Theory and Practice of Science and Technology
2025, VOL. 6, NO. 9, 184-191
DOI: 10.47297/taposatWSP2633-456936.20250609

Analysis of Diamond Price Prediction Models — Based on Regression Models

Yue Lin

Department of Mathematics and Statistics, York University, Toronto, Canada

ABSTRACT
Diamond prices are influenced by multiple factors such as carat, cut, color, clarity, and dimensions. Traditional pricing methods based on the 4Cs framework rely on expert judgment, which can be subjective and inefficient. This study applies machine learning techniques to predict diamond prices more accurately and automatically. We compare several regression models, including Linear Regression, Ridge Regression, and Stepwise Regression, using the Kaggle Diamonds dataset. After data preprocessing and model diagnostics, we find that while linear models offer interpretability, they may not capture complex patterns well. To address this, we further implement Decision Tree and Random Forest Regression, which show improved prediction performance. The study highlights the effectiveness of combining feature engineering with appropriate model selection in diamond price forecasting.

KEYWORDS
Diamond pricing; Regression models; Machine learning; Random forest; Decision trees

1 Introduction

Diamond pricing is a complex task influenced by multiple interdependent factors such as carat weight, cut, color, clarity, and physical dimensions. Traditional pricing methods based on the 4Cs framework often rely heavily on human expertise and subjective judgment, resulting in inconsistencies and inefficiencies in large-scale or online transactions.

With the growing availability of open datasets and the advancement of predictive modeling techniques, machine learning (ML) offers a promising alternative for automated, data-driven price estimation. Recent studies have demonstrated that ML algorithms can effectively model the complex and nonlinear relationships between diamond attributes and their prices, providing faster and more accurate predictions while reducing human bias.

This research investigates the use of several regression-based ML models — including Linear Regression, Ridge Regression, Stepwise Regression, Decision Tree Regression, and Random Forest Regression — for diamond price prediction. Using the publicly available Kaggle Diamonds dataset, we conduct thorough data preprocessing, including Likert encoding, normalization, and outlier handling.

Through model training and evaluation, we aim to compare predictive performance and interpretability across models, and identify the most suitable approach for price forecasting tasks. Ultimately, the study contributes to the development of fair, transparent, and automated pricing tools in the gemstone market.

2 Data

2.1 Data source

The data used in this study come from the publicly available "Diamonds" dataset on the Kaggle platform. This dataset, provided by Blue Nile, a well-known online diamond retailer in the United States, contains 53,940 records of diamond sales, with each record corresponding to the characteristics and price information of a single diamond.

The main variables included in the dataset are:

The original format of the dataset is a CSV file. Since it is sourced from an e-commerce platform, both the price and feature records reflect real market sales data, providing high reference value. The rationale for using this dataset in the present study lies in its large scale, the completeness of its variable information, and the clear relationship between

Table 1

Variable	Type	Description
carat	numerical variable	Carat weight of the diamond
cut	categorical variable	Cut grade of the diamond
color	categorical variable	Color grade of the diamond
clarity	categorical variable	Clarity grade of the diamond
depth	numerical variable	Depth percentage of the diamond
table	numerical variable	Table percentage of the diamond
price	numerical variable	Price of the diamond (in U.S. dollars)
x, y, z	numerical variable	Length, width, and height of the diamond (in millimeters)

features and prices, which make it well-suited for constructing and comparing various regression models.

2.2 Data preprocessing

Before constructing the models, the following preprocessing steps were applied to the raw data to ensure the reliability and accuracy of the analysis results:

2.2.1 Encoding of categorical variables using Likert coding

The categorical variables cut, color, and clarity were transformed into ordered numerical values, preserving their ordinal ranking information.

lcut: Fair(1), Good(2), Very Good(3), Premium(4), Ideal(5)

lcolor: J(1) to D(7) (in ascending order of color quality)

lclarity: I1(1) to IF(8) (in ascending order of clarity grade)

2.2.2 Outlier detection and treatment

l a) Retention of genuine high-priced diamonds: The boxplot of price shows a large number of data points above the upper limit, which are marked as outliers. However, after verification with attributes such as carat weight, cut, clarity, and color, these high-priced diamonds were confirmed to be valid transactions. Therefore, they were retained to ensure the applicability of the model in the high-end market.

l b) Removal of unreasonable records: Observations with x, y, or z equal to zero were deleted, since a dimension of zero is physically impossible.

l c) Missing value treatment: The dataset contains no missing values.

