“Data Exploration” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Howdy! My name is Austin Hambrick, and I'm a machine learning engineer on the One AI team at One Model.

In previous modules, such as "Introduction to Machine Learning" and "Supervised Learning", we covered the overall machine learning process.

One critical step in this process is data exploration, where model creators analyze and understand the structure, patterns, and relationships within a dataset before building models, ensuring that the data is suitable for use.

That's what we'll explore in-depth in this module.

We will discuss an introduction to data exploration and its importance, data statistics, including descriptive statistics, distribution analysis, measures of dispersion, and correlation analysis, visual exploration, including charts in the EDA report, such as histograms, bar charts, and violin plots, and data quality checks to ensure the dataset is suitable for machine learning modeling.

After completing this module, you will understand what data exploration is and its critical role in the machine learning process, serving as a foundational step before model building. You will respect the role of descriptive distribution and correlation statistics and uncovering data patterns and relationships.

You will become familiar with the diverse array of visualization tools, including histograms, bar charts, and violin plots available in the exploratory data analysis report.

And finally, you will understand the importance of data quality checks around missing values, anomalies, and unreliable features to ensure clear and consistent data for modeling.

Chapter 2

Intro to Data Exploration

We will now give an introduction to data exploration.

Data exploration is an important early step in the machine learning process.

It involves digging into your dataset to understand its structure, relationships, and patterns before moving on to data preprocessing and model building.

This step is useful for identifying data quality issues, informing preprocessing strategies, and guiding feature selection, as the model dataset ultimately trains the model and determines which features will act as drivers, driving predictions.

Bad datasets result in irrelevant or biased drivers and inaccurate predictions.

By exploring the data before modeling, you can prevent errors, reduce biases, and ultimately maximize the model's predictive power.

However, data exploration should also continue after the model is created and run with the exploratory data analysis or EDA report providing valuable insights each time the model is run.

I will talk about the EDA report throughout this module as it's an exceptional exploratory tool, but I also recommend checking out the EDA module to learn how to leverage this report.

Data exploration is also important for interpreting your model because it helps explain model behavior and the features that drive predictions.

Without a clear understanding of your data, it's hard to know why certain features affect the model's decisions and whether biases or quality issues could lead to incorrect predictions.

If we can't understand what the model is doing or the features it's using, it's not very helpful because understanding what drives the behaviors is just as important as the predictions themselves. To highlight the importance of data exploration, consider the case where a model produced unexpected results due to insufficient examination of the dataset.

Amazon once created an AI tool to screen resumes.

They trained it on 10 years worth of resumes submitted to Amazon.

Unfortunately, the model heavily favored male candidates and even penalized resumes with terms like "women's", such as "women's chess club captain", and the bias occurred due to the significant gender imbalance in the past applicants' data.

If careful data exploration was performed, the imbalance would have been obvious from analyzing the gender distribution.

Data exploration involves utilizing data statistics, visual exploration, and data quality checks, which we will cover in detail in the next few sections.

Chapter 3

Data Statistics

Now we will talk about data statistics.

Data statistics are numerical summaries that reveal important characteristics of your dataset.

They show averages, variations, and relationships between features, which are important for making smart modeling decisions.

We will walk through the following subsets of data statistics used and how One AI supports this analysis.

Keep in mind that some categories overlap, so not everything fits neatly into just one type.

Descriptive statistics, distribution analysis, measures of dispersion, and correlation analysis are some of these types. We will now take a deeper dive into what the descriptive statistics actually are.

Chapter 4

Data Statistics: Descriptive Stats

Descriptive statistics offer a quick summary of the dataset, describing key characteristics by input feature.

Here are some tips on how to use this data.

Variable type indicates whether a feature is numerical or non-numerical.

Non-null count shows the number of non-missing values, helping you assess data quality and decide whether to drop or fill missing values.

Unique count tells you how many distinct values exist in a feature, which is particularly useful for categorical data. This can help you decide whether to treat a feature as binary or if you need to update your category size threshold.

Most frequent identifies the most common values in a feature, which highlights dominant classes in a categorical feature and possible data imbalances.

Mean is the average value of a numerical feature, providing a central tendency.

