Model Evaluations and Classification Metrics

Compiled by

Dr. Shashank Shetty

Evaluation Metrics

- Evaluation metrics in machine learning are used to assess the performance of a model on a particular task.
- The choice of evaluation metric depends on the nature of the problem (e.g., classification, regression, clustering) and the specific goals of the analysis.

Confusion Metrix

- A confusion matrix is a table that summarizes the performance of a classification model.
- It allows us to visualize the performance of a classification algorithm by comparing predicted labels with true labels.
- A confusion matrix has rows representing the actual classes and columns representing the predicted classes.
- Each cell in the matrix represents the count of instances that fall into a particular combination of actual and predicted classes.

Consider a binary classification problem where we are predicting whether patients have a certain disease (positive class) or not (negative class). Let's say we have a dataset of 100 patients, and after running our classification algorithm, we obtain the following confusion matrix:

Actual/Predicted	Predicted Negative	Predicted Positive
Actual Negative	60	5
Actual Positive	10	25

- **True Positive (TP)**: 25
  - These are the cases where the model correctly predicted positive (disease present) when the actual condition was positive.
- **False Positive (FP)**: 5
  - These are the cases where the model incorrectly predicted positive (disease present) when the actual condition was negative. Also known as Type I error.
- **True Negative (TN)**: 60
  - These are the cases where the model correctly predicted negative (disease not present) when the actual condition was negative.
- **False Negative (FN)**: 10
  - These are the cases where the model incorrectly predicted negative (disease not present) when the actual condition was positive. Also known as Type II error.

- **Accuracy**:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{25+60}{25+60+5+10} = \frac{85}{100} = 0.85$$

  - The accuracy of the model is 0.85 or 85%.

- **Precision**:

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{25}{25+5} = \frac{25}{30} = 0.83$$

  - The precision of the model is 0.83 or 83%.

- **Recall**:

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{25}{25+10} = \frac{25}{35} = 0.71$$

  - The recall of the model is 0.71 or 71%.

- **F1 Score**:

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\times0.83\times0.71}{0.83+0.71} = \frac{1.18}{1.54} = 0.77$$

  - The F1 score of the model is 0.77.

Accuracy vs Error rate:

Accuracy:

- **Definition**: Accuracy measures the proportion of correctly classified instances out of the total instances.
- **Calculation**: $\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$
- **Interpretation**: Accuracy indicates the overall correctness of the model's predictions.

Error rate:

- **Definition**: Error rate (Misclassification rate) measures the proportion of incorrectly classified instances out of the total instances.
- **Calculation**: $\text{Error rate} = \frac{\text{Number of incorrect predictions}}{\text{Total number of predictions}}$
- **Interpretation**: Error rate indicates the overall incorrectness of the model's predictions.

Accuracy Vs Error Rate

- For example, if a model correctly predicts 90 out of 100 instances, its accuracy would be 90%.
- While accuracy is a straightforward measure, it may not always be the best metric, especially when dealing with imbalanced datasets where one class is much more prevalent than others

- Using the same example as above, if the model incorrectly predicts 10 out of 100 instances, its error rate would be 10%.
- Error rate provides a complementary view to accuracy and is particularly useful when classes are imbalanced. For instance, in scenarios where the cost of misclassification differs between classes, minimizing the error rate might be more important than maximizing accuracy.

Comparison:

- **Accuracy** focuses on correct predictions, giving equal weight to both classes.
- **Error rate** focuses on incorrect predictions, giving equal weight to both classes.
- While accuracy gives a clear picture of correct predictions, the error rate gives a clear picture of incorrect predictions. Both measures are inversely related; as accuracy increases, the error rate decreases, and vice versa.

Precision

- Precision is a metric used in classification problems to evaluate the accuracy of positive predictions made by a model.
- It answers the question: "Of all the instances predicted as positive, how many are actually positive?“
- In other words, precision measures the proportion of true positive instances among all instances predicted as positive by the model.

