



COMPARATIVE STATISTICAL EVALUATION OF REGRESSION MODELS FOR IMBALANCED CONTINUOUS DATA USING ROBUST PERFORMANCE MEASURES

*Milee Brahmakshatriya

*Research Scholar, GLS University, Email: mileemadhu45@gmail.com

Corresponding author: mileemadhu45@gmail.com

ARTICLE INFO

Received: 14/09/2025

Revised: 07/12/2025

Accepted: 11/01/2026

Published: 25/01/2026

DOI: <https://doi.org/10.69980/edvdh578>

Volume:12, Issue:01, January 2026

Abstract

A great difficulty with regression modelling of imbalanced continuous outcomes is that traditional assessment methods may overestimate the accuracy of prediction in well-explored response zones and underestimate the error in poorly-explored rare extremes. The present study compared the performance of seven statistical and machine-learning regression models using powerful and specific performance measures. The models experimented with were Multiple Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Random Forest Regression, Gradient Boosting Regression and XGBoost Regression. 10-fold cross-validation was used, and errors were measured in terms of mean absolute error, root mean square error, median absolute error, R^2 , and errors specific to the normal and rare parts of the response space. The distribution of responses was negatively skewed and heavy-tailed, which is indicative of meaningful continuous-target imbalance. Support Vector Regression performed best in general with the minimum MAE, RMSE, median absolute error and maximum R^2 values and in addition had the smallest rare-region error values. Linear models had significantly larger errors especially extreme observations compared to the performance of Random Forest and XGBoost. Friedman tests showed that there were significant differences between models on all the evaluation measures. The results show the importance of using strong and region-specific assessment for practical applications to compare regression models when the target is continuous and the imbalance is high.

Keywords: Imbalanced regression; Continuous-target imbalance; Robust performance measures; Support Vector Regression; Machine learning; Rare-value prediction

1. Introduction

Many real-world problems have continuous outcomes instead of categorical labels, so regression modeling is a key aspect of statistical learning. Traditional regression methods have typically been designed to assume that the response distribution is well characterized within the range of the response. However, in the real world, continuous targets are frequently extremely skewed with concentrations of observations for certain values and few cases for important or extreme regions. Unlike class imbalance, it is fundamentally different in that rarity is not measured by classes but by a continuous response scale (Ribeiro & Moniz, 2020). This imbalance may lead to the predictive models being accurate in the areas with a high number of values but failing to perform well in areas with low numbers of values, which reduces the trustworthiness of global performance summaries.

Recent efforts have turned more to approaches specifically for imbalanced continuous outcomes. To give more weight to less frequently occurring response values when fitting a model, density-based weighting has been proposed (Steininger et al., 2021). Other works have investigated deep learning methods that explicitly consider the non-uniform density of the target and enhance representation in rare response regions (Yang et al., 2021). Another approach is to develop balanced loss functions to overcome the dominance of the more common target values in the optimization process, for visual regression (Ren et al., 2022). The results show that traditional regression goals can be inappropriate when the importance of the prediction is not necessarily equal to the frequency of observations.

Resampling has now emerged as another big focus in imbalanced regression studies. Targeted preprocessing to improve learning from rare continuous values (Arteaga et al., 2023) has been demonstrated through advanced soft-computing methods. Regressor-classifier ensembles have also been employed to pre-separate normal and rare areas prior to making continuous predictions, enhancing the sensitivity to the under-represented outcomes (Orhobor et al., 2023). In addition, there are synthetic data generation techniques which make it possible to boost the representation of rare target regions without replicating observations (Tian et al., 2023). Recently, affine combination-based oversampling has been proposed for creating informative synthetic regression samples (Li et al., 2024), and WSMOTER is an extension of relevance-based oversampling for difficult continuous-target distributions (Camacho & Bacao, 2024). Based on the findings of Belhaoui et al. (2024), oversampling strategies can be shown to significantly change the predictive behavior when the targets are highly imbalanced.

Even with these developments, model evaluation is still a methodological challenge of methodology. A model can produce large errors in a few places and small errors at the majority of points. It is possible that the model can have large errors in a few discrete places and small errors in the majority of places. Based on this, research has focused on the assessment of prediction quality in different target regions, rather than on overall prediction quality. Evidence from surveys indicates that the performance of resampling is highly sensitive to both the distribution of underlying data and the model class and evaluation criterion (Avelino et al., 2024). Feature relevance may also impact the prediction of rare values, leading to methods that leverage sparse feature selection and imbalance-aware learning (Guan & Fu, 2024). Gradient boosting has been proposed when it comes to imbalanced regression, with the clear potential of ensemble learners to boost accuracy across imbalanced target ranges (Zhang et al., 2024).

