Diamond Price Prediction Using Machine: A Comparative Analysis of Regression Models

Jannah Cleine T. Glodo
College of Computing and Information Technology
National University
Metro Manila, Philippines
glodojt@students.national-u.edu.ph

Lawrence Renante A. Guelos
College of Computing and Information Technology
National University
Metro Manila, Philippines
guelosla1@students.national-u.edu.ph

Ronaldo D. Rico Jr.
College of Computing and Information Technology
National University
Metro Manila, Philippines
ricord@students.national-u.edu.ph

***Abstract* – Understanding the variables that affect diamond pricing is critical for stakeholders, as precise forecasts enable informed decision-making in a market influenced by variable demand and economic factors. This study examines the predictive accuracy of linear regression models in estimating diamond prices, emphasizing the fundamental characteristics that impact valuation. Utilizing a comprehensive dataset sourced from the diamond market, the analysis focuses on key factors such as quality—evaluated through color, clarity, cut, and carat weight—as well as size and demand. The research reveals the relationships between these features and diamond prices. The results indicate that linear regression is more suitable for the precise prediction of diamond prices. The methodology illustrates how advanced machine learning techniques can effectively estimate diamond values, thereby enhancing decision-making processes within the market.**

Keywords – Diamond, prices, machine learning, regression, techniques, Kaggle.

I. Introduction

The diamond mining industry plays a crucial role in driving economic growth for several countries, generating $16 billion in annual net economic benefits globally. This considerable financial contribution sustains the livelihoods of millions, including approximately 77,000 people employed directly within the mining sector. For stakeholders, the accuracy of diamond price predictions is essential, as it allows investors to make informed decisions in a market influenced by shifting factors. Diamonds are prized gemstones known for their brilliance and durability, with values affected by changes in demand and economic conditions. Accurate price forecasting supports businesses in managing inventory, setting competitive prices, and anticipating market shifts, all of which are essential for profitability. Understanding diamond price trends is particularly valuable for investors, mining companies, and jewelry retailers to ensure variety in asset holdings and mitigate risk, as diamonds are often viewed as alternative investments. Research by Smith et al. (2020) demonstrates the effectiveness of machine learning models for predictive tasks, especially in finance and pricing. Integrating these techniques for diamond price prediction can be highly beneficial, given the complexity of diamond valuation, which is based on factors like carat, cut, color, depth, and clarity. This study aims to showcase the potential of advanced algorithms to enhance accuracy and efficiency in diamond price estimation by handling complex, non-linear relationships, managing high-dimensional data, and learning from extensive historical data to improve predictive outcomes.

II. Review of Related Literature

Machine learning algorithms have long been a foundation for accurate financial predictions, demonstrating their efficacy in recent studies focused on diamond price forecasting. Techniques such as multiple regression, decision trees, support vector machines, and neural networks adeptly manage the non-linear relationships and interactions characteristic of diamond pricing. This study adopts linear regression as a baseline model, establishing a straightforward relationship between independent variables (diamond features) and the dependent variable (price).

Linear regression is praised for its interpretability, simplicity, and computational efficiency, facilitating insights into how factors like cut, clarity, color, depth, and carat weight influence pricing. Research by Mamonov and Triantoro identified weight, color, and clarity as primary determinants of diamond prices. Additionally, Kigo et al. assessed various machine learning algorithms for diamond price

prediction, noting that Alsuraihi et al. achieved a mean absolute error (MAE) of 112.93 and a root mean square error (RMSE) of 241.97 with Random Forest. Other studies reported different accuracies, with Mamonov and Triantoro's Decision Forest yielding an MAE of 5.8% and Sharma et al. achieving an R^2 of 0.9793 with Random Forest. Notably, Mihir et al. reported an R^2 of 0.9872 with CatBoost Regression, while Chu's multiple linear regression model attained an R^2 of 0.972.

The pricing of diamonds is intricately influenced by several key factors, which include quality—typically assessed through color, clarity, cut, and carat weight—along with size and demand. Diamonds of superior quality, characterized by their colorlessness, lack of inclusions, and excellent cut, command the highest prices in the market. Additionally, carat weight plays a significant role in determining cost, as larger diamonds tend to be more expensive due to their rarity. In this study, we utilize a dataset sourced from Kaggle, comprising 53,940 data points to investigate these factors further. Within this dataset, carat weight is generally recognized as the feature with the highest predictive impact on price, exhibiting a strong correlation with diamond value. Following carat weight, other attributes such as cut quality, color, and clarity also play important roles. Employing machine learning techniques to analyze these features enables us to derive insights into their relative importance in predicting diamond prices. To explore these relationships, we conducted a correlation analysis that examines the connections between various features and the target variable, which is the price. This analysis highlights the need to select the most appropriate model, taking into account preprocessing steps and correlation assessments to enhance predictive accuracy.

