OneModel Academy

One AI

Global Settings

Machine Learning

Hayley Bresina
One AI Client Enablement

10 Jun 2024

Adjusting Global Settings is Optional

- Configuring global settings can be advanced
  - Ensure you are comfortable with the basics of model building & interpretation first
  - One AI automatically selects default settings based on best practices that perform well on most models in One Model
- Informed configuration enhances model performance & improves insight visualization on storyboards

Topics Covered

- Continuous strategy & null indicators
- Correlation type & general correlation threshold
- Leakage & suspicious performance threshold
- Category size threshold
- Null drop threshold
- Random state
- Global settings overview

Global Settings Overview

Overview & Purpose

- **Global settings** refer to overarching parameters that impact the entire model's behavior and performance
  - Dictate the rules for automatically dropping columns from the model dataset
  - Influences preprocessing
  - Impacts how the model insights & results can be visualized in storyboards
- Unique to each model & do not apply across all models on your site
- Purpose
  - Give the user control over model configuration
  - Improve performance, feature selection, & data handling

Navigation in One Model

OneModel Academy

Thanks for watching!

OneAI

OneModel Academy

One AI

Continuous Strategy & Null Indicators

Global Settings

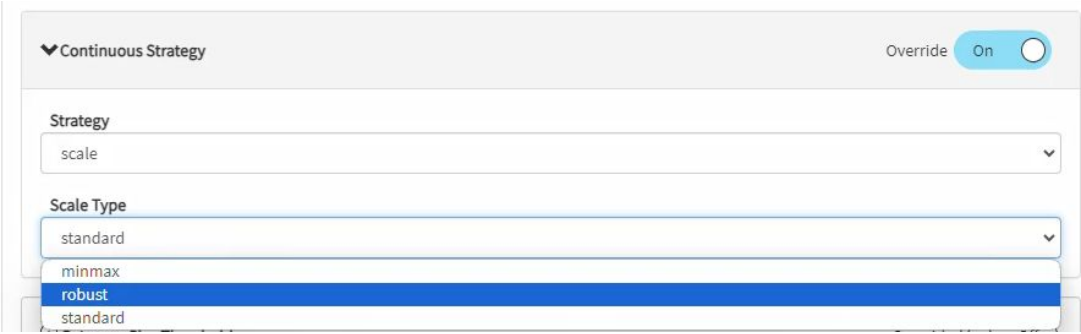
Hayley Bresina
One AI Client Enablement

10 Jun 2024

Continuous Strategy: Scaling

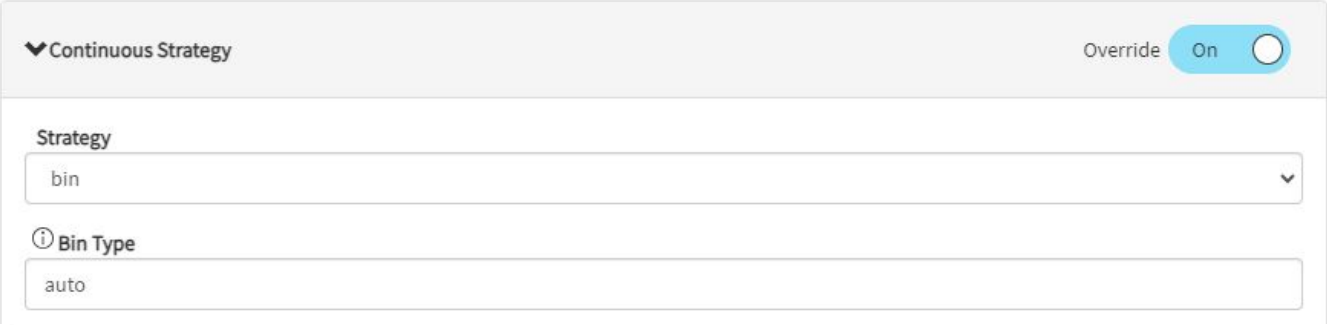
How One AI handles numerical & date input variables during preprocessing

- **Default strategy:** linear scaling
  - Puts all features on the same scale, preventing incorrect weighting & bias
- **Default scale type:** standard
  - Other scale type options: MinMax; Robust



