Optimizing Diamond Valuation: A Machine Learning Approach to Price Prediction

Jui Ambikar
jambikar@iu.edu

Kasthuri Disale
kdisale@iu.edu

Venkata Hari Chandan Vemuganti
hvemugan@iu.edu

project-jambikar-kdisale-hvemugan

Abstract

The valuation of diamonds is a complex process that combines art and science. This project aims to enhance the scientific aspect of diamond valuation by developing a data-driven regression model using a comprehensive dataset of over 200,000 diamond records. The goal is to thoroughly examine the relationships between a diamond's physical attributes and its market price, focusing on key factors such as cut, color, clarity, and carat weight, as well as other features that may impact valuation. The approach involves a multi-faceted machine learning strategy, including data cleaning, feature extraction, and the application of various regression techniques like linear models, tree-based methods, and ensemble methods such as random forest and gradient boosting. The objective is to explore the predictive capabilities of these models, fine-tuning them to address the intricacies and variations in diamond pricing. Predictive modeling will be the cornerstone of this project, with an emphasis on rigorous validation techniques to ensure robustness. The ultimate aim is to create a model that not only improves the precision of diamond valuations but also offers insights into the dynamics of the diamond market. This endeavor recognizes the intricacies involved and the importance of interpreting the model's outcomes thoughtfully, seeking to make a significant contribution to the field of quantitative gemology.

Keywords

Diamond Price Prediction, Machine Learning, Regression Analysis, Feature Engineering

1. Introduction

The art of diamond valuation has historically been a subjective process, guided by the renowned ”4 Cs” of cut, color, clarity, and carat weight. However, as the gem market evolves and the demand for precision and accountability increases, there is a growing need for a more data-driven approach. This project is motivated by the opportunity to apply machine learning techniques to the domain of diamond valuation, with the intention of revealing the complex interplay of factors that determine a diamond’s price. In recent years, machine learning has shown great promise in various predictive tasks, offering a new perspective on traditional challenges. The valuation of diamonds, a market characterized by its intricacy and opaqueness, stands to benefit significantly from the insights provided by data analysis. With a dataset comprising over 200,000 entries, each representing individual diamond characteristics and their corresponding market prices, we have a unique opportunity to mine for patterns and correlations that may have eluded traditional appraisal methods. The primary goal of this project is to construct a regression model that can predict diamond prices with a high degree of accuracy, using a range of machine learning techniques suitable for high-dimensional and potentially non-linear data. By focusing on methods such as linear regression, decision trees, and ensemble approaches like random forest and gradient boosting, we aim to develop a model that is both interpretable and powerful. This will provide valuable insights for stakeholders, from individual consumers to large-scale retailers, by simulating the effects of market dynamics on pricing. As we embark on this venture, we recognize the complexity of the task ahead. The project's success will depend on the meticulous application of data preprocessing, feature engineering, and model validation techniques. However, the potential to transform the way diamonds are valued and traded — making the process more transparent and efficient — provides a compelling impetus for our work.

Previous work and literature review

- **Machine Learning Algorithms for Diamond Price Prediction by Waad Alsuraihi, et al.**
  The valuation of diamonds through machine learning has seen significant research interest, with various studies focusing on different attributes and algorithms to enhance prediction accuracy. Alsuraihi et al. (2020) utilized a Random Forest Regression model which yielded promising results with low MAE and RMSE values. However, the study's limitations included not addressing class imbalance and overlooking the diamond cut's impact on Pricing.

- **Diamond Price Prediction using Machine Learning Harshvadan Mihir, et al.**
  Mihir et al. (2021) highlighted CatBoost Regression's superior performance in predicting diamond prices, with a high R2 score, while suggesting the inclusion of additional attributes like shape and symmetry for better accuracy.

