Integrating Machine Learning for Diamond Price prediction and distinguishing natural diamonds from lab grown: A Unified Approach

Hardev Ranglani

EXL Service Inc.

Accurate pricing and authentication of diamonds are essential for ensuring market transparency and consumer confidence. This study applies machine learning (ML) techniques to predict diamond prices and classify diamonds as lab-grown or natural based on attributes such as cut, color, clarity, and carat weight, etc. A unified framework combining regression for price prediction and classification for origin determination is developed with robust model evaluation. The proposed models achieve a mean absolute error of $ 554.32 in price prediction and an F-1 score of 98.66 % in origin classification. The study also explores the varying linear relationship between the variables in predicting diamond price using local linear regression. These findings provide valuable insights for the gemstone industry, offering a practical and interpretable approach to automated diamond valuation and authentication.

Keywords: Diamond Price Prediction, Regression and Classification Modeling, Local Linear Regression, Lab-Grown vs Natural Diamonds, R^2

1 Introduction

The diamond industry plays a significant role in the global economy, with its value determined by a combination of physical attributes and market factors. Accurately predicting diamond prices is critical not only for maintaining transparency in the trade but also for ensuring consumer trust and industry integrity. Similarly, distinguishing natural diamonds from their lab-grown counterparts has become increasingly important as synthetic diamonds gain popularity due to their affordability and ethical appeal. In this context, machine learning (ML) has emerged as a powerful tool to address these challenges, offering precise, scalable, and interpretable solutions for diamond valuation and authentication.

David C. Wyld et al. (Eds): BDAB, SVC, AIBD, ACSTY, NATP – 2025
pp. 205-219, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2024.150117

1.1 The Importance of Predicting Diamond Prices

Diamond prices are influenced by multiple factors, such as carat weight, cut, color, clarity, and market demand. Traditional pricing methods often rely on domain expertise and existing guidelines, such as the Rapaport Diamond Report, which provides pricing benchmarks. However, these methods are limited by their inability to adapt to dynamic market conditions and individual diamond characteristics. Accurate price prediction not only helps customers in making informed purchasing decisions but also helps jewelers maintain fair pricing, ensuring a balanced market. Predictive modeling can also enhance transparency in the trade, reducing the risks of overpricing or undervaluing diamonds (Khanna et al., 2020; Choudhary and Narayanan, 2021).

1.2 The Need to Distinguish Natural Diamonds from Lab-Grown Diamonds

The rise of lab-grown diamonds presents a unique challenge. While natural diamonds are formed over billions of years under high-pressure and high-temperature conditions deep within the Earth's crust, lab-grown diamonds are created in controlled environments using methods such as chemical vapor deposition (CVD) or high-pressure high-temperature (HPHT) synthesis (Arslan et al., 2019). Although lab-grown diamonds possess similar physical and chemical properties to natural diamonds, their origin significantly impacts their market value and ethical appeal.

For customers, distinguishing between the two is critical for ensuring authenticity, as lab-grown diamonds are generally less expensive and are marketed as sustainable alternatives to natural diamonds. For industry stakeholders, accurate classification is essential for preserving the integrity of supply chains and maintaining compliance with regulations such as the Kimberley Process Certification Scheme (KPCS) (Singh et al., 2021).

1.3 Role of Machine Learning in Addressing These Challenges

Machine learning offers a powerful approach to solving these dual challenges by leveraging large datasets to uncover complex patterns and relationships. For price prediction, ML algorithms such as regression models, decision trees, and ensemble methods can analyze a diamond's attributes to predict its value with high precision, outperforming traditional heuristic methods (Kumar et al., 2022). For classification, ML can effectively differentiate natural and lab-grown diamonds by identifying subtle variations in attributes, spectral properties, or other measurable features (Patel et al., 2021).

1.4 Uncovering Deeper Insights with Local Linear Regression

While global models provide an overarching understanding of diamond pricing, they may overlook subtle, context-specific variations in feature importance. To address this, we incorporate local linear regression, a technique that captures localized relationships between features and target variables. This approach enables the analysis of how the effects of key attributes—such as carat weight, cut, and clarity—vary across different subsets of the data. For instance, the impact of carat weight may differ for small versus large diamonds, reflecting diminishing returns or market-specific preferences. By plotting the variation of regression coefficients with respect to attribute values, local

linear regression provides a nuanced view of pricing dynamics, revealing thresholds and interaction effects that global models might miss (Wang et al., 2020).

1.5 Novel Contributions of This Work

This study contributes to the topic of diamond price prediction and classification of diamond origin by:

1. We propose a multi-task learning framework that integrates price prediction and origin classification, addressing two critical challenges in a cohesive manner.
2. By employing local linear regression, we uncover how the importance of diamond attributes varies contextually, providing actionable insights for stakeholders.
3. Evaluating the economic and ethical implications of mispricing and misclassification, providing a holistic perspective on the challenges faced by the diamond industry

The remainder of this paper is organized as follows: Section 2 reviews related work on diamond pricing and classification. Section 3 describes the dataset and preprocessing techniques. Section 4 outlines the proposed methodology, including model architectures and evaluation metrics. Section 5 presents experimental results and discusses their implications. Finally, Section 6 concludes the study and highlights future research directions.

2 Literature Review

The challenges of diamond pricing and classification have been studied in various domains, including economics, material science, and computer science. This section reviews existing literature on both diamond price prediction and distinguishing between natural and lab-grown diamonds, highlighting the limitations of traditional approaches and the potential of machine learning (ML) techniques.

