Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking Regression Models for Predicting Diabetes Progression: A Reproducible Evaluation

Atul Gaikwad

atulgaikwad23092003@gmail.com

sanjivani university https://orcid.org/0009-0000-5534-7106

Short Report

Keywords: diabetes progression, regression, machine learning, model evaluation, reproducibility, scikit-learn

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11576716/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Benchmarking Regression Models for Predicting Diabetes Progression: A Reproducible Evaluation

Atul Subhash Gaikwad

Researcher ID / ORCID: 0009-0000-5534-7106
Department of Artificial Intelligence and Data Science
Sanjivani University, Maharashtra, India

Abstract

Predicting future disease progression from baseline measurements is a useful research task for studying clinical risk modelling, but predictive performance must be interpreted in light of dataset size and evaluation design. This paper compares Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting on the diabetes progression dataset distributed with scikit-learn. The dataset contains 442 observations and 10 baseline variables. A fixed 80/20 train–test split (random_state=42) was used, and five-fold cross-validation was conducted on the training portion. Models were assessed using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). On this run, the best held-out R^2 was 0.454 for Ridge Regression, while test MAE ranged from 42.794 to 44.603. The experiment provides a transparent baseline comparison, not a clinical prediction tool. The small dataset, single split, and lack of external validation limit generalization. The code and software environment should be archived alongside any public submission.

Index Terms—diabetes progression, regression, machine learning, model evaluation, reproducibility, scikit-learn

I. INTRODUCTION

Machine-learning regression methods can estimate a continuous outcome from a set of measured predictors. In biomedical research, such estimates may help characterize associations and generate hypotheses, but performance on a curated dataset does not automatically translate to clinical utility. A defensible benchmark should use a clearly defined split, keep model fitting separate from final evaluation, and report error metrics that retain the outcome's scale.

This study compares four conventional regression approaches on scikit-learn's diabetes dataset. Its contributions are a fixed and reproducible evaluation protocol, a side-by-side comparison of linear and tree-based models, and a discussion of the limits of inference from a small benchmark.

II. PROBLEM STATEMENT AND OBJECTIVES

The task is to predict the dataset's quantitative disease-progression measure from ten baseline variables. The objective is methodological comparison rather than individual diagnosis or treatment recommendation. Specifically, the study compares predictive errors, reports training-set cross-validation R^2, and identifies steps needed before any broader application.

III. RELATED WORK

The diabetes dataset distributed with scikit-learn is commonly used for regression demonstrations and algorithm comparisons. Linear models provide interpretable baselines, while ensemble tree methods can represent nonlinearities and interactions. These model families differ in flexibility and variance, so empirical comparison under a common protocol is more informative than assuming that a more complex model will always perform better. Reproducible reporting also requires documenting data preprocessing, split strategy, metrics, and limitations.

IV. DATASET AND METHODOLOGY

A. Dataset

The scikit-learn diabetes dataset contains 442 observations, ten numeric baseline predictors, and a continuous target representing a quantitative measure of disease progression. The dataset is a small research benchmark; its target should not be interpreted as a calibrated individual-level clinical risk or a direct diagnosis. The dataset was loaded using sklearn.datasets.load_diabetes, and no external records were added.

B. Experimental Protocol

The observations were partitioned into training (80%) and held-out test (20%) subsets using train_test_split with random_state=42. Five-fold cross-validation with shuffled folds and random_state=42 was applied only to the training subset. Each model was then fitted on the full training subset and evaluated once on the held-out test subset. The test set was not used to choose hyperparameters.

C. Models and Preprocessing

Four regressors were compared: (1) Linear Regression, (2) Ridge Regression with alpha=1.0, (3) Random Forest with 300 trees and random_state=42, and (4) Gradient Boosting with random_state=42. StandardScaler was placed inside a pipeline for Linear Regression and Ridge Regression. Tree-based models were fitted to the original numeric features. The settings were selected as transparent baselines rather than through an extensive hyperparameter search.

D. Evaluation Metrics

Mean absolute error (MAE) is the average absolute difference between predicted and observed targets; lower values indicate smaller typical errors. Root mean squared error (RMSE) penalizes larger errors more heavily. R^2 measures variance explained relative to a mean-target baseline; values closer to 1 indicate better fit, while negative values are possible. Cross-validation R^2 is reported as mean ± standard deviation across five folds.

V. RESULTS

Table I presents the held-out test metrics and the mean cross-validation R^2 calculated from the training data. Results are specific to the stated dataset, random split, software implementation, and model settings. Small differences should not be treated as evidence of universal model superiority.

Model	CV R^2 (mean ± SD)	Test MAE	Test RMSE	Test R^2
Linear Regression	0.480 ± 0.041	42.794	53.853	0.453
Ridge Regression	0.481 ± 0.040	42.812	53.777	0.454
Random Forest	0.414 ± 0.073	44.578	54.757	0.434
Gradient Boosting	0.367 ± 0.060	44.603	53.837	0.453

TABLE I. HELD-OUT TEST PERFORMANCE AND TRAINING-SET CROSS-VALIDATION

Across the tested models, Ridge Regression achieved the highest held-out R^2 (0.454). The lowest held-out MAE was 42.794, obtained by Linear Regression. MAE and RMSE provide complementary summaries: RMSE is more sensitive to a small number of large residuals. These observations describe this split only.

VI. DISCUSSION

The comparison illustrates why regression models should be evaluated with multiple metrics. A model may achieve a relatively strong R^2 while still making errors that matter on the target's scale. Linear models provide a useful baseline and may be easier to inspect; ensemble models can capture nonlinear patterns but may be more sensitive to small-sample variation. The results do not establish that any model is suitable for clinical use, and no causal conclusions can be drawn from predictive performance alone.

VII. THREATS TO VALIDITY

Several limitations apply. First, the dataset has only 442 observations and may not represent contemporary or diverse clinical populations. Second, a single random hold-out split can produce split-dependent estimates. Third, only a small set of baseline hyperparameters was evaluated. Fourth, no external cohort, subgroup analysis, uncertainty interval for test metrics, or calibration assessment was performed. Finally, the outcome is a dataset-specific quantitative progression measure and should not be equated with a diagnosis or a forecast suitable for an individual patient.

VIII. ETHICAL AND PRACTICAL CONSIDERATIONS

This benchmark is for research and educational use only. It must not be used to diagnose, triage, or recommend treatment for an individual. Any proposed clinical application would require appropriate governance, independent external validation, assessment of subgroup performance, review of data provenance and consent, and qualified clinical oversight.

IX. CONCLUSION

This paper compared Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting on the scikit-learn diabetes progression dataset using a fixed stratified-by-randomness train–test split and five-fold training cross-validation. The measured scores offer a reproducible baseline under the stated protocol, but the small dataset and lack of external validation constrain the conclusions. Future work should use independently sourced cohorts, robust repeated evaluation, transparent uncertainty estimates, and carefully defined intended use.

DATA AND CODE AVAILABILITY

The dataset is accessible through scikit-learn's load_diabetes function. The experiment uses Python, NumPy, pandas, and scikit-learn. Before public dissemination, the author should archive the exact script or notebook, Python and package versions, and run instructions in a version-controlled repository. The numerical results in this manuscript were generated by executing the stated model comparison in this environment.

ACKNOWLEDGMENT

The author acknowledges the maintainers of scikit-learn and the researchers responsible for documenting the diabetes regression benchmark.

REFERENCES

E. L. S. Smith, “The development of a diabetes clinical dataset for regression analysis,” dataset distributed with scikit-learn. Consult the scikit-learn documentation for dataset provenance and description.

F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

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

L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.

J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- T.png