III. Methodology

This section explains the dataset, methods, tools, and approaches for predicting diamond prices using linear regression and Random Forest Regression models as seen in **Figure 1**. It offers a high-level review of the study strategy, followed by precise information on data collection, preprocessing, experimental design, and assessment.

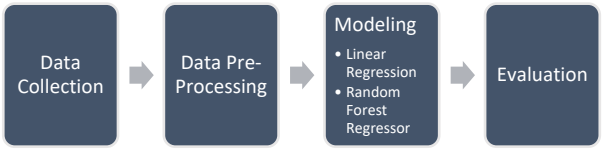

Figure 1. Model Classification Framework

A. Data Collection

The dataset is retrieved from Shivam Bansal's "Diamonds" from Kaggle which originates from Hadley Wickham's GitHub datasets, which includes 53,940 entries. Features including carat, cut, color, clarity, depth, table, and price are all included in each record. These attributes provide important details regarding the quality and physical attributes that usually affect a diamond's market value. The data for the dataset was compiled and arranged by the author and was retrieved from the Blue Nile website which is known for its Jewelry Education.

B. Data Pre-Processing

The dataset needed minimum cleaning since there were no missing values or outliers. Preprocessing focuses on altering and encoding characteristics to make them appropriate for model input:

- **Categorical Encoding:** One-hot encoding was used for categorical characteristics (e.g., cut, color, clarity) to render them interpretable by the model.
- **Feature Scaling:** Standard scaling was used on numerical features (e.g., carat, depth, table) to maintain a uniform range of values, enhancing model training stability and performance. Feature extraction was unnecessary since each characteristic contributed pertinent information for the target prediction.

C. Experimental Setup

The experiment was conducted in Python using the following tools:

- pandas – for data handling,
- scikit-learn (version 0.24) – for modeling and preprocessing,
- NumPy – for numerical operations.
- A standard CPU environment was utilized to execute the model.
- **Data Split:** The dataset was divided into training and testing sets to assess model performance on unseen data, with 80% for training and 20% for testing.

D. Algorithm

Two models were used to predict diamond prices:

1. **Linear Regression:** Selected for its simplicity and interpretability, Linear Regression gives insights into the linear correlations between attributes (e.g., carat, clarity) and price.
2. **Random Forest Regressor:** Chosen for its resilience and accuracy, Random Forest is an ensemble model that integrates numerous decision trees, minimizing

overfitting and capturing complicated connections in the data. This paradigm is especially useful for problems involving both numerical and categorical data.

E. Training Procedure

An 80/20 train-test split was utilized to assess the model. Cross-validation was not implemented since the dataset size was adequate, and linear regression offered a clear training framework.

- **Number of Trees (n_estimators):** Set to 100 for an optimal balance between accuracy and computational efficiency.
- **Maximum Depth (max_depth):** Limited to avoid overfitting; values were obtained by testing. Cross-validation was not done owing to the dataset's appropriate size, however future research may integrate it for further accuracy.

F. Evaluation Metrics

Both models were evaluated using the following metrics:

- **Mean Squared Error (MSE):** Calculates the average squared differences between expected and actual prices. Lower numbers indicate more accuracy.
- **R-Squared (R^2):** Indicates the percentage of the target variable's variation that the model can account for; values nearer 1 indicate a better fit. For regression tasks, these indicators are typical. R2 gives information on how effectively each model accounts for price variance, whereas MSE gives an overall error measure.

G. Baselines and Comparative Models

While no complicated baselines were originally employed, the performance of both linear regression and Random Forest Regressor was compared to understand their different strengths:

- **Linear Regression:** Provided a basic, interpretable baseline with insights on feature relevance but was projected to underperform with complicated relationships.
- **Random Forest Regressor:** Expected to exceed linear regression in capturing non-linear correlations and offering greater accuracy.