2.2.3 Standardization of numerical variables

Z-score standardization was applied to the numerical variables (carat, depth, table, x, y, z):

$$z = \frac{x - \mu}{\sigma}. \quad (1)$$

This procedure ensures that features with different scales are assigned equal weight during modeling, thereby avoiding bias caused by differences in numerical ranges.

2.2.4 Dataset partitioning

The dataset was randomly divided in an 8:2 ratio into:

- Training set: used for model training and parameter fitting.
- Test set: used to evaluate the generalization performance of the final model.

3 Assumptions of linear egression (A0-A5)

In constructing and evaluating the regression models, this study follows the six classical regression assumptions (A0–A5), which are specified as follows:

A0-linearity

A1-zero mean of errors

A2-homoscedasticity

A3-independence of errors

A4-normality of errors

A5-no Multicollinearity

4 Linear regression model

4.1 Simple linear regression

At the initial stage of the study, a complete linear regression model (Complete Linear Model) was employed to model diamond prices, incorporating all independent variables:

$$price = \beta_0 + \beta_1 *carat + \beta_2 *cut\_num + \beta_3 *color\_num + \beta_4 *clarity\_num + \beta_5 *depth + \beta_6 *table + \beta_7 *x + \beta_8 *y + \beta_9 *z + \varepsilon \quad (2)$$

Multicollinearity Analysis and Principal Component Analysis (PCA) for Dimensionality Reduction

Variance Inflation Factor (VIF) analysis revealed that the variables carat, x, y, and z had VIF values significantly exceeding the threshold (VIF > 10), indicating severe multicollinearity.

To address this issue, Principal Component Analysis (PCA) was applied. The results showed that the variance contribution rate of the first principal component was 0.97, which is well above the threshold of 0.95. This indicates that it captures nearly all the information contained in the original four variables. Therefore, replacing carat, x, y, and z with prin1 is both reasonable and effective.

After introducing prin1, a reduced linear regression model was reconstructed. The VIF values of all variables were below 5, indicating that multicollinearity had been effectively eliminated.。

4.2 Model assumptions testing (A0-A5)

In constructing and evaluating the regression models, this study adheres to the six classical regression assumptions (A0 –A5), which are outlined as follows:

4.2.1 A0-linearity

● Requirement: The relationship between the independent variables and the dependent variable in the model should be linear.

● Testing method: Scatter plots of residuals against independent variables and predicted values should display a random distribution without any clear trend.

● Result in this study:

n When using prin1 (the first principal component extracted from PCA), the residuals exhibited a downward trend, indicating that the linearity assumption was not satisfied.

n After introducing the quadratic term prin1_sq, the trend was somewhat alleviated but still present.

n With the addition of the cubic term prin1_3, the residual distribution showed no obvious trend, thereby satisfying assumption A0.

4.2.2 A1-zero mean of errors

● Requirement: The expected value of the residuals should be zero.

● Testing method: Calculate the mean of the residuals; if it is close to zero, the assumption holds.

● Result in this study: 10-14，which is sufficiently close to zero, indicating that assumption A1 is satisfied.

4.2.3 A2-homoscedasticity

● Requirement: The variance of the residuals should remain constant across all values of the independent variables.

● Testing method: Scatter plots of residuals against predicted values and sample indices should display a random distribution without a funnel-shaped pattern.。

● Result in this study: The residual distribution showed no obvious trend, and the influence of a few outliers was negligible.

4.2.4 A3-independence of errors

● Requirement: The residuals of different observations should be uncorrelated.

● Testing method: The Durbin–Watson test is applied; if the first-order autocorrelation coefficient is close to zero, the assumption holds.

● Result in this study: The Durbin – Watson statistic indicated that the first-order autocorrelation coefficient was approximately 0.921. Thus, assumption A3 was not satisfied; however, since the data are not time series, this influence is considered acceptable.

