Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Diamond Price Prediction Based on Regression Model: Comparison and Analysis of Different Algorithms

Qifeng Xiao

Xi'an High Tech No.1 Middle School, South Campus, Xi'an, China

XQFzzzRyan@outlook.com

Abstract. Diamonds, as a valuable commodity, have investment value. Through price prediction, investors can determine the trend of diamond prices and decide when to buy or sell diamonds to achieve maximum investment returns. For some consumers who plan to purchase diamonds as collectibles or investments, price forecasting can help them determine whether the purchased diamonds have appreciation potential and make more informed consumption decisions. The government or relevant industry associations can formulate corresponding industrial policies based on the trend and prediction of diamond prices, regulate market order, and promote the healthy development of the diamond industry. In order to compare and analyze the performance of several regression models in diamond price prediction, and which one used, this study preprocessed multiple features, including standardization and feature selection, on the diamond dataset on Kaggle. Subsequently, this paper compares the performance of linear regression, random forest, XGBoost, and Deep Neural Network (DNN) in price prediction tasks. This study indicates that the random forest model has superior performance in diamond price prediction and can be used for actual market pricing.

Keywords: Diamond price prediction; regression model; machine learning.

1. Introduction

Diamonds have a typical cubic crystal structure, with carbon atoms strongly bonded via covalent interactions, forming a highly stable lattice structure [1, 2]. This structure grants diamonds with many unique physical properties. It is the hardest naturally occurring material, with a Moho hardness of 10. It also has extremely high refractive index and dispersion index, which can make diamonds shine brightly under light, which is also one of the important reasons why diamonds are loved. In addition, diamonds have high thermal conductivity and excellent thermal conductivity. Due to its high hardness, diamonds are widely used in industrial fields such as cutting, grinding, drilling and other tools. In the field of jewelry, diamonds are one of the most popular gemstones, often made into various jewelry such as diamond rings, necklaces, earrings, etc., symbolizing love and eternity.

Based on price prediction [3-5], financial institutions can more accurately evaluate the value and potential risks of collateral when providing loans or investing in diamond related businesses, and reasonably control credit scale and investment risks through price prediction. At the same time, consumers can choose to purchase diamonds at relatively low prices based on price predictions, avoiding buying during peak price periods, thus saving expenses and achieving more rational consumption. Traditional models are difficult to predict the price of diamonds due to many additional nonlinear factors and some characteristics of the data set itself.

Nowadays, Artificial Intelligence (AI) technology has a very powerful force [6-8], such as ChatGPT, DeepSeek etc. Artificial intelligence technologies have been widely applied in various fields such as chemistry, biology, medicine, and have received a lot of attention, especially in the field of financial and business analysis. For example, regression models can be used to predict stock prices or gold prices etc. From past works, it can be found that machine learning models have been widely applied to various price prediction tasks and have achieved good performance. Therefore, this article intends to predict diamond prices and analyze the importance of features influencing prediction performance.

2. Method

2.1. Data Set Preparation

The data set is sourced from the Kaggle website [9], with nearly 53, 000 data points and a total of nine features, such as carat and purity. The target of prediction is diamond prices, which is a regression task.

2.2. Machine Learning Models

2.2.1. Linear regression.

Linear regression is a simple model that assumes a linear relationship between input features and the target variable [10]. It has several advantages: 1) The model is simple in form, easy to understand and explain. 2) There are mature calculation methods, have low computational complexity, and can relatively efficiently obtain results when processing large-scale data. 3) Being able to intuitively see the contribution of each independent variable to the results, facilitating the analysis of the relationship between variables. It has several disadvantages: 1) It is sensitive to outliers in the data, which may affect the accuracy and stability of the model. 2) If there are multiple local extremum points or highly curved data relationships, linear regression models may not fit well and may experience underfitting. 3) If the data volume is insufficient or the data distribution is uneven, it may be difficult to accurately estimate the model parameters, resulting in poor generalization ability of the model and poor prediction performance on new data.