IV. Results and Discussion

The analysis provides valuable insights into the performance of both the linear regression and random forest models for predicting diamond prices. In particular, the linear regression model exhibited high r^2 values across both the training and test datasets, suggesting a robust fit to the underlying data and a strong capacity to capture the variability associated with diamond pricing. Additionally, the model's low mean squared error (MSE) values imply that its predictions align closely with the observed prices, enhancing confidence in its ability to make accurate predictions within this dataset.

On the other hand, while the random forest model also demonstrated predictive capabilities, its slightly lower r^2 values indicate a weaker fit to the data when compared to linear regression. This is further reflected in its higher MSE scores, which suggest a greater discrepancy between predicted and actual values. The increased MSE points to a larger error margin in the random forest predictions, potentially making it less suitable for applications where precision is critical for pricing.

Table 1. Model Comparison

Model	Training MSE	Training R2	Test MSE	Test R2
Random Forest	1521.284655	0.999904	1773.704222	0.999888
Linear Regression	1268160.927452	0.92034	1282237.940982	0.91934

Although the study did not calculate p-values to establish statistical significance explicitly, the cross-validation approach used in this analysis still provides valuable evidence for the model comparison. The observed improvements in r^2 suggest that linear regression outperforms random forest in this context. However, further research is needed to validate these findings rigorously. Future studies could employ statistical significance tests, such as t-tests, on the cross-validation results to confirm these observed differences more formally. This would allow for greater confidence in the comparative performance of the models, offering a statistical foundation to support the selection of the more suitable model.

Figure 2. Actual Price vs. Predicted Price

In this figure, the positive correlation between the actual and predicted prices is apparent. The overall trend shows that as the actual price increases, the predicted price also tends to rise based on diamond features. This is encouraging, as it indicates that the model is effectively capturing the underlying relationship between features and price.

An advantage of the dataset chosen by the researcher is its reliability, supported by a history of related studies, making it suitable for large-scale diamond price prediction tasks. The accuracy of the machine learning model and the selected algorithm support the goal of identifying the most suitable regression model. However, the researcher also notes certain limitations, including: (1) Unseen Data – predictions may be less accurate for diamonds with unique characteristics that are not well-represented in the training set; and (2) Ethical Considerations – ethical concerns in using AI in decision-making, such as potential biases and fairness in algorithms, should be addressed.

V. Conclusion

This study demonstrates the effectiveness of predictive models, particularly linear regression and Random Forest, in achieving precise diamond price predictions. It aims to identify a suitable regression algorithm that can enhance the accuracy of diamond price estimation. The analysis highlights carat weight as the most influential factor, consistent with prior research that identifies it as a primary determinant of diamond prices. Linear regression offers clear interpretability and insights into linear relationships, whereas Random Forest excels in capturing complex, non-linear interactions, making it particularly well-suited for the intricate dynamics of the diamond market. Given the significant economic value of diamonds, previous studies have emphasized the importance of accurate pricing for supporting economic growth. By utilizing machine learning models, this research demonstrates how enhanced price prediction methods can lead to more efficient market practices and improved investment strategies. Ultimately, this study provides valuable insights to inform stakeholders' decision-making processes, whether for investment, inventory management, or profit maximization. Additionally, it seeks to expand the foundational research on diamond valuation, contributing to a more comprehensive understanding of effective decision-making within the industry.

References:

[1] Slauter, A. (2024). *How diamond mining uplifts communities and Fuels economic growth: A Deeper look at the impact of the diamond industry*. International Diamond Center. https://shopidc.com/blogs/news/how-diamond-mining-uplifts-communities-and-fuels-economic-growth-a-deeper-look-at-the-impact-of-the-diamond-industry

[2] Mihir, H., Patel, M. I., Jani, S., & Gajjar, R. (2021). *Diamond price prediction using machine learning. In 2021 2nd International Conference on Communication, Computing and Industry*. 4.0 (C2I4) (pp. 1-5). IEEE. Retrieved from https://www.ijcrt.org/papers/IJCRTAG02013.pdf

[3] Zhang, H. (2023). Prediction and feature importance analysis for diamond Price based on machine learning models. Advances in Economics Management and Political Sciences, 46(1), 254–259. https://doi.org/10.54254/2754-1169/46/20230347

[4] Wickham, H., et al. (2019). *Welcome to the Tidyverse*. The Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

[5] Agrawal, S. (2017). *Diamonds*. https://www.kaggle.com/datasets/shivam2503/diamonds/data