Invited Review: A Guide on How to Avoid Common Pitfalls in Model Performance Metrics and Validation

# Chun-Peng James Chen

# Introduction

Modeling is an indispensable tool for hypothesis formulation and decision-making. It functions as a structured framework that validates system understanding through the analysis of empirical data, and further extends this understanding by enabling the extrapolation of results to novel trials and conditions. The advancement of research is fundamentally built upon the accumulation of prior knowledge within the scientific community. Therefore, the evaluation of model performance becomes particularly critical, necessitating a rigorous and standardized approach that allows for both reproducibility and comparability. The failure to adhere to these standards, by reporting model performance through ill-defined metrics or non-rigorous procedures, has the potential to introduce misinterpretations and miscommunications. Such lapses not only impede scientific progress but can also compromise the integrity of the collective body of research in the field.

This review aims to scrutinize four common pitfalls in model evaluation. The first pitfall is the failure to validate model performance on unseen or test datasets; the second involves indiscriminate feature selection across the full dataset; the third emerges when tuning hyperparameters with the same data designated for model evaluation; the fourth is neglecting block effects in model validation. In each section, current relevant practices in dairy science are examined to offer practical examples for clearer illustration. To underscore the importance of these pitfalls, a minimally reproducible simulation is presented to demonstrate their impact on model evaluation.

This review tried to answered commonly asked questions. For example, how to choose between leave-one-out cross validation and five-fold cross validation? How many samples are enough to unbiasedly estimate a model performance? There is no single approach can applied to all research problem, this review aims to provide a comprehensive perspective of evaluating models with different types of metrics and validation strategy.

# Performance Metrics: Regression

Performance metrics serve as quantitative indicators for evaluating model performance. They are essential tools for benchmarking various modeling approaches and validating hypotheses. However, it's important to note that the choice of metrics can significantly influence the evaluation results. Depending on the specific hypothesis being tested, an overly optimistic assessment may be concluded if inappropriate metrics are selected. This section aims to introduce commonly used performance metrics in the field of dairy science and discuss potential pitfalls that researchers should be cautious of.

A regression model aims to predict a continuous variable and is commonly employed in various applications, such as estimating milk composition, yield, and feed efficiency, as well as assessing environmental impacts in livestock production. This section delves into three widely used metrics for evaluating regression models: Pearson's Correlation Coefficient ($r$), Root Mean Squared Error (RMSE), and the Coefficient of Determination ($R^2$).

A group of graphs with numbers and points

Description automatically generated with medium confidence

In the hypothetical example depicted in Figure 1, 100 observations are generated from two separate normal distributions. The first 50 observations are drawn from a normal distribution with a mean of -3 and a standard deviation of 1, denoted as \( \mathcal{N}(-3, 1) \). The remaining 50 observations are generated from another normal distribution, \( \mathcal{N}(3, 1) \). Utilizing two distinct distributions serves to simulate experimental block effects, preset at a magnitude of 6 units for this experiment.

Based on these simulated observations, four sets of predictions are generated:

* **First Prediction**: Each observation is multiplied by 0.3 to establish a correlation, followed by the addition of random noise \( \mathcal{N}(0, 0.7) \) to introduce prediction errors.
* **Second Prediction**: Values from the first prediction are multiplied by 5, simulating predictions with a larger variance while maintaining the same relative ordering with the original predictions.
* **Third Prediction**: Each value from the first prediction is raised to the power of 5. This transformation serves to compress values within the range of -1 to 1 towards zero, without affecting their relative order. Additionally, values greater than 1 or less than -1 are pushed farther from zero, simulating outlier predictions.
* **Fourth Prediction**: Values sampled from \( \mathcal{N}(-3, 1) \) and \( \mathcal{N}(3, 1) \) are added to the first and second blocks of observations, respectively. This amplifies the block effects, simulating a model that effectively distinguishes between different blocks but is less capable of predicting individual variations within each block.

This quartet of predictions serves to simulate potential challenges and complexities encountered in real-world modeling scenarios, thereby providing a foundation for evaluating different performance metrics used in regression problems.