However, studies based on deep learning have shown that the imbalance can be reduced by changing the learning structure. For underrepresented target intervals, hierarchical classification adjustment can be employed to boost deep regression (Xiong & Yao, 2024), while implicit data augmentation can be used to augment the training set by exposing the model to rare response regions without re-sampling (Zhu et al., 2024). Uncertainty-aware learning has also become an attractive research area since the uncertainty of prediction is typically high under sparse training observations (Dolar et al., 2025). In sum, these results support the need for the development of evaluation frameworks that provide a comparison of several regression paradigms, including standard and powerful measures.

In this regard, the present study addresses the question of behavior of the statistical and machine-learning regression models on an imbalanced and heavy-tailed continuous-response distribution. Model quality is not evaluated by a single criterion but by the use of complementary criteria that reflect the average error, the sensitivity to large residuals, robustness to extreme errors, explained variance, and the accuracy in rare regions. This is done to get an accurate estimation of model behaviour under distributional imbalance.

Objectives:

- The first is to compare the predictive performance of several statistical and machine-learning regression models for an imbalanced continuous outcome.
- To assess the accuracy of the model with classical, robust and rare region performance measures.
- To see if the differences in model performance are statistically significant between repeated validation runs.

2. Methodology

2.1 Research Design

In this study, a quantitative comparative research design was used to compare the predictive accuracy of different multiple regression models in imbalanced continuous data. The analysis mainly centered on the comparison of conventional statistical and machine-learning regression methods adopting both standard and robust performance measures. Special attention was given to the question of whether the performance of the model was different for the most common response range and the relatively rare extreme response values.

2.2 Dataset and Variables

The dataset `ReEntrant_ML_DataSets_` was used, and the subset of the five-predictor regression was taken, which was configured as `tdata_Big_miu_5p`. There were 1000 observations in the analytical data set, and six numeric variables. L,

H, h, tL, tH were used as predictor variables and Response was used as the continuous dependent variable. The preliminary assessment indicated that there were no missing values or duplicate observations. The response variable had significant negative skewness and heavy-tailedness, justifying its use for the investigation of continuous-target imbalance (Tabacu, 2025).

2.3 Data Preprocessing

The data set was subjected to checking for missing data, repetitive data, outlying data, and inconsistencies in numbers. There were no missing or duplicate records identified so the full dataset was kept. For some algorithms like Ridge, Lasso, and Support Vector Regression, which are sensitive to scaling, the parameters were estimated on the training set and then used to predict with the test set. The models were fitted using the original predictor scale with trees. Extreme response observations were kept as these were the rare areas that were central to the study.

2.4 Assessment of Continuous-Target Imbalance

Descriptive statistics, histogram, density distribution, boxplot, skewness, kurtosis and empirical quantiles were used to examine the distribution of the Response variable. The negative skewness (-1.165) and excess kurtosis (7.644) of the response were significant, suggesting the presence of heavy tails and a fair amount of asymmetry. About 7.9% of observations were determined to be extreme using the interquartile-range-based assessment. These rare and normal response regions were then created using thresholds based on the distribution of the training data, allowing for assessment of model performance independently for the underrepresented and dominant regions.

2.5 Model Development and Validation

Seven regression models were tested: Multiple Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Random Forest Regression, Gradient Boosting Regression and XGBoost Regression. To control for dependency on a train-test split, 10-fold cross-validation was used to get consistent performance estimates. To avoid any information leakage, hyperparameters were optimized within the training folds of regularized, kernel-based and ensemble models by utilizing cross-validated searching methods.

2.6 Performance Evaluation

The performance of model was evaluated by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Median Absolute Error (MedAE) and R2 (coefficient of determination). More robust measures were added to the data analysis, specifically MAE and MedAE, which use the absolute value of the residual instead of squaring the residual as squared-error measures do. Overall performance was also calculated in normal and rare response regions to evaluate if models that performed well globally also had good performance at extreme responses that are underrepresented.

2.7 Statistical Comparison

The regression models were statistically compared based on repeated cross-validation performance, and the Friedman test was used. Pairwise Wilcoxon signed-rank tests were performed when the omnibus test showed that there were significant differences. Holm correction was used to adjust for multiple comparisons, and $p < 0.05$ was used to determine statistical significance. In addition to consistency within the response region, the model differences were interpreted as a result of consistency across multiple metrics.

2.8 Robustness Analysis

The robustness was evaluated using rankings by MAE, RMSE, MedAE, and R2 between model runs, and between model runs on average and rare target regions. Sensitivity analyses of various definitions of rare response values were also explored to see if the comparison results were sensitive to a single imbalance threshold.

2.9 Software

All analyses were performed in Python. Data preprocessing and manipulation were done with pandas and NumPy, model development and validation with scikit-learn, and the xgboost package for XGBoost modeling, while statistical testing was done with SciPy. To enhance reproducibility, a fixed random seed was used.

2.10 Ethical Considerations

The study was based on a publicly available data set, ReEntrant_ML_DataSets_, and did not require direct involvement of humans, personally identifiable information, clinical intervention or confidential information. So informed consent and institutional ethical approval were not relevant. The data set has been used only for secondary analysis, and the original data source must be cited accordingly in the final manuscript.

