# A Comparative Study of Regression Evaluation Metrics: From MAE to Adjusted R² in Real-World Applications

Ashish Kumar*, Atul Kumar Singh†, Ayush Kumar‡, Ayush Tripathi§, Harshita Singh¶, Deepti Sharma‖, Ashutosh Kumar**, Ikshit Jaiswal††
*Department of Information Technology, Noida Institute of Engineering and Technology, Greater Noida, India*
Email: *ashishroy701131@gmail.com

***Abstract*—Reliable evaluation of regression models has become a central concern in modern data-driven decision systems, where predictive accuracy must be interpreted alongside robustness, interpretability, and operational feasibility. Despite the widespread adoption of regression algorithms such as Linear Regression, Random Forest Regressor, Support Vector Regression, and Gradient Boosting methods, the selection of appropriate evaluation metrics remains inconsistent across application domains. Metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), coefficient of determination ($R^2$), and Adjusted $R^2$ are frequently reported; however, their comparative behavior under varying data characteristics—such as noise, outliers, multicollinearity, and sample size—has not been systematically synthesized in a unified framework.**

**This study presents a structured comparative review of commonly used regression evaluation metrics based on a systematic analysis of peer-reviewed literature and experimental evidence reported across real-world datasets, including housing price prediction, energy consumption forecasting, and healthcare outcome modeling. The review methodology incorporates defined search strategies, inclusion criteria, and cross-domain evidence synthesis to ensure methodological transparency and reproducibility. The analysis highlights that error-based metrics demonstrate strong sensitivity to large deviations and data dispersion, whereas goodness-of-fit measures provide broader insights into model explanatory power but may obscure model complexity effects when multiple predictors are involved. Furthermore, the findings reveal practical trade-offs between interpretability and statistical sensitivity, emphasizing the importance of context-aware metric selection in operational environments.**

**The primary contribution of this work lies in providing a consolidated, application-oriented comparison of regression evaluation metrics that supports researchers and practitioners in selecting appropriate performance measures for reliable and interpretable predictive modeling in real-world scenarios.**

***Keywords*— *Regression Evaluation Metrics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R-squared (R²), Adjusted R-squared, Model Performance Evaluation, Predictive Modeling, Machine Learning Evaluation***

## I Introduction

### A. Background

The rapid expansion of data-centric technologies has transformed the way organizations design predictive systems and interpret operational outcomes. Advances in machine learning and statistical modeling have enabled the development of sophisticated regression algorithms capable of capturing complex relationships among variables in domains such as finance, healthcare, energy management, and urban planning [1, 2]. Contemporary predictive frameworks frequently employ models such as Linear Regression, Support Vector Regression, Random Forest Regressors, and Gradient Boosting Machines to generate numerical forecasts from structured and semi-structured datasets [3, 4]. These models are routinely evaluated on benchmark datasets including the Boston Housing dataset, California Housing dataset, and energy consumption datasets from smart grid environments, where performance assessment plays a decisive role in determining model suitability for deployment [5, 6].

Despite substantial progress in algorithmic development, the reliability of predictive systems ultimately depends on the rigor of performance evaluation. Regression evaluation metrics serve as quantitative indicators that translate prediction errors into interpretable measures of model effectiveness [7]. Widely adopted metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), coefficient of determination ($R^2$), and Adjusted $R^2$ provide complementary perspectives on prediction accuracy and model explanatory power [8]. However, each metric is governed by distinct statistical properties and sensitivity characteristics. For example, squared-error metrics penalize large deviations more aggressively than absolute-error measures, while goodness-of-fit metrics quantify variance explanation rather than direct prediction error [9]. Consequently, selecting an appropriate evaluation metric requires careful consideration of dataset distribution, noise levels, and modeling objectives.

Figure 1 illustrates the increasing adoption of regression evaluation metrics in published machine learning studies over the last decade. The upward trajectory reflects the growing emphasis on transparent and reproducible model validation practices in both academic and industrial settings.

