OneModel Academy

One AI

Data Preprocessing

ML Models

Hayley Bresina
One AI Client Enablement

21 May 2024

Topics Covered

- Overview of data preprocessing
- How data preprocessing works in One AI machine learning models
  - Data scaling & one hot encoding
  - Data cleaning
  - Dimensionality reduction
- Data preprocessing for individual variables

Learning Outcomes

You will:

- Grasp the concept of data preprocessing & its significance in preparing raw data for machine learning models
- Appreciate the importance of data quality assurance in identifying & addressing errors, missing values, multicollinearity, & noise in the dataset to enhance performance & reduce overfitting
- Understand the default preprocessing steps carried out by One AI to make educated decisions on manual configuration & per column interventions

Data Preprocessing Overview

Overview

- **Data preprocessing** occurs before machine learning begins & is the process of transforming raw data into a suitable format for training ML models
- Quality data preprocessing results in better ML outcomes:
  - Data quality assurance
  - Improved model performance
  - Categorical variable handling
  - Reduces overfitting
- The EDA Report is a great window into this process

Data Preprocessing in One AI

Data Scaling & One Hot Encoding (OHE)

Scaling

- The mathematical transformation of numerical features to a common scale so all continuous features will be on the same scale & thus won't get incorrectly weighted by the algorithm
    - One AI Default: standard linear scaling
- **Date** features are separated out & scaled with a date-suited technique
    - More recent dates have a higher value
    - Earlier dates have a lower value
- **Scaled** features are labeled on the EDA report 

One Hot Encoding (OHE)

- Involves splitting out each node of a categorical variable into its own binary column with a value of 1 or 0 to be interpreted by ML algorithms & avoid bias

Feature (Color)
Red
Green
Yellow
Green
Red

One Hot Encoding →

One Hot Encoded Vector
[1,00]
[0,1,0]
[0,0,1]
[0,1,0]
[1,00]

Red	Green	Yellow
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

- **One hot encoded** features are labeled on the EDA report

One Hot Encoded

Data Cleaning

Data Cleaning

- Dropped Missing — **Missing data handling**: variables containing a certain percentage of null data will be automatically dropped & labeled on the EDA
- Dropped Constant / Dropped Unique — **Constant & unique data handling**: variables containing nearly all the same value or completely different values (categorical variables only) will be automatically dropped & labeled on the EDA
- Dropped Leakage — **Leaking data detection**: variables with data leakage (based on ROC-AUC scores) will be automatically dropped & labeled on the EDA
- Dropped Correlated — **Correlation feature reduction**: variables that are highly correlated with other predictor variables will be automatically dropped and labeled on the EDA except for the most performant one

Dimensionality Reduction

Dimensionality Reduction

- Technique to reduce the number of features in the model dataset while preserving the most relevant information
  - Less is more
  - One AI optimizes dimensionality through the use of filter & wrapper methods
  - Default settings result in the model selecting 5-15 features
  - Configuration can be viewed in the Results Summary Report

Data Preprocessing for Individual Variables

Per Column Interventions

- Making changes related to preprocessing specifically for individual columns or features in a dataset
- This approach recognizes that different columns have distinct characteristics (data types, scales, degrees of outliers, etc.)
- Per Column Interventions in One AI
  - Droppability
  - Null fill strategy
  - Type-specific intervention

OneModel Academy

Thanks for watching!

OneAI