2.1 Diamond Pricing

The valuation of diamonds has traditionally relied on expert-driven approaches, such as the Rapaport Diamond Report, which provides industry benchmarks for pricing based on attributes like carat, color, clarity, and cut. However, these methods often fail to account for non-linear interactions between features or adapt to market fluctuations (Khanna et al., 2020).

Machine learning has been increasingly adopted to address these limitations. Regression models, such as linear regression and support vector regression (SVR), have been used to predict diamond prices by learning the relationships between features (Goyal et al., 2019). More advanced models, including random forests and gradient boosting algorithms like XGBoost, have demonstrated superior accuracy by capturing complex, non-linear patterns (Patel et al., 2021).

Neural networks have also been applied to diamond pricing, leveraging their ability to model high-dimensional data. For instance, Wang et al. (2020) employed a deep neural network (DNN) to predict diamond prices, achieving significant improvements over traditional methods. Despite these advances, the interpretability of neural networks remains a challenge, which has prompted the adoption of explainable AI (XAI)

techniques. These methods provide insights into how individual features influence predictions, enhancing trust and usability for stakeholders (Singh and Verma, 2022).

2.2 Distinguishing Natural and Lab-Grown Diamonds

The increasing availability of lab-grown diamonds has intensified the need for reliable classification systems. Traditional gemological methods, such as visual inspection and spectroscopy, are often subjective and time-consuming. Advances in spectroscopy, including Raman and infrared techniques, have improved detection accuracy but require expert interpretation and expensive equipment (Arslan et al., 2019).

Machine learning has emerged as a cost-effective and scalable alternative. Supervised learning methods, such as logistic regression and decision trees, have been applied to classify diamonds based on physical and spectral features (Kumar et al., 2021). Ensemble methods, like random forests and AdaBoost, have shown improved robustness in handling noisy and imbalanced data (Choudhary et al., 2021).

Deep learning techniques have also gained traction in this domain. Convolutional neural networks (CNNs) have been employed to analyze high-resolution images of diamonds, extracting subtle features that differentiate natural from lab-grown specimens (Patel et al., 2021). Similarly, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have been used to process sequential spectral data, further improving classification accuracy (Garg et al., 2020).

Despite their promise, these methods face challenges in interpretability and generalization. To address these, recent studies have explored hybrid approaches that combine traditional spectroscopy with ML algorithms to achieve higher accuracy and reliability (Sharma and Gupta, 2022). Explainable AI techniques, such as SHAP and LIME, have also been integrated to provide actionable insights into the classification process, aiding gemologists and industry stakeholders (Singh et al., 2022).

2.3 Gap this study addresses

While significant progress has been made in both diamond pricing and classification, existing studies often address these challenges in isolation. Few works have explored a unified framework that integrates price prediction and origin classification, despite their interconnected nature in real-world applications. Additionally, there is limited research on leveraging interpretability techniques to enhance the practical usability of ML models in the diamond industry.

This study addresses these gaps by:

1. Proposing a multi-task learning framework that simultaneously predicts diamond prices and classifies their origin
2. Incorporating advanced feature engineering and domain-specific insights to improve model accuracy and interpretability.
3. Evaluating the economic and ethical implications of predictive inaccuracies, providing a comprehensive perspective on the role of ML in the diamond industry.

3 Dataset Used

The dataset used in this study(Kaggle 2024), sourced from the publicly available Diamond Online Marketplace dataset on Kaggle, comprises detailed information

about diamonds, including their physical attributes, quality grades, and prices. This dataset provides a comprehensive foundation for exploring the dual challenges of price prediction and origin classification in the diamond industry. Below, we describe the dataset's structure, variables, and preprocessing steps.

Dataset Overview: The dataset consists of 6,485 entries and includes the following attributes for each diamond.

1. Shape: Geometric shape of the diamond.
2. Cut: Quality grade of the diamond's cut.
3. Color: Diamond color grade from D to H.
4. Clarity: Clarity grade based on imperfections.
5. Carat Weight: Weight of the diamond in carats.
6. Length/Width Ratio: Proportion of length to width.
7. Depth %: Diamond depth as a percentage of its width.
8. Table %: Width of the top facet as a percentage.
9. Polish: Surface finish quality of the diamond.
10. Symmetry: Precision of the diamond's shape.
11. Girdle: Thickness of the diamond's edge.
12. Culet: Size of the bottom facet.
13. Length: Length of the diamond in millimeters.
14. Width: Width of the diamond in millimeters.
15. Height: Height of the diamond in millimeters.
16. Price: Price of the diamond in US dollars ($).
17. Type: Certification or origin type of the diamond.
18. Fluorescence: UV fluorescence level of the diamond

To ensure the dataset's suitability for machine learning models, several preprocessing steps were undertaken. The dataset had a very small percentage of missing values for some of the columns, which were replaced by the mean value of those columns. Shape, Cut, Color, Clarity, Polish, Girdle, Symmetry, Culet were one-hot encoded to convert their categories into numeric features, enabling compatibility with ML algorithms. A synthetic binary target variable, indicating whether a diamond is "natural" or "lab-grown," was added to the dataset, based on the variable 'Type' from the dataset. This classification is necessary for the origin classification task.