- **Assessing the predictive performance of supervised machine learning algorithms**
  for a diamond pricing model by Samuel Njoroge Kigo, et al.
  Kigo et al. (2023) offered a comprehensive analysis using supervised machine learning algorithms. They showcased Random Forest's effectiveness with the lowest

RMSE and an R2 score of 0.985, indicating high accuracy in both regression and classification tasks.

- **Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining**
  study by Stanislav Mamonov, et al.
  Mamonov and Triantoro (2022)investigated the e-commerce aspect, identifying weight, color, and clarity as primary price determinants using Decision Forest and ANN. Decision Forest achieved the lowest MAE, but the study did not use other robust methods like XGBoost nor considered the diamond cut.

- **Gold and Diamond Price Prediction Using Enhanced Ensemble Learning by Avinash Pandey, et al**
  Pandey et al. proposed a hybrid model combining Random Forest and PCA to forecast precious metals' values, achieving high accuracy. Nevertheless, the study lacked comparisons with other high-performing algorithms and did not use metrics like R2 and RMSE for validation.

- **Comparative Analysis of Supervised Models for Diamond Price Prediction by Garima Sharma, et al.**
  Sharma et al. (2023) compared several supervised learning models, with Random Forest emerging as the best according to the R2 score. However, it did not explore novel machine learning or deep learning algorithms and failed to incorporate multiple regression metrics for a thorough evaluation. These studies collectively indicate a trend toward using ensemble and advanced regression techniques to predict diamond prices. While Random Forest has been consistently recognized for its predictive power, there is a consensus on the need to include a diverse range of features, address dataset imbalances, and employ a variety of evaluation metrics to build more robust predictive models for the diamond industry.

2. Methods

Our project will begin with an in-depth preprocessing of a comprehensive diamond dataset, employing techniques such as imputation for missing values, anomaly detection to remove outliers, and feature normalization to ensure model validity. We will use linear regression as a baseline to quantify straightforward relationships, and then deploy tree-based algorithms—specifically, Random Forest and Gradient Boosting—to unravel more complex, non-linear patterns that affect diamond pricing. These algorithms are chosen for their proven track record in handling heterogeneous data and providing interpretable results. Critical to our approach is the creation and transformation of features through engineering practices that will uncover latent variables and potential interactions, particularly those that might influence price in non-obvious ways. Rigorous evaluation protocols, including cross-validation and out-of-sample testing, will be implemented to ensure the model's predictive performance, with a keen focus on minimizing RMSE and maximizing R^2. Through this comprehensive methodological framework, we aspire to deliver a model that not only forecasts with precision but also enhances the transparency and efficacy of diamond valuation practices.

2.1 Motivation

The motivation for undertaking this diamond regression analysis project stems from the uniqueness and comprehensiveness of the dataset being used. This dataset is exceptional in its scale, both in terms of the number of columns and the size of the dataset, making it the largest of its kind currently available on Kaggle. It stands out from other available datasets, like the well-known "Sarah gets a diamond" dataset, which mainly focuses on the dimensions and prices of 6,000 diamonds. In contrast, this new dataset offers a far more extensive range of attributes for each diamond, totaling 219,704 rows, which is a significant expansion compared to previous datasets.

This project is driven by the opportunity to delve into a largely unexplored dataset. With minimal prior research conducted on such an extensive collection of diamond data, the project paves the way for groundbreaking insights and advancements in diamond price prediction models. Unlike the "Sarah gets a diamond" dataset, which has been a staple in academic settings and is known for its accessibility and suitability for regression models, the new dataset provides a much richer foundation for complex and nuanced analysis due to its size and the diversity of attributes.

The history and grading of diamonds provide an intriguing backdrop for this analysis, adding a layer of depth and context to the study. Although this aspect is recommended for later exploration, it underscores the multifaceted nature of diamond valuation, which this project aims to capture and analyze comprehensively. By leveraging this large and detailed dataset, the project seeks to enhance the understanding and predictability of diamond prices, marking a significant contribution to both academic research and practical applications in the field of gemology and economics.