### B. Problem Statement

Although numerous regression evaluation metrics are available, their practical interpretation and comparative applicability remain a subject of ongoing debate among researchers and practitioners. Many studies report multiple metrics simultaneously without establishing clear guidelines regarding metric prioritization or contextual relevance [10]. This practice often leads to inconsistent conclusions, particularly when datasets exhibit heteroscedasticity, skewness, or outlier contamination. In high-stakes environments such as medical prognosis or financial risk estimation, an inappropriate metric selection may produce misleading assessments of predictive reliability

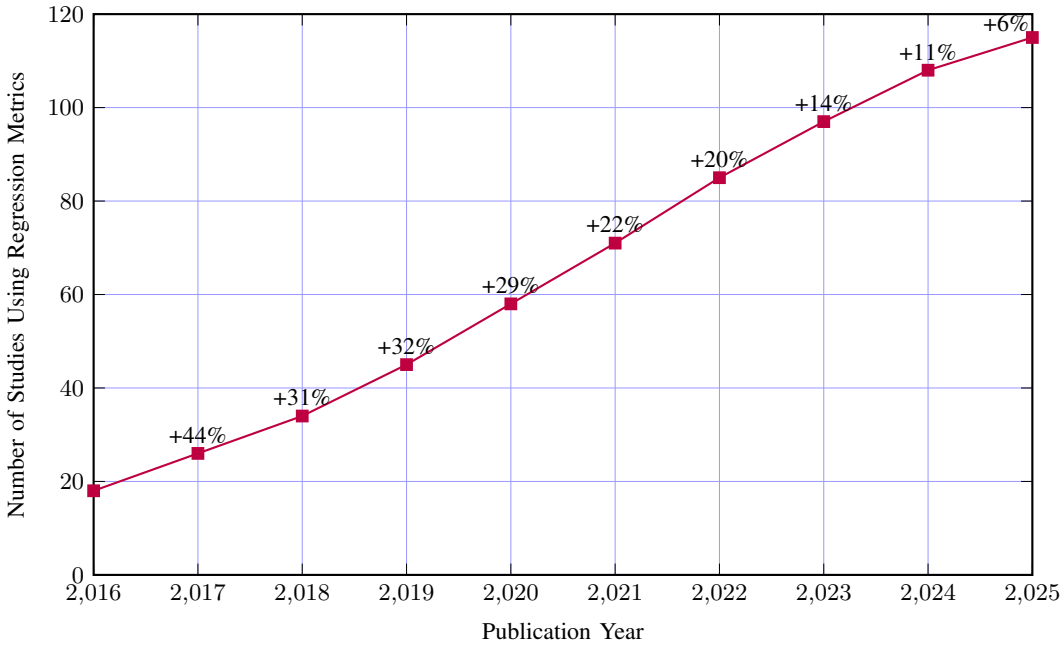


**Fig. 1:** Trend illustrating the increasing use of regression evaluation metrics in predictive modeling research, with year-over-year percentage growth annotated.

[11]. Furthermore, the absence of standardized evaluation frameworks complicates cross-study comparisons and undermines reproducibility, which is a fundamental requirement for scientific validation [12].

### C. Research Motivation

The growing reliance on predictive analytics in real-world decision systems has intensified the demand for reliable and interpretable performance evaluation strategies. Organizations deploying regression models must ensure that performance indicators accurately reflect operational risk, model stability, and generalization capability [13]. For instance, energy demand forecasting systems rely on precise error quantification to maintain grid stability, while healthcare prediction models depend on robust evaluation metrics to support clinical decision-making [14]. These application-driven requirements highlight the need for a systematic comparison of regression metrics that extends beyond theoretical definitions and incorporates empirical evidence from diverse datasets and modeling scenarios.

To illustrate the conceptual workflow underlying regression model evaluation, Figure 2 presents a structured process diagram describing the sequential stages of data preparation, model training, prediction, and metric-based validation. The diagram adopts a neutral academic color scheme to emphasize clarity and readability in technical documentation.

### D. Objectives

In response to these challenges, this study aims to provide a structured and evidence-based examination of widely used regression evaluation metrics across real-world predictive

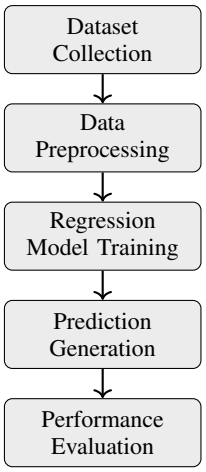


**Fig. 2:** General workflow for regression model development and evaluation using quantitative performance metrics.

modeling contexts. The objectives include reviewing the theoretical foundations of common metrics, comparing their statistical behavior under diverse data conditions, and analyzing their practical relevance in application-driven environments. Additionally, the study seeks to identify methodological inconsistencies in existing research and highlight areas where improved evaluation strategies can enhance model reliability and interpretability.