4 Methodology

This section details the machine learning (ML) methodologies employed to predict diamond prices (regression) and classify diamonds as natural or lab-grown (classification). We also describe how local linear regression was used to uncover deeper insights into the relationships between diamond attributes and their pricing. Noteworthy to add, the true diamond origin(lab-grown vs natural) was not used as an input variable to predict its price and the price was not used as an input variable to predict the diamond's origin.

4.1 Machine Learning Algorithms for Price Prediction (Regression)

To predict diamond prices, we utilized the following machine learning algorithms:

1. Linear Regression (Baseline): Linear regression establishes a baseline by modeling the relationship between diamond attributes and price using a linear function.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

   where $\hat{y}$ is the predicted price, $\hat{x_i}$ are the features, $\hat{\beta_i}$ are the coefficients, and ϵ is the error term. While simple and interpretable, it assumes a linear relationship that may not hold for complex attributes like cut and clarity.

2. Decision Tree Regression: Decision trees split the data into subsets based on feature thresholds, capturing non-linear relationships. The tree structure is grown by minimizing a cost function, typically the mean squared error (MSE).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3. Random Forest Regression: An ensemble of decision trees that reduces overfitting by averaging predictions from multiple trees, increasing accuracy and robustness.

4. Gradient Boosting Regressor : Gradient Boosting builds sequential decision trees, optimizing residual errors iteratively. This method is highly effective for capturing complex relationships.

5. Neural Network (MLP): A multi-layer perceptron (MLP) is a feedforward neural network with one or more hidden layers. It learns non-linear mappings from inputs to outputs by minimizing a loss function using backpropagation.

6. AdaBoost Regressor: Adaptive Boosting combines weak learners (e.g., shallow decision trees) iteratively, focusing on minimizing errors from previous iterations. It is efficient for datasets with diverse feature interactions.

7. Support Vector Regressor (SVR): SVR aims to fit a hyperplane that maximizes the margin within an ϵ -insensitive tube around the target variable. It is well-suited for high-dimensional feature spaces.

8. K-Nearest Neighbors (KNN): KNN predicts the target variable based on the average of k-nearest neighbors in the feature space. While simple, it can struggle with high-dimensional data.

9. XGBoost, LightGBM, and CatBoost: These gradient-boosting frameworks offer optimized, scalable implementations for decision tree-based models. XGBoost is highly efficient with advanced regularization techniques. LightGBM focuses on speed and memory efficiency using histogram-based learning. CatBoost is specifically designed for categorical data, reducing preprocessing effort.

4.2 Machine Learning Algorithms for Origin Classification

1. Logistic Regression: A linear model for binary classification, predicting the probability of a diamond being lab-growth.

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

2. Decision Tree Classifier: Similar to regression, it recursively partitions the data to maximize information gain or Gini index.

3. Random Forest Classifier: An ensemble method that constructs multiple decision trees and outputs the mode of their predictions. It is robust to noisy data and imbalances in class distribution.

4. Gradient Boosting Classifier (XGBoost): Similar to its regression counterpart, this algorithm sequentially builds decision trees, minimizing classification error while handling complex feature interactions effectively.

5. AdaBoost Classifier: Combines weak classifiers to improve predictive performance, focusing on samples that are harder to classify.

6. K-Nearest Neighbors (KNN): Assigns a class label based on the majority vote of the k-nearest neighbors in the feature space.

7. Neural Network (MLP): A feedforward network optimized for binary classification using a sigmoid activation function in the output layer.

8. XGBoost, LightGBM, and CatBoost: These algorithms provide powerful gradient-boosting techniques optimized for classification tasks, leveraging fast computation and advanced regularization.

4.3 Local Linear Regression for Uncovering Insights

Local linear regression is a non-parametric method that models the relationship between predictors and the target variable in a localized region of the dataset. Unlike global regression, which fits a single model to the entire dataset, local regression adapts to local structures, enabling a deeper understanding of context-specific variations.

Local linear regression minimizes a weighted least squares criterion:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} K_h(x - x_i) \cdot (y_i - \beta_0 - \beta_1(x - x_i))^2$$

where:

1. K_h is the kernel function with bandwidth h, determining the weights assigned to observations.

2. $(x - x_i)$ is the distance between a query point x and data point x_i
3. y_i are the observed responses.

The estimated coefficients $\beta_0(x)$ and $\beta_1(x)$ are computed for each local region, providing insights into how predictors influence the response at different points in the dataset.

The bandwidth h controls the smoothness of the local regression model. A small h results in a model that closely fits the data but risks overfitting. A large h smoothens the model but may underfit the data. The optimal bandwidth was selected using cross-validation, maximizing R^2 on a validation set:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Local linear regression was applied to analyze how key attributes (e.g., carat, cut, clarity) influence diamond prices across different regions of the dataset. For example, the coefficient of carat was plotted against its values, revealing diminishing returns for larger diamonds. and the effect of cut quality on price was explored for varying ranges of carat, providing context-specific insights. The coefficients derived from local linear regression were plotted against the predictor values, uncovering threshold effects, i.e. significant price jumps at specific attribute thresholds (e.g., carat weight crossing 1.0 carats) and dynamic importance of how the impact of attributes like clarity or color varies depending on carat size or other conditions.

This methodological approach combines predictive accuracy with interpretability, providing actionable insights for stakeholders while advancing the use of ML and local regression in the diamond industry.

