Performance Metrics

JESÚS CALDERÓN

Learning Objectives

- Define different performance measures.
- Explain each measure's strengths and applications.

Performance Metrics

An Overview of Performance Metrics

Regression

- Root Mean Squared Error (rmse)
- Mean Absolute Error (mae)
- Mean Absolute Percentage Error (mape)
- Mean Percentage Error (mpe)
- R Squared (rsq)

Classification

- Class metrics (hard predictions)
  - Accuracy
  - Kappa
  - Precision and Recall
  - F-measure
  - Sensitivity and Specificity
- Class probability metrics (soft predictions)
  - Log Loss
  - Area Under the Receiver Operating Characteristic Curve (ROC AUC)
  - Area Under the Precision-Recall Curve (PR AUC)

Regression

Errors, RMSE, and MAE

- When we talk about "errors", we generally refer to differences between true and predicted values.
- Names of errors are indicative of the operations that we perform on them.
- We use squared and absolute errors to avoid the cancellation of positive and negative errors.
- RMSE is differentiable over all values, and MAE is not.
- RMSE and MAE are expressed in the same units as the response variable.

Errors

$$error = (y - \hat{y})$$

Root Mean Squared Error

$$rmse = \sqrt{\frac{\sum_n (y_i - \hat{y}_i)^2}{n}}$$

Mean Absolute Error

$$mae = \frac{\sum_n |y_i - \hat{y}_i|}{n}$$

Percentage Error

- Percent errors reduce the scale effect, as they are all expressed in percentage.
- Typically, the denominator is the true value.
- MAPE avoids error cancellation.
- MAE does not avoid error cancellation, so positive and negative forecast errors can offset each other. As a result, MAE can be used as a measure of bias in prediction.
- Disadvantage of percent errors: not defined when the true value is 0.

Mean Absolute Percentage Error

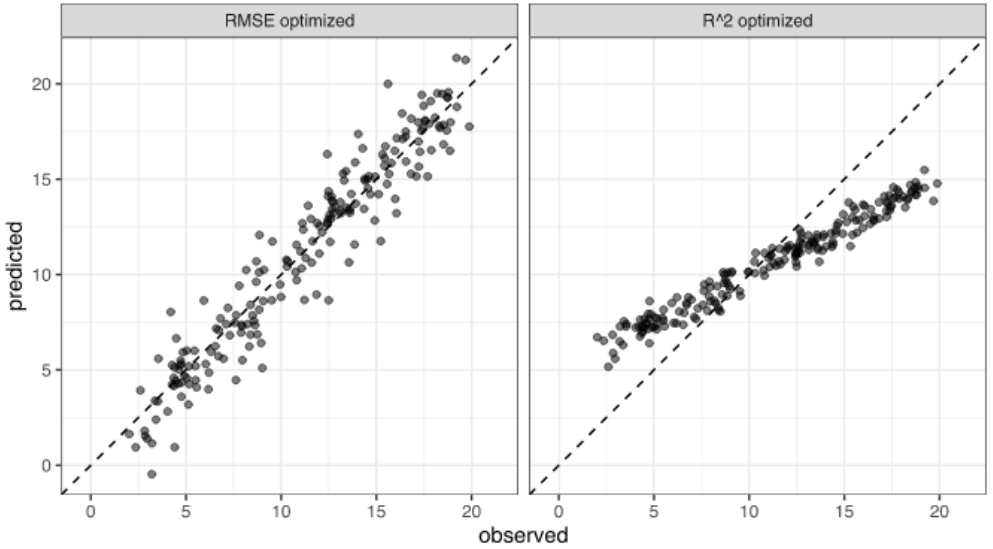
$$mape = \frac{1}{n}\sum_{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Mean Percent Error

$$mape = \frac{1}{n}\sum_{n}\frac{y_i - \hat{y}_i}{y_i}$$

R-Squared

- Measure of association or correlation.
- Does not measure predictive accuracy.
- A model optimized for RMSE has more variability but has relatively uniform accuracy across the range of the outcome.
- A model optimized for R-sq has a tighter correlation between true and predicted values but may perform poorly in the tails.



(Khun and Silge, 2021)

Classification: Class Metrics

Class Metrics

- Class Metrics apply when the predicted value is a hard class (e.g., 1 or 0).
- Class Metrics can be derived from the Confusion Matrix.
- These definitions are presented in the two-class case but can be extended to multiple classes.

