“EDA” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Howdy, folks. My name is Austin, and I'm an ML engineer on the One AI team here at One Model. In the past modules, we've discussed topics like data exploration, model performance, and ethical AI, which highlight the Exploratory Data Analysis, or EDA, report as a critical tool for interpreting machine learning models. This module will examine the EDA report and demonstrate how it complements the Results Summary report to help you understand your models.

The concepts we will cover build upon several previous modules. They'll be linked in the description under prerequisites for easy access, so you can refer to them if you need additional information. We will cover an overview of EDA and the significance of the EDA report, navigating to the EDA report in One Model, and key takeaways from each section of the EDA report. This would include the overview, variable status, variable analysis, correlations, missing values, and sample.

After completing this module, you will understand the importance of the EDA process and report in promoting transparency and understanding the model dataset and feature selection. You will identify areas for improving the model dataset based on findings, such as modifying model configuration, adjusting feature selection, and addressing missing data. And you will confidently use EDA insights to determine if the model is suitable for deployment, understanding variable treatments and decisions made during model creation.

Chapter 2

Exploratory Data Analysis Overview

We will now give an overview of exploratory data analysis. Exploratory data analysis, or EDA, is a critical step in machine learning. Its primary goal is to help understand the model datasets main characteristics, summarize key features, gain insights into its underlying structure, and identify patterns and relationships among variables. As covered in the "Supervised Learning" module, One AI models are built with labeled datasets containing input and target variables.

EDA helps in understanding the data used to train and validate models, identifying data quality issues, discovering relationships among variables, and ensuring the model settings and configurations meet expectations.

This process can also be used independently to understand a dataset without applying it to a predictive model. One AI provides a robust, transparent EDA report with various data visualization methods to analyze the model dataset for predictive modeling.

The EDA report clearly displays which variables were selected, not selected, or automatically dropped due to model settings.

It also provides insights into variable preprocessing, advanced variable analysis, and correlation information. The EDA report empowers you to identify areas for improving the model dataset.

For instance, you may need to adjust global settings, configure null filling, or address missing data. These changes enable the model to use variables with good predictive power and exclude those detrimental to performance or interpretability.

Ultimately, EDA facilitates a deeper comprehension of the model dataset, ensuring you understand how each variable was treated and why the model made specific decisions. This understanding is crucial for making informed decisions about whether the model should be shared to be used for strategic decision-making purposes or if the model needs more refinement.

Chapter 3

Navigating to the EDA Report in One Model

Now let's talk about how to find an EDA report in One Model. A unique EDA report is automatically generated by One AI for each iteration, or run, of a machine learning model that is pending, deployed, ignored, or deployed and persisted.

Runs that error or are canceled before the run is complete will not have an EDA report because it cannot be generated for incomplete runs. You can access this report by clicking on the 'One AI' tab in the main ribbon menu and scrolling to the machine learning model you wish to view the EDA report for. Click the 'Runs' button and then the status label for the iteration you are interested in. In this case, 'Pending'.

This window will automatically open to the EDA report tab. Please note that it may take a moment to load. The One AI EDA report is made up of 6 sections, all of which help us interpret our machine learning models. We will go through each of these sections in detail now.

Chapter 4

Section Takeaways of the EDA Report

Now let's talk about each section in the EDA report.

Chapter 5

Overview Section

First, we have the Overview section. The Overview section contains high-level information about the structure of the model dataset and the variables.

The number of observations is the number of instances we're making predictions for.

Variables are the attributes about each employee or instance we're making predictions for. You can see a breakdown of the types of variables on the right. Please note that all information contained in the EDA report is based on the train-test dataset and not the predict data set.

Chapter 6

Variable Status Section

Next, we have the Variable Status section. The Variable Status section provides information about handling of each variable in the model dataset. This section utilizes colored labels, so you can easily identify: If the variable was processed or selected. If the variable was automatically dropped by One AI, and if so, why? If the variable was marked as suspicious. And how the variable was processed.

Let's go over what each of the color-coded variable statuses mean. The green 'Selected' label indicates that the variable was selected by One AI to be used in the model to make its predictions.

For example, as you can see, date of birth was selected by this model for its predictive power in predicting new hire failures.

If desired, you can manually exclude selected columns from the Core Attribute step of the recipe. This may be done when the column that isn't validated, is causing data leakage, or makes the model difficult to interpret and needs to be removed.

The blue 'Processed' label indicates that the variable was tried by One AI and may or may not be selected.

These variables did not violate a global setting or the model configuration and, therefore, were not force dropped. If it has a green 'Selected' label next to it, it was selected.

If not, it was not selected in the final version of the model because One AI did not find that it had enough predictive power to be included. Other variables were found to be more important and resulted in better model performance.

You may perform a per column intervention to force the model to include processed columns if desired. Processed variables will have an additional orange label that identifies the treatment applied to the variable during the data cleaning stage. We go through these labels in detail in our "Data Preprocessing" module, but let's do a quick recap.

'Scaled' indicates that the variable is numerical or a date -so One AI transforms the continuous feature to a common scale so that all continuous features will be on the same scale and thus won't get incorrectly weighted by the algorithm.

'One Hot Encoded' means that the variable is categorical.

One AI splits each grouping into its own binary column with a value of 1 or 0 so that categorical variables can be put into a format that the machine learning algorithm can interpret and treat without bias.