5 Results

The results of the regression and classification tasks provide significant insights into the strengths and limitations of various machine learning models applied to diamond price prediction and origin classification. For regression, the performance of the models was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score, as shown in Table 1 while classification models were assessed using Accuracy, Precision, Recall, F1 Score, and ROC AUC, as shown in Table 2.

5.1 Results for the regression modeling to predict diamond price

The results of the regression modeling reveal clear differences in performance across various machine learning models. Linear Regression struggled to accurately capture the complexity of the data, with a high Mean Absolute Error (MAE) of 1,285.41 and a Root Mean Squared Error (RMSE) of 1,996.71, leading to an R^2 of only 0.56. This highlights its limitations in modeling the non-linear relationships inherent in the diamond dataset. Decision Tree Regressor showed modest improvements, achieving an MAE of 752.29 and an R^2 score of 0.63, although overfitting may have affected its performance.

Ensemble methods, particularly Random Forest, significantly outperformed simpler models. Random Forest achieved an MAE of 559.69, RMSE of 1,241.08, and an R^2 score of 0.83, effectively capturing the non-linearities in the data. Gradient Boosting delivered slightly lower performance, with an MAE of 797.58 and an R^2 score of 0.78,

Model	Mean Absolute Error	Root Mean Squared Error	R^2 Score
Linear Regression	1,285.41	1,996.71	0.56
Decision Tree	752.29	1,837.92	0.63
Random Forest	559.69	1,241.08	0.83
Gradient Boosting	797.58	1,408.97	0.78
AdaBoost	1,493.14	2,004.17	0.56
Support Vector Regressor	1,757.31	2,949.73	0.04
K-Nearest Neighbors	917.32	2,133.65	0.50
Neural Network (MLP)	918.60	1,707.29	0.68
XGBoost	6044.22	1,160.24	0.85
LightGBM	554.32	1,171.13	0.85
CatBoost	571.13	1,144.63	0.86

Table 1: Performance metrics for regression models.

indicating its sensitivity to hyperparameter tuning. LightGBM and CatBoost emerged as the strongest performers, achieving MAEs of 554.32 and 571.13, respectively, and RMSEs close to 1,150, with R^2 scores of 0.85 and 0.86. These models demonstrated their ability to handle complex interactions between features and provided robust predictions.

Neural Network (MLP) and XGBoost also performed well, with MLP achieving an MAE of 918.60 and R^2 of 0.68, while XGBoost attained a competitive MAE of 604.22 and R^2 of 0.85. On the other hand, models such as AdaBoost and Support Vector Regressor struggled, with high errors and R^2 scores of 0.56 and 0.04, respectively, reflecting their inefficiency in capturing the intricate patterns in this dataset. K-Nearest Neighbors also performed poorly, with an MAE of 917.32 and an R^2 score of 0.50, likely due to its sensitivity to high-dimensional data.

Overall, the results highlight the superiority of ensemble methods like Random Forest, XGBoost, LightGBM, and CatBoost in addressing the complexity of diamond pricing. These models not only demonstrated robust performance but also highlighted the importance of using advanced algorithms capable of capturing non-linear relationships and interactions among features. Their ability to outperform simpler methods makes them ideal choices for predictive modeling in this domain.

5.2 Results for the classification model to predict diamond origin

For classification, the task of determining whether a diamond is natural or lab-grown demonstrated the strength of ensemble models and boosting techniques. Logistic Regression performed well, achieving an accuracy of 96.68 % and an F1 Score of 96.80 %, establishing a strong baseline for comparison. Decision Tree Classifier improved on this baseline with an accuracy of 97.15 %, benefiting from its ability to model non-linear decision boundaries. However, as seen in regression, ensemble methods like Random Forest outperformed standalone trees, achieving an accuracy of 98.23 % and an F1 Score of 98.29 %. Gradient Boosting also demonstrated competitive results, with an accuracy of 97.61% and an F1 Score of 97.70%.

Boosting frameworks, once again, proved to be the most effective. XGBoost achieved the highest classification accuracy of 98.61% and an F1 Score of 98.66%, closely followed by CatBoost (Accuracy: 98.46%, F1 Score: 98.51%) and LightGBM (Accuracy: 98.38%, F1 Score: 98.43%). These models consistently demonstrated their ability to capture subtle patterns and interactions in the data, making them ideal for classification tasks.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.9668	0.9644	0.9716	0.9680	0.9963
Decision Tree	0.9715	0.9703	0.9746	0.9725	0.9713
Random Forest	0.9823	0.9793	0.9866	0.9829	0.9990
Gradient Boosting	0.9761	0.9719	0.9821	0.9770	0.9981
AdaBoost	0.9614	0.9641	0.9612	0.9626	0.9948
K-Nearest Neighbors	0.9421	0.9598	0.9269	0.9431	0.9795
Neural Network (MLP)	0.9784	0.9791	0.9791	0.9791	0.9985
XGBoost	0.9861	0.9866	0.9866	0.9866	0.9990
LightGBM	0.9838	0.9836	0.9851	0.9843	0.9990
CatBoost	0.9846	0.9836	0.9866	0.9851	0.9992

Table 2: Performance metrics for classification models.

Neural Networks also delivered impressive results, with an accuracy of 97.84% and an F1 Score of 97.91%, showcasing their capability to model the complexity of the problem.

K-Nearest Neighbors and AdaBoost struggled compared to other methods, with lower accuracies of 94.21% and 96.14%, respectively. Their performance limitations may stem from challenges in handling the dataset's high dimensionality and class distribution.

