# Predictions in Energy Economics - Which Error Metrics Are Suitable?

Accurate predictions play an increasingly important role in the energy sector, as the volatility of renewable energy presents challenges to the energy system. To avoid bottlenecks and network overload, the flexibilities within the energy system need to be effectively utilized. This requires not only accurate predictions for energy production, demand, and prices but also for the flexibilities themselves. It is crucial to predict how long an electric vehicle will remain connected to utilize its flexibility optimally or when the heat from a heat pump will be needed to use it at the right time. In this regard, one can use supervised machine learning methods (among other applications in the energy industry).

Evaluating the accuracy of predictions is essential in this context. Specific metrics are used to describe the deviations between forecasts, and actual events and to estimate future deviations. Different metrics are suitable for various accuracy assessments. In the following document, we provide an overview of some metrics, their advantages and disadvantages, and their specific application areas. There are two types of tasks that a prediction can fulfill: regression and classification. Regression involves predicting continuous values, while classification involves predicting group memberships. In a subsequent article, we will discuss the selection of the appropriate method for time series forecasting.

## Metrics for Regression

Typical regression tasks in the energy industry include predicting PV or wind energy generation, energy prices, or load profiles. In the following, we will refer to the target variable as $y_i$ and its prediction as $\hat{y}_i$.

### Root-Mean-Square Error (RMSE)

The root-mean-square error (RMSE) is the most commonly used metric for regression problems. It is the square root of the mean squared error of the predictions. Therefore, the RMSE is always positive, and the smaller it is, the better the prediction. In mathematical notation, for $N$ predictions:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(\hat{y}_i - y_i)^2}\,.$$

The RMSE has the same unit as the target variable $y_i$. The Mean Square Error (MSE) is often used as the loss function to be minimized during the training of machine learning models, as its derivative is zero at its minimum, making it easier to find local minima. However, the RMSE is strongly influenced by the scale and the dataset. It can be normalized to make predictions with different scales more comparable. One way to normalize the RMSE is

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{y_{max} - y_{min}}.$$

The RMSE and NRMSE are strongly influenced by outliers, as they are heavily weighted in the summation due to squaring. On the other hand, squaring can result in an underestimation of errors smaller than one. Therefore, the RMSE should be used as a metric when outliers must be heavily penalized. [1,2]

### Mean Absolute Error (MAE)

The mean absolute error (MAE) is the average magnitude of the prediction error. For N predictions, the MAE can be calculated by

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=0}^{N}|\hat{y}_i - y_i|\,.$$

By utilizing the absolute value, the MAE (Mean Absolute Error) is robust against outliers, unlike the RMSE. Similar to the RMSE, it shares the same unit as the variable $y_i$ and is even easier to interpret since no square is included. The MAE can be used as a loss function; however, it is not differentiable at zero and remains constant when approaching zero, which makes finding minima more challenging. Like the RMSE, the MAE can be normalized to ensure better comparability across different scales or datasets. One approach is to compare the MAE of the prediction with the MAE of a naive forecast, resulting in the calculation of the mean absolute scaled error (MASE)

$$\mathrm{MASE} = \frac{\mathrm{MAE}}{\mathrm{MAE}_{naive}}.$$

The better the predictions, the smaller the MASE value. Predictions that outperform the naive method have a MASE value smaller than 1. An example of a naive prediction is using the value from the previous day. Other options include using average values or linear regression as a simple benchmark model. An alternative method for normalizing the MAE is to use the percentage deviation error, known as the mean absolute percentage error (MAPE)

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=0}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|.$$

The MAPE is scale-independent and easy to interpret. However, it produces very large values for $y_i$ close to zero and is asymmetric, as errors with $\hat{y}_i > y_i$ are penalized more heavily than for $\hat{y}_i < y_i$. [1 3,4,5]

## Huber-Loss

As a loss function, the MAE is suitable when outliers occur in the data but should not be crucial for the prediction. On the other hand, the RMSE has the advantage of being differentiable at zero, making it better suited for finding minima. To combine the benefits of RMSE and MAE, one can use the Huber Loss $\mathrm{HL}_\delta$. This loss function is linear concerning the MSE for deviations more minor than $\delta$ and linear concerning the MAE for more significant deviations:

$$\mathrm{HL}_\delta(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2}(\hat{y}_i - y_i)^2, & \text{if } |\hat{y}_i - y_i| \le \delta \\ \delta|\hat{y}_i - y_i| - \frac{1}{2}\delta^2, & \text{else.} \end{cases}$$

However, one drawback of the Huber Loss is the additional hyperparameter $\delta$ that can be optimized. [6]