2.2.2. Decision tree.

Decision tree is a tree-based model that splits data into branches based on feature thresholds to predict continuous values. It has several advantages: 1) Easy to understand, able to clearly demonstrate how independent variables affect dependent variables, facilitating analysis of data characteristics and decision-making processes. 2) There is no need to normalize or standardize the data, and there are no strict requirements for the distribution and characteristics of the data. 3) Nonlinear relationships in data can be fitted through the branching structure of trees. It has several disadvantages: 1) Decision trees may overfit training data and be sensitive to noise and outliers in the training data. 2) Minor changes in data may lead to significant alterations in the decision tree structure, thereby affecting the predictive results of the model. 3) As the number of features increases, the complexity of the decision tree rapidly increases, the computational complexity increases, and overfitting problems are prone to occur, resulting in lower efficiency in processing high-dimensional data.

2.2.3. Random forest.

Random Forest is an ensemble of decision trees that improves prediction accuracy and reduces overfitting through averaging. It has several advantages: 1) Capable of handling various types of data, including numerical and categorical data, without strict requirements for data distribution and characteristics, and without the need for complex preprocessing of the data. 2) It can also work effectively in high-dimensional data and is not easily affected by dimensional disasters. 3) It can conveniently evaluate the importance of each feature to the prediction results, helping to understand the data and make feature selections. It has several disadvantages: 1) The training and prediction process involves multiple decision trees, with high computational complexity and long training time. 2) The overall decision-making process is quite complex, making it difficult to intuitively understand how the predictions for each sample are derived. 3) There is a lot of noise in the training data, which may affect the construction of decision trees and thus affect the performance of random forests.

2.2.4. Kneighbors.

KNeighbors is a non-parametric model that predicts a value based on the average of its k nearest neighbors in the training set. It has several advantages: 1) Simple and intuitive. 2) Strong adaptability to data distribution. 3) It can avoid bias caused by incorrect model assumptions. It has several advantages: 1) The large amount of computation results in slower prediction speed. 2) Affected by

data noise. 3) There is a strong correlation between features, which may lead to inaccurate distance calculation and affect the performance of the model.

2.2.5. XGB Regressor.

Based on the gradient boosting framework, a powerful regression model is constructed by iteratively training multiple weak regressors (such as decision trees). In each iteration, it adjusts the weights of samples based on the prediction error of the previous model, paying more attention to those samples that were incorrectly predicted, and then fits a new weak regressor to correct these errors, continuously optimizing the performance of the model. It has several advantages: 1) Being able to quickly train models on large-scale datasets greatly improves training efficiency. 2) Can achieve high prediction accuracy. 3) Good robustness, with a certain tolerance for noise and outliers in the data, not easily affected by individual extreme data points, and the model has good stability and generalization ability. It has several advantages: 1) The model complexity is high. 2) The demand for computing resources is high. 3) Overall, it is not as intuitive as simple linear models or decision tree models, making it difficult to fully understand how the model makes predictions.

3. Results and Discussion

The predictive performance of the random forest model shown in Table 1 is slightly better than that of the XGBoost model under these evaluation metrics, but overall both models have excellent performance, with fitting effects close to 1 and relatively small error metrics.

Table 1. Performance comparison of models

	Random forest	XGB regression	Linear Regression	KNeighbors Regression	Decision Tree Regression
R^2	0.98	0.98	0.88	0.95	0.96
Adjusted R^2	0.98	0.98	0.88	0.95	0.96
MAE	280.17	270.03	849.35	402.92	357.23
MSE	307728.16	296832.11	1741183.66	628471.70	556636.74
RMSE	554.73	544.82	1319.53	792.76	746.08

Variable distribution: The graph along the diagonal is a histogram or density plot of each variable shown in Fig. 1, showing the data distribution of a single variable. For example, the distribution of "price" shows a right skewed state, indicating that there are relatively few samples with high prices.