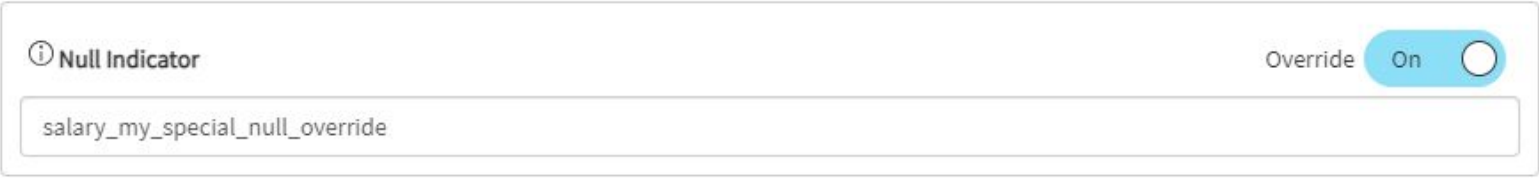
Continuous Strategy: Binning

- **Alternative strategy option:** binning
  - Grouping numerical data into discrete intervals or bins
  - **Bin type default:** auto
    - Alternative bin types: manually specified numerically



Null Indicator

- Only applicable if the continuous strategy is set to **binning**
  - Binning results in numerical values being one hot encoded, creating categories or intervals (bins)
  - By default, binning sets aside null rows & doesn't create bins for nulls
  - To bin nulls, use the designated field & input free text value to label the null bin



Null Indicator

- No null indicator override:

id	salary_0	salary_1	salary_2	salary_null
0	0	0	1	0
1	0	1	0	0
2	0	0	0	1
3	1	0	0	0

- Null indicator override:

id	salary_0	salary_1	salary_2	salary_my_special_null_override
0	0	0	1	0
1	0	1	0	0
2	0	0	0	1
3	1	0	0	0

OneModel Academy

Thanks for watching!

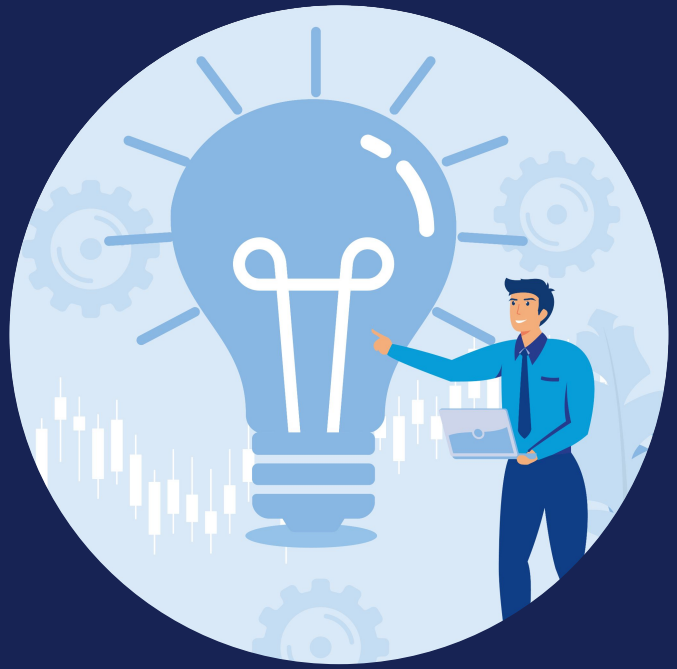
OneAI

OneModel Academy

One AI

Correlation Type & General Correlation Threshold

Global Settings



Hayley Bresina
One AI Client Enablement

10 Jun 2024

Correlation Type

During a One AI model run, a correlation check is performed to determine how correlated each input feature column is with the target column

- Checking for **data leakage**
- **Default behavior**: Cramér's correlation test for categorical variables & Pearson's correlation test for continuous variables
  - **Cramér's V**: correlation coefficient ranging 0 to 1 that assesses the association between a pair of categorical variables
  - **Pearson's** correlation coefficient: "r", quantifies the strength & direction of a linear relationship between two continuous variables ranging -1 to 1

General Correlation Threshold

One AI also runs a correlation test to check the correlation between each input variable

- **General correlation threshold**: how correlated 2+ predictor variables must be for the less performant variable(s) to be automatically dropped
  - Model datasets should not have multiple variables that are effectively the same thing
  - Avoid multicollinearity
  - **Default threshold: +/- 0.65**, which indicates a moderately strong linear relationship between 2 variables

OneModel Academy

Thanks for watching!