Explain the intension of each case

## Coefficient of Determination ()

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

The Coefficient of Determination, commonly known as R-squared \( R^2 \), provides a similar measurement of the SSE as RMSE, but it is normalized by the total sum of squares (SST) to yield a more versatile metric for comparing results across different experiments. As defined in Equation \(3\), \( n \) represents the number of observations, \( y \) signifies the actual observed values, \( \hat{y} \) indicates the predicted values, and \( \bar{y} \) is the mean of the observed values. A higher \( R^2 \) value suggests a lower prediction error. Importantly, if \( R^2 \) falls below zero, the predictions are deemed inferior to a naive approach that predicts the mean of the observed values for all samples, particularly in terms of absolute prediction errors. Similar to RMSE, \( R^2 \) is influenced by the variance of the predictions. Specifically, when predictions exhibit a high degree of variance that deviates from the actual observations, the \( R^2 \) value can be adversely impacted, dropping from 0.47 in Prediction 1 to -0.27 in Prediction 2, for example. Additionally, due to the presence of squared terms in its calculation, \( R^2 \) is sensitive to outliers, as evidenced in Prediction 3.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

The fourth prediction scenario offers an intriguing case study for \( R^2 \) to identify misleading model performance due to block effects. According to Equations \(4\) and \(5\), the SST can be decomposed into the Regression Sum of Squares (SSR), which represents the variation explained by the model, and the SSE, which accounts for the unexplainable errors as discussed in the RMSE section. From this standpoint, \( R^2 \) quantifies the proportion of variation that the model is able to explain. In the case of the fourth prediction, the model effectively captures the differences between various blocks, which contributes substantially to the SST. Consequently, the \( R^2 \) value rises to as high as 0.80. However, this elevated metric can be misleading. The hypothetical model in the fourth prediction does not possess the ability to capture individual variations within each block. The strength of \( R^2 \) lies in its capability to differentiate between sources of variation. Upon closer inspection of the data within each block, the \( R^2 \) values plummet to -0.71 and -1.10, respectively, highlighting that the model fails to account for intra-block variability. In summary, while both RMSE and \( R^2 \) aim to measure prediction errors, \( R^2 \) offers additional statistical insights that facilitate a more nuanced evaluation of model performance.

Current practices in dairy science usually report both R2 and RMSE, and sometimes r correlation. as exemplified in <example *1*>, <example *2*>.

## Pearson Correlation Coefficient (r)

|  |  |  |
| --- | --- | --- |
|  |  | () |

The Pearson Correlation Coefficient (\(r\)) is a widely used metric for assessing the linear relationship between two variables, as defined by Equation \(1\). In this equation, the observed values are denoted by \(x\) and the predicted values are represented by \(y\). To calculate the correlation coefficient, the equation first computes the covariance between \(x\) and \(y\), which captures how the coordinates of data points deviate from their means. This value is then normalized by the product of the standard deviations of \(x\) and \(y\). The coefficient can be either positive or negative, reflecting a positive or negative correlation between the two variables, respectively. Owing to the normalization term in the denominator, the coefficient is scale-invariant and will always fall within the range of -1 to 1. This attribute is illustrated in the first two scenarios: despite the second scenario having a variance five times larger than the first, the coefficient remains consistent.

However, the metric has its limitations. It is sensitive to the presence of outliers, as shown in the third scenario where most data points cluster near zero but a few outliers yield a high correlation coefficient. Similarly, the metric is influenced by block effects, leading to inflated correlation values, as observed in the fourth scenario. When calculated within individual blocks, the coefficients drop to 0.11 and 0.06, respectively. This emphasizes the importance of either visually inspecting regression results through scatter plots or examining them within individual blocks. If the objective is to examine the model predictability of individual variation rather than block effects, caution should be exercised when applying this metric, especially in scenarios where block effects are evident.

In practice, this coefficient is often utilized to evaluate a model capability to correctly order or rank individual observations based on a specific trait of interest. For instance, <example *1*>, <example *2*>, among others [ref 3-6].