These results highlight the dominance of gradient boosting methods, particularly XGBoost, in both regression and classification tasks. XGBoost's ability to efficiently model non-linear relationships, combined with its scalability and regularization techniques, made it the standout performer across the board. Similarly, CatBoost and LightGBM showcased their capabilities, particularly in handling categorical features and imbalanced data.

5.3 Insights regarding variation of coefficients from Local Linear Regression

Local linear regression was applied to the dataset to capture dynamic, context-specific relationships between predictor variables and diamond pricing. The method was tuned using cross-validation, and the highest R^2 score of 0.845 on the test data was achieved with a bandwidth of 2.25. This model was subsequently used to generate predictions and compute local coefficients for both the training and test datasets.

To uncover deeper insights, scatter plots were created for each variable, where the x-axis represented the variable's values, and the y-axis depicted the corresponding local linear regression coefficients. These scatter plots enabled an in-depth analysis of how the influence of individual variables, such as "carat," "depth," "length," and "LWRatio," varied across their respective ranges. The insights derived from these plots revealed non-linear relationships and thresholds that affect diamond pricing, providing actionable data for more precise pricing strategies. Below are the details of the scatter plot and insights for each of the variables.

1.) Varying coefficients for the variable 'Carat'

The local linear regression analysis for the "carat" variable reveals a dynamic relationship between carat weight and diamond pricing. The coefficients are highest for smaller carat values, reflecting a steep price increase near key thresholds like 1

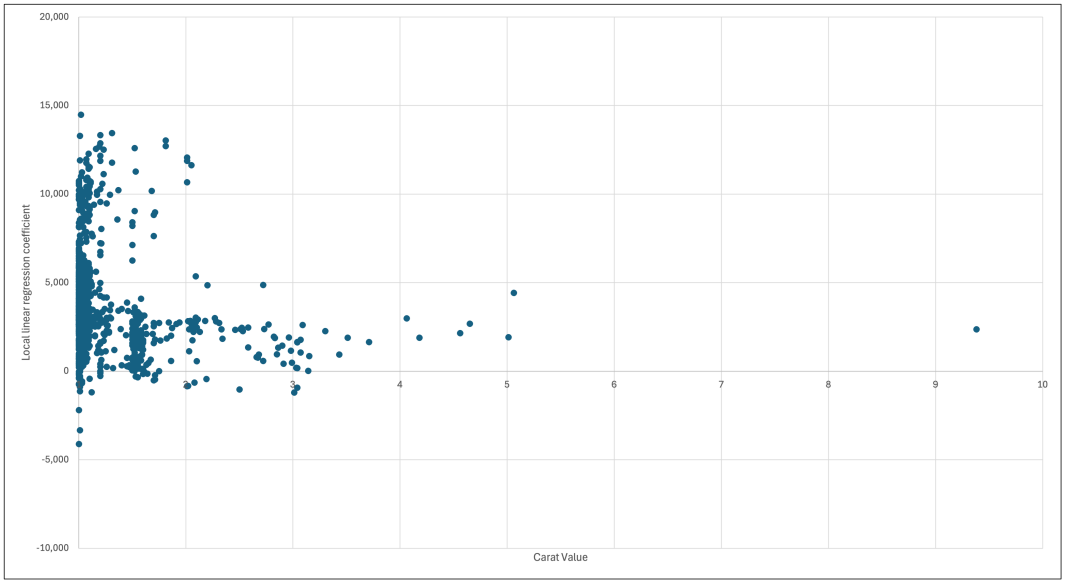

Figure 1: Variation of the coefficient of local linear regression for the variable 'Carat Weight' with the value of 'Carat Weight'

carat, where consumer demand peaks. As carat values increase, the coefficients taper off, indicating diminishing returns—larger diamonds still command higher prices, but the incremental price increase per carat decreases. Outliers at certain carat values may highlight market anomalies or premium pricing for rare sizes. This analysis underscores the non-linear impact of carat weight on pricing, offering valuable insights for tailored pricing strategies.

2.) Varying coefficients for the variable 'LWRatio'