Precision Formula:

Precision is calculated as the ratio of true positive instances to the sum of true positive and false positive instances.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Interpretation:

- Precision focuses on the relevance of retrieved items. It indicates how many of the positively predicted instances are relevant out of all instances predicted as positive. A high precision means that the model is making fewer false positive predictions, which is desirable in many applications, especially when false positives are costly or undesirable.
- **Interpretation of Precision:**
- If precision = 1: All instances predicted as positive are true positives, indicating perfect precision.
- If precision = 0: None of the instances predicted as positive are true positives, indicating no precision.
- Precision is a crucial metric in cases where the cost of false positives is high, such as medical diagnosis or fraud detection.

Conclusion:

- Precision is a vital metric for evaluating the performance of a classification model, particularly in scenarios where the focus is on minimizing false positives.
- It provides insights into the model's ability to make accurate positive predictions and is often used alongside other evaluation metrics to gain a comprehensive understanding of the model's performance.

Recall

- Recall, also known as sensitivity, is a metric used in classification problems to evaluate the ability of a model to correctly identify all relevant instances of a particular class.
- It answers the question: "Of all the actual positive instances, how many were identified correctly by the model?"
- In other words, recall measures the proportion of true positive instances among all actual positive instances.

Recall Formula:

Recall is calculated as the ratio of true positive instances to the sum of true positive and false negative instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Interpretation

- Recall focuses on the completeness of retrieved items. It indicates how many of the actual positive instances were correctly identified out of all instances that are actually positive.
- A high recall value means that the model is good at capturing most of the positive instances, which is desirable in many applications, especially when missing positive instances is costly or undesirable.
- If recall = 1: All actual positive instances are correctly identified by the model, indicating perfect recall.
- If recall = 0: None of the actual positive instances are correctly identified by the model, indicating no recall.
- Recall is particularly important in scenarios where missing positive instances is costly, such as disease diagnosis or anomaly detection.

Recall

- Recall is a critical metric for evaluating the performance of a classification model, especially when the focus is on capturing as many positive instances as possible.
- It provides insights into the model's ability to identify relevant instances of a particular class and is often used alongside other evaluation metrics to gain a comprehensive understanding of the model's performance.

Precision Vs Recall

- **Focus**:
  Precision: Precision focuses on minimizing false positive predictions.
  Recall: Recall focuses on capturing as many true positive instances as possible.

- **Context**:
  Precision: Precision is important when the cost of false positive predictions is high.
  Recall: Recall is important when missing positive instances is costly.

•**Trade-off**:
There is often a trade-off between precision and recall. Improving one metric may degrade the other. For example, increasing the threshold for positive predictions may increase precision but decrease recall, and vice versa.

•**Interpretation**:
Precision and recall provide complementary insights into the model's performance. While precision tells us how accurate the positive predictions are, recall tells us how comprehensive the predictions are in capturing all positive instances.

F1 Score

- The F1 score is a single metric that combines both precision and recall into a single value.
- It provides a balance between precision and recall, giving equal importance to both metrics.
- The F1 score is particularly useful when you have an uneven class distribution or when false positives and false negatives have different costs.

Formula:

The F1 score is calculated as the harmonic mean of precision and recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Or equivalently:

$$\text{F1 Score} = \frac{2 \times \text{True Positives}}{2 \times \text{True Positives} + \text{False Positives} + \text{False Negatives}}$$

Interpretation:

