

Exploratory Data Analysis (EDA) Report

ML Models



Austin Hambrick
Machine Learning Engineer



07 Jun 2024

Topics Covered

- An overview of EDA & the significance of the EDA report
- Navigating to EDA report in One Model
- Key takeaways from each section of the EDA report
 - Overview
 - Variable status
 - Variable analysis
 - Correlations
 - Missing values
 - Sample

Learning Outcomes

You will:

- Understand the importance of the EDA process & report in promoting transparency & understanding the model dataset & feature selection
- Identify areas for improving the model dataset based on findings, such as modifying model configuration, adjusting feature selection, variables, & addressing missing data
- Confidently use EDA insights to determine if the model is suitable for deployment, understanding variable treatments & decisions made during model creation



Overview & Significance



Overview & Significance

- **Exploratory data analysis (EDA)** is a process that helps in understanding the main characteristics of the model dataset
 - Summarizes its **key features** & provides insights into its underlying structure
 - Assists in identifying **relationships** among variables
- **One AI EDA report**
 - Several data **visualizations** to aid analysis
 - Displays which variables made up the model dataset
 - Insights into variable **preprocessing**
 - Advanced variable analysis & correlations
- EDA empowers you to identify areas to **improve** the model dataset



Navigation in One Model



Navigation

- A unique EDA report is automatically generated for each model iteration in the following statuses:
 - Pending, Ignored, Deployed, Deployed & persisted
- Navigation: One AI tab > Runs > Status Label

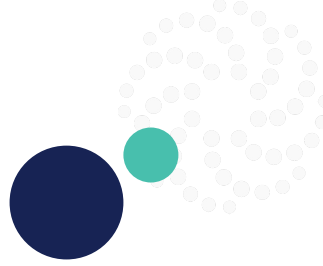
Status	Run Initiated	Result Received
Pending	2024/02/09 3:03:25 PM by Adrian Barrera	2024/02/09 3:12:07 PM

- The EDA report opens by default

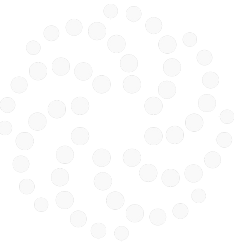


Section Takeaways of the EDA Report





Overview Section



Overview Section

Overview

Dataset info

Number of variables	150
Number of observations	2647
Missing cells	117424 (29.6%)
Duplicate rows	0 (0.0%)
Total size in memory	3.1 MiB
Average record size in memory	1.2 KiB

Variables types

Numeric	31
Categorical	109
Date	10

- Contains high-level information about the structure of the model dataset & variables
- All information contained in the EDA report is based on the **Train/Test dataset**



Variable Status



Variable Status

- Provides information about the handling of each variable in the model dataset; utilizes **colored labels** so you can easily identify:
 - If the variable was **processed** and/or **selected**
 - If the variable was **automatically dropped** by One AI, & if so, **why**
 - If the variable was marked as **suspicious**
 - How the variable was **preprocessed**

Variable Status

Suspicious **Selected!** **Processed** **Scaled** **Hire Cohort Terminations (same as base query)** has been processed as a numeric feature and scaled. This column may be leaking target data. A random forest using only this column achieved a ROC-AUC score of 0.72 against the target.

Selected! **Processed** **Scaled** **Promotions for All Time** has been processed as a numeric feature and scaled.

Selected! **Processed** **Scaled** **Team Promotions Past Year** has been processed as a numeric feature and scaled.

Selected! **Processed** **One Hot Encoded** **one.dim_age.level3name** has been processed as a categorical feature and one-hot-encoded with cardinality 8.

Dropped **Missing** **Demotions** has 2087 / 99% missing values.

Processed **Scaled** **Diverse Team Hires - Null Filled (same as base query)** has been processed as a numeric feature and scaled.

Variable Status Label Guide

- **Selected** indicates that the variable was selected by One AI to be used in the model to make its predictions

Selected! **Processed** **Scaled** `one.employee.date_of_birth` has been processed as a date feature and numerically scaled.

- **Processed** indicates that the variable was tried by One AI, & may or may not be selected; these variables conform to the model's global settings
 - **Scaled** indicates that the variable is numerical or a date so One AI transformed the variable to a common scale so that all continuous features will be on the same scale & won't get incorrectly weighted by the algorithm
 - **One Hot Encoded** indicates that the variable is categorical so One AI split each grouping into its own binary column with a value of 1 or 0 so it can be put into a format that the ML algorithm can interpret & treat without bias.

