Supervised Learning-Regression

Metrics for Regression

Regression is a type of Machine learning algorithm which helps in finding the relationship between independent and dependent variables.

Before learning about precise metrics, let's familiarize ourselves with a few essential concepts related to regression metrics:

1. True Values and Predicted Values:

In regression, we've got two units of values to compare: the actual target values (authentic values) and the values expected by the model (predicted values). The performance of the model is assessed by means of measuring the similarity among these sets.

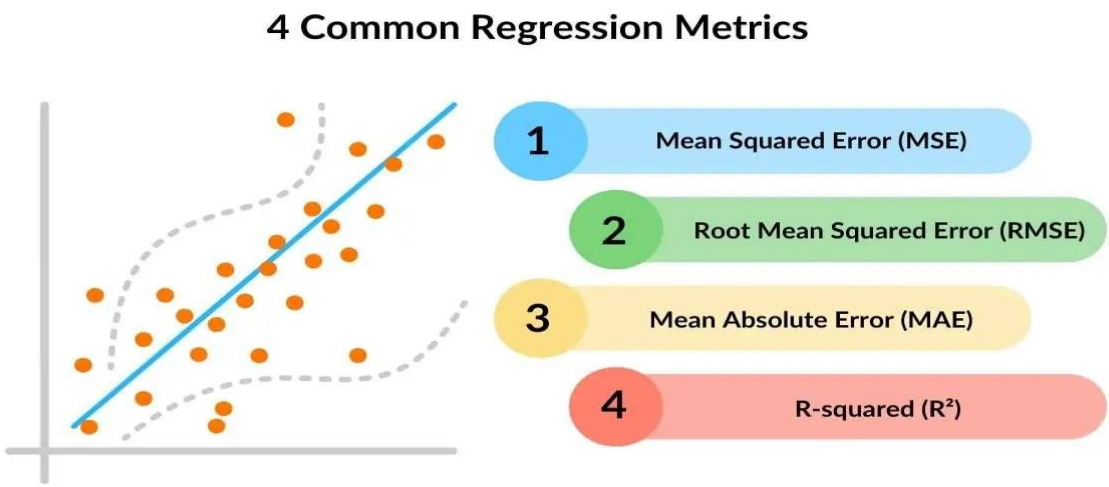
2. Evaluation Metrics:

Regression metrics are used to evaluate the performance of regression models by quantifying how well a model's predictions match actual values. Evaluation Metrics for regression are essential for assessing the performance of regression models specifically. These metrics help in measuring how well a regression model is able to predict continuous outcomes.

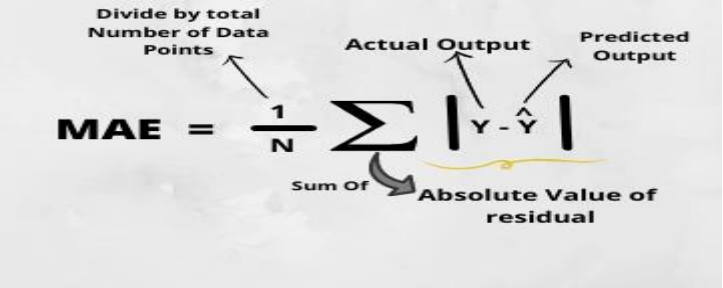
Types of Regression Metrics

Some common regression metrics in scikit-learn with examples

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2) Score
- Adjusted R-Squared



1. **Mean Absolute Error (MAE)**



- **Description**: MAE is the average of the absolute differences between the predicted and actual values. It measures how close predictions are to the actual outcomes.
- **Interpretation**: Lower MAE values indicate better model performance, as it implies smaller errors on average.
- **Sensitivity**: It is less sensitive to outliers than other metrics because it doesn't square the errors. However, it may undervalue large errors.

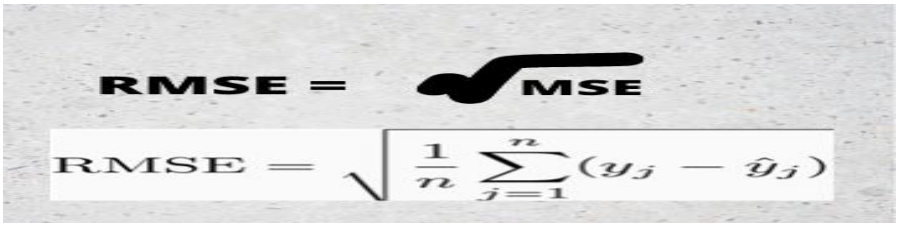
2. **Mean Squared Error (MSE)**

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

- **Description**: MSE is the average of the squared differences between actual and predicted values. Squaring the errors penalizes larger errors more than smaller ones.
- **Interpretation**: Lower MSE values are preferred. It emphasizes larger errors, which can be useful if large errors are particularly undesirable in your application.
- **Sensitivity**: Highly sensitive to outliers because errors are squared, causing models with high variance to have higher MSE values.

3. Root Mean Squared Error (RMSE)



- **Description**: RMSE is the square root of MSE. It provides the error in the same units as the dependent variable y, making it more interpretable.
- **Interpretation**: Lower RMSE indicates better model performance. It is especially useful when large errors should be heavily penalized.

- **Sensitivity**: Like MSE, it is sensitive to outliers. However, RMSE is more interpretable than MSE because it is in the original units of the outcome variable.

4. R-Squared (Coefficient of Determination)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Description**: R^2 represents the proportion of variance in the dependent variable explained by the model. It ranges from 0 to 1.
- **Interpretation**: Higher values (closer to 1) indicate that the model explains a large proportion of the variance. An R^2 of 0 means the model does not explain any variance, while 1 indicates perfect prediction.
- **Sensitivity**: Sensitive to data variability. While useful, it does not provide a measure of error magnitude, and its value can be misleading, especially with high-dimensional or overfitted models.

5. Adjusted R-Squared

Formula:

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\,\frac{n-1}{n-k-1}$$

- **Description**: Adjusted R^2 modifies R^2 by adjusting for the number of predictors in the model. This avoids overestimating the model's performance when additional predictors are added.
- **Interpretation**: Higher adjusted R^2 values suggest a better model fit while penalizing for unnecessary predictors.
- **Sensitivity**: Adjusted R^2 is particularly useful when comparing models with different numbers of predictors, as it accounts for added complexity.

<u>Choosing the Right Metric</u>

- **Interpretability**: Metrics like RMSE and MAE provide errors in actual units, making them interpretable.
- **Outliers**: If the data has outliers, MAE or MSLE may be preferable as they are less sensitive than MSE and RMSE.
- **Percentage-Based Accuracy**: Use MAPE or SMAPE for models where percentage accuracy is meaningful and the dataset doesn't contain zero or near-zero values.
- **Model Comparison**: Adjusted R2R^2R2 and Explained Variance Score are valuable when comparing models, especially with different numbers of predictors.

Selecting the best metric will depend on your specific goals, data properties, and the application context.