Relationship between variables: A non-diagonal graph is a scatter plot between two variables, which can be used to determine the correlation between variables. For example, there is a clear positive correlation between "cargo" and "price", meaning that the larger the carat weight, the higher the price; The relationship between "depth" and "table" is relatively scattered and the correlation is not obvious.

Outliers: Scatter plots can reveal some outliers, such as points in the relationship between "price" and other variables where prices are extremely high or low, which may have an impact on data analysis.

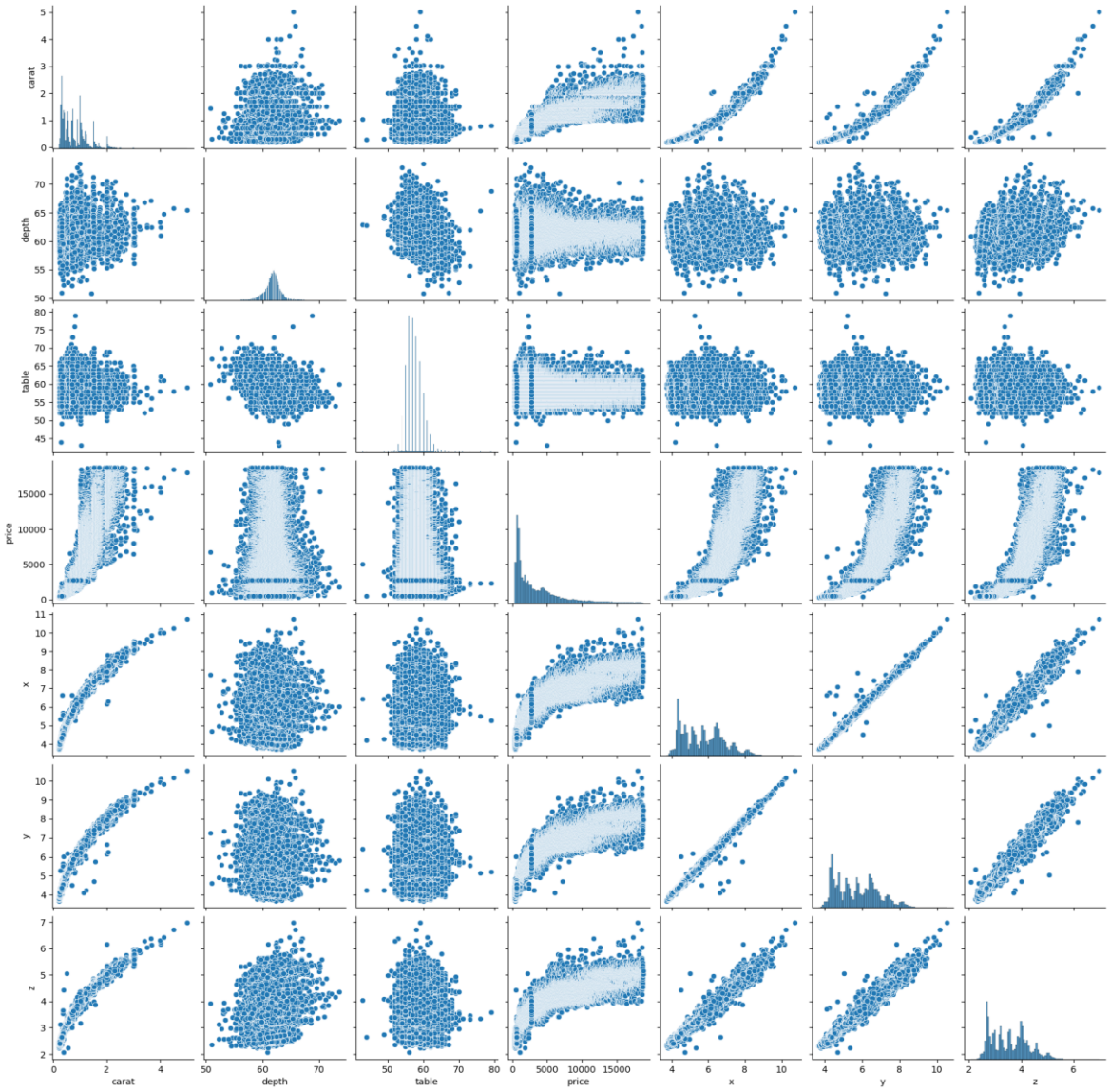

Figure 1. Relationship between variables (Picture credit: Original).

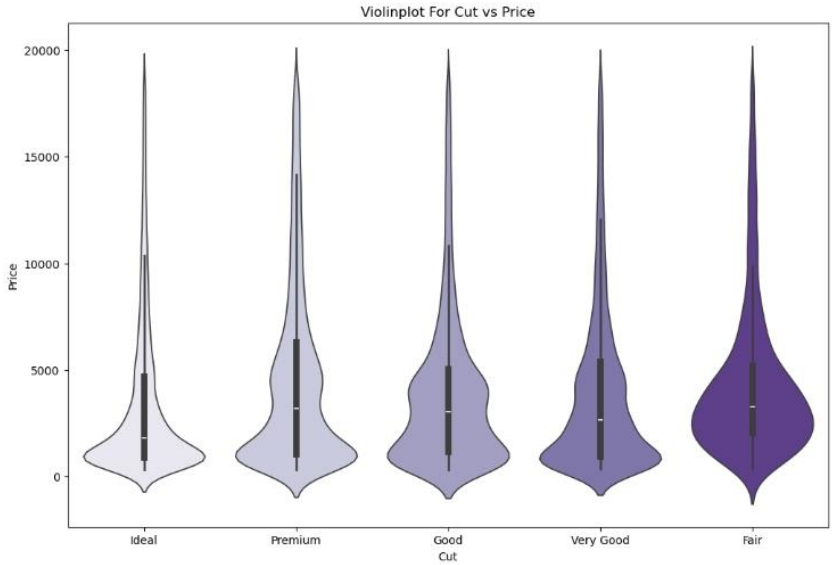

Figure 2. Violin plot for Cut vs Price (Picture credit: Original).

The graph illustrates the relationship between diamond cutting and price shown in Fig. 2. The horizontal axis represents different cutting levels (from ideal to fair), and the vertical axis represents

price. The higher the cutting grade (such as ideal), the more the median price distribution and the higher the number of diamonds in the high price range, indicating that diamonds with better cutting have higher prices.

