

Predicting a Simulated Outcome Variable Using Ensemble Regression Methods

Hissan Omar — K23136072 — 5CCSAMLf (CW1)

February 2026

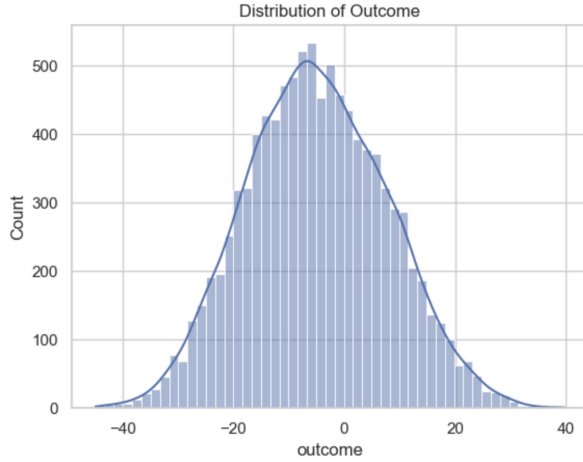
This paper investigates a supervised regression problem using a simulated tabular dataset to predict continuous outcome variables from numerical and categorical features using several model families.

1 Exploratory Data Analysis

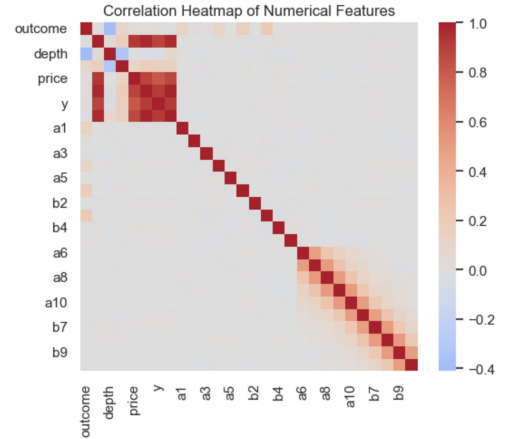
The training dataset consists of 10000 observations and 31 variables including continuous target variables (`outcome`), 27 numerical features and 3 categorical features (`cut`, `color`, and `clarity`). All variables were fully observed with no missing values present in the dataset. In addition to visual inspection, several preprocessing considerations were brainstormed. As no missing values were present, no change was required. Categorical variables (`cut`, `color`, and `clarity`) were treated as nominal features and encoded using 'one hot' encoding to allow their inclusion in regression based models. Numerical features were standardised to zero mean and unit variance to ensure comparable feature scales and to stabilise optimisation for regularised and ensemble based methods.

No features were excluded, as exploratory analysis did not reveal redundant variables. Preprocessing was implemented within a unified pipeline to prevent data leakage and to ensure consistent transformations across cross validation folds.

Outcome variable analysis



(a) Distribution of the outcome variable.



(b) Correlation heatmap of numerical features.

Figure 1: Exploratory analysis of the outcome variable and feature relationships.

Figure 1 summarises key exploratory findings. The outcome variable exhibits an approximately unimodal and symmetric distribution with moderate tails and a slight skew, suggesting a well-behaved continuous target suitable for regression modelling without transformation. No extreme skewness or heavy outliers were observed, reducing the need for robust loss functions or target transformations.

The correlation heatmap reveals that most numerical predictors exhibit weak to moderate pairwise linear correlations, both with each other and with the outcome variable. This suggests that predictive signal is likely distributed across multiple features rather than dominated by a small subset of strongly correlated

variables. Consequently, models capable of capturing non linear effects and feature interactions are expected to outperform purely linear approaches.

2 Model Selection

Given the tabular nature of the dataset and the mixture of numerical and categorical features, many regression model families were considered and compared directly. The objective was to identify a model that maximised out-of-sample predictive performance.

Firstly, linear models (ordinary least squares, Ridge, Lasso, and Elastic Net) were evaluated as baselines due to their simplicity and interpretability. However their performance was limited with cross-validated R^2 scores around 0.28, indicating that linear assumptions were insufficient to capture the relationship between features and the target variable.

Secondly, tree based ensemble methods were explored to model non linear types of relationships and feature interactions. Random Forest and Extremely Randomised Trees (ExtraTrees) substantially improved performance, achieving cross validated R^2 scores of approximately 0.44–0.46 which suggests better alignment with the data structure.

Finally, gradient boosting methods, including Gradient Boosting Regression and XGBoost, were evaluated. These models iteratively refine predictions by fitting weak learners to residuals. Gradient Boosting Regression achieved the highest and most stable performance, with a mean cross-validated R^2 of approximately 0.47, while XGBoost achieved comparable but slightly lower results.

Overall, Gradient Boosting Regression was selected as the final model. The convergence of performance across several ensemble methods suggests that there exists irreducible noise in the simulated dataset, making Gradient Boosting the best choice, offering predictive performance, stability, and model complexity.

3 Model Training and Evaluation

Following model selection, Gradient Boosting Regression was trained using a pipeline combining preprocessing and model fitting. Categorical variables were one-hot encoded, while numerical features were standardised to ensure stable optimisation and consistent treatment across folds.

Model performance was evaluated using 5-fold cross-validation with the coefficient of determination (R^2) as the primary metric. Cross-validation was chosen to provide a robust estimate of generalisation performance and to reduce variance associated with a single train–test split. **The final Gradient Boosting model achieved a mean cross-validated R^2 of approximately 0.47**, with low variability across folds, indicating stable and reliable performance.

Key hyperparameters were selected empirically to balance bias and variance:

1. **Number of estimators**: set to 300 to allow sufficient model capacity.
2. **Learning rate**: set to 0.05 to ensure gradual refinement of predictions.
3. **Maximum tree depth**: limited to 3 to reduce overfitting by constraining individual weak learners.
4. **Random seed**: fixed to ensure reproducibility.
5. **Subsample**: value of 1.0, showing standard Gradient boosting was chosen instead of stochastic.

In conclusion, despite the experimentation with alternative model families and hyperparameters, performance improvements beyond this level were marginal. The convergence of results across ensemble-based approaches suggests that the remaining error is largely irreducible, consistent with the simulated nature of the dataset. As such, the selected model is considered to be the closest model to achieve optimal performance given the available information.

4 Code Supplement

Here is the link to the repository. After cloning the repository, please take a look at the `README.md` file for instructions on installation and setup: <https://github.com/Hissan7/data-science-regression-model>