3. Results

3.1 Descriptive Characteristics of the Dataset

There were 1000 complete observations in the final data set, with five predictor variables and one continuous response variable. There were no missing values or duplicate observations found. The predictor variables were presented in Table 1 to show the relatively wide range of numbers. The values of the Response variable ranged from -10.0691 to 5.4841 with a mean of -0.3834 and a standard deviation of 1.3151 . There was a significant difference between the minimum, median and IQR values, which indicated that there was an asymmetric lower tail.

Table 1. Descriptive statistics of the analytical variables

Variable	Mean	SD	Minimum	Q1	Median	Q3	Maximum
L	24.6650	8.4973	10.0157	17.0764	24.6823	31.9204	39.9847
H	25.3006	8.5585	10.0102	18.0906	25.5226	32.5284	39.9843
h	20.3797	8.6795	5.0198	12.8179	20.5589	28.0491	34.9938
tL	3.0022	1.1712	1.0009	2.0027	3.0016	4.0508	4.9937
tH	3.0460	1.1468	1.0059	2.0334	3.0382	4.0367	4.9991
Response	-0.3834	1.3151	-10.0691	-0.8731	-0.2533	0.2106	5.4841

The distribution of the continuous response is shown in Figure 1. The distribution had a heavy lower tail and was skewed towards the negative values, with the majority of the distribution flattened around the central response range.

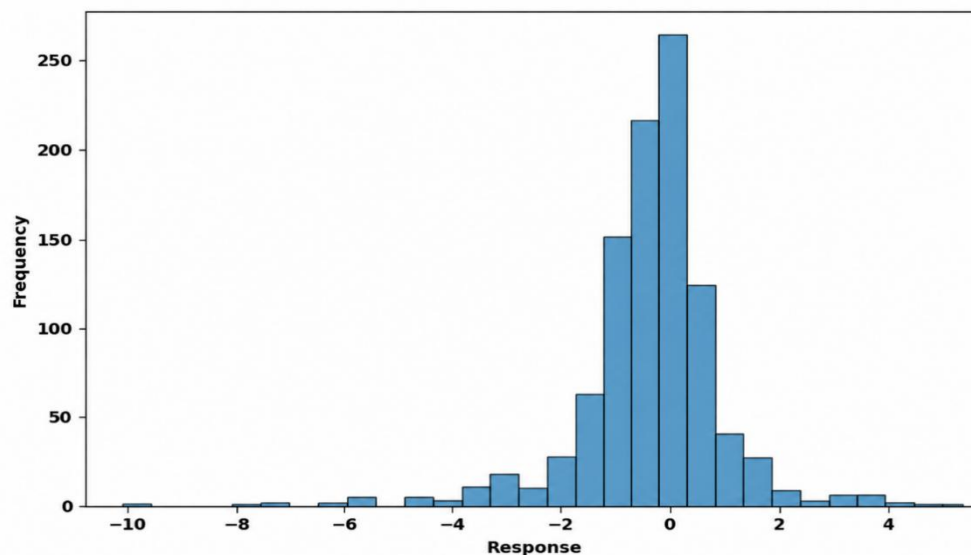


Figure 1. Distribution of the continuous Response variable showing concentration around the central range and a pronounced negative tail

3.2 Continuous-Target Imbalance

The response variable was strongly skewed, as confirmed by a formal distributional assessment. From Table 2, skewness and kurtosis were -1.1647 and 7.6441 , respectively, suggesting negative asymmetry and significant heavy-tailedness. The criterion used was the IQR method (extreme response observations were defined as those less than -2.4987 or greater than 1.8362). The 79 observations, representing 7.9% of all observations, were found in these thin tails.

Table 2. Continuous-target imbalance characteristics

Characteristic	Value
Skewness	-1.1647
Excess kurtosis	7.6441
5th percentile	-2.6652
10th percentile	-1.6174
90th percentile	0.7762
95th percentile	1.3996
Lower IQR threshold	-2.4987
Upper IQR threshold	1.8362
Lower-tail rare observations	51 (5.1%)

Upper-tail rare observations	28 (2.8%)
Total rare observations	79 (7.9%)
Normal-region observations	921 (92.1%)

Figure 2 also shows the boxplot, which further highlights the non-uniform response distribution. The lower tail is significantly extended and there are relatively few observations falling outside the IQR, hence, the data set is considered an imbalanced continuous-target regression problem.