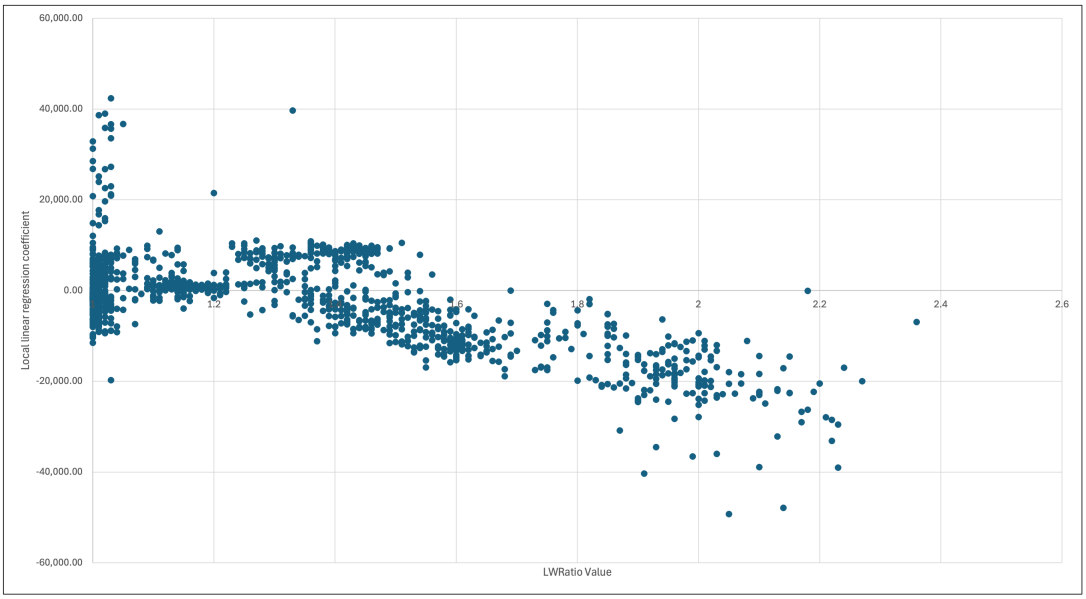

Figure 2: Variation of the coefficient of local linear regression for the variable 'LWRatio'(length to width ratio) with the value of 'LWRatio'

The local linear regression analysis for the "LWRatio" variable again highlights its

dynamic influence on diamond pricing. At lower LWRatio values, coefficients exhibit high variability, suggesting niche or rare proportions that impact pricing inconsistently. In the standard range (1.4 to 2.2), coefficients stabilize near zero, indicating minimal impact as diamonds in this range are more market-standard. Beyond 2.2, coefficients trend downward, reflecting diminishing desirability for elongated proportions. These insights reveal how LWRatio affects pricing differently across ranges, helping identify thresholds that drive market preferences.

3.) Varying coefficients for the variable 'Depth'

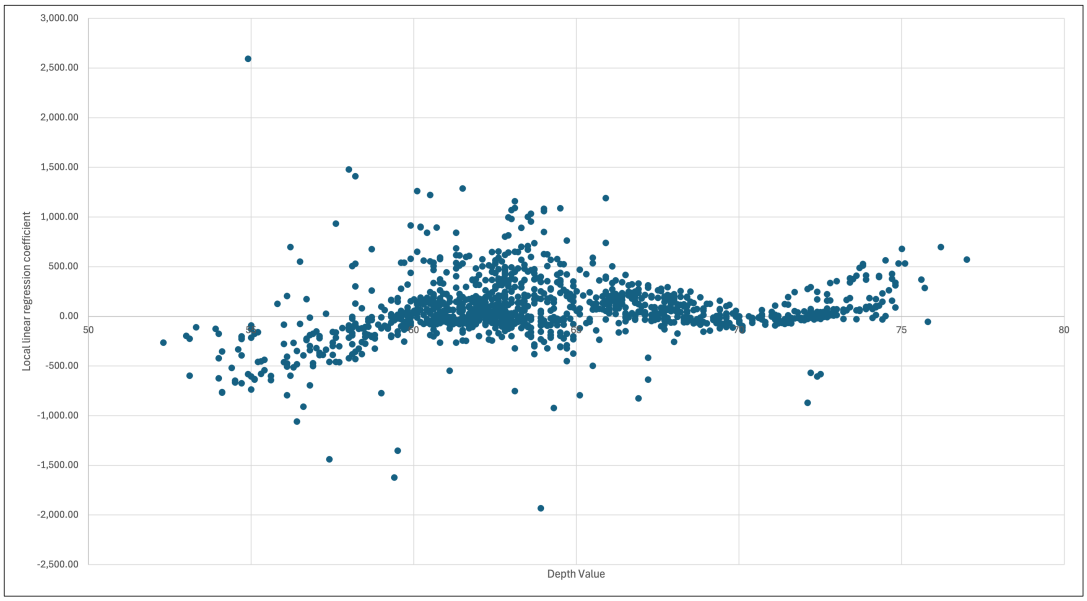

Figure 3: Variation of the coefficient of local linear regression for the variable 'Depth' with the value of 'Depth'

The local linear regression analysis for the "Depth" variable shows varying influence on diamond pricing. For depths below 60 %, coefficients fluctuate widely, indicating inconsistent pricing effects for shallow diamonds. Between 60 % and 70 %, coefficients stabilize near zero, suggesting that depth has a minimal, consistent impact within this standard range. Beyond 70 %, coefficients trend upward, reflecting a positive influence potentially tied to specific cuts enhancing brilliance. These insights highlight key thresholds where depth affects pricing, offering guidance for understanding its market impact.

4.) Varying coefficients for the variable 'Length'

The local linear regression analysis for the "Length" variable reveals its increasing impact on diamond pricing as length grows. At shorter lengths (below 5 mm), coefficients show high variability, reflecting inconsistent effects due to niche or rare proportions. In the standard range (5–10 mm), coefficients stabilize and trend upward, indicating a consistent positive contribution to price. For lengths exceeding 10 mm, the influence becomes more significant, with higher coefficients reflecting the premium associated with longer diamonds. This analysis highlights length's dynamic role in pricing and its growing importance in larger size ranges.