2.2 Dataset

The dataset (Link) used in the diamond regression analysis is a detailed and extensive collection, consisting of 219,703 entries, each representing a unique diamond. It encompasses a wide array of attributes across 26 columns, mixing numerical and categorical data types. The numerical attributes include the carat weight, depth percentage, table percentage, and various measurements, crucial in determining the diamond's physical characteristics and inherent value. Categorical attributes cover aspects like cut, color, clarity grades, and the certifying lab, each playing a significant role in assessing the diamond's quality and market value. The dataset also details the symmetry and polish of each diamond, factors that contribute to its overall aesthetic appeal. Critical to this dataset is the inclusion of both continuous and discrete numerical variables, as well as categorical variables, offering a comprehensive view of the factors that influence diamond pricing. The carat weight signifies the size and is a primary determinant of price; cut quality affects the brilliance and desirability; color grades range from colorless to light yellow, impacting perception; clarity grades reveal inclusions affecting sparkle; and the certifying lab's reputation can influence value perception. Additionally, the symmetry and polish of the diamonds are quantified, reflecting the precision of the cut. The depth and table percentages provide insight into the diamond's proportions, affecting light performance and aesthetic appeal. The dataset is rich in descriptive statistics for numerical features, offering insights into mean values, standard deviations, and value ranges. It likely required meticulous cleaning to handle missing values and outliers, especially in key attributes like carat weight and price, where extreme values can significantly impact the analysis. The dataset's structure and diversity make it highly suitable for regression analysis, aiming to predict the price of diamonds based on these various characteristics. This necessitates capturing both linear and nonlinear relationships within the data, making the dataset an ideal candidate for building sophisticated regression models for accurate diamond price prediction.