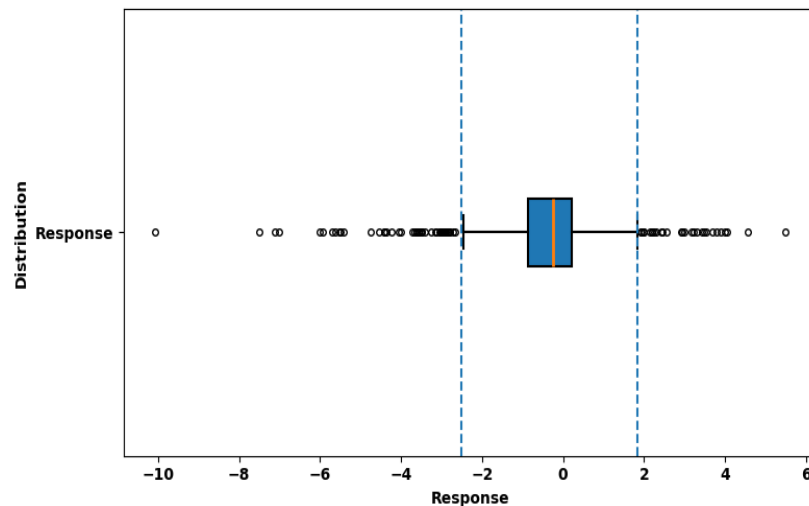


Figure 2. Boxplot of the Response variable showing the central distribution and relatively sparse extreme response values

3.3 Overall Predictive Performance of Regression Models

The seven regression methods were found to have significant differences between them. The average predictive performance is presented in Table 3 over 2 runs of 10-fold cross-validation. Support Vector Regression not only gained the lowest average MAE of 0.0718 and RMSE of 0.1344 but also the smallest median absolute error of 0.0418. The mean R^2 was 0.9883, which means that the model captured the majority of the variance in the response variable between the validation folds.

Table 3. Overall regression performance under repeated 10-fold cross-validation

Model	MAE Mean $\pm$ SD	RMSE Mean $\pm$ SD	MedAE Mean $\pm$ SD	R^2 Mean $\pm$ SD
SVR	0.0718 $\pm$ 0.0143	0.1344 $\pm$ 0.0602	0.0418 $\pm$ 0.0047	0.9883 $\pm$ 0.0093
Random Forest	0.1820 $\pm$ 0.0231	0.3718 $\pm$ 0.0969	0.0808 $\pm$ 0.0134	0.9152 $\pm$ 0.0316
XGBoost	0.1836 $\pm$ 0.0264	0.3462 $\pm$ 0.0760	0.0893 $\pm$ 0.0167	0.9263 $\pm$ 0.0242
Gradient Boosting	0.2261 $\pm$ 0.0293	0.4195 $\pm$ 0.0777	0.1177 $\pm$ 0.0151	0.8910 $\pm$ 0.0361
Lasso Regression	0.4551 $\pm$ 0.0608	0.7974 $\pm$ 0.1501	0.2475 $\pm$ 0.0306	0.6215 $\pm$ 0.0471
Ridge Regression	0.4810 $\pm$ 0.0568	0.7838 $\pm$ 0.1329	0.2901 $\pm$ 0.0361	0.6317 $\pm$ 0.0465
Linear Regression	0.4837 $\pm$ 0.0566	0.7840 $\pm$ 0.1317	0.2916 $\pm$ 0.0351	0.6313 $\pm$ 0.0468

Random Forest and XGBoost were the next best ensemble models, but they were still significantly higher than those of the SVR. The RMSE, R^2 and MAE values were slightly lower for XGBoost than for the Random Forest, while the MAE was slightly lower for Random Forest than for XGBoost. The three linear or regularized linear models had significantly

higher errors, indicating that there was a nonlinear relationship between the five predictors and the continuous response that could not be captured by a simple linear functional form. The validation results of the MAE distributions over the folds are presented in Figure 3, where SVR showed low prediction error over all validation folds. The error distributions of the linear regression, ridge regression and Lasso regression were significantly higher, while the ensemble models based on trees fell somewhere in between.