Selected! **Processed** **One Hot Encoded** `one.employee.is_future_manager` has been processed as a categorical feature and one-hot-encoded with cardinality 2.

Variable Status Label Guide

- **Dropped** indicates that the variable did not conform to a global setting so One AI had objections about the variable & left it out of the predictive model

Dropped **Missing** `Demotions` has 2087 / 99% missing values.

- **Missing** indicates that the variable contains too large of a percentage of null data as compared to the null drop threshold
- **Constant & Unique** indicates that the variables contain nearly all the same value or completely different values (categorical variables only)

Dropped **Constant** `one.dim_manager.level1name` has constant value Ronaldo Roy.

- **Correlated** indicates that the variable is too correlated, or related, to at least one other variable in the dataset

Dropped **Correlated** `Team Transfers (same as base query)` is highly correlated with `Team Promotions Past Year` ($\rho = 0.6935800789448381$).

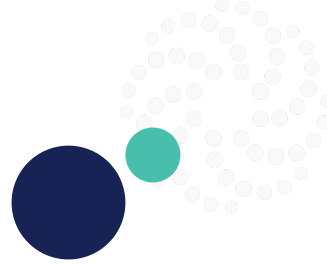
Variable Status Label Guide

- **Leakage** indicates that the target for the model (the thing the model is predicting) is likely “leaking” data into the variable, meaning it’s a variable that predicts the outcome too well to be plausible so it is considered a cheat column

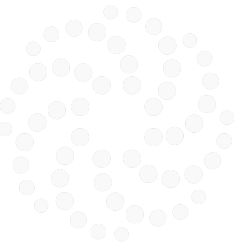
Dropped **Leakage** `one.dim_is_future_termination.level1name` has been dropped due to data leakage. A random forest using only this column achieved a ROC-AUC score of 0.98 against the target. The most predictive value within the column was No, comprising 51% of the feature importance.

- **Suspicious** indicates that there is possible data leakage
 - It is a less stringent version of the test performed to check for data leakage.
 - The default threshold is 0.75 vs. the 0.85 for data leakage, & these variables are **not** automatically dropped, but simply flagged for further analysis

Suspicious **Selected!** **Processed** **Scaled** `Hire Cohort Terminations (same as base query)` has been processed as a numeric feature and scaled. This column may be leaking target data. A random forest using only this column achieved a ROC-AUC score of 0.72 against the target.