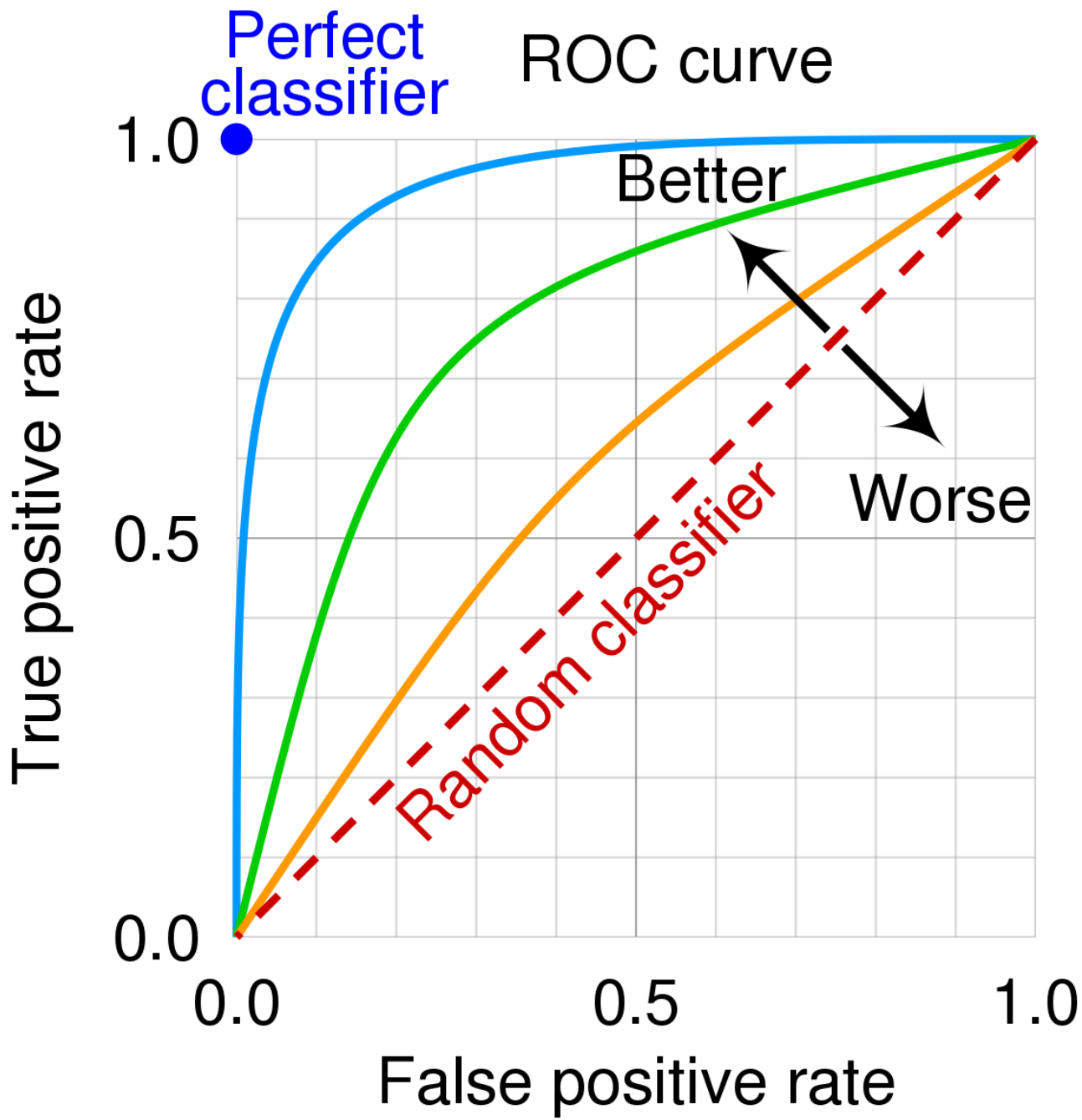
- The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates poor performance.
- A high F1 score indicates that the model has both high precision and high recall.
- F1 score reaches its best value at 1 and worst at 0.
- F1 score provides a single metric that balances the trade-off between precision and recall.

Conclusion:

- The F1 score is a useful metric for evaluating the overall performance of a classification model, particularly in situations where you want to balance both precision and recall.
- It provides a single value that summarizes the trade-off between precision and recall, making it easier to compare and interpret the performance of different models.

AUC

- The AUC (Area Under the Curve) is a metric used to evaluate the performance of a binary classification model based on its Receiver Operating Characteristic (ROC) curve.
- The ROC curve is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The AUC represents the area under the ROC curve, which provides a single scalar value summarizing the model's performance across all possible threshold settings.