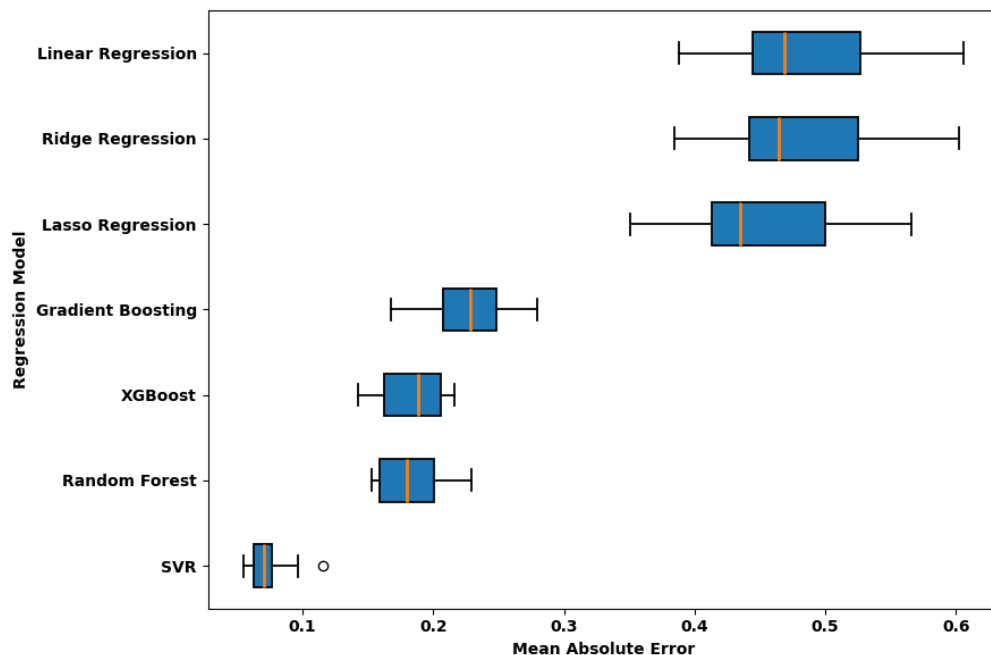


Figure 3. Distribution of mean absolute error across repeated cross-validation folds for the seven regression models

3.4 Predictive Performance in Normal and Rare Response Regions

The aperiodic-target imbalance effect was further observed when the errors in normal and rare observations were analyzed separately. As seen in Table 4, all models had significantly higher errors in the rare response region than in the dominant normal response region. However, SVR achieved the lowest rare-region MAE and RMSE values of 0.2224 and 0.3385, respectively.

Table 4. Model performance across normal and rare response regions

Model	Normal MAE	Normal RMSE	Rare MAE	Rare RMSE
SVR	0.0598	0.0973	0.2224	0.3385
Random Forest	0.1279	0.2121	0.8331	1.1035
XGBoost	0.1339	0.2148	0.7727	0.9844
Gradient Boosting	0.1644	0.2596	0.9593	1.2012
Lasso Regression	0.3064	0.4127	2.2057	2.4963
Ridge Regression	0.3544	0.4755	1.9808	2.2972
Linear Regression	0.3589	0.4817	1.9632	2.2820

Figure 4 displays the difference in MAE between the normal region and the rare region. The linear and regularized linear

models showed a higher increase in prediction error. For instance, Linear Regression went from 0.3589 in the normal region to 1.9632 in the rare region. An even larger rare-region MAE of 2.2057 was achieved by Lasso Regression. Overall performance measures can significantly underestimate the difficulty of prediction with sparsely represented response values as is shown here.

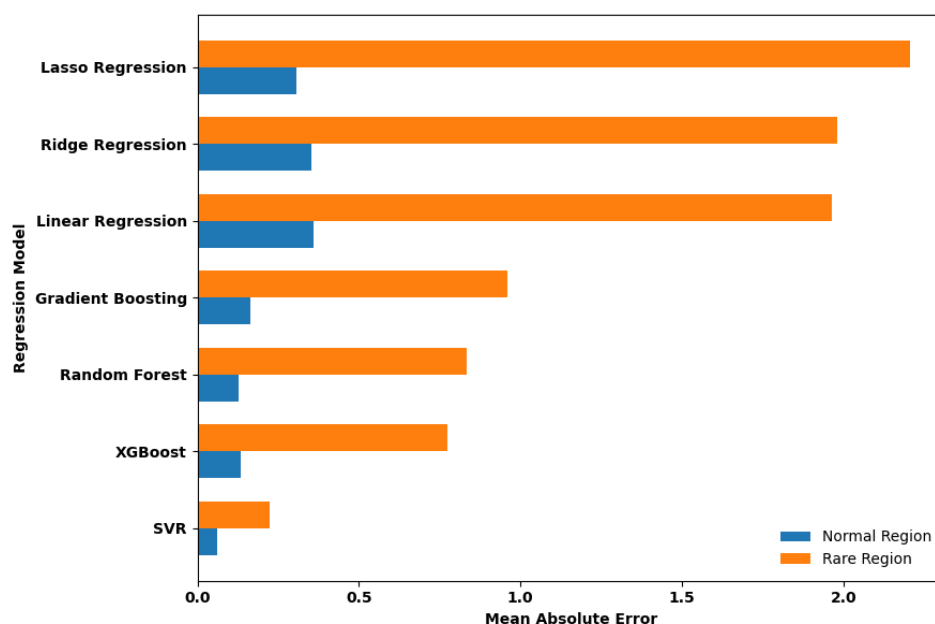


Figure 4. Comparison of mean absolute error in normal and rare continuous-response regions across regression models

3.5 Observed and Predicted Responses

The out-of-fold predictions were compared to the actual response values, since SVR yielded the smallest errors, both overall and in the rare regions. As shown in Figure 5, most of the predictions were clustered near the 45-degree line, encompassing the entire central range of the responses. The larger deviations were found to the extreme response values which in turn was associated with higher error estimates for rare observations.