Standard deviation measures the spread of the data around the mean. A high standard deviation means the data points are more spread out, while a low standard deviation indicates that they are closely clustered around the mean. This is useful for identifying outliers.

One AI supports this analysis by allowing users to generate a data statistics report in the One AI Query Builder in the validation step. Additionally, you can view quantile statistics and descriptive statistics in the Variable Analysis section of the EDA report.

Chapter 5

Data Statistics: Distribution Analysis

Now, we will talk about distribution analysis.

Distribution analysis helps you understand the spread and shape of features, revealing skewness, unusual patterns in the data, or outliers.

The EDA reports variable analysis section allows users to view distribution globally for the whole dataset or by class label. This analysis can help you find if certain groups are over or underrepresented in the data. For categorical data, bar charts illustrate common values, composition, and encoding details like cardinality or grouped other values.

Histograms visualize the frequency distribution of numerical values.

Chapter 6

Data Statistics: Measures of Dispersion

Let's talk about measures of dispersion. Measures of dispersion use statistics to help understand the data's variability, which is how spread out or dispersed the values in the dataset are.

It measures the extent to which data points differ from each other and from their average average value. It's important for assessing the reliability of the data, determining the statistical significance, and informing decision making processes. Here are some measures that help with this analysis.

Variance and standard deviation measure how far data points are from the mean. Range is the difference between the minimum and maximum values.

Interquartile range is the range between the 25th and the 75th percentile. This data can be found in the EDA report in the Variable Analysis section under 'Statistics', as seen in the Data Statistics section.

Chapter 7

Data Statistics: Correlation Analysis

Let's talk a bit about correlation analysis. Correlation analysis identifies relationships between different features or variables.

One AI's EDA report includes downloadable correlation information, showing how each input variable is correlated with others and the target variable, which is essential for feature selection.

Chapter 8

Visual Exploration

Visual exploration uses graphs to reveal patterns, anomalies, relationships, and data quality issues that may not be easily noticeable in the raw data. Graphical representations offer an alternative view of the data, often providing a clearer understanding of complex relationships.

The EDA report in One AI provides various visualization tools for exploring your model dataset. We're going to hop into the One Model site so I can show you the various options.

The variable analysis section includes bar charts displaying common values and compositions for both numerical and categorical variables. For numerical variables, you can explore data through histograms, overlay histograms, violin plots, and other options.

Additionally, these visualizations can be viewed globally or per class label for deeper insights. In the missing value section, you can use a matrix, you can use a heatmap, or a dendrogram.

And these graphs show patterns of missing data globally or by class label. This can help assess if you need to employ null filling strategies, adjust your null drop threshold, or exclude null columns from the model dataset.

Chapter 9

Data Quality Checks

Now let's talk about how to perform data quality checks.

Throughout the data exploration process, you should be getting to know your model data set and assessing data quality. This is crucial to ensure that your model is trained in reliable and good data. As you identify issues, develop plans to address them before proceeding to model building.

There are some key steps in this process:

Handling missing values - these can be handled by fixing the data using null filling strategies or removing the features if necessary.

By default, One AI will automatically drop features of more than 5% missing values. This can be modified by changing the null drop threshold in the global settings.

Identify anomalies - when encountering anomalies in the data, try to determine their source. Is it a data entry error, a system issue, or a genuine value? Based on your findings, decide whether to remove the anomalies, transform the data, or use an algorithm designed specifically to handle them.

Drop unreliable features - when encountering features with low variance, high missing data, high correlation, or irrelevant data with no predictive power, like a random ID column, decide if you want to include or exclude it from the dataset.

One AI automatically drops variables that are highly correlated with others, have excessive missing data, or have a constant value.

By leveraging data in the One AI Query Builder and the EDA report for comprehensive data exploration, you can develop data-driven insights that directly influence your feature engineering and model building strategies, ultimately improving predictive performance.

Chapter 10

Conclusion & Thanks

In conclusion, data exploration is a vital early step in the machine learning process that ensures your dataset is well understood, reliable, and ready for model training. By using data statistics, visualization tools, and quality checks, you can identify patterns, relationships, and issues that impact predictive power.

With these insights, you'll make informed decisions on the feature selection and preprocessing strategies that optimize model performance. Approach data exploration thoughtfully as it lays the groundwork for successful machine learning outcomes. Happy modeling!