Class Metrics

- Accuracy
- Kappa
- Precision and Recall
- F-measure
- Sensitivity and Specificity

Confusion Matrix

	Actual Negative	Actual Positive
Predicted Negative	True Negative (TN)	False Negative (FN)
Predicted Positive	False Positive (FP)	True Positive (TP)

Class Metrics: Accuracy and Kappa

Accuracy: the fraction of instances that were correctly classified.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Kappa: similar to accuracy, but adjusted or normalized by the accuracy that would be expected by chance alone and is very useful when one or more classes have large frequency distributions.

$$\kappa = \frac{acc - p_e}{1 - p_e}$$

Where p_e is the hypothetical probability of classification by chance.

Class Metrics: Precision, Recall, and F-Measure

Recall: measures the ability of the classifier to find all positive samples.

$$recall = \frac{TP}{TP + FN}$$

Precision: measures the ability of the classifier not to label as positive a sample that is negative.

$$precision = \frac{TP}{TP + FP}$$

F-Measure: combines precision and recall, but does not consider true negatives.

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

Sensitivity and Specificity

Sensitivity (or True Positive Rate) measures the proportion of postives that are correctly identified.

$$sensitivity = \frac{TP}{TP + FN}$$

Specificity (or True Negative Rate) measures the proportion of negatives that are correctly identified.

$$specificity = \frac{TN}{TN + FP}$$

Class Probability Metrics

Log Loss

Mean log loss: (aka Binary Cross-Entropy, logistic loss or cross-entropy loss) is the logistic model's negative log-likelihood.

$$L_{log}(y,p) = -\frac{1}{n}\sum_{n}\left(y_i log(p_i) + (1-y_i) log(1-p_i)\right)$$

From tidymodels documentation:

- Compared with `accuracy()`, log loss takes into account the uncertainty in the prediction and gives a more detailed view into the actual performance.
- For example, given two input probabilities of .6 and .9 where both are classified as predicting a positive value, say, "Yes", the accuracy metric would interpret them as having the same value.
- If the true output is "Yes", log loss penalizes .6 because it is "less sure" of it's result compared to the probability of .9.

Receiver Operating Characteristic

- Class probabilities offer more information about model predictions than the simple class value.
- Given class probabilities, one could decide to predict a class by comparing them to a threshold.
- An ROC curve shows the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for various thresholds.

True Positive Rate (TPR), also called Sensitivity or Recall.

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate or (1 - Specificity).

$$FPR = \frac{FP}{FP + TN}$$

The ROC Curve

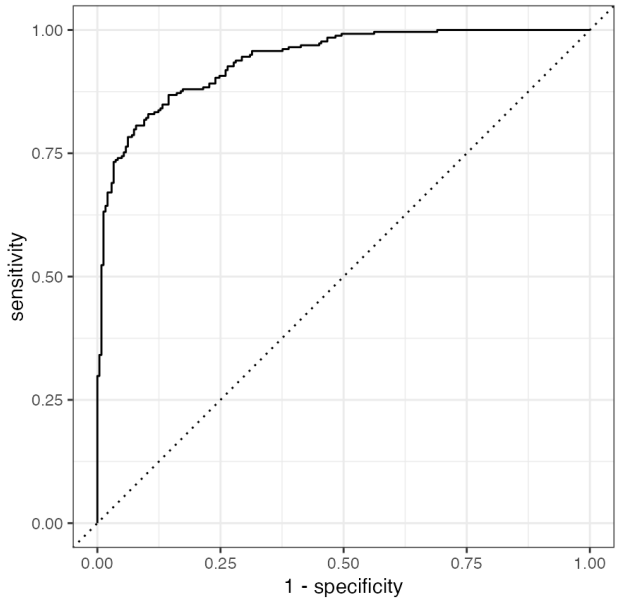
- When a classifier provides a *score* or *probability* as a prediction, we can choose a threshold above which we predict a positive case.
- Learning algorithms set the threshold to 50% as it minimizes overall error, but there may be cases in which we want to minimize error with respect to one class (e.g., imbalanced class problems like credit or fraud).
- In these cases, we may want to use a lower threshold. This will reduce the error in the minority class, but will increase error in the majority class.
- The ROC curve helps us measure performance across all possible thresholds.

Construct an ROC curve:

- Sort all predictions by decreasing probability of positive class.
- Set a high threshold (say, after first observation).
- Repeat for all ordered observations:
  - Calculate and store TPR and FPR.
  - Set a lower threshold (e.g., after second observation).
- Plot all combinations of TPR and FPR.

ROC AUC: Area Under ROC

- The performance of a classifier over all possible thresholds is given by the Area Under the (ROC) Curve (AUC).
- An ideal classifier would produce a ROC that "tracks" the top left corner of the chart and that has a rotated L-shaped curve.
- We expect that a classifier that performs no better than chance to have an AUC of 0.5 (when evaluated on an independent test set not used in model training).



ROC Curve (tidymodels.org)