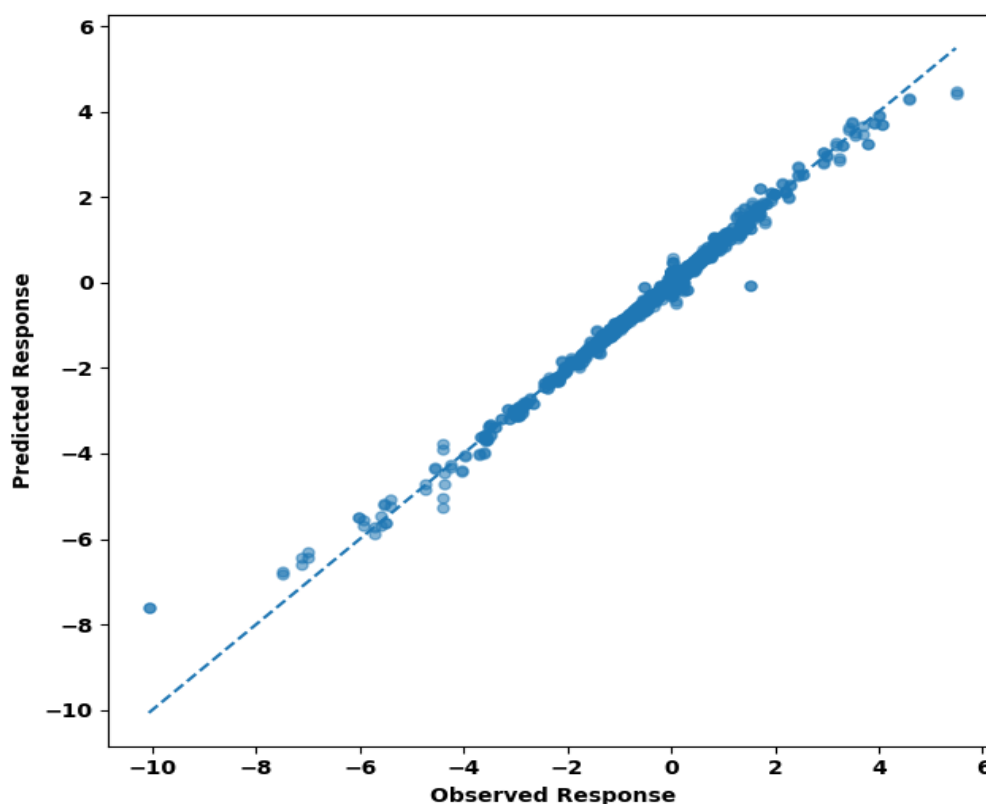


Figure 5. Observed versus cross-validated predicted Response values for Support Vector Regression with the 45-degree line indicating perfect prediction

3.6 Statistical Comparison of Model Performance

Statistically, the differences in performance at the various folds were tested with the Friedman test. As seen in Table 5, there were significant differences between the models for all four evaluation criteria. The Friedman test yielded a statistic of 116.66 with a p-value of <0.001 for MAE and a p-value of <0.001 for MedAE. There were also considerable between-model differences for RMSE and R^2 .

Table 5. Friedman tests for differences in regression-model performance

Performance measure	Friedman statistic	p-value	Interpretation
MAE	116.66	<0.001	Significant model differences
RMSE	108.49	<0.001	Significant model differences
MedAE	114.88	<0.001	Significant model differences
R^2	108.49	<0.001	Significant model differences

Holm-adjusted Wilcoxon signed-rank comparisons of fold-level MAE revealed that SVR was significantly different from all of the other six regression models (adjusted $p < 0.001$). The majority of the other model pairs were also significantly different. However, for Random Forest and XGBoost, the RMSE and R^2 differences were not large enough to indicate that the MAE distributions were significantly different ($p = 0.927$) after adjustment. Together, these results demonstrate that overall regression error would not be an adequate measure of performance in the presence of continuous-target imbalance. Ensemble models significantly outperformed linear models but the error rose rapidly in the rare response area. SVR exhibited the lowest values of MAE, RMSE, MedAE and R^2 and comparatively moderate rare-region errors among the evaluated models.

4. Discussion

The present study clearly illustrated the differences between regression algorithms for processing an imbalanced continuous target. Support Vector Regression has the lowest mean absolute error (MAE), root mean square error (RMSE), median absolute error (MedAE) and the highest coefficient of determination (R^2) and has performed the best. The same holds true for support vector methods which are able to model the nonlinear relationships using kernel based transformation but can be controlled in order to keep the model complexity low; this can be beneficial when the response structure is also uneven, with sparse extreme values (Moguerza et al., 2020). The relatively low rare region error (RRE) achieved by SVR indicates that the epsilon-insensitive nature of SVR was effective in fitting the nonlinear predictor-response relationships found in this study.