## Root Mean Squared Error (RMSE)

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  |  |

The Root Mean Squared Error (RMSE) serves as a quantitative measure to gauge the average magnitude of prediction errors between observed values (\(y\)) and their predicted values (\(\hat{y}\)). It gives the error in the same units as the observation (\(y\)), and a lower RMSE value indicates a better model performance. Defined by Equation \(2\), \(n\) stands for the number of observations. Distinct from the correlation coefficient, RMSE is sensitive to scale, implying that achieving predictions with a variance akin to the observed values takes precedence over maintaining their order or trend. This is particularly pertinent when the focus is on the absolute magnitude of the error. Take for instance Scenario 2, where the predictions have been scaled by a factor of 5 compared to Scenario 1. The RMSE inflates from 2.41 to 3.63, underscoring that even if both scenarios rank the observations identically, RMSE effectively captures the discrepancies in the absolute errors. Another notable characteristic of RMSE is its sensitivity to outliers. In Prediction 3, where certain predictions deviate substantially from the majority, the squaring operation within RMSE accentuates these outliers, culminating in a RMSE value of 25.56. It's also worth mentioning that RMSE is impervious to block effects, unlike the correlation coefficient. In Prediction 4, both the complete set of predictions and the intra-block predictions yield comparable RMSE values—1.49, 1.46, and 1.52, respectively.

In case 4 RMSE decreased which is not an good indicator

Due to its emphasis on absolute error, RMSE is frequently employed in contexts where the magnitude of the error is a critical consideration. For example, <example *1*> and <example *2*>.

# Performance Metric: Classification

Classification models aim to predict categorical outcomes such as 'healthy' or 'sick', 'susceptible' or 'resistant', and 'high yield' or 'low yield'. This section presents a hypothetical example to highlight how the choice of performance metrics can lead to misleading model evaluations. Consider a binary classification model predicted on ten samples, of which four are positive (+) and six are negative (-), and the model produces a list of probabilities between 0 and 1, indicating the likelihood that each sample belongs to the positive class. The prediction can be expressed as the following joint probability distribution:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Where $X$ is a random variable representing the predicted probabilities, and $Y$ representing the ground truth labels. $*\text*{Uniform}(a, b)$ denotes a uniform distribution between $a$ and $b$. Table 1 shows ten samples drawn from this distribution to simulate the model predictions.

A table with numbers and labels

Description automatically generated

This example illustrates a scenario where the positive outcome is rare, which is commonly encountered in abnormality detection tasks. Before judging the prediction outcomes, a threshold that defines how the prediction probability is converted to a binary outcome is needed. If a prediction probability is greater than the threshold, the sample is classified as positive. In this example, the threshold is set to 0.5 for simplicity, the following confusion matrix summarizes the model performance (Figure 2a). Noted that the threshold can be adjusted for specific applications and change the confusion matrix accordingly.

This illustrate the variant outcomes from the same prediction, when the metric is different.

Assigning the positive to the rare category can result in different conclusion

A screenshot of a computer

Description automatically generated

## Accuracy

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Accuracy is a straightforward metric in classification problems, as defined in Equation 6. Here, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Based on this definition (Equation 6) and the confusion matrix, the model performance from the example is 0.60, which may appear to be better than random guesses (0.50). However, this metric could be misleading when applied to imbalanced datasets, warranting the use of additional metrics for a comprehensive evaluation. However, caution must be exercised when applying this metric to imbalanced datasets. In such cases, a model could show an accuracy that is higher than random guessing by predicting all samples as negative in an imbalanced dataset where negatives are predominant. This demonstrates that solely relying on accuracy is insufficient for evaluating a classification model, particularly when dealing with imbalanced datasets. Therefore, it is crucial to consider additional metrics for a more comprehensive and robust evaluation.

## Precision, Recall, and Precision-Recall Curve

|  |  |  |
| --- | --- | --- |
|  |  | (7)  (8) |

Precision evaluates the proportion of true positive predictions among all positive predictions. It effectively measures how reliable a positive prediction is. Recall, or sensitivity, measures the proportion of true positives among all actual positives. It gauges how effectively the model identifies positive samples. For instance, if the threshold is set as low as 0.1, the model is prone to making false positives, resulting in low precision. A high rate of false positives could be particularly costly in applications like <example *1*>, where unnecessary treatments could be administered based on these incorrect results. Conversely, the same low threshold can yield high recall as the model is less likely to miss actual positives. In situations where failing to identify a positive instance can have severe consequences, such as <example *2*> failing to detect a disease in its early stages, high recall could be more valuable.

