NeuroTune

Machine Learning Tutorials

Data Preparation: Handling Numeric and Categorical Variables

Proper preprocessing ensures models are stable, interpretable, and able to generalize to new data, which is essential for accurate prediction and reliable feature selection.

Before applying regression or regularization methods, it is important to properly clean, transform, and split the data.

The steps performed in this tutorial include:

1. **Data Cleaning**
   - Remove rows with missing values to avoid errors during modeling.
   - Check dimensions before and after cleaning.

2. **Target Transformation**
   - Log-transform the Salary variable to reduce skewness and stabilize variance.

3. **Feature Engineering**
   - Convert categorical variables (e.g., League, Division, NewLeague) into numeric dummy variables using one-hot encoding.
   - Combine all features into a single matrix for modeling.

4. **Data Splitting**
   - Randomly split the data into training (60%), validation (20%), and test (20%) sets.
   - The training set is used to fit the models, the validation set to tune hyperparameters (for Lasso, Ridge, Elastic Net), and the test set to evaluate final performance.

5. **Scaling and Centering**
   - Scale features using the mean and standard deviation from the training set to ensure consistent units and improve numerical stability.
   - Center the target variable using the training mean, which is useful for regularization methods that assume zero-centered response.

6. **Output**
   - The resulting matrices (X_train, X_valid, X_test) and vectors (y_train, y_valid, y_test) are ready to be used with linear regression, Lasso, Ridge, Elastic Net, or other machine learning models.