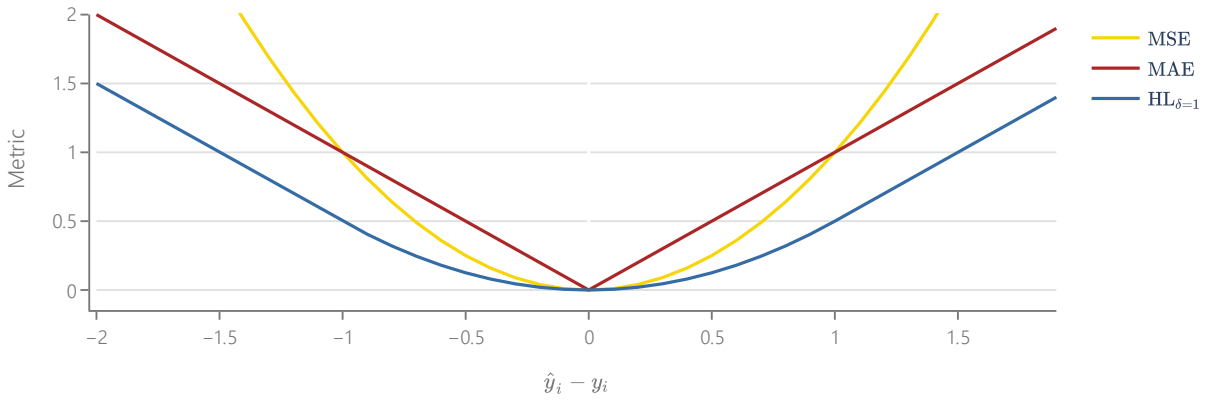


Figure 1: Comparison between MSE, MAE, and Huber Loss.

### $R^2$

The $R^2$ score compares the MSE of the predictions with the MSE that would be obtained if the average value $\overline{y}$ were used as the prediction:

$$R^2 = 1 - \frac{\sum_{i=0}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=0}^{N}(\overline{y} - y_i)^2}.$$

The $R^2$ score assigns a value of 0 to the model that always predicts the average value, and the closer the $R^2$ score is to 1, the better the model performs. Very poor predictions can yield negative values for the $R^2$ score. So far, none of the metrics have considered the number of additional variables (referred to as features) used to make a prediction. Each additional feature adds complexity to the model, which should be penalized if the predictions do not improve. This is done in the adjusted $R^2$ score ($R^2_{adj}$),

$$R^2_{adj} = 1 - \frac{N-1}{N-k-1}(1 - R^2),$$

where $N$ is once again the number of predictions, and k is the number of features. Therefore, if features are added that do not significantly improve $R^2$, the $R^2_{adj}$, will deteriorate. This allows for the inclusion of only those features that actually enhance the model. [1,3]

### Quantile-Metrics

In many cases, it is not crucial for a prediction to be exact, but rather it is essential to determine boundaries within which the target variable should fall. Quantiles are used to establish these boundaries. To evaluate the accuracy of the quantiles, appropriate metrics are needed. One approach is to determine the percentage of the target variable that falls within the predicted quantiles. This metric is here referred to as II („inside interval")

$$II = \frac{q_{in}}{N},$$

where $q_{in}$ is the number of measurements with $y_i \in [\hat{q}_{i,low}, \hat{q}_{i,up}]$ where $\hat{q}_{i,low}$ represents the lower quantile and $\hat{q}_{i,up}$ represents the upper quantile. However, evaluating the II is not straightforward. For example, an II of 1 can be achieved for a positive target variable by using the quantile interval $[0, \infty)$, without providing a meaningful constraint. For a large number of $N$, the II of a good model should correspond to the quantile range. Let $q_{low=5\%}$ be the 5th percentile and $q_{up=95\%}$ be the 95th percentile; the II should yield a value of 90%. A more interpretable metric is the Pinnball-Loss $PL_\tau$, which evaluates a quantile prediction separately. For a specific quantile prediction $\hat{q}_{i,\tau}$, it is calculated as

$$PL_\tau(y_i, \hat{q}_{i,\tau}) = \begin{cases} |\hat{q}_{i,\tau\, y_i} - y_i|\tau, & \text{if } y_i \geq \hat{q}_{i,\tau} \\ |\hat{q}_{i,\tau} - y_i|(1-\tau), & \text{if } \hat{q}_{i,\tau} \geq y_i, \end{cases}$$

where $\tau$ represents the target quantile. A lower Pinnball-Loss indicates a better quantile prediction. In this metric, deviations towards the median are penalized less. The Pinnball-Loss is commonly used as a loss function for quantiles. [7]

## Metrics for Classification

Classification tasks in energy economics include, for example, determining whether a network bottleneck occurs or whether a vehicle is connected to a charging station. In binary classification, a confusion matrix can be used to differentiate between different types of errors.

| | Reality | |
|---|---|---|
| | Positive | Negative |

| | | | |
|---|---|---|---|
| Prediction | Positive | True Positive (TP) | False Positive (FP) |
| | Negative | False Negative (FN) | True Negative (TN) |

Table 1: Confusion Matrix of a binary classification.