A screenshot of a graph

Description automatically generated

When applying the metrics of precision and recall to the hypothetical example, the precision and recall values are 0.5 and 0.25, respectively. Both metrics yield lower values than the calculated accuracy of 0.6, underscoring the necessity of utilizing multiple metrics for a thorough evaluation of model performance. However, it is worth noting that both metrics are primarily focused on the classification of positive samples. This can introduce bias and can be particularly problematic when dealing with imbalanced datasets. As an example, if there is a scenario where the sample labels get reversed (i.e. the positives become negatives and vice versa) but the model parameters remain unchanged, the precision and recall values will shift to 0.625 and 0.833, respectively, as shown in Figure 2b. This outcome suggests that relying solely on precision and recall may not provide a complete understanding of model performance in certain contexts. Although incorporating metrics that focus on negative samples, such as specificity, can partially mitigate this issue, there remains a clear need for more robust and label-invariant metrics for an unbiased and comprehensive evaluation.

The trade-off between precision and recall is presented by adjusting the threshold. As the threshold increases, the model becomes more conservative and predicts fewer positive samples, yielding higher precision and lower recall. Since the precision and recall values reported from one single confusion matrix can only represent one specific threshold, the precision-recall (PR) curve is introduced to provide a more comprehensive view of the model performance. In the PR curve, the x-axis represents the recall values and the y-axis represents the precision values. The curve is derived by calculating the precision and recall values at different thresholds (Figure 2b). The area under the curve (AUC) is a common metric to summarize the performance of the PR curve. Still, the AUC is label-dependent, showing 0.76 for the original labels and 0.94 for the reversed labels.

Serve as a visual tool to optimize the threshold if one seek for a higher recall with the constant precision level.

## Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve represents the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It is a label-invariant and threshold-invariant alternative to the PR curve. The x-axis is the FPR and the y-axis is the TPR, which is equivalent to recall.

|  |  |  |
| --- | --- | --- |
|  |  | (7)  (8) |

The ROC curve can be interpreted as how much cost is needed to capture true positives. If the curve climbs steeply from the left side, it means that the model can capture most true positives with a low cost of false positives. A random guess, which yield a 50% chance of making a correct prediction, is represented by a diagnoal line in the ROC curve. This curve is widely used in reporting genetics markers in Genome-Wide Association Studies (GWAS) [ref 7-9], as whether the top-associated markers (i.e., prediction with high positive probability) can be identified is more important than considering the entire list of prediction quality. In the hypothetical example, the ROC curve show a labe-invariant AUC of 0.875, which is the same for both the original labels and the reversed labels. However, such high metric may still mislead the conclusion of model evalution, as it does not reflect the poor quality that the model has in predicting positive samples.

At the same level, when 0.25, inverted is worse, where 0.50, inverted get a better result.

Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) has been proposed as a more robust metric for the evaluation of binary classification models, particularly in the context of imbalanced datasets [refs 10, 11]. The MCC is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

This metric comprehensively incorporates all four elements of the confusion matrix, rendering it invariant to changes in label assignments. The MCC ranges from -1 to 1, with 1 denoting perfect classification devoid of false positives and false negatives, and 0 suggesting performance equivalent to random guessing. In a hypothetical case study using a threshold of 0.5, the MCC value was found to be 0.10, signifying suboptimal classification quality. This evaluation aligns more closely with the expected model performance compared to metrics like accuracy, ROC AUC, or precision and recall, particularly when the dataset skews toward positive samples. The characteristics of MCC make it a valuable tool for identifying optimized classification thresholds. By evaluating the MCC across different thresholds, one can pinpoint the threshold that maximizes the MCC value, thereby enhancing the model's overall performance. For instance, in the aforementioned example, the MCC reached its maximum value of 0.82 at a threshold of 0.2. This threshold yielded accuracy, precision, and recall values of 0.90, 0.75, and 1.00, respectively. Interestingly, even when labels were swapped, the metrics remained high, with values of 0.90 for accuracy, 1.00 for precision, and 0.83 for recall. This serves to underscore the MCC's balanced consideration of both positive and negative samples, solidifying its role as a versatile metric for a well-rounded model evaluation.

It should carefully consider which label to be positive. Usually, if the event is of interest, it should be prioritized to be the positive label as both precision and recall focus on the predictability of positive samples.