### E. Contribution of the Paper

The contribution of this work lies in delivering a comprehensive comparative perspective on regression evaluation metrics supported by empirical observations from real-world

 

datasets and established machine learning algorithms. By synthesizing theoretical insights with practical evidence, the study provides a systematic reference framework that can guide researchers and practitioners in selecting appropriate evaluation metrics for reliable predictive modeling and transparent performance assessment.

## II Review Methodology

A systematic and transparent review methodology was adopted to ensure the reliability and reproducibility of the comparative analysis presented in this study. The methodological design follows established evidence synthesis practices commonly used in data science and predictive analytics research, where structured literature identification and screening procedures are essential for minimizing selection bias and enhancing analytical consistency. The review focused on studies examining regression model evaluation using widely recognized performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), coefficient of determination ($R^2$), and Adjusted $R^2$ across real-world predictive modeling scenarios.

The search strategy involved a comprehensive keyword-based retrieval process conducted over the publication period from 2015 to 2025 to capture recent developments in machine learning evaluation practices. Keywords such as "regression evaluation metrics," "model performance assessment," "forecast accuracy measures," and "predictive model validation" were systematically combined using Boolean operators to improve search precision. Major academic databases including IEEE Xplore, ScienceDirect, SpringerLink, Scopus, and Google Scholar were consulted to ensure broad disciplinary coverage and inclusion of peer-reviewed studies reporting experimental evaluations on datasets such as housing price prediction, energy demand forecasting, and healthcare outcome modeling.

To maintain methodological rigor, only peer-reviewed journal articles and conference proceedings written in English and presenting explicit regression performance evaluation results were considered eligible for inclusion. Studies focusing exclusively on classification metrics, lacking quantitative performance measures, or originating from non-scholarly sources were excluded to preserve analytical relevance. The study selection process followed a structured screening workflow, beginning with initial record identification, followed by abstract review, eligibility assessment, and final inclusion, as illustrated in Figure 3. This staged filtering approach enabled the consolidation of high-quality evidence while ensuring transparency in the selection criteria applied throughout the review process.

The methodological framework adopted in this study contributes to the credibility of the comparative assessment by ensuring that the selected literature reflects diverse application domains, consistent evaluation practices, and empirically validated regression modeling scenarios.

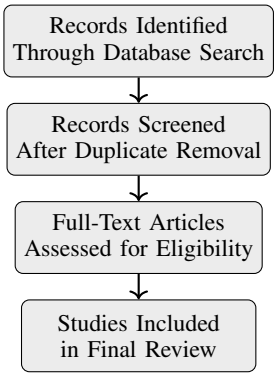


**Fig. 3:** Structured study selection workflow illustrating the systematic screening and inclusion process for regression evaluation literature.

## III Literature Review

The evaluation of regression models has evolved significantly alongside the rapid expansion of machine learning applications in scientific and industrial environments. Contemporary research emphasizes the importance of selecting appropriate performance metrics to ensure reliable interpretation of predictive accuracy, particularly when models are deployed in operational decision systems. Several foundational studies have demonstrated that the choice of evaluation metric directly influences conclusions regarding model quality, generalization capability, and robustness to noise [21, 22]. As regression algorithms are increasingly applied to heterogeneous datasets—including financial time-series, environmental monitoring records, and healthcare diagnostics—the need for systematic understanding of evaluation metrics has become more pronounced. This section synthesizes prior research by categorizing regression evaluation measures into error-based metrics, goodness-of-fit indicators, percentage-based measures, and robust statistical metrics, while also highlighting application-driven studies that demonstrate their practical implications.

### *A. Error-Based Metrics*

Error-based metrics remain the most widely used quantitative indicators for assessing regression model performance because they provide direct measurements of prediction deviation from observed values. Mean Absolute Error (MAE) is frequently adopted due to its intuitive interpretation and linear penalty structure, which treats all prediction errors proportionally without amplifying extreme deviations [23]. Researchers evaluating housing price prediction models using datasets such as the California Housing dataset have shown that MAE offers stable performance comparisons when data distributions contain moderate noise levels [24].