You can also see how many category groupings a variable has based on the cardinality, which is specified in the sentence following the one hot encoded label. For example, if a variable was one hot encoded with a cardinality of 8, that means there are 8 groupings within that category. Please keep in mind that categories that contain less than 5% of the records are binned together in an 'Other' grouping.

This is controlled by the Category Size Threshold, which can be configured in the global settings.

The red 'Dropped' label indicates that the variable did not conform to a global setting, and therefore, One AI had objections about the variable and left it out of the predictive model. These variables are not cleaned and, therefore, do not have an orange label next to them. Instead, each will have a grey label with the reason for being dropped to explain exactly why the variable was left out. We also go through these labels in detail in our "Data Preprocessing" module, but let's do a quick recap.

'Missing' indicates that the variable had too many null values compared to the null drop threshold, so it was too empty for One AI to use. The null drop threshold is defaulted to 5%.

The following sentence tells you how null the column was as a number and percentage.

You can perform a per column intervention or adjust the null drop threshold if you want the column to be processed.

'Too Few OHE Values' indicates that the variable is categorical, so it was one hot encoded or OHE, but no category groupings made up over 5% of the records. So the entire column was categorized as other and dropped due to the category size threshold.

This is different from being unique because there's still groupings, just a bunch of small ones; for example, job ID or title. You can adjust the category size threshold from the global settings if desired.

'Constant' indicates that the variable contained all or nearly all of the same values and was therefore dropped. For example, level 1 of an Org Unit where every employee's value is the CEO.

'Unique' indicates that nearly every value of a categorical variable is different; for example, first name.

'Correlated' indicates that the variable is being dropped because it is too correlated or related to at least one other variable in the data set. We don't want multiple columns with essentially the same information to be included in the same model, so we only want to process the best fit. For example, date of birth and age are 100% correlated, so the model should drop one of them. The default threshold is 0.65, but this can be manually configured by adjusting the general correlation threshold if desired.

'Leakage' indicates that the target for the model, the thing that the model is predicting, is likely leaking data into the variable. In simple terms, it's a variable that predicts the outcome too well to be plausible, which indicates that the column is "cheating". An example would be a flag in the data that indicates whether someone is a future termination when that's also the thing that the model is predicting, like here. The model won't learn well if we give it the answer upfront, and we want it to be able to make good predictions in the absence of a cheating column, which is why One AI automatically drops these columns. The default threshold is set to 0.85, but this can be manually configured by adjusting the Leakage Performance Threshold if desired.

The purple 'Suspicious' label indicates that there is a possible situation of data leakage. This is simply a less stringent version of the test performed to check for data leakage. The default threshold is 0.7 versus the 0.85 for data leakage. These variables are not automatically dropped, but instead are flagged and should be validated by you to ensure that there is not leakage present and that the column is not allowing the model to cheat. If you determine that there is leakage present, you should exclude the column. If you determine data leakage is not present, no action is needed. It is important to go through each variable that was automatically dropped and make sure you understand why. Then you need to decide if you want to keep the column out of the model or if you want to reconfigure the model settings to allow the column to be processed and tried again.

Chapter 7

Variable Analysis Section

Now, let's talk about the Variable Analysis section. The Variable Analysis section provides detailed information about each variable in the dataset, which is especially helpful if you're not fully familiar with the data. You can access this section by scrolling or clicking the red variable hyperlink to navigate directly to the specific variable section.

First, you'll see whether the variable is categorical, numerical, or a date. The report allows you to analyze the variable globally, showing data for the entire model or by individual outcomes, such as new hire failure or new hire success. This analysis helps you identify how the data differs across different outcomes, which is super important because it's how the model learns. For categorical variables, the report provides high level information such as the distinct count of categories, percentage of unique values, and the percentage and number of missing values.

You can also view the categorical groupings and distributions, including which categories did not meet the size threshold and were grouped into the "Other" category. Additionally, composition and encoding details are available.

For numerical and date variables, you'll see similar high-level data along with the mean, minimum, maximum, and percentage of zeros. Detailed views offer various visualization options, including histograms, quantile and descriptive statistics, common and extreme values, and violin plots. You can toggle between global analysis or specific outcomes.

Chapter 8

Correlations Section

Correlations measure the degree to which at least two variables are associated. You can download a zip file of all your models correlations at the bottom of the EDA report. Each file indicates how each variable is correlated with the others, and this is particularly useful when each variable is correlated with multiple columns. Understanding correlations in your model dataset is crucial because it reveals how changes in one variable can affect another. It helps detect and address multicollinearity and informs feature selection.

Chapter 9

Missing Data Section

Let's talk about how missing data can be identified in the EDA report. The missing value section of the EDA report highlights null values and variables throughout various views. These views can help identify areas where additional data preparation may be needed. Similar to other sections in the EDA report, multiple tabs are available, allowing you to view results for the entire dataset or filtered by individual outcomes.

Chapter 10

Sample Section

The sample section provides a view of the first and last five rows in the dataset. As you scroll, you can view each variable's value for each instance. This can be useful for quick data inspection to identify obvious issues.

Chapter 11

Conclusion & Thanks

Mastering the One AI Exploratory Data Analysis report is essential for gaining deep insights into your models and datasets. Paired with the Results Summary, this report helps you understand the model dataset, feature selection, and model construction.

Analyze it carefully after each run to refine global settings, attributes, and configurations. The goal of EDA is to ensure that your models are built on solid foundations and align with your objectives.

In the next module, we'll explore the Results Summary report and see how it complements the EDA report for full transparency that you can trust in your models. Happy modeling!