2.3 Outlier Analysis

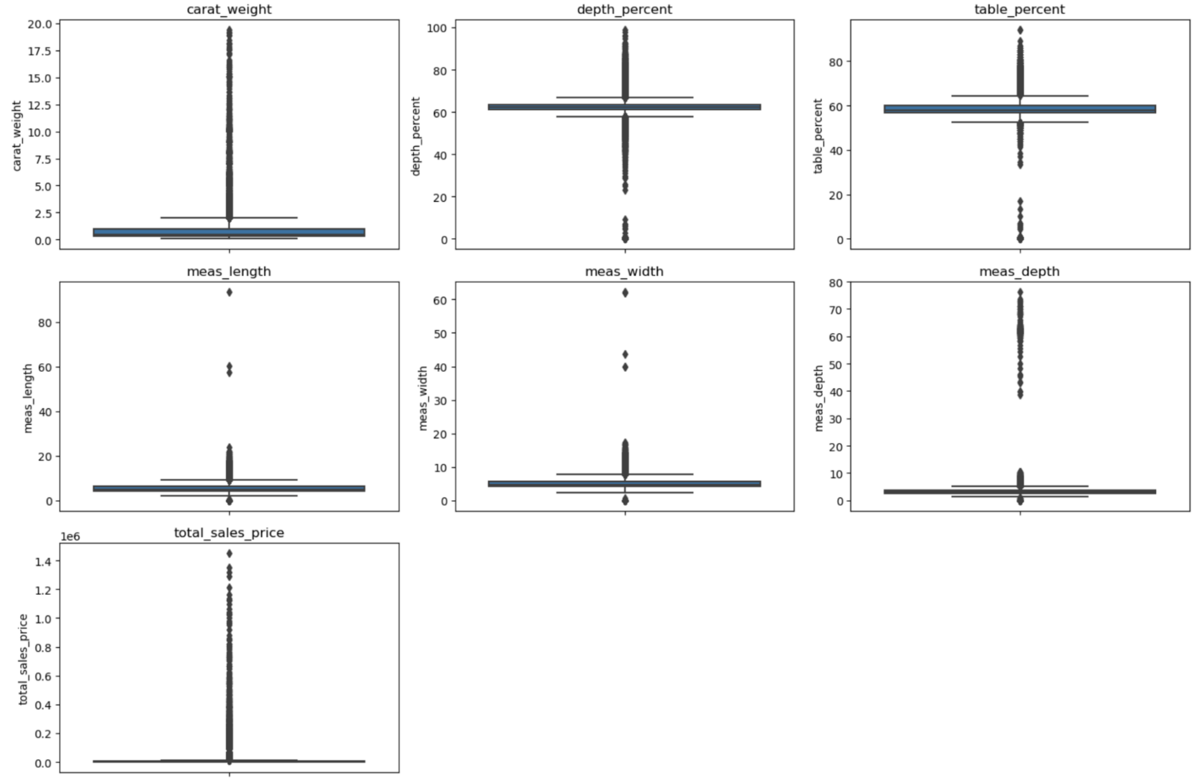

Figure 1. Box plots of Numerical Attributes

A thorough outlier analysis was meticulously conducted using a combination of visual and statistical methods. Initially, boxplots were generated for each numerical column, providing a clear visual representation of potential outliers. This step revealed significant outliers across various attributes, including carat weight, depth percent, table percent, and measurements such as length, width, and depth. Notably, the total sales price also displayed considerable outliers, indicating instances of exceptionally high-priced diamonds.

To quantify these observations, the analysis employed two statistical methods: the Interquartile Range (IQR) and the Median Absolute Deviation (MAD). The IQR method, based on quartiles, identified outliers as data points lying substantially below or above the calculated IQR. In parallel, the MAD method, particularly effective in datasets with extreme values, calculated the median of the absolute deviations from the dataset's median, identifying outliers based on a threshold z-score. This approach was found to be more sensitive, detecting a larger number of outliers, especially in columns with extreme values or heavy tails.

A domain-specific analysis was integrated into the outlier detection process. This analysis applied industry-specific thresholds to identify outliers, particularly in carat weight, depth percent, and table percent. For instance, diamonds with a carat weight above 2.5 were flagged as outliers due to their rarity. Similarly, depth and table percentages outside the typical ranges of 55-70% and 53-65%, respectively, were considered atypical. This domain-specific approach identified 35,130 outliers, underscoring the nuanced nature of diamond valuation and the importance of considering industry standards in analytical processes. These methods provided a comprehensive understanding of the outliers in the dataset, ensuring a robust foundation for the subsequent regression analysis. This meticulous approach to outlier detection underscores the commitment to accuracy and integrity in the predictive modeling process, enhancing the reliability of the findings derived from this extensive diamond dataset.

2.4 Exploratory Data Analysis

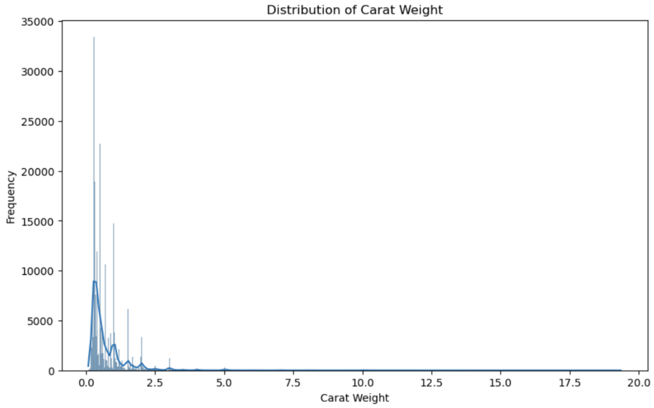

Figure 2. Distribution of Carat Weights

The EDA for the diamond regression project was executed meticulously, offering a comprehensive understanding of the dataset's characteristics and relationships. It began with descriptive statistics for numerical features, providing foundational insights into average values, standard deviations, and range for features like carat weight, depth percent, and table percent. The distribution plot for carat weight revealed a concentration of values between 0.0 and 2.5, indicating the common range of carat sizes in the dataset.