Tree based ensemble models also on average achieved significantly better performance than linear models, but the accuracy deteriorated significantly in the rare response region. Random Forest achieved low overall prediction error and was also competitive across the folds of the validation set, thus corroborating the results of Fawagreh & Gaber (2020) that ensemble aggregation can enhance the stability of a regression function by combining several decorrelated trees. XGBoost had a slightly smaller RMSE and a larger explained variance than Random Forest, suggesting better control of larger residuals. Gradient-boosted trees can be used to achieve effective regression performance in a sequence of error pattern correction (Zhang et al., 2020). However, the errors in the rare regions were much higher for both ensemble methods than for SVR, demonstrating that the ensemble methods can be trained to be strong on the global accuracy but not on the accuracy of rare outcomes.

Linear Regression, Ridge Regression and Lasso Regression had the poorest results. The larger MAE and RMSE values indicate that the model was not capturing the entire relationship between the predictor and response with a linear model only. It is also worth noting that the heavy-tailed response distribution raises the chance of extreme observations having an undue influence on the conventional estimation. A strong emphasis in regression research is on minimizing sensitivity to extreme observations and large errors when the data violate the normality of the distribution assumptions (Björklund et al., 2022). In sparse regression, outliers could also affect the estimation of the regression coefficients and variable selection if robustness is not taken into account (Bottmer et al., 2022). Good modeling methods are therefore applicable for prediction in regions of central and extreme outcomes (Mishra & Müller, 2022).

One key result was the significant difference between the errors of prediction in the normal region and rare region. All models had higher MAE and RMSE for rare observations, which indicates that the aggregate measures may mask local lack of ability when dealing with imbalanced regression. Predictive behavior cannot be completely captured by a single metric. Evaluation research highlights that the choice of metrics should not only be based on one measure of the error structure of prediction failure, but on the practical implications of such failure (Hicks et al., 2022). Broadly speaking, results can also vary according to the focus of the evaluation whether on average error, squared error, explained variance, or robustness (Miller et al., 2024). The current findings thus indicate the use of combined metrics of MAE, RMSE, MedAE, R^2 and regional errors.

The comparative interpretation was enhanced by the statistical analysis. Friedman test revealed significant differences for all the measures examined between the seven models, and Holm adjusted Wilcoxon comparison demonstrated that SVR was significantly different from each alternative in fold-level MAE. Random Forest and XGBoost did not significantly differ after adjustment on MAE, which is a numerical difference, and not a statistical difference. When evaluating

algorithms across multiple samples, the following algorithm evaluation guidelines suggest the use of more than one metric and to perform formal statistical testing (Rainio et al., 2024). The use of repeated cross-validation further minimized the risk of overreliance on a single partition and also yielded a more consistent estimate of the performance across partitions, in line with existing theory that recommends cross-validation for model assessment and selection (Wilson et al. 2020).

Collectively, the results show that continuous-target imbalance is indeed an evaluation and modelling challenge. An algorithm that works well in the bulk of the data may be poor in the tails of the distribution. An algorithm that does well overall may not do well in the sparsely covered tails. The results were consistent across overall and rare regions with the best performance being recorded by the SVR methods, intermediate performance was obtained using ensemble methods and the linear methods were prone to the errors of extreme responses. Larger, more varied data sets, other rarity measures, targeted resampling strategies, and uncertainty-based assessments could be used in future studies to see if the relative patterns are representative of other domains. These findings help to support a strong regression.

5. Conclusion

This paper compared and contrasted the statistical and machine-learning regression models for predicting imbalanced continuous outcomes with the use of conventional performance measures as well as robust performance measures. The results indicated that the imbalance of the targets had a significant effect on prediction accuracy, especially in the extreme response areas with a small number of targets. SVR showed the best overall performance with the lowest MAE, RMSE, and MedAE values and the highest R^2 value; in addition, it also showed relatively low prediction errors for some extreme values. The nonlinear alternatives, such as Random Forest and XGBoost, were competitive, but showed some degradation in rare regions. On the other hand, Linear, Ridge, and Lasso Regression yielded much higher errors suggesting poor linear model performance for complex and heavy-tailed response structures. Notably, all the models evaluated had higher errors in the rare region than in the dominant response region, meaning that overall model performance indicators can mask flaws in predicting rarer responses. The Friedman test showed significant differences between the performance of the regression methods, and the post-hoc comparisons further revealed that SVR has the best consistency among the regression methods. In general, the results indicate that when assessing imbalanced regression models, MAE, RMSE, MedAE, R^2 , region-specific errors and repeated cross-validation, in addition to statistical testing, are essential. To validate these results, future research is needed with larger data sets, different rarity thresholds, imbalance-aware resampling methods and uncertainty-based prediction frameworks in various application areas and real-world scenarios.