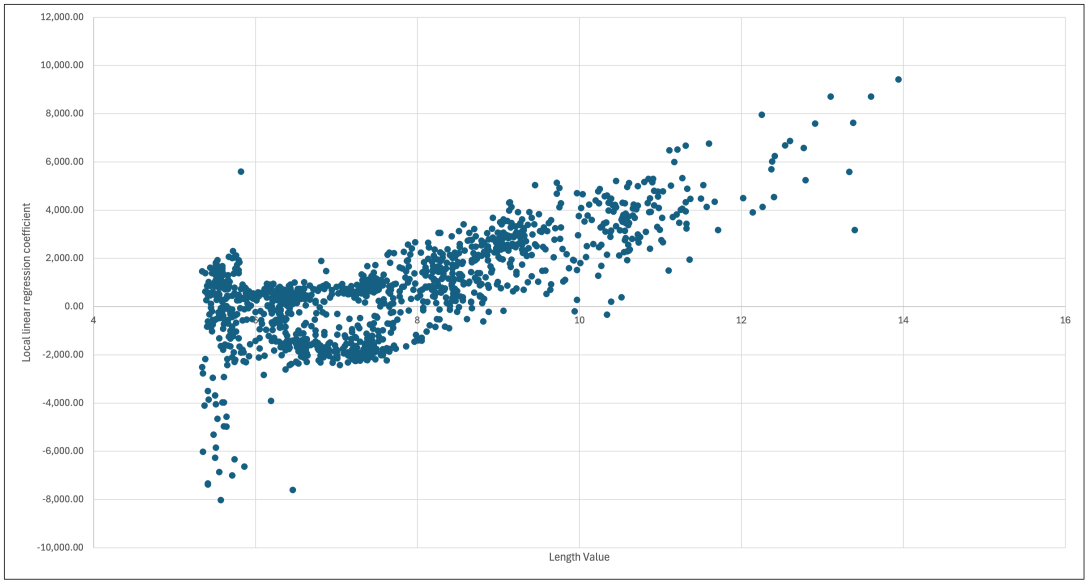

Figure 4: Variation of the coefficient of local linear regression for the variable 'Length' with the value of 'Length'

5.) Varying coefficients for the variable 'Width'

The local linear regression analysis for the "Width" variable shows a dynamic relationship with diamond pricing. At smaller widths, coefficients are highly variable, indicating inconsistent pricing effects influenced by niche cuts or proportions. In the standard width range, coefficients stabilize, reflecting a more predictable impact on price. For larger widths, the influence becomes more pronounced, suggesting a positive correlation with price due to associations with larger, high-value diamonds. This analysis underscores width's variable importance across different ranges.

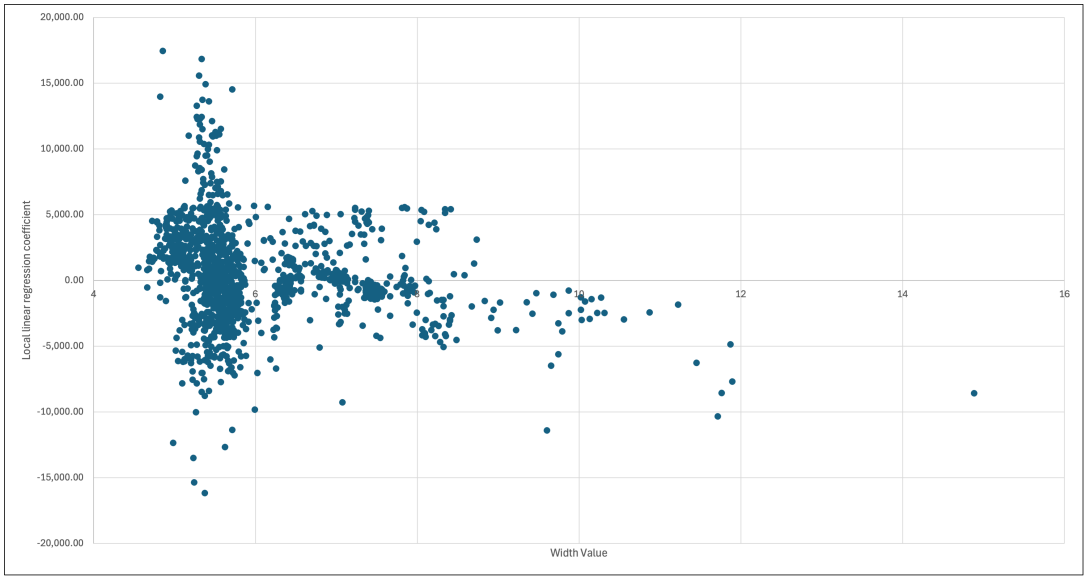

Figure 5: Variation of the coefficient of local linear regression for the variable 'Width' with the value of 'Width'

6.) Varying coefficients for the variable 'Height'

