How to evaluate the performance of a regression model?

Evaluating the performance of a regression model is crucial to determine how well it can predict continuous numeric values. There are several metrics and techniques you can use to assess the performance of your regression model. Here are some common methods:

Mean Absolute Error (MAE):

MAE calculates the average absolute difference between the predicted and actual values. It gives you a measure of the model's average prediction error.

MAE = (1 / n) * Σ |actual - predicted|

Mean Squared Error (MSE):

MSE calculates the average squared difference between the predicted and actual values. It penalizes larger errors more heavily than MAE.

MSE = (1 / n) * Σ (actual - predicted)^2

Root Mean Squared Error (RMSE):

RMSE is the square root of the MSE. It provides a measure of the typical magnitude of the prediction error in the same units as the target variable.

RMSE = sqrt(MSE)

R-squared (R^2) / Coefficient of Determination:

R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating better fit.

$$R^2 = 1 - (SSR / SST)$$

where SSR is the sum of squared residuals (MSE * n) and SST is the total sum of squares (variance of the target variable).

Mean Absolute Percentage Error (MAPE):

MAPE calculates the average percentage difference between the predicted and actual values. It is often used in scenarios where you want to measure prediction accuracy as a percentage of the actual values.

$$MAPE = (1 / n) * \Sigma (|actual - predicted| / |actual|) * 100\%$$

Mean Percentage Error (MPE):

MPE calculates the average percentage difference between the predicted and actual values. It can be used to understand the direction of prediction errors (overestimation or underestimation).

$$MPE = (1 / n) * \Sigma ((actual - predicted) / actual) * 100\%$$

Median Absolute Error (MedAE):

MedAE is the median of the absolute differences between the predicted and actual values. It is less sensitive to outliers compared to mean-based metrics like MAE.

Coefficient of Variation (CV):

CV measures the ratio of the standard deviation of the predicted values to their mean. It helps assess the stability and dispersion of predictions.

$$CV = (\sigma\_predicted / mean\_predicted) * 100\%$$

Residual Analysis:

Plotting the residuals (the differences between actual and predicted values) can provide insights into the model's performance. Look for patterns or trends in the residuals, which can indicate model deficiencies.

Cross-Validation:

Use techniques like k-fold cross-validation to assess how well your model generalizes to unseen data. Cross-validation helps detect overfitting and provides a more robust estimate of model performance.

Visualizations:

Visualizing the actual vs. predicted values, as well as residual plots, can help you gain a better understanding of your model's behavior.

The choice of evaluation metric should align with the specific goals of your regression task. For example, if you want to prioritize minimizing large errors, you might focus on RMSE. If you want to assess the overall fit of the model, R-squared can be informative. Additionally, it's a good practice to consider multiple metrics to get a comprehensive view of your model's performance.