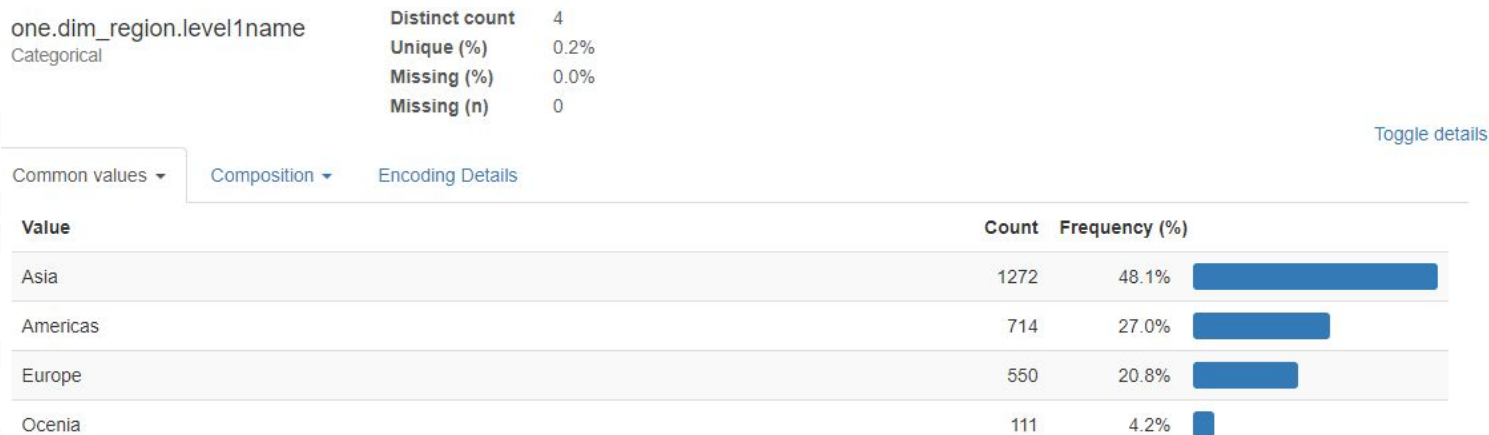


Variable Analysis



Variable Analysis

- Provides more information about the data the variable contains
- You can analyze the variable globally (data for the entire model) or by individual outcomes (i.e. new hire failure or success) to see how data differs in instances in each outcome









Correlations



Correlations

- Correlation is the degree to which 2+ variables are associated or related
- Can download a zip file of all your model's correlations
 - Each file shows how correlated each variable is with each of the other variables that it can be compared to

Name	Type
 crammers_one.dim_age.level3name	Microsoft Excel Comma S...
 crammers_one.dim_annual_salary_range.level1name	Microsoft Excel Comma S...
 crammers_one.dim_country.level1name	Microsoft Excel Comma S...
 crammers_one.dim_eeo_job.level1name	Microsoft Excel Comma S...

1		Promotions - Null Filled
2	Terminations	0
3	Promotions for Previous 1 Year_cm_Binary	1
4	one.employee.annual_salary_range	0
5	one.employee.city	0.010055793
6	one.employee.corrective_action	0.017568985
7	one.employee.country	0.026485944
8	one.employee.diverse	0



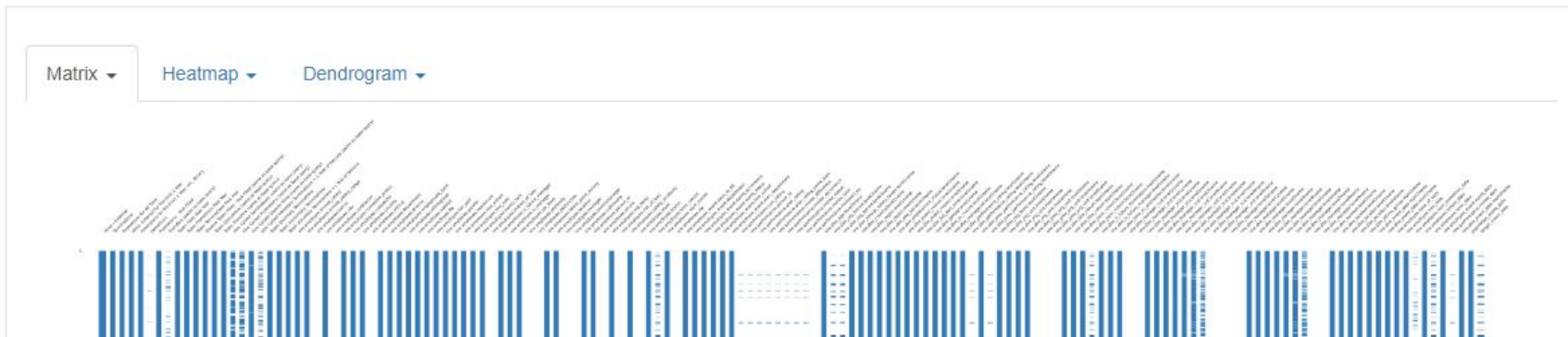
Missing Data



Missing Values

- Provides a number of views highlighting NULL values in variables
 - Helpful in determining areas where additional data preparation would be beneficial

Missing values





Sample

Sample

- Provides a view of the first & last 5 rows in the dataset
- Useful for quick data inspection to identify obvious issues

First rows

one.dim_manager_scd.level3name	one.dim_manager_scd.level4name	one.dim_manager_scd.level5name	one.dim_manager_scd.level6name	one.dim_manager_scd.level7name
Kylee Cope	NaN	NaN	NaN	NaN
Charley Davenport	Samara Rader	Samir Sarah	Gordon Sprague	NaN
Iris Boyer	Jade Sellers	Daisy Goad	Edwin Hines	Zoe Tompkins
Emanuel Reed	Lorelai Dickerson	Laylah Austin	Maggie Harrell	NaN



Thanks for watching!