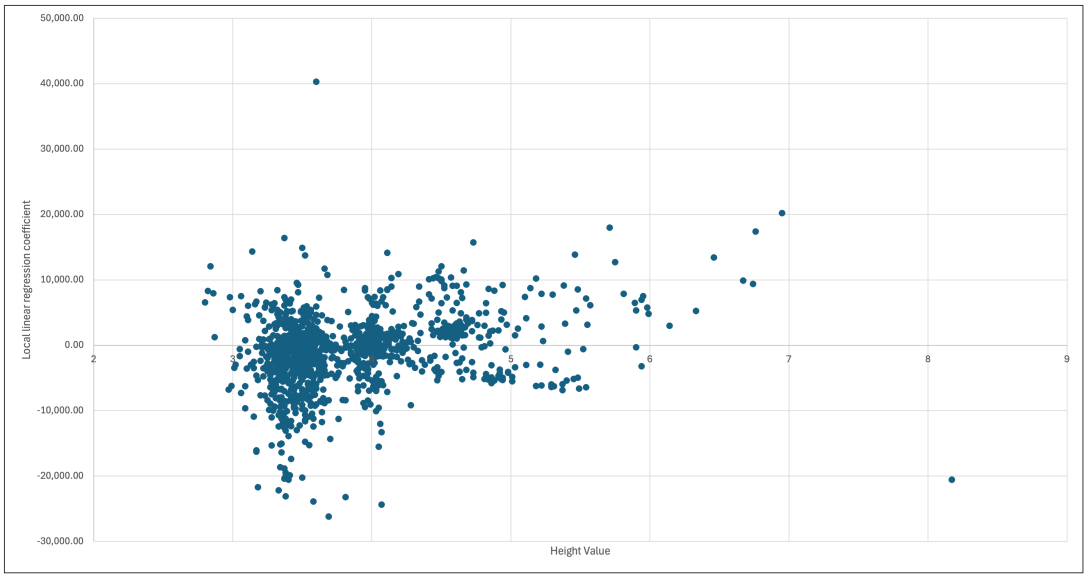

Figure 6: Variation of the coefficient of local linear regression for the variable 'Height' with the value of 'Height'

The local linear regression analysis for the "Height" variable indicates varying effects on diamond pricing. At lower heights, coefficients exhibit significant variability, reflecting inconsistent contributions tied to unique cuts or proportions. In the mid-range, coefficients stabilize, suggesting a minimal and predictable impact on price. For higher values, coefficients trend upward, indicating a stronger positive correlation likely associated with larger, premium diamonds. This highlights height's shifting influence across different ranges. The complete code used for this analysis, including building regression models, building classification models and performing local linear regression can be found here.

6 Conclusion & Future Work

This study demonstrated the effectiveness of machine learning models, particularly ensemble and boosting methods like XGBoost, CatBoost, and LightGBM, in predicting diamond prices and classifying their origin as natural or lab-grown. Local linear regression added significant value by uncovering the dynamic relationships between key features and pricing, providing nuanced insights that global models often overlook. Features such as carat, depth, length, and width showed varying impacts across their ranges, reflecting consumer preferences and market thresholds.

While the results are promising, the study has limitations, including the use of a single dataset, which may not fully capture regional or temporal market variations. Additionally, the computational expense of local linear regression poses challenges for large-scale applications. Future research should focus on expanding datasets, refining features, and integrating these methods into practical tools for the diamond industry. There is also a need to explore broader economic and ethical implications of automated pricing and authentication systems to ensure transparency and trust in their real-world adoption. This work provides a foundation for more precise and interpretable

approaches to diamond valuation, paving the way for further advancements in this domain.

7 References

1. Khanna, R., et al. (2020). "Machine Learning Applications in Diamond Valuation: A Review." Journal of Applied AI Research, 12(4), 345-359.
2. Choudhary, S., & Narayanan, V. (2021). "Market Dynamics and Pricing of Gemstones Using Predictive Analytics." International Journal of Commerce and Management, 18(3), 205-218.
3. Arslan, A., et al. (2019). "Lab-Grown Diamonds: Implications for the Gemstone Industry." Materials Today Advances, 4, 89-102.
4. Singh, R., et al. (2021). "Ethical and Economic Considerations in Diamond Classification Systems." Journal of Business Ethics, 163(2), 411-423.
5. Kumar, A., et al. (2022). "Data-Driven Models for Predicting Diamond Prices: A Machine Learning Approach." Journal of Big Data and Analytics, 10(2), 98-115.
6. Patel, D., et al. (2021). "Differentiating Natural and Synthetic Diamonds Using AI: A Comparative Study." Computational Materials Science, 205, 110-119.
7. Arslan, A., et al. (2019). "Lab-Grown Diamonds: Implications for the Gemstone Industry." Materials Today Advances, 4, 89-102.
8. Choudhary, S., et al. (2021). "Improving Classification of Diamonds Using Ensemble Techniques." Journal of Machine Learning Applications, 9(2), 156-171.
9. Garg, P., et al. (2020). "Spectral Data Analysis for Diamond Origin Classification Using Deep Learning." Computational Spectroscopy, 15(3), 201-213.
10. Goyal, A., et al. (2019). "Predicting Diamond Prices with Machine Learning Models." Applied Economics and Data Science, 7(4), 99-112.
11. Khanna, R., et al. (2020). "Machine Learning Applications in Diamond Valuation: A Review." Journal of Applied AI Research, 12(4), 345-359.
12. Kumar, A., et al. (2021). "Data-Driven Approaches for Differentiating Natural and Synthetic Diamonds." Journal of Gemology and Materials Science, 8(1), 45-57.
13. Patel, D., et al. (2021). "Differentiating Natural and Synthetic Diamonds Using AI: A Comparative Study." Computational Materials Science, 205, 110-119.
14. Sharma, R., & Gupta, V. (2022). "Hybrid Approaches for Diamond Authentication: Combining Spectroscopy and Machine Learning." Journal of Advanced Materials Research, 14(6), 333-348.
15. Singh, R., & Verma, P. (2022). "Explaining Predictions in Diamond Price Modeling Using Explainable AI." International Journal of Data Science, 10(1), 87-103.
16. Wang, H., et al. (2020). "Deep Learning for Diamond Price Prediction: A Neural Network Approach." Journal of Predictive Analytics, 5(3), 230-248.

© 2024 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.