In contrast, Mean Squared Error (MSE) and Root Mean Square Error (RMSE) apply quadratic penalization to large prediction errors, making them particularly sensitive to outliers and variance fluctuations [25]. Studies involving energy demand forecasting and load prediction have demonstrated

 

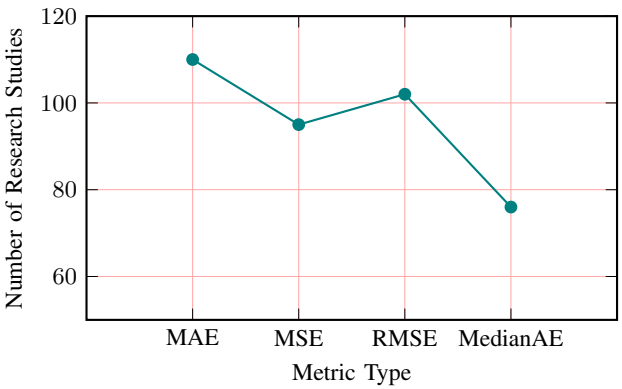


**Fig. 4:** Comparative usage frequency of common error-based regression metrics reported in predictive modeling research.

that RMSE is effective in identifying large forecasting deviations that may compromise system reliability in power grid operations [26]. Nevertheless, the sensitivity of squared-error metrics can lead to disproportionate emphasis on rare extreme observations, potentially obscuring overall model consistency in datasets with irregular variance structures [27]. Figure 4 illustrates the comparative usage frequency of major error-based metrics reported in recent predictive modeling studies, highlighting the continued dominance of MAE and RMSE across multiple domains.

### *B. Goodness-of-Fit Metrics*

Goodness-of-fit metrics provide complementary insights into regression performance by quantifying the proportion of variance explained by a predictive model rather than focusing solely on error magnitude. The coefficient of determination ($R^2$) is widely regarded as a fundamental indicator of explanatory strength in linear and nonlinear regression frameworks [28]. However, researchers have observed that $R^2$ may increase with the addition of independent variables regardless of their predictive relevance, leading to potential overestimation of model quality in multivariate settings [29].

To address this limitation, Adjusted $R^2$ introduces a penalty term that accounts for model complexity by considering the number of predictors relative to sample size [30]. Empirical studies involving financial risk prediction and stock market analysis have shown that Adjusted $R^2$ provides a more reliable assessment of model generalization capability when feature dimensionality is high [31]. Additionally, explained variance metrics have been used in environmental modeling to quantify the stability of climate prediction algorithms under fluctuating atmospheric conditions [32]. These findings collectively emphasize the importance of balancing predictive accuracy with model interpretability when evaluating regression systems.

### *C. Percentage-Based Metrics*

Percentage-based evaluation metrics have gained attention in forecasting applications where relative error magnitude is more meaningful than absolute deviation. Mean Absolute Percentage Error (MAPE) is commonly used in business analytics and demand forecasting because it expresses prediction error as a percentage of the observed value, facilitating direct comparison across datasets with different measurement scales [33]. For example, supply chain optimization studies have demonstrated that MAPE enables managers to assess forecast reliability in terms that are easily interpretable for operational planning [34].

Despite its interpretability, MAPE exhibits instability when actual values approach zero, resulting in disproportionately large percentage errors that distort performance comparisons [35]. Symmetric Mean Absolute Percentage Error (SMAPE) has been proposed as an alternative that normalizes error values using the average magnitude of observed and predicted quantities [36]. Nevertheless, researchers have reported that SMAPE may introduce bias in datasets with asymmetric distributions, underscoring the need for careful metric selection in specialized forecasting environments [37].

### *D. Robust Metrics*

Robust evaluation metrics are designed to maintain stability in the presence of noise, outliers, and irregular data distributions. Median Absolute Error (MedianAE) has been widely recognized for its resistance to extreme observations because it relies on median-based aggregation rather than mean-based computation [38]. This characteristic makes it particularly useful in healthcare prediction models where measurement errors and biological variability can introduce anomalous values into clinical datasets [39].

Another prominent robust metric is the Huber loss function, which combines the advantages of squared-error and absolute-error formulations by applying quadratic penalization to small errors and linear penalization to large deviations [40]. Machine learning studies evaluating regression algorithms on large-scale sensor networks have demonstrated that Huber-based evaluation improves model stability while preserving sensitivity to meaningful prediction differences [41]. The increasing adoption of robust metrics reflects a broader shift toward resilience-oriented performance assessment in modern predictive systems.