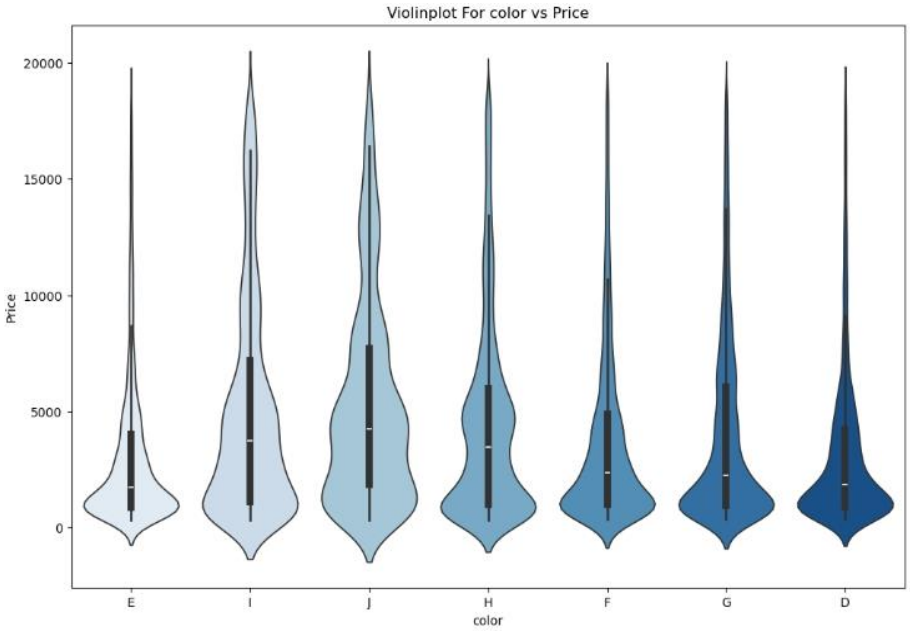

Figure 3. Violin plot for Color vs Price (Picture credit: Original).

The graph presents the relationship between diamond color and price shown in Fig. 3. The horizontal axis represents different color grades (from E to D), and the vertical axis represents price. As the color level changes from E to D, the color becomes increasingly white, and the median and overall price range of the price distribution also shows an upward trend, indicating that diamonds with whiter colors tend to have higher prices.

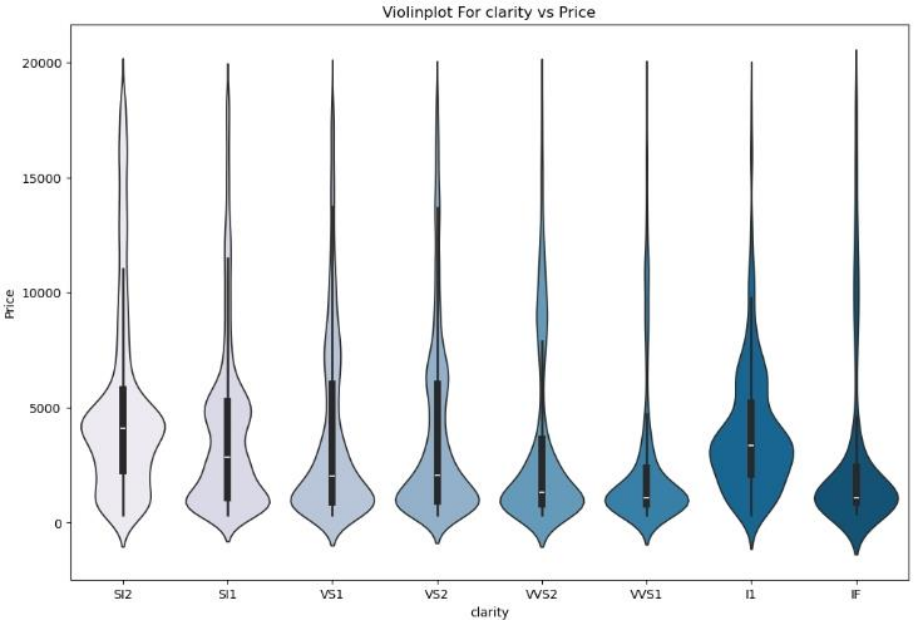

Figure 4. Violin plot for Clarity vs Price (Picture credit: Original).

The graph shows the relationship between diamond clarity and price shown in Fig. 4. The horizontal axis represents different cleanliness levels (from SI2 to IF), and the vertical axis represents price. The width of the graph represents the distribution density of prices at that clarity, with wider areas indicating a higher number of diamonds within that price range. It can be seen that overall, as the clarity level increases (from SI2 to IF), the median of the price distribution increases, indicating that diamonds with higher clarity usually have higher prices.

Fig. 5 shows the correlation between multiple attributes of diamonds. The horizontal and vertical axes list attributes such as carat weight, cut, color, etc. The number in each square represents the correlation coefficient between two attributes. The closer the value is to 1 or -1, the stronger the correlation. Red represents positive correlation, and green represents negative correlation. It can be seen that the correlation coefficient between carat weight and price of diamonds is as high as 0.92, indicating a strong positive correlation between carat weight and price; The correlation coefficient between clarity and price is relatively low, indicating that the impact of clarity on price is relatively small compared to carat weight.