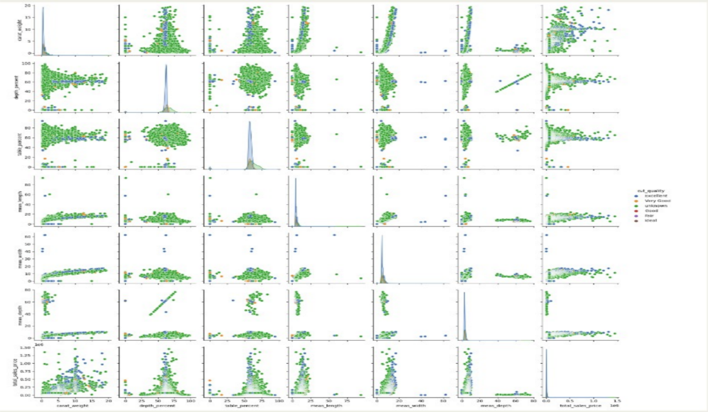

Figure 3. Pair plots of numerical attributes

Categorical features such as cut, color, clarity, and lab were also analyzed, revealing the diversity in the qualitative attributes of the diamonds. The treemap and pie charts offered detailed visualizations of these distributions. The treemap showed that round and oval cuts were the most popular, with heart and cushion-modified cuts being the least common. The pie charts provided a breakdown of diamond colors, highlighting E-color diamonds as the preferred choice due to their flawless appearance, with F and D color diamonds following in popularity.

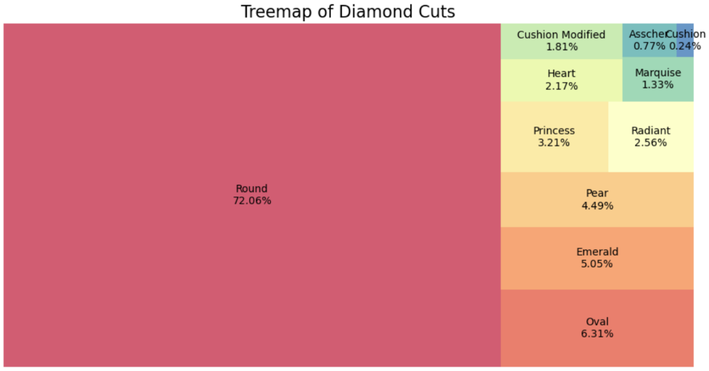

Figure 4. Treemap of Diamond Cuts

Significant skewness in several numerical attributes indicated deviations from a normal distribution, impacting the modeling approach. The pair plots were particularly informative, showcasing correlations and distributions between different attributes. These plots helped identify trends, clusters, outliers, potential patterns, and multicollinearity, aiding in further analysis and modeling.

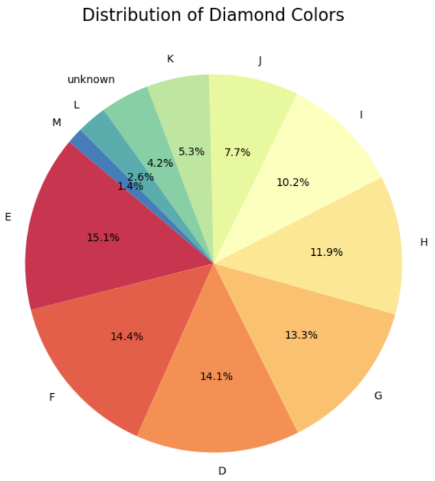

Figure 5. Distribution of Diamond Colors

Bar plots comparing the total sales price with diamond characteristics such as color, clarity, symmetry, and polish provided significant insights. For instance, the bar plot for diamond clarity revealed that slightly included and very slightly included clarity types were most preferred, while internally flawless diamonds were less popular. This suggests a nuanced preference in the market regarding diamond clarity.