Perfect classifier
ROC curve
1.0
Better
Worse
0.5
Random classifier
True positive rate
0.0
0.0
0.5
1.0
False positive rate

Interpretation:

- AUC ranges from 0 to 1, where 0 indicates a poor model (classifies all instances incorrectly) and 1 indicates a perfect model (classifies all instances correctly).
- AUC measures the model's ability to distinguish between positive and negative classes. Higher AUC values indicate better model performance.
- AUC can be interpreted as the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative instance.

Mean Squared Error (MSE)

- **Definition**: MSE is the average of the squared differences between predicted and actual values in a regression model.
- **Formula**: $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of instances.
- **Interpretation**: MSE measures the average squared deviation between the predicted and actual values. It penalizes larger errors more heavily due to squaring.

2. Root Mean Squared Error (RMSE):

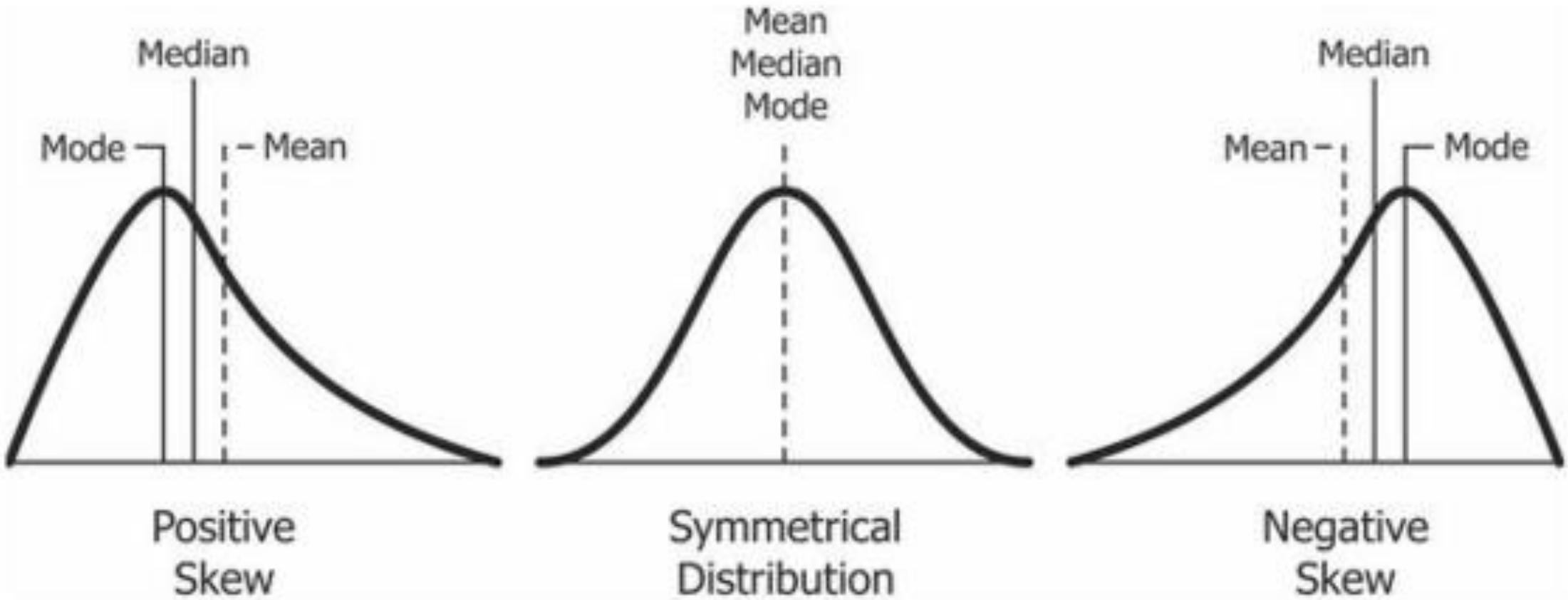
- **Definition**: RMSE is the square root of the average of the squared differences between predicted and actual values in a regression model.
- **Formula**: $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$
- **Interpretation**: RMSE provides a measure of the spread of errors in the same units as the target variable. It's more interpretable than MSE since it's in the same units as the target variable.