4.2.5 A4-normality of errors

● Requirement: The residuals should approximately follow a normal distribution.

● Testing method: Residual histograms and normal Q–Q plots are used.

● Result in this study: The residual distribution was close to normal, with most points in the Q–Q plot lying along the reference line.

4.2.6 A5-no multicollinearity

● Requirement: The independent variables should not exhibit high correlations.

● Testing method: Calculate the Variance Inflation Factor (VIF), with the common criterion being VIF < 10.

● Result in this study: In the A0 test, it was confirmed that the cubic term (prin1_3) needed to be introduced. Based on the improved model, VIF analysis showed that all variables had VIF values below 5, indicating that the multicollinearity problem had been effectively eliminated. Compared with the initial model, where carat, x, y, and z had VIF values greater than 10, the stability of the model was significantly improved.

4.3 Simple linear regression model fitting and result analysis

4.3.1 Model specification

During the data preprocessing stage, VIF analysis revealed severe multicollinearity among carat, x, y, and z. In addition, the A0 test indicated that the first principal component (prin1) alone could not fully satisfy the linearity assumption.

To address this issue, Principal Component Analysis (PCA) was applied to extract prin1, and its quadratic term (prin1_sq) and cubic term (prin1_3) were introduced to improve the linear relationship.

The final form of the Reduced Linear Model is as follows:

$$price = \beta_0 + \beta_1 \times prin1\_3 + \beta_2 \times cut\_num + \beta_3 \times color\_num + \beta_4 \times clarity\_num + \beta_5 \times depth + \beta_6 \times table + \varepsilon \quad (3)$$

Upon examination, all independent variables were found to have VIF values below 5, indicating that the multicollinearity problem had been effectively eliminated.

4.4 Model fitting results (Training Set)

Using the training set (80% of the data), the reduced model was fitted, and the following key results were obtained:

Table 2

variable	numerical value
Root MSE	0.95378
R-Square	0.0904
Adj R-Sq	0.0903

From the results, although the R^2 value is low, the model shows poor goodness of fit and very weak explanatory power.

4.5 Model fitting results (Test Set)

To evaluate the generalization ability of the model, the final cubic model (prin1_3) fitted on the training set was applied to the test set for prediction, and various performance metrics were calculated.

The main results are as follows:

● Root MSE: 0.95378 - This indicates that the root mean square error on the test set is close to that of the training set, suggesting no significant overfitting.

● R-Square: 0.0904 - The model explains approximately 9.04% of the price variation in the test set, reflecting a low degree of fit but a high consistency with the training set.

● Adj R-Sq: 0.0903 - The adjusted coefficient of determination is close to R^2, further confirming the model's limited explanatory power.

● ASE (Test): 0.93983- The small difference compared with ASE (Train): 0.90957 indicates stable performance on new data and good generalization ability.

5 Ridge regression

Multicollinearity Detection and Variable Transformation

In the complete linear regression model (Complete Linear Model), Variance Inflation Factor (VIF) analysis revealed that the VIF values of carat, x, y, and z were significantly above the threshold (VIF > 10), indicating severe multicollinearity. To address this issue, Ridge Regression was employed by introducing an L2 regularization term into the objective function of the ordinary least squares estimation, which effectively reduced the variance inflation of the variable coefficients.

Prior to modeling, residual – variable plots were examined and showed a downward trend between carat and the residuals, suggesting a nonlinear relationship. To improve the linearity assumption (A0), the following variable transformations were tested:

● Quadratic term ($carat^2$): The trend showed no significant improvement.

● Logarithmic transformation (log(carat)): As shown in the plots, the relationship between the residuals and the variable was largely trend-free, with only a few outliers.

Ultimately, carat_log was selected to replace carat in the Ridge Regression model。

5.1 Model assumptions testing (A0-A5)

In constructing and evaluating regression models, this study follows the six classical regression assumptions A0 – A5, specifically as follows:

5.1.1 A0-linearity

After applying carat_log, the relationships between the residuals and each independent variable, as well as the predicted values, showed no obvious trend, with only a few outliers present. This indicates that the linearity assumption was satisfied.

