“Data Preprocessing” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hey there. My name is Hayley Bresina and I'm part of the One AI team here at One Model. In this module, we will build upon the concepts you learned in the introduction to machine learning module and examine data preprocessing for machine learning models, which will help you understand what happens to your data before any machine learning occurs.

We will cover an overview of data preprocessing, how data preprocessing works in One AI machine learning models to include data scaling and one hot encoding, data cleaning, and dimensionality reduction. And then we will finish with data preprocessing for individual variables.

After watching, you will grasp the concept of data preprocessing and its significance in preparing raw data for machine learning models. Appreciate the importance of data quality assurance in identifying and addressing errors, missing values, multicollinearity, and noise in the dataset to enhance performance and reduce overfitting. And you will understand the default preprocessing steps carried out by One AI, which will guide you to make educated decisions in manual configuration and per column interventions.

Chapter 2

Data Preprocessing Overview

Section 2 - Data Preprocessing Overview

Data preprocessing is the process of transforming raw data into a format that is suitable for training machine learning models. This is critical because the data format used in the front end of One Model or from other HRISs where you can export data does not typically follow the format necessary to use in machine learning.

Quality data preprocessing before model training leads to the following better outcomes. Data quality assurance. Raw data often contains errors, missing values, outliers, and noise. Preprocessing identifies and handles these issues, ensuring high quality data is used to train your models.

Also, improved model performance. Preprocessing techniques like scaling and one hot encoding help in standardizing the data. This prevents any one feature from dominating training and leads to more stable and efficient model training.

Handling missing and correlated data is also important as some algorithms do not handle null values and multicollinearity well.

Next, we have categorical variable handling. Machine learning typically relies on numerical data, but real world datasets often contain categorical variables. Techniques like one hot encoding convert categorical variables into numerical representations, allowing for more robust and realistic datasets to be used without manual work.

And finally, it reduces overfitting. Overfitting occurs when a model learns the training data too well, capturing noise and irrelevant patterns. Techniques such as dimensionality reduction simplify the data while preserving important information, which reduces the risk of overfitting.

The EDA report is a powerful window into the preprocessing of the model dataset. It shows how features were treated - numerical, date, or categorical - and whether they were scaled or one hot encoded (OHE). If a feature was one hot encoded, the report also indicates the cardinality, which is the number of categories the the feature was broken into.

Additionally, you can see which features were dropped and their drop reasons, including containing a constant value, being too correlated with another column, or containing too many null values. Each dropped feature will have a gray drop reason label next to the red dropped label with text explaining why it was dropped. We will cover this more in a later section.

Chapter 3

Data Preprocessing in One AI - Scaling & OHE

Section 3 - Data Preprocessing in One AI

Data Scaling and One Hot Encoding

One AI groups all features in the model dataset into either a numerical or categorical bucket. Numerical variables are scaled while categorical variables are one hot encoded.

Scaling is the mathematical transformation of numerical features to a common scale. This ensures all continuous features are on the same scale and prevents features with larger scales from dominating the learning process, which can lead to biased results. For example, salaries contain much larger numbers than performance ratings, so scaling ensures everything is weighted equally.

Scaling is also known as normalization. By default, One AI performs linear scaling with a standard scale type on numerical variables. This process uses subtraction and division to replace the original value with a number either between 1- and +1 or between 0 and 1. Users can manually override the scale type from standard to min max or robust or switch from a scaling strategy to a binning strategy in the continuous strategy section of the global settings configuration.

Check out the global settings module, specifically the continuous strategy section, if you would like more information.

One AI treats dates separately from other numerical features, applying a scaling technique best suited for dates. Date variables are converted to the difference from the sample date and then scaled as numeric variables.

This results in more recent dates having higher values and earlier dates having lower values, which is helpful to know when interpreting models with SHAP in storyboards. Scaled features, both numerical and date, can be identified in the Exploratory Data Analysis(EDA) report by the orange scaled label in the variable status section. One hot encoding involves splitting each node of a categorical variable into its own binary column with a value of 1 or 0. This allows machine learning algorithms to interpret categorical data without ordinality, ensuring values are treated without bias as separate variables.

This image shows how one hot encoding works. One hot encoded variables can be identified in the EDA report by the orange one hot encoded label in the variable status section.

Chapter 4

Data Preprocessing in One AI - Data Cleaning

Data cleaning

Data cleaning is the process of identifying and addressing issues in the dataset that could negatively impact the performance or reliability of the model.

One AI performs several data cleaning tasks during preprocessing, which we will discuss now.

Missing data handling. Variables containing a certain percentage of null or missing values, as defined by the null drop threshold in the global settings, are dropped and not used by the model. The default threshold is set to 5%(0.05), meaning any column with 5% or more missing values will be automatically dropped. These variables can be identified in the EDA report in the variable status section by the gray missing label next to the red dropped label with text following detailing the number of null values and the null percentage.

Constant and unique data handling

Variables containing all or nearly all the same value will be automatically dropped because they provide no discriminatory information as they do not vary across observations.

