5CCSAMLF Data Challenge Report

Yingtao Zheng K23158987

18th Feburary 2026

1 Introduction

The objective of this coursework is to develop a regression model that predicts the continuous target variable *outcome* using the provided training dataset and to generate predictions for an unseen test dataset. Model performance is evaluated using out-of-sample R^2, and this report summarises the data exploration, preprocessing, model selection, and training pipeline used to construct the final predictor.

2 Exploratory Data Analysis

The dataset contains both numeric and categorical variables, with *outcome* as a continuous regression target. Categorical features include *cut, color, clarity*, while remaining variables are numeric measurements and engineered features.

The data contained minimal missing values and no significant duplication, so simple imputation was sufficient. The target distribution showed moderate skew but did not require transformation. Numeric features varied in scale, and some engineered variables contained negative values, making logarithmic transformation unsuitable.

Categorical variables demonstrated predictive value, as boxplots of *outcome* by category showed clear differences between levels. All categorical levels present in the training data also appeared in the test data, enabling safe one-hot encoding with unknown categories ignored.

Correlation analysis indicated that several numeric variables had meaningful relationships with the target variable, confirming the dataset contains predictive structure. A comparison of training and test distributions showed no significant distribution shift, suggesting that the trained model should generalise well.

Based on this analysis, preprocessing included median imputation, one-hot encoding, and feature standardisation, with no explicit outlier removal.

3 Model Selection

Multiple regression algorithms were evaluated using the same preprocessing pipeline. Model choice was guided by both commonly used approaches in public machine learning practice and direct empirical comparison on this dataset. The models tested included Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, and Gradient Boosting.

A fixed train–validation split was used to estimate out-of-sample performance using R^2, allowing a fair comparison between models under identical conditions. While linear and regularised models provided a useful baseline, ensemble tree-based models captured non-linear feature interactions more effectively and achieved higher predictive performance.

Based on validation R^2, **Gradient Boosting** and **Random Forest** were selected as the best-two-performing model for further tuning.

4 Model Training and Hyperparameter Tuning

Hyperparameters of the strongest models were tuned using cross-validation on the training split. For Random Forest, parameters such as the number of trees, maximum depth, and minimum samples per split were explored. For Gradient Boosting, the number of estimators, learning rate, tree depth, and subsampling rate were tuned.

After testing all 48 different combinations of hyperparameters, the best configuration achieved was a validation R^2 of **0.4719**. Compared with the baseline model, this represented a significant improvement, demonstrating the effectiveness of model selection and hyperparameter optimisation.

The final model was **Gradient Boosting Regression** with hyperparameters **n_estimators = 200, learning_rate = 0.1, max_depth = 2, subsample=1.0**. This model was retrained on the full training dataset before generating predictions for the test set. Consistent preprocessing and feature alignment ensured no mismatch occurred between training and test features.

5 Final Model and Prediction

The final model was trained on the complete training dataset and used to produce predictions for the unseen test dataset. The output was formatted as a single-column CSV file containing predicted values (*yhat*), as required by the coursework specification.

The modelling pipeline was designed to ensure:

- Consistent preprocessing between training and test data
- No data leakage by fitting preprocessing only on training data
- Reproducibility through fixed random seeds
- Stable generalisation performance

6 Code Availability

The full implementation, including preprocessing, model selection, hyperparameter tuning, and prediction generation, is available at:

`https://github.com/sdsdsdasa/5CCSAMLF-CW1`

7 Conclusion

This coursework developed a complete regression pipeline including data exploration, preprocessing, model selection, and hyperparameter tuning. Ensemble tree models outperformed linear baselines, and tuning further improved performance. The final model was trained on the full dataset and generates valid predictions for unseen data. Overall, the pipeline is robust, reproducible, and achieves strong out-of-sample performance.