5.1.2 A1-zero mean of errors

The mean of the residuals was close to zero (-6.13×10^{-14}), indicating that the assumption holds.

5.1.3 A2-homoscedasticity

The scatter plots of residuals versus predicted values and residuals versus sample indices exhibited a band-like distribution without a funnel-shaped pattern, indicating that the assumption holds.

5.1.4 A3-independence of errors

The first-order autocorrelation coefficient was approximately 0.399, indicating that the independence assumption was not fully satisfied. However, since the data are cross-sectional rather than time series, this situation is considered acceptable.

5.1.5 A4-normality of errors

The residual histogram and Q – Q plot indicated that the residual distribution was approximately normal, with most points lying along the reference line, thus satisfying the assumption.

5.1.6 A5-no multicollinearity

After introducing carat_log, all variables in the Ridge Regression model had VIF values below 5, representing a significant improvement in the multicollinearity problem compared with the initial model.

5.2 Model specification

The final Ridge Regression model selected log(carat), cut_num, color_num, clarity_num, depth, table, x, y, and z as independent variables. The ridge parameter λ was searched within the range of 0 to 5, and the value that yielded the best model performance was chosen.

5.3 Model fitting results (Training Set）

Using the training set (80% of the data), the Ridge Regression model was fitted, and the following key results were obtained:

Main Conclusions:

● R-Square = 0.8418 - The model explains approximately 84.18% of the price variation, representing a significant improvement compared with the unregularized linear regression model.

● Adj R-Sq = 0.8417 - The adjusted coefficient of determination is nearly identical to R^2, indicating that the model did not overly rely on specific independent variables during fitting.

● Root MSE = 0.39407 - The root mean square error is relatively low, reflecting high fitting accuracy.

5.4 Model fitting results (Test Set)

The Ridge Regression model trained on the training set was applied to the test set for prediction, yielding the following key metrics:

- ASE(Train): 0.18352
- ASE (Test): 0.15775
- The mean squared error of the test set was slightly lower than that of the training set, indicating no obvious overfitting.
- R^2(Test) = 0.8129, suggesting that the model's explanatory power on new data was consistent with that on the training set, demonstrating good generalization performance.

5.5 Model comparison and conclusion

Compared with the Reduced Model obtained through PCA dimensionality reduction, Ridge Regression outperformed the PCA-based model in terms of R^2, adjusted R^2, and ASE (Test), while being equally effective in addressing multicollinearity. Therefore, when multicollinearity is severe and retaining the original variable information is preferred, Ridge Regression represents the superior choice.

Table 3

Model Type	Dataset	Root MSE	R^2	Adj R^2	ASE
PCA	Train	0.95378	0.0904	0.0903	0.90957
	Test	0.95378	0.0904	0.0903	0.93983
Ridge Regression	Train	0.39407	0.8418	0.8417	0.18352
	Test	0.42844	0.8129	0.8129	0.15775

Conclusion

- Ridge Regression outperformed the PCA cubic model across all metrics, including R^2, Adjusted R^2, Root MSE, and ASE.
- While the PCA model improved the linearity assumption (A0), its fitting ability was limited. Ridge Regression, on the other hand, effectively eliminated multicollinearity while retaining the original variable information, and significantly enhanced the model's explanatory power.
- In terms of generalization performance, Ridge Regression showed only minor differences between the training and test sets, demonstrating greater stability.

6 Decision tree regression

To further enhance the nonlinear fitting capability of the model, this study introduced a Decision Tree Regression model to predict diamond prices. This method offers strong feature interpretability and the ability to automatically capture nonlinear relationships among variables.

6.1 Model construction

In the model construction stage, nine variables — carat, cut_num, color_num, clarity_num, depth, table, x, y, and z — were selected as input features. A regression tree was built using a recursive splitting approach, and the Cost-Complexity Pruning method was applied to control model complexity. The final tree model had a depth of 10, and after pruning, 800 leaf nodes were retained, ensuring both strong representational capacity and effective avoidance of overfitting.

6.2 Variable importance analysis

The variable importance results indicated that carat was the most critical factor affecting price, followed by x, clarity_num, and color_num, while z and table had the least importance. This aligns with the pricing logic of the diamond industry, suggesting that the model possesses strong explanatory power.

