5CCSAMLF Data Challenge Report

Yingtao Zheng K23158987

18th Feburary

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 training dataset contains a mixture of numeric and categorical variables. The target variable *outcome* is continuous, confirming a supervised regression problem. The categorical features are *cut*, *color*, and *clarity*, while the remaining variables are numeric, including physical measurements (*carat, depth, table, x, y, z*), price-related information, and engineered variables (*a1–a10, b1–b10*).

Initial analysis revealed minimal missing values and no significant duplicate observations, therefore only simple imputation was required for robustness. The distribution of the target variable was moderately skewed but not extreme, so no transformation was applied. Histograms of numeric features showed variation in scale and distribution, and several 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 these observations, the preprocessing pipeline included:

- Median imputation for numeric features and most-frequent imputation for categorical features
- One-hot encoding for categorical variables
- Standardisation of numeric features to improve optimisation stability
- No explicit outlier removal, as tree-based and regularised models are robust to moderate outliers

3 Model Selection

Multiple regression algorithms were evaluated using the same preprocessing pipeline. The models tested included Linear Regression (baseline), Ridge, Lasso, and ElasticNet (regularised linear models), as well as ensemble tree methods including 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. The baseline Linear Regression model provided a reference performance but was limited in capturing non-linear relationships. Regularised linear models improved stability slightly, but ensemble tree models achieved substantially higher validation performance due to their ability to capture complex non-linear feature interactions.

Among the tested models, **Gradient Boosting Regression** achieved the highest validation R^2 and was therefore selected 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.

The best configuration achieved 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 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.