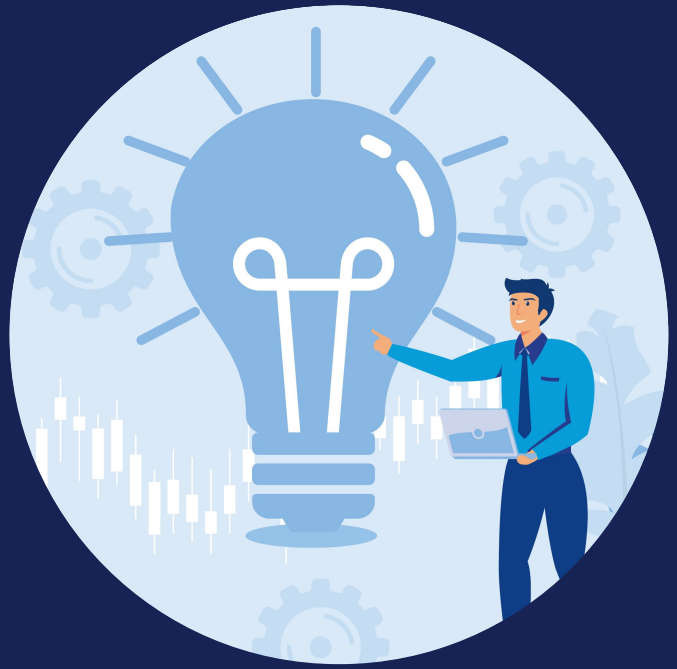
OneAI

OneModel Academy

One AI

Leakage & Suspicious Performance Threshold

Global Settings



Hayley Bresina
One AI Client Enablement

10 Jun 2024

Leakage Performance Threshold

- Data leakage check is performed during preprocessing
  - **Data leakage**: when the training dataset contains information about the target variable that won't be available when making predictions on new, unseen data
    - Cheat column
  - Identified by generating a random forest model against the target using only the input feature in question & measuring the performance with an ROC-AUC score
  - Default threshold: 0.85;
    - Variables exceeding this threshold will be automatically dropped by One AI

Suspicious Performance Threshold

Less stringent version of the data leakage performance threshold

- Informs users of possible leakage
- Default threshold: 0.7
  - Variables exceeding this threshold will **not** be automatically dropped by One AI & instead will be flagged in the EDA report for the user to investigate

Suspicious Selected! Processed Scaled `one.employee.date_of_birth` has been processed as a date feature and numerically scaled. This column may be leaking target data. A random forest using only this column achieved a ROC-AUC score of 0.76 against the target.

- If leakage is present, column should be manually dropped
- If leakage is not present, no action is needed

OneModel Academy

Thanks for watching!

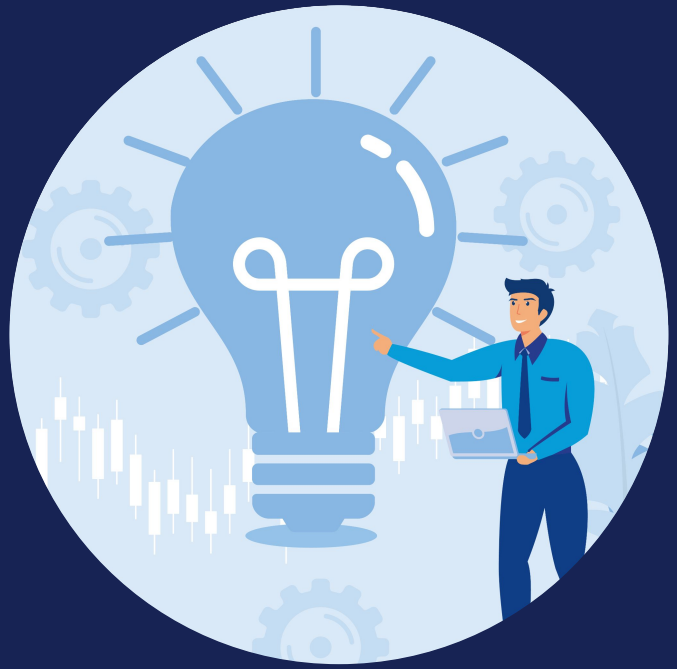
OneAI

OneModel Academy

One AI

Category Size Threshold

Global Settings



Hayley Bresina
One AI Client Enablement

10 Jun 2024

Category Size Threshold

Determines how categorical variables are handled

- In One AI, specifies the minimum size a categorical grouping must have before being grouped into an "Other" category
  - **Default threshold: 0.05**
    - Any categorical grouping representing less than 5% of the total will be placed in the "Other" grouping
- Reasonable category size thresholds are important for model performance interpretability, & efficiency

OneModel Academy

Thanks for watching!