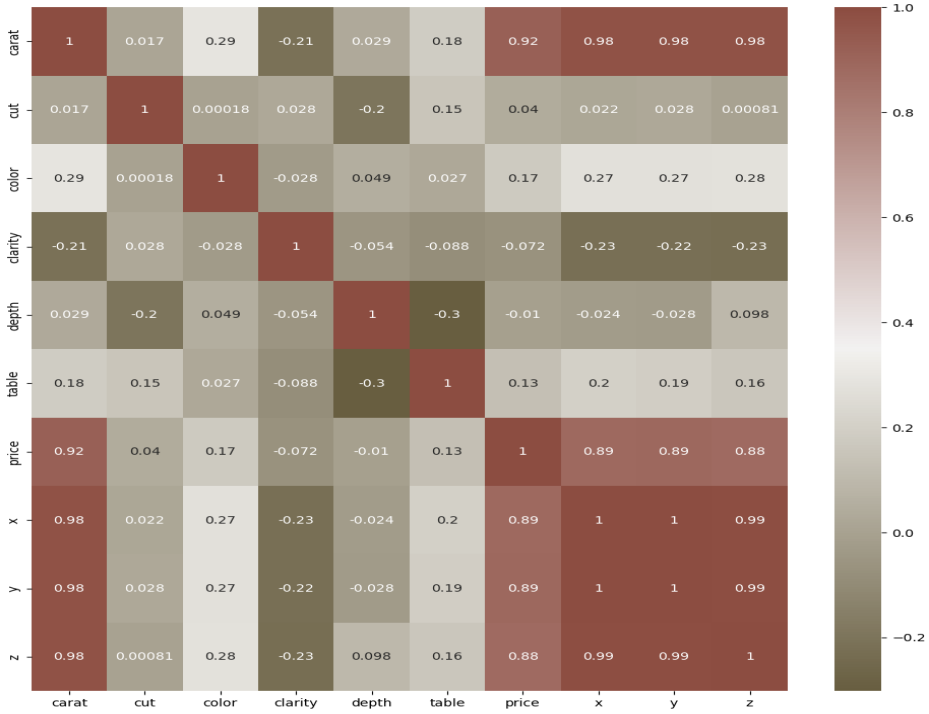

Figure 5. Correlation map (Picture credit: Original).