References

1. Arteaga, M., Gacto, M. J., Galende, M., Alcalá-Fdez, J., & Alcalá, R. (2023). Enhancing soft computing techniques to actively address imbalanced regression problems. *Expert Systems with Applications*, 234, 121011.
2. Avelino, J. G., Cavalcanti, G. D., & Cruz, R. M. (2024). Resampling strategies for imbalanced regression: a survey and empirical analysis: JG Avelino et al. *Artificial Intelligence Review*, 57(4), 82.
3. Belhaouari, S. B., Islam, A., Kassoul, K., Al-Fuqaha, A., & Bouzerdoum, A. (2024). Oversampling techniques for imbalanced data in regression. *Expert systems with applications*, 252, 124118.
4. Björklund, A., Henelius, A., Oikarinen, E., Kallonen, K., & Puolamäki, K. (2022). Robust regression via error tolerance. *Data Mining and Knowledge Discovery*, 36(2), 781-810.
5. Bottmer, L., Croux, C., & Wilms, I. (2022). Sparse regression for large data sets with outliers. *European Journal of Operational Research*, 297(2), 782-794.
6. Camacho, L., & Bacao, F. (2024). WSMOTER: a novel approach for imbalanced regression: L. Camacho and F. Bacao. *Applied intelligence*, 54(19), 8789-8799.
7. Dolar, T., Chen, J., & Chen, W. (2025). Uncertainty quantification driven machine learning for improving model accuracy in imbalanced regression tasks. *Expert Systems with Applications*, 261, 125526.
8. Fawagreh, K., & Gaber, M. M. (2020). Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach: K. Fawagreh, MM Gaber. *Computing*, 102(5), 1187-1198.
9. Guan, Y., & Fu, G. H. (2024). Sparse feature selection and rare value prediction in imbalanced regression. *Information Sciences*, 680, 121145.
10. Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1), 5979.
11. Li, Z. Z., Huang, N., Yi, L. Z., & Fu, G. H. (2024). Affine combination-based over-sampling for imbalanced regression. *Journal of Chemometrics*, 38(3), e3537.
12. Miller, C., Portlock, T., Nyaga, D. M., & O'Sullivan, J. M. (2024). A review of model evaluation metrics for machine learning in genetics and genomics. *Frontiers in Bioinformatics*, 4, 1457619.
13. Mishra, A., & Müller, C. L. (2022). Robust regression with compositional covariates. *Computational Statistics & Data Analysis*, 165, 107315.
14. Moguerza, J. M., Muñoz, A., & Venturini, G. M. (2020). Support Vector Regression.
15. Orhobor, O. I., Grinberg, N. F., Soldatova, L. N., & King, R. D. (2023). Imbalanced regression using regressor-classifier ensembles. *Machine Learning*, 112(4), 1365-1387.
16. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific reports*, 14(1), 6086.
17. Ren, J., Zhang, M., Yu, C., & Liu, Z. (2022, June). Balanced mse for imbalanced visual regression. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7916-7925). IEEE.
18. Ribeiro, R. P., & Moniz, N. (2020). Imbalanced regression and extreme value prediction. *Machine Learning*, 109(9), 121011.

19. Steininger, M., Kobs, K., Davidson, P., Krause, A., & Hotho, A. (2021). Density-based weighting for imbalanced regression. *Machine Learning*, 110(8), 2187-2211.
20. Tabacu, S. (2025). ReEntrant_ML_DataSets_ (Version 2) [Data set]. Mendeley Data. <https://doi.org/10.17632/z4m93zsb2s.2>
21. Tian, H., Tian, C., Li, K., & Jia, W. (2023). Unbalanced regression sample generation algorithm based on confrontation. *Information Sciences*, 642, 119157.
22. Wilson, A., Kasy, M., & Mackey, L. (2020, June). Approximate cross-validation: Guarantees for model assessment and selection. In *International conference on artificial intelligence and statistics* (pp. 4530-4540). PMLR.
23. Xiong, H., & Yao, A. (2024, June). Deep imbalanced regression via hierarchical classification adjustment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 23721-23730). IEEE.
24. Yang, Y., Zha, K., Chen, Y., Wang, H., & Katabi, D. (2021, July). Delving into deep imbalanced regression. In *International conference on machine learning* (pp. 11842-11851). PMLR.
25. Zhang, W., Shi, P., Jia, P., & Zhou, X. (2024). A novel gradient boosting approach for imbalanced regression. *Neurocomputing*, 601, 128091.
26. Zhang, X., Yan, C., Gao, C., Malin, B. A., & Chen, Y. (2020). Predicting missing values in medical data via XGBoost regression. *Journal of healthcare informatics research*, 4(4), 383-394.
27. Zhu, W., Wu, O., & Yang, N. (2024). IRDA: Implicit data augmentation for deep imbalanced regression. *Information Sciences*, 677, 120873.

EPH – International Journal of Mathematics and Statistics

<https://ejjms.com/index.php/ms/index>



This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article, You are free to: Share — copy and redistribute the material in any medium or format for any purpose, even commercially. To see the complete license contents,

please visit <http://creativecommons.org/licenses/by/4.0/>.