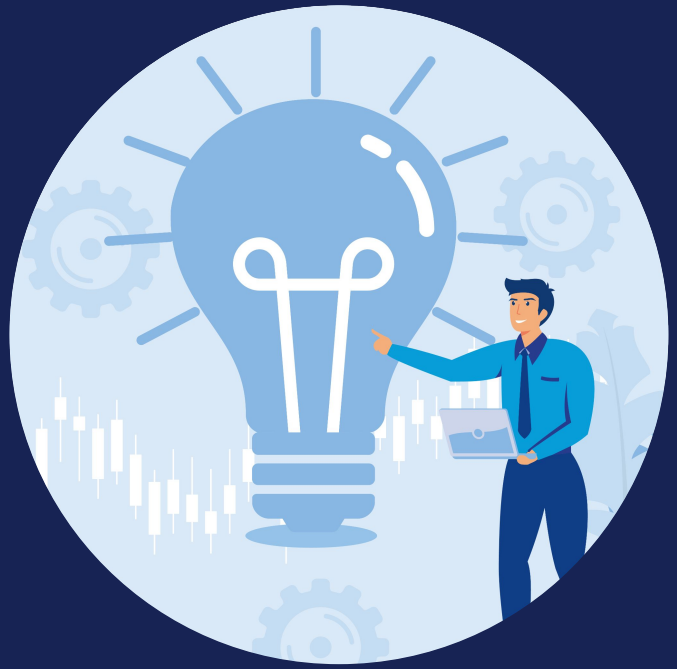
OneAI

OneModel Academy

One AI

Null Drop Threshold

Global Settings



Hayley Bresina
One AI Client Enablement

10 Jun 2024

Null Drop Threshold

Determines if input features in the model dataset should be dropped based on the proportion of missing values they contain

- In One AI, specifies the percentage of null data in an input feature before it is automatically dropped and excluded from the model
  - Zeroes are not nulls
  - **Default threshold: 0.05**
    - Input features with 5% or more null values will automatically be dropped during preprocessing
- Alternative options: droppability or null-filling **per column interventions**

OneModel Academy

Thanks for watching!

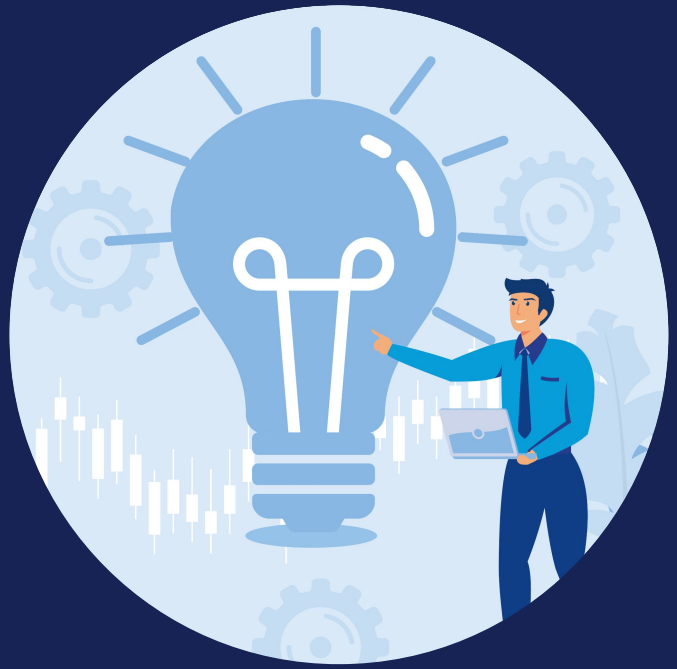
OneAI

OneModel Academy

One AI

Random State

Global Settings



Hayley Bresina
One AI Client Enablement

10 Jun 2024

Random State

Initializes the random number generator for tasks involving randomness:

- Splitting datasets into training & test sets, initializing model parameters, or conducting random sampling
- In One AI, a pseudo-random number parameter that allows you to reproduce the same train test split each time the model is run
  - Default value: **43**
  - Random state range: any value **ranging 0 to 4,294,967,295**
- Not a hyperparameter related to model performance
- **This setting should be left alone in most cases**

OneModel Academy

Thanks for watching!

OneAI