This means they offer little to no predictive power to the model. These variables can be identified in the EDA report by the gray constant label next to the red dropped label with text stating the constant value.

Conversely, categorical variables containing nearly completely unique values will be automatically dropped as well because patterns cannot be found, so they are not predictive.

These variables will have a gray unique label next to the red dropped label.

Leaking data detection

Data leakage occurs when the training data contains information about the target, which is the outcome the model is predicting, that will not be available when the model is used for predictions on new unseen data. In simpler terms, it's a cheating variable that predicts the outcome too well to be realistic.

One AI identifies leakage by generating a random forest model against the target using only the suspect feature and measuring its performance with an ROC-AUC score. A score above 0.85 is considered leakage, and the variable will be automatically dropped.

This threshold can be adjusted in the global settings by overriding the leakage performance threshold.

Variables dropped for this reason can be identified in the EDA report by the gray leakage label next to the red dropped label with text stating the ROC-AUC score and the most predictive value in the column.

Correlation feature reduction

One AI runs a correlation test to identify how correlated each predictor variable is with every other predictor variable. This process helps automatically drop all but one of the similar features to avoid multicollinearity in the model. The test detects if two or more predictor variables are too similar to exist in the same model. For example, both date of birth and age are often available to models but typically shouldn't both be selected because they are effectively the same thing and highly correlated.

The default correlation threshold is set to 0.65, but it can be adjusted in the global settings.

Each predictor variable should contribute unique information to the model. When multiple variables provide similar information, reducing dimensionality and noise by dropping all but the best-performing variable results in better interpretability for the model.

Variables dropped for this reason can be identified in the EDA report by the gray correlated label next to the red dropped label with text indicating which other variable it is highly correlated with and the degree of correlation.

Chapter 5

Data Preprocessing in One AI - Dimensionality Reduction

Dimensionality reduction

Dimensionality reduction is a technique to reduce the number of features in the model dataset while preserving the most relevant predictive information.

There's an optimal number of features a machine learning model should use, and beyond that, the performance and interpretability of the model degrade and overfitting becomes a concern.

Although it might seem logical that more features would always be better, as dimensionality increases, the need for more data points and computational power increases exponentially.

This is known as the curse of dimensionality.

Dimensionality reduction addresses these challenges by transforming the original data into a lower dimensional representation with the goal of only presenting the model with the best predictive features for selection.

One AI optimizes dimensionality through the use of filter and wrapper methods.

First, the number of features is limited using filter methods, which is defaulted to mutual information and initially tries 5, 10, and 15 features.

This approach examines the general characteristics of the features and selects the best ones using a univariate test. One AI then applies a wrapper method to the filtered features, further refining them using recursive feature elimination(RFE) with a minimum of 5 features as the default.

A wrapper method leverages a predictive model, specifically a random forest, to score each combination of features and select the best combination.

Essentially, it tests each variable selected by the filter method by excluding one at a time from the model. If the model's performance does not decrease without a particular variable, it is dropped. This method is more computationally expensive than the filter method, hence the strategy of filtering first.

If no overrides are applied, the end result will be a model with the number of selected features falling between 5 and 15.

The result summary report provides information about the specific dimensionality reduction methods used and the number of features selected by the model. It can be configured in the advanced One AI settings.

Chapter 6

Data Preprocessing for Individual Variables

Section 4 - Data Preprocessing for Individual Variables

One AI allows for configuration of preprocessing for individual variables with per column interventions.

Different columns within the same dataset may have distinct characteristics, such as varying data types, scales, distributions, and degrees of missing data or outliers.

Applying the same preprocessing to all features may not be optimal. This approach recognizes these differences between columns and optimizes the model's ability to learn and generalize patterns from the data, resulting in a more performant and understandable model.

Various treatments can be applied in One AI, including the following.

First, modifying the column or column's droppability with the choices of droppable, where One AI determines if it will be dropped based on predictive power, not droppable, where violating global settings will not result in the automatic dropping of a column, and always droppable, where the column will always be dropped.

Next, configuring how you want One AI to handle nulls for the selected column or columns. We have several null fill strategies to include mean, median, mode, backwards fill, forwards fill, pad, or you can choose a custom value to fill all nulls with.

And finally, performing type specific interventions to include categorical interventions that allow users to specify nodes to exclude from being considered as features and or specify nodes to force select as features. Additionally, continuous interventions allow users to force select variables as features and change the continuous strategy for a specific column to bin or scale. Reminder that the default is scale unless changed in the global settings.

If changed to bin here, you can specify the bin type as well. We go into more detail about these interventions to include the how to in the per column interventions module.

Chapter 7

Conclusion & Thanks

Effective preprocessing turns raw data into a format ready for model training, ensuring high quality data and better model performance. By scaling numerical features, encoding categorical features, and applying solid data cleaning techniques, you can avoid common issues and improve your models. Additionally, dimensionality reduction

and tailored preprocessing for individual variables help optimize model accuracy and interpretability. With these insights and strategies, you are now equipped to enhance data preprocessing efforts and create more accurate and reliable machine learning models. Happy modeling!