### Accuracy

The accuracy $\mathrm{A}$ indicates the proportion of correct predictions:

$$\mathrm{A} = \frac{n_{TP} + n_{TN}}{N},$$

where $N$ s the total number of predictions, and $n_{TP}/n_{TN}$ is the number of true positives/true negatives. However, even predictions with high accuracy may not be helpful. For example, if we want to predict the connection times of a vehicle that is only connected 10% of the time, a prediction that the vehicle is never connected would also have 90% accuracy. One way to improve the interpretation of accuracy is to differentiate it by classes. In this case, the previous example would have 100% accuracy for the disconnected time but 0% for the connected time. [1,8]

### Sensitivity and Precision

Another approach is to use two metrics: sensitivity $\mathrm{S}$ (often referred to as recall) and precision $\mathrm{P}$ , which are defined by:

$$\mathrm{S} = \frac{n_{TP}}{n_{TP} + n_{FN}}, \mathrm{P} = \frac{n_{TP}}{n_{TP} + n_{FP}}.$$

Sensitivity represents the proportion of true positive cases correctly predicted as positive. On the other hand, precision indicates the proportion of positive predictions that are actually true positives. The choice between sensitivity and precision depends on the goal of the prediction. Achieving a high sensitivity is more important for preventing shortages, as it is better to intervene too often than too little. However, in scenarios where planning is based on the flexibility of a vehicle, high precision may be more crucial. For example, a vehicle must be connected when its flexibility has been taken into account. Often, a compromise between high sensitivity and precision is desired. [1,8]

### F1

The F1 score represents the compromise between sensitivity and precision and ranges from 0 to 1, with 1 indicating a perfect prediction. The F1 score is calculated as the harmonic mean of sensitivity and precision:

$$\mathrm{F1} = 2\,\frac{\mathrm{S} * \mathrm{P}}{\mathrm{S} + \mathrm{P}}.$$

By maximizing the F1 score, a good balance between high sensitivity and precision is achieved. Using the example of predicting whether a vehicle is connected or disconnected, a high F1 score indicates that the prediction accurately captures the times when the vehicle is connected. When it predicts the vehicle as connected, it rarely makes incorrect predictions. [8]

### Cross-Entropy

Most machine learning methods provide probabilities for different events in classification tasks. In binary classification, where the event occurred $y_i \in \{0,1\}$ and $\hat{y}_i$ represents the predicted probability of $y_i = 1$ occurring. One way to evaluate these probabilities is through the use of log-loss, which is often referred to as binary cross-entropy loss:

$$\text{LogL} = \frac{1}{N}\sum_{i=0}^{N} -y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i),$$

where $\hat{y}_i$ represents the predicted probability of event 1 occurring. For each individual prediction, the log-loss is the negative logarithm of the likelihood function. The logarithm is used here to avoid working with very small numbers. The negative sign is used so that smaller values indicate better predictions, resulting in a minimization problem. The binary cross-entropy loss can also be generalized to multiple classes, known as cross-entropy loss. Therefore, it can serve as a metric for classification tasks with more than two classes. [1,9]

### Brier-Score

The Brier score also measures the agreement between predicted probability $\hat{y}_i$ and if the class occurred. For binary classification, it is calculated in a similar way as the MSE for regression problems:

$$\text{Brier} = \frac{1}{N}\sum_{i=0}^{N} (\hat{y}_i - y_i)^2.$$

A smaller Brier score indicates a better prediction, and similar to log-loss, it can be used as the loss function to minimize during training of machine learning methods. [10]

## Literature


[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

[2] Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting. Springer.

[3] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

[4] R. Hyndman (2006). Another look at forecast accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting, vol. 4, pp. 43–46.

[5] Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. International Journal of Forecasting, 9(4), 527-529.

[6] Huber, P.J. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35(1), 73-101.

[7] Koenker, R., & Bassett, G. (1978). Regression quantiles. Econometrica, 46(1), 33-50.

[8] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[9] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[10] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3.