5CCSAMLF Machine Learning Coursework 1: Data Processing and Model Training Report

Jason Immanuel
K23169531@kcl.ac.uk

Abstract

This report outlines the end-to-end machine learning pipeline developed to predict the target variable 'outcome' in the synthetic Diamonds dataset. Through detailed EDA, non-linearities involving 'depth' and skewed numerical features were identified. Four regression models were evaluated: Linear Regression, Random Forest, XGBoost, and GBM. The GBM model, optimized by grid search, achieved an R^2 of 0.4575 and RMSE of 9.18, confirming the need for non-linear ensemble techniques. Finally, we discuss limitations and potential directions for further improvement.

1 Introduction

The objective is to predict a continuous target from 10,000 records. The dataset contains standard diamond features (e.g., carat, cut, price) and synthetic numerical features. This report details preprocessing, model selection, evaluation, and insights extracted from feature importance.

2 Exploratory Data Analysis

The dataset contains 31 features (28 numerical, 3 categorical) with no missing values.

2.1 Distributions and Observations

The target, *the outcome*, is approximately normal ($\mu \approx -4.9, \sigma \approx 12.7$), which does not require transformations. Several predictors exhibit a large skew. For instance, *carat* and *price* are right-skewed, while synthetic features *b1* and *b3* show minor tails. Visual inspections of the graphs confirm these distributions, highlighting extreme values and asymmetry. Such characteristics indicate potential challenges for models sensitive to skewed inputs, especially linear methods.

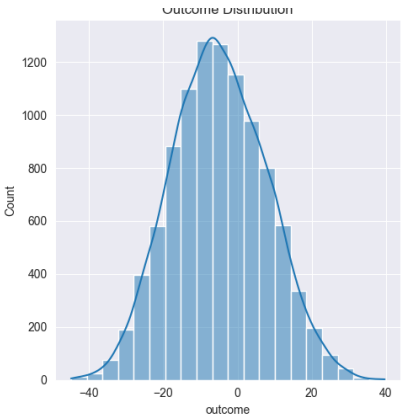

Figure 1: Target Variable Distribution

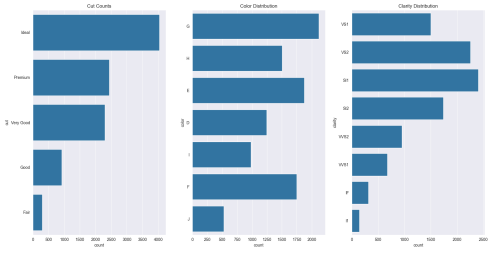

Figure 2: Cut, Color, Clarity Distribution

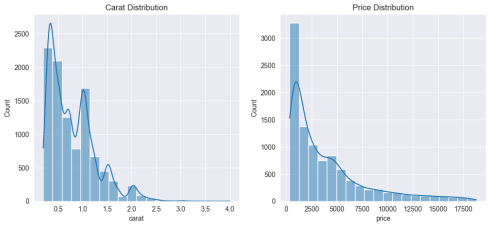

Figure 3: Feature Skewness

2.2 Feature Correlations

Pearson correlation analysis shows weak linear correlations for most features, except 'depth' ($r = -0.41$). Pairwise scatterplots reveal non-linear and noisy relationships, especially between 'depth' and 'outcome'.

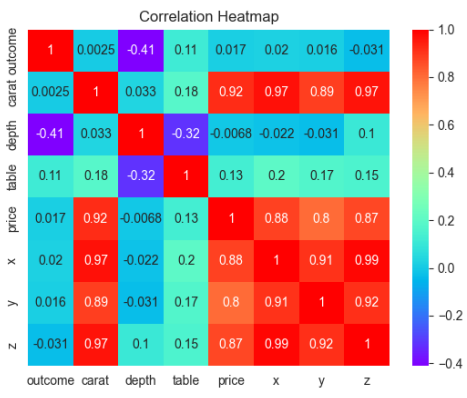

Figure 4: Correlation Matrix of Features

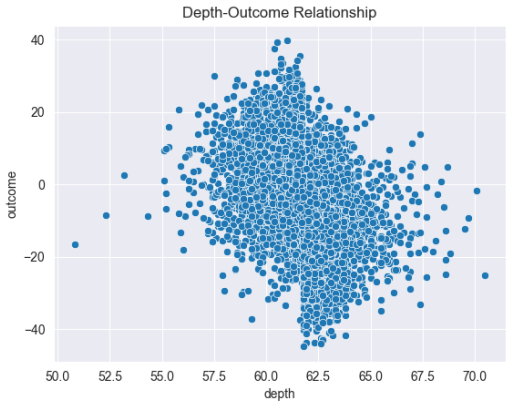

Figure 5: Depth vs Outcome Relationship

3 Methodology

3.1 Preprocessing Pipeline

A `Scikit-Learn` pipeline was implemented:

- **Categorical Encoding:** One-Hot Encoding for *cut*, *color*, *clarity*.
- **Feature Scaling:** Standardization (`StandardScaler`) for numerical features.
- **Imputation:** `SimpleImputer` (median) for strength in unseen data.

3.2 Model Selection

Evaluated models:

- **Linear Regression:** baseline for non-linear comparison.
- **Random Forest:** ensemble of trees, robust to noise and non-linear patterns.
- **XGBoost:** gradient boosting with L1/L2 regularization to prevent overfitting.
- **GBM:** iterative boosting optimizing squared error loss.

3.3 Model Evaluation Strategy

Data was split 80/20 for training/testing. 5-fold cross-validation was used during hyperparameter tuning. Metrics:

- R^2: proportion of variance explained, intuitive for regression.
- RMSE: measures prediction error in the original scale, sensitive to large deviations.

4 Results and Discussion

4.1 Model Performance

Table 1: Test Set Performance

Model	R^2	RMSE
Linear Regression	0.2762	10.82
Random Forest	0.4324	9.45
XGBoost	0.4460	9.32
GBM	**0.4575**	**9.18**

4.2 Hyperparameter Tuning

Grid search for GBM explored n_estimators, learning rate, and max depth. Optimal: 100 estimators, 0.1 learning rate, depth 3.

Table 2: GBM Hyperparameter Search

Parameter	Values
n_estimators	[100, 300]
learning_rate	[0.05, 0.1]
max_depth	[3, 5]

4.3 Feature Importance

Table 3: Top Feature Importances (GBM)

Feature	Importance	Feature	Importance
num__depth	0.623	num__a1	0.065
num__b3	0.114	num__a4	0.034
num__b1	0.079	num__a2	0.028

5 Limitations

- Synthetic variables limit interpretability; domain knowledge could improve feature engineering.
- Feature interactions (polynomial or tree-based) could improve predictions.
- Alternative boosting frameworks such as LightGBM or CatBoost may improve performance and speed.
- Larger hyperparameter search or Bayesian optimization could refine tuning.
- The high noise floor of the synthetic features suggests an upper bound on predictive accuracy

6 Conclusion

Non-linear ensemble models, especially GBM, outperform linear methods on this dataset due to strong non-linear patterns. Thoughtful preprocessing, hyperparameter tuning, and feature importance analysis were key to achieving an R^2 of 0.4575. Future work could focus on feature engineering and testing alternative boosting algorithms.

Code Supplement

Full code: `https://github.com/Jason-Immanuel1/5CCSAMLF_Machine_Learning_Coursework_1`