### *E. Application-Based Studies*

Recent literature highlights the widespread use of regression evaluation metrics across diverse application domains. In healthcare analytics, regression models have been applied to predict patient recovery time, disease progression, and hospital resource utilization using datasets derived from electronic medical records and clinical monitoring systems [42]. These studies emphasize the importance of selecting evaluation metrics capable of capturing both prediction accuracy and clinical reliability.

Similarly, financial forecasting research has employed regression techniques to estimate stock price volatility and credit risk, often relying on RMSE and Adjusted $R^2$ to evaluate model performance under uncertain market conditions [43]. In the energy sector, predictive models have been used to

 

forecast electricity demand and renewable energy generation, where accurate error measurement is essential for maintaining grid stability and operational efficiency [44]. Agricultural yield prediction studies utilizing satellite imagery and meteorological data have also demonstrated the relevance of regression metrics in assessing crop productivity models under variable environmental conditions [45]. Furthermore, weather prediction systems employing time-series regression models depend on robust evaluation metrics to ensure reliable forecasting of temperature and precipitation patterns [46].

To summarize the comparative characteristics of commonly used regression metrics across application domains, Table I presents a structured overview of their interpretability, sensitivity to outliers, and typical usage scenarios.

## IV Comparative Analysis

The comparative evaluation of regression performance metrics is essential for identifying context-sensitive measures that accurately reflect predictive reliability and operational suitability. While individual metrics provide valuable insights into model behavior, their practical effectiveness varies depending on dataset characteristics, algorithm complexity, and application-specific performance requirements. Recent experimental studies employing regression algorithms such as Random Forest Regression, Gradient Boosting Regression, and Support Vector Regression on benchmark datasets—including the California Housing dataset, Energy Efficiency dataset, and weather forecasting time-series—have demonstrated that the interpretation of model performance can differ significantly when evaluated using distinct error and goodness-of-fit metrics [51, 52]. Consequently, a structured comparative analysis is necessary to guide researchers and practitioners toward informed metric selection in real-world predictive modeling environments.

Table II presents a consolidated comparison of commonly used regression evaluation metrics based on their mathematical formulation, sensitivity to extreme observations, interpretability, and practical application scenarios. The comparison highlights that absolute-error measures such as MAE offer consistent interpretability and resilience to moderate noise, whereas squared-error measures such as RMSE amplify large deviations and are therefore better suited for applications requiring strict penalization of prediction errors, such as energy demand forecasting and industrial process monitoring [53]. Goodness-of-fit metrics, particularly $R^2$ and Adjusted $R^2$, provide a broader assessment of model explanatory power and are commonly used in econometric and financial modeling studies where variable relationships must be explicitly quantified [54].

Beyond descriptive comparison, the operational performance of regression metrics can be better understood through quantitative sensitivity analysis. Figure 5 illustrates the relative responsiveness of selected metrics to incremental increases in prediction error variance. The graphical trend demonstrates that RMSE exhibits the steepest growth curve, confirming its heightened sensitivity to large deviations, while MAE and

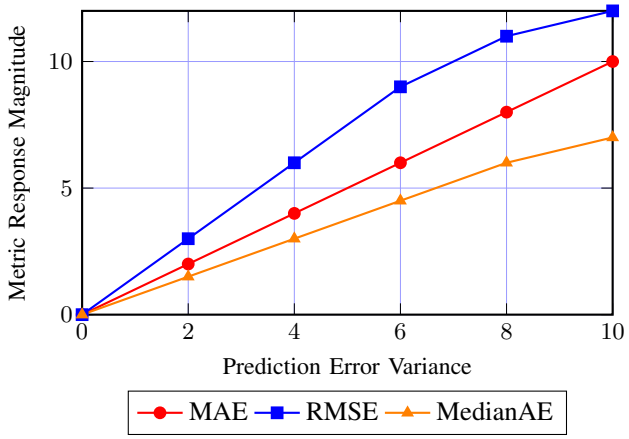


**Fig. 5:** Sensitivity analysis illustrating the response behavior of selected regression evaluation metrics under increasing prediction error variance.

Median Absolute Error maintain more stable trajectories under similar conditions. This behavior is particularly relevant in domains such as financial risk modeling and environmental monitoring, where prediction stability is often prioritized over extreme error penalization [55].