3. Mean Absolute Error (MAE):

- **Definition**: MAE is the average of the absolute differences between predicted and actual valu in a regression model.
- **Formula**: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **Interpretation**: MAE measures the average absolute deviation between the predicted and actual values. It's less sensitive to outliers compared to MSE since it doesn't square the differences.

4. Root Mean Squared Log Error (RMSLE):

- **Definition**: RMSLE is the square root of the average of the squared differences between the natural logarithm of predicted and actual values in a regression model.
- **Formula**: $\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2}$
- **Interpretation**: RMSLE is useful when the target variable has a wide range of values. It penalizes underestimation more than overestimation and is often used in competition scenarios where predictions are evaluated on a logarithmic scale.


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Problem 1

- Consider a binary classification problem where we are predicting whether transactions are fraudulent (positive class) or not fraudulent (negative class). Suppose the dataset consists of 1000 transactions, out of which only 10 are fraudulent.
- True Positives (TP) = 5
- False Positives (FP) = 10
- True Negatives (TN) = 970
- False Negatives (FN) = 15

Problem 2

- Suppose we have a spam email detection system. In this scenario, we want to prioritize precision (minimizing false positives) while still maintaining a reasonable level of recall (capturing most of the spam emails).
- True Positives (TP) = 90
- False Positives (FP) = 30
- True Negatives (TN) = 850
- False Negatives (FN) = 30

House	Actual Price (y_i)	Predicted Price ($\hat{y}_i$)
1	$300,000	$310,000
2	$400,000	$390,000
3	$250,000	$260,000
4	$600,000	$580,000
5	$350,000	$340,000

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

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\text{MAE} = \frac{|300{,}000-310{,}000|+|400{,}000-390{,}000|+|250{,}000-260{,}000|+|600{,}000-580{,}000|+|350{,}000-340{,}000|}{5}$$

$$\text{MAE} = \frac{10{,}000+10{,}000+10{,}000+20{,}000+10{,}000}{5} = \frac{60{,}000}{5} = 12{,}000$$

2. **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

$$\text{RMSE} = \sqrt{\frac{(300{,}000-310{,}000)^2+(400{,}000-390{,}000)^2+(250{,}000-260{,}000)^2+(600{,}000-580{,}000)^2+(350{,}000-340{,}000)^2}{5}}$$

$$\text{RMSE} = \sqrt{\frac{100{,}000{,}000+100{,}000{,}000+100{,}000{,}000+400{,}000{,}000+100{,}000{,}000}{5}}$$

$$\text{RMSE} = \sqrt{\frac{800{,}000{,}000}{5}} = \sqrt{160{,}000{,}000} \approx 12{,}650$$

3. **Root Mean Squared Logarithmic Error (RMSLE):**

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i+1)-\log(\hat{y}_i+1))^2}$$

$$\text{RMSLE} = \sqrt{\frac{(\log(300{,}000+1)-\log(310{,}000+1))^2+(\log(400{,}000+1)-\log(390{,}000+1))^2}{5}} + \dots$$

$$\text{RMSLE} = \sqrt{\frac{(12.611-12.621)^2+(12.899-12.878)^2+\dots}{5}}$$

$$\text{RMSLE} = \sqrt{\frac{(0.01)^2+(0.021)^2+\dots}{5}}$$

$$\text{RMSLE} = \sqrt{\frac{0.0001+0.000441+\dots}{5}}$$

$$\text{RMSLE} = \sqrt{\frac{0.002681}{5}} \approx \sqrt{0.000536} \approx 0.02315$$

These are the calculated values for MAE, RMSE, and RMSLE for this regression problem. Each metric provides insights into different aspects of the model's performance, allowing us to assess its accuracy and suitability for predicting house prices.