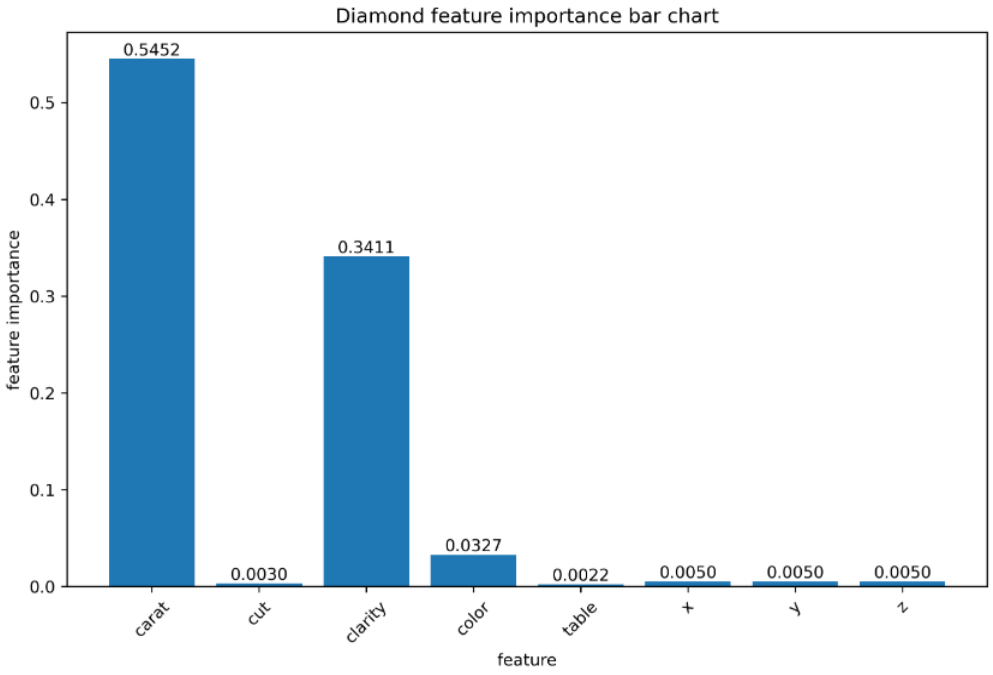

Figure 6. Feature importance (Picture credit: Original).

The higher the numerical value of feature importance, the greater the impact of the feature on diamond price prediction. From general experience and common analysis results shown in Fig. 6, carat is

usually the most important factor affecting diamond prices, as diamond weight is generally strongly positively correlated with price, with larger weights often leading to higher prices. Other features such as cut, color, and clarity also have a significant impact on price, with high-quality cuts, high color grades, and high clarity diamonds priced higher. And features such as depth, table, and x, y, z, although also related to price, have a relatively weaker impact compared to carat.

4. Conclusion

In conclusion, this study compared the performance of five different regression models—Linear Regression, Decision Tree, Random Forest, KNeighbors, and XGBoost—in predicting diamond prices based on a real-world dataset. Among them, Random Forest and XGBoost showed the best results, with high R^2 scores (0.98) and low prediction errors, making them suitable for actual pricing tasks. Carat weight was found to be the most important feature, showing a strong positive correlation with price. Other factors such as cut, color, and clarity also had noticeable effects, while features like depth and table had relatively smaller influence. Visualizations like violin plots and correlation maps further supported these findings and helped explain how different features affect diamond value. This study highlights the effectiveness of using machine learning models, especially ensemble methods, in price prediction tasks. In future work, combining more data sources or introducing deep learning could further improve prediction accuracy and support better decision-making for investors and consumers.

References

[1] Bundy FP, Hall HT, Strong HM, Wentorfjun RH. Man-made diamonds. nature. 1955 Jul 9; 176 (4471): 51-5.

[2] Shelah S. Diamonds. Proceedings of the American Mathematical Society. 2010 Jun; 138 (6): 2151-61.

[3] Sharma G, Tripathi V, Mahajan M, Srivastava AK. Comparative analysis of supervised models for diamond price prediction. In2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2021 Jan 28, 1019-1022.

[4] Mihir H, Patel MI, Jani S, Gajjar R. Diamond price prediction using machine learning. In2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) 2021 Dec 16, 1-5.

[5] Pandey AC, Misra S, Saxena M. Gold and diamond price prediction using enhanced ensemble learning. In 2019 Twelfth International Conference on Contemporary Computing (IC3) 2019 Aug 8, 1-4.

[6] Jiang Y, Li X, Luo H, Yin S, Kaynak O. Quo vadis artificial intelligence?. Discover Artificial Intelligence. 2022 Mar 7; 2 (1): 4.

[7] Korteling JH, van de Boer-Visschedijk GC, Blankendaal RA, Boonekamp RC, Eikelboom AR. Human-versus artificial intelligence. Frontiers in artificial intelligence. 2021 Mar 25; 4: 622364.

[8] Zhang C, Lu Y. Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration. 2021 Sep 1; 23: 100224.

[9] Kaggle, diamond price prediction https://www.kaggle.com/datasets/shubhankitsirvaiya06/diamond-price-prediction/code, 2020.

[10] James G, Witten D, Hastie T, Tibshirani R, Taylor J. Linear regression. InAn introduction to statistical learning: With applications in python 2023 Jul 1, 69-134.