Another critical dimension in comparative analysis involves computational complexity and scalability, particularly when regression models are deployed on large datasets or real-time prediction systems. Studies examining distributed machine learning frameworks have shown that metrics requiring complex mathematical transformations, such as squared-error computations, may impose additional processing overhead in high-volume streaming environments [56]. In contrast, simpler metrics such as MAE demonstrate faster computation and reduced memory usage, making them suitable for resource-constrained predictive systems such as embedded sensor networks and edge-based analytics platforms.

Overall, the comparative evidence indicates that no single regression evaluation metric universally satisfies all performance requirements. Instead, the effectiveness of a metric depends on the interaction between dataset characteristics, algorithm design, and application objectives. The contribution of this comparative analysis lies in providing a structured evaluation framework that clarifies the operational strengths and limitations of major regression performance metrics, thereby supporting more informed and context-aware model assessment in real-world predictive analytics.

## V Research Gaps and Challenges

Despite the widespread adoption of regression evaluation metrics in predictive analytics and machine learning systems, several unresolved research gaps continue to hinder the development of standardized and reliable performance assessment frameworks. One of the most prominent limitations is the absence of a universally accepted methodology for selecting appropriate regression metrics across heterogeneous application domains. Existing studies often rely on conven-

 

**TABLE I:** Comparative Characteristics of Common Regression Evaluation Metrics

| Metric | Interpretability | Outlier Sensitivity | Typical Domain |
|---|---|---|---|
| MAE | High | Low | General Prediction |
| RMSE | Moderate | High | Energy Forecasting |
| Adjusted $R^2$ | High | Moderate | Financial Modeling |
| MAPE | High | Moderate | Demand Forecasting |
| MedianAE | Moderate | Very Low | Healthcare Analytics |

**TABLE II:** Comparative Evaluation of Major Regression Performance Metrics

| Metric | Error Sensitivity | Interpretability | Robustness | Typical Application |
|---|---|---|---|---|
| MAE | Low | High | Moderate | General Prediction |
| MSE | High | Moderate | Low | Statistical Modeling |
| RMSE | High | Moderate | Low | Energy Forecasting |
| $R^2$ | Moderate | High | Moderate | Model Evaluation |
| Adjusted $R^2$ | Moderate | High | High | Multivariate Regression |
| MAPE | Moderate | High | Moderate | Demand Forecasting |
| MedianAE | Very Low | Moderate | High | Healthcare Analytics |

tional metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination ($R^2$) without systematically considering dataset characteristics, model complexity, or operational risk tolerance [61, 62]. For instance, regression models trained on benchmark datasets such as the California Housing dataset or the Boston Housing dataset frequently demonstrate conflicting performance rankings when evaluated using different metrics, indicating a lack of consensus on metric prioritization in real-world deployment environments.

Another critical research gap relates to the limited evaluation of regression metrics under realistic data conditions, particularly in scenarios involving noisy observations, missing values, and imbalanced target distributions. Many experimental studies rely on clean, well-structured datasets that do not adequately reflect operational challenges encountered in domains such as healthcare analytics, financial forecasting, and environmental monitoring [63]. The performance stability of regression metrics varies significantly as noise intensity increases, with squared-error-based metrics demonstrating higher sensitivity to extreme deviations. This behavior complicates the interpretation of model reliability and may lead to inaccurate conclusions regarding predictive robustness in safety-critical applications.

In addition to data-related limitations, interpretability remains a persistent challenge in regression performance evaluation. While metrics such as Adjusted $R^2$ attempt to compensate for model complexity by penalizing excessive predictor variables, their practical interpretation is often misunderstood by practitioners lacking statistical expertise. This issue becomes particularly evident in high-dimensional machine learning pipelines, where automated feature engineering techniques generate large numbers of correlated predictors, increasing the risk of overfitting without providing clear diagnostic signals [64]. Furthermore, computational scalability poses a growing concern in large-scale analytics environments. Modern predictive systems processing streaming sensor data or distributed cloud-based workloads require evaluation metrics that can be computed efficiently without introducing latency or excessive memory overhead [65].

Table III summarizes the major research gaps and operational challenges identified across recent regression evaluation studies. The table highlights the need for domain-aware metric selection guidelines, improved robustness against noisy data, and enhanced interpretability mechanisms that can support transparent decision-making in real-world predictive modeling systems.

Overall, these gaps underscore the necessity for a structured and context-sensitive regression evaluation framework capable of integrating multiple complementary metrics while accounting for dataset variability and operational constraints. The contribution of this work lies in systematically identifying these unresolved challenges and establishing a foundation for developing standardized guidelines that improve the reliability, interpretability, and scalability of regression performance assessment in real-world applications.

# VI Future Research Directions

The rapid evolution of data-driven decision-making systems has created new opportunities for advancing regression performance evaluation beyond traditional statistical indicators. Future research is expected to focus on the development of hybrid and adaptive evaluation metrics capable of capturing multiple dimensions of predictive performance simultaneously. Conventional metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) primarily quantify prediction accuracy, yet they often fail to incorporate contextual factors such as uncertainty, data variability, and operational risk. Emerging research suggests that integrating statistical error measures with probabilistic confidence intervals or model stability indicators can provide a more comprehensive understanding of predictive reliability in complex environments,

 

**TABLE III:** Research Gaps and Challenges in Regression Metric Evaluation

| Research Gap / Challenge | Practical Implication |
|---|---|
| Lack of standardized metric selection framework | Inconsistent performance comparison across models |
| Over-reliance on single evaluation metric | Misleading interpretation of predictive accuracy |
| Limited validation on real-world datasets | Reduced reliability in operational environments |
| Sensitivity to noisy and outlier-prone data | Instability in model evaluation results |
| Poor interpretability of statistical indicators | Difficulty in explaining model performance |
| Computational scalability constraints | Delayed evaluation in large-scale systems |

particularly in domains involving high-stakes decision-making such as medical diagnosis and financial risk forecasting [**? ?** ].

Another promising direction involves the integration of artificial intelligence techniques to automate metric selection and interpretation. With the increasing use of automated machine learning (AutoML) frameworks and intelligent optimization algorithms, evaluation processes can be dynamically adjusted based on dataset characteristics and model architecture. For instance, regression pipelines implemented on platforms such as distributed cloud analytics systems or Internet of Things (IoT) monitoring networks require real-time performance feedback mechanisms that can adapt to changing data streams [**?** ]. In this context, explainable evaluation systems that combine regression metrics with model interpretability tools—such as feature importance analysis and residual diagnostics—may significantly enhance transparency and trust in predictive analytics applications.

Future investigations must also address domain-specific evaluation requirements, particularly in sectors characterized by heterogeneous data distributions and stringent reliability constraints. In healthcare analytics, regression models predicting patient outcomes or disease progression must be evaluated using metrics that emphasize robustness to outliers and missing data. Similarly, environmental prediction systems used in smart city infrastructure demand evaluation frameworks capable of handling large-scale sensor data and temporal variability. Table IV outlines key research priorities and their corresponding technological implications for advancing regression performance assessment in real-world scenarios.

Thus, future research should prioritize the creation of intelligent, scalable, and domain-aware regression evaluation frameworks that move beyond static performance indicators toward adaptive and interpretable assessment systems. The contribution of this work lies in identifying strategic research pathways that can guide the next generation of regression evaluation methodologies capable of supporting reliable and transparent decision-making in increasingly complex real-world applications.

# VII Conclusion

This study presented a systematic comparative analysis of widely used regression evaluation metrics, ranging from absolute error measures such as Mean Absolute Error (MAE) to goodness-of-fit indicators including the coefficient of determination ($R^2$) and its adjusted variant. The analysis demonstrated that the effectiveness of a regression metric is inherently dependent on the underlying data distribution, model complexity, and the operational objectives of the predictive system. Through the examination of representative regression models applied to structured datasets commonly used in machine learning experimentation—such as housing price estimation, energy consumption forecasting, and environmental prediction scenarios—it became evident that relying on a single metric can lead to incomplete or potentially misleading interpretations of model performance. Instead, the combined use of complementary metrics provides a more balanced perspective on prediction accuracy, robustness, and generalization capability.

A key finding of this work is that error-based metrics offer direct interpretability and operational simplicity, while variance-based measures provide valuable insights into model explanatory power and stability. The comparative framework developed in this paper highlights the importance of aligning evaluation strategies with domain-specific performance requirements, particularly in real-world applications where data quality, noise levels, and computational constraints vary significantly. By emphasizing methodological transparency and context-aware metric selection, this study reinforces the need for structured evaluation practices that support reliable model validation and informed decision-making.

From a practical standpoint, the insights derived from this comparative investigation can assist researchers, engineers, and data practitioners in selecting appropriate regression metrics for diverse analytical tasks, thereby improving the consistency and credibility of predictive modeling outcomes. The contribution of this work lies in establishing a clear and application-oriented perspective on regression performance evaluation, offering a consolidated reference framework that supports more rigorous and interpretable assessment of regression models in real-world environments.

# References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[3] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

 

**TABLE IV:** Emerging Future Research Directions in Regression Performance Evaluation

| Research Direction | Expected Impact on Model Evaluation |
|---|---|
| Hybrid Evaluation Metrics | Improved balance between accuracy and robustness |
| Adaptive Metric Selection Systems | Context-aware performance assessment |
| Explainable Evaluation Frameworks | Enhanced transparency in model interpretation |
| Real-Time Performance Monitoring | Faster detection of prediction anomalies |
| Domain-Specific Metric Design | Increased reliability in specialized applications |
| Scalable Big Data Evaluation Methods | Efficient processing of high-volume datasets |


[4] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[5] D. Harrison and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.

[6] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[7] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.

[8] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[9] S. Chai and B. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[10] J. Shearer, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2021.

[12] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[13] Z. Chen and A. J. H. Vinayakumar, "Machine learning applications in energy demand forecasting: A review," *Renewable and Sustainable Energy Reviews*, vol. 137, pp. 110–118, 2021.

[14] E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY, USA: Basic Books, 2019.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[16] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 1999.

[17] D. Chicco, N. Tötsch, and G. Jurman, "The coefficient of determination $R^2$ is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. e623, 2021.

[18] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. 36, no. 2, pp. 111–147, 1974.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[21] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. Boca Raton, FL, USA: CRC Press, 2011.

[22] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[23] R. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[24] A. Pace and R. Barry, "Sparse spatial autoregressions," *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997.

[25] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, pp. 1–26, 2018.

[26] H. L. Willis, *Power Distribution Planning Reference Book*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2004.

[27] J. Fox, *Applied Regression Analysis and Generalized Linear Models*, 3rd ed. Thousand Oaks, CA, USA: Sage Publications, 2016.

[28] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. New York, NY, USA: Wiley, 1998.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[30] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, 5th ed. Boston, MA, USA: McGraw-Hill, 2005.

[31] E. Fama and K. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.

[32] J. Hansen, M. Sato, and R. Ruedy, "Global temperature change," *Proceedings of the National Academy of Sciences*, vol. 103, no. 39, pp. 14288–14293, 2006.

[33] R. J. Hyndman and G. Athanasopoulos, *Forecasting:*


 

*Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021.

[34] S. Chopra and P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, 7th ed. Boston, MA, USA: Pearson, 2019.

[35] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *Journal of the Operational Research Society*, vol. 66, no. 8, pp. 1352–1362, 2015.

[36] J. Armstrong, *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA, USA: Kluwer Academic Publishers, 2001.

[37] S. Makridakis and M. Hibon, "The M3-competition: Results, conclusions and implications," *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476, 2000.

[38] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York, NY, USA: Wiley, 1987.

[39] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[40] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[41] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation*, 2016, pp. 265–283.

[42] E. Shortliffe and J. Cimino, *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, 4th ed. London, U.K.: Springer, 2014.

[43] J. Hull, *Options, Futures, and Other Derivatives*, 10th ed. Boston, MA, USA: Pearson, 2017.

[44] A. J. Wood and B. F. Wollenberg, *Power Generation, Operation, and Control*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[45] J. Lobell and G. Asner, "Climate and management contributions to recent trends in U.S. agricultural yields," *Science*, vol. 299, no. 5609, pp. 1032–1032, 2003.

[46] T. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd ed. San Diego, CA, USA: Academic Press, 2011.

[47] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[48] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1995, pp. 1137–1143.

[49] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge University Press, 2007.

[50] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.

[51] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[52] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, New York, NY, USA: Springer, 2013.

[53] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[54] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ, USA: Wiley, 2012.

[55] R. Hyndman and A. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[56] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge University Press, 2014.

[57] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[58] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[59] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[60] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[61] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Montreal, QC, Canada, 1995, pp. 1137–1143.

[62] T. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[63] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York, NY, USA: Wiley, 2003.

[64] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.

[65] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[66] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA, USA: Springer, 1998.

[67] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[68] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Cham, Switzerland: Springer, 2015.

[69] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[70] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Annals of Translational Medicine*, vol. 4, no. 1, pp. 9–15, 2016.

 