6.3 Model performance evaluation

Although the Decision Tree model does not satisfy the traditional linear regression assumptions (A0 – A5), its fitting

Table 4

Variable	Relative Significance
carat	1
x	0.5745
clarity_num	0.3112
color_num	0.2162
cut_num	0.0446
y	0.0301
depth	0.0293
table	0.0228
z	0.0063

performance can still be evaluated using error metrics. The results on the training and test sets are as follows:

Table 5 Training Set Evaluation Results

Metric	Value
R^2	0.8539
RMSE	0.3822
MAE	0.0753
MSE	0.0191

Table 6 Test Set Evaluation Results

Metric	Value
R^2	1
RMSE	0.1382
MAE	0.0753
MSE	0.0191

The results indicate that the Decision Tree model demonstrated good fitting performance on both the training and test sets, with particular superiority in capturing nonlinear patterns.

7 Random forest regression

To further enhance the nonlinear fitting ability of the model and mitigate overfitting, this study introduced the Random Forest Regression model, an ensemble learning method. By constructing multiple regression trees and aggregating their predictions, Random Forest effectively improves model robustness and generalization capability, making it well-suited for handling complex nonlinear relationships among variables.

7.1 Model construction

A Random Forest Regression model was implemented in Python and trained using the default parameter settings. The dataset was split into training, validation, and test sets in a 6:2:2 ratio, ensuring comparability by using the same data partitions.

7.2 Model performance evaluation

The model ' s performance was evaluated on the test set, and the results are as follows:

Table 7

R^2	0.9848
MAE	0.0651
RMSE	0.1222

The model achieved an R^2 of 0.9848 on the test set, which is substantially higher than that of Linear Regression (R^2 = 0.0904), Ridge Regression (R^2 = 0.8418), and Decision Tree Regression (R^2 = 0.9644), indicating its exceptional fitting capability in predicting diamond prices. In addition, the relatively low MAE and RMSE values further confirm the model ' s superiority in error control.

7.3 Model advantage analysis

Random Forest is capable of handling high-dimensional features, performing automatic feature selection, and

showing robustness to outliers, which explains its outstanding performance in this study. Its ensemble structure effectively avoids the overfitting problem common in single-tree models and enables the capture of complex nonlinear relationships among features.

8 Conclusion

The experimental results of this study demonstrate that the Random Forest model achieves the highest prediction accuracy, with an R^2 of 0.9848 on the test set, significantly outperforming both the linear regression model and the single decision tree model. This finding is consistent with existing literature, as noted by Alsuraihi et al. (2020) and Nagargoje et al. (2023), who reported that Random Forest not only excels in metrics such as R^2, RMSE, and MAE, but also exhibits stronger robustness and generalization ability [5-6].

References

[1] A. Iftikhar, H. Alsuraihi, and F. Saeed, "Diamond price prediction using machine learning," in Procedia Computer Science, vol. 170, 2020, pp. 725–730.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] H. Alsuraihi, A. Iftikhar, and F. Saeed, "Diamond price prediction using machine learning," in Proceedings of the 2020 International Conference on Machine Intelligence and Smart Systems, ACM, 2020, pp. 1–5.

[3] D. Thakur, M. Sharma, and N. Agarwal, "Enhancing predictive modeling of diamond prices using ML and meta-ensemble techniques," in Proceedings of the 13th Innovations in Software Engineering Conference, ACM, 2020, pp. 1–7.

[4] Kaggle, "Diamonds Dataset," [Online]. Available: https://www.kaggle.com/datasets/shivam2503/diamonds

[5] H. Alsuraihi, A. Iftikhar, and F. Saeed, "A comparative study on machine learning techniques for diamond price prediction," Procedia Computer Science, vol. 170, pp. 725–730, 2020.

[6] V. Nagargoje, M. Hannure, V. Kokane, K. Kondekar, G. Parate, and V. Deotare, "Diamond Price Prediction using Machine Learning Algorithm," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol. (IJSRCSEIT), vol. 9, no. 6, pp. 164–167, Nov.–Dec. 2023.