshold is set to 0.65 as this value traditionally indicates a moderately strong linear relationship between two variables.

Chapter 5

Adjusting the General Correlation Threshold in One Model

This threshold can be adjusted by toggling the Override slider to 'On' and entering a new value ranging from 0 to 1 in the designated field. For example, if you want the threshold to be more conservative, meaning the variables need to be even more correlated for all but one to be automatically dropped, you can increase it to 0.8, like so. Don't forget to save any general correlation threshold changes before rerunning your model.

Chapter 6

Conclusion & What's Next

Thanks for joining me to examine correlation type and general correlation threshold settings. In the next mini-module of the Global Settings Series, we will explore leakage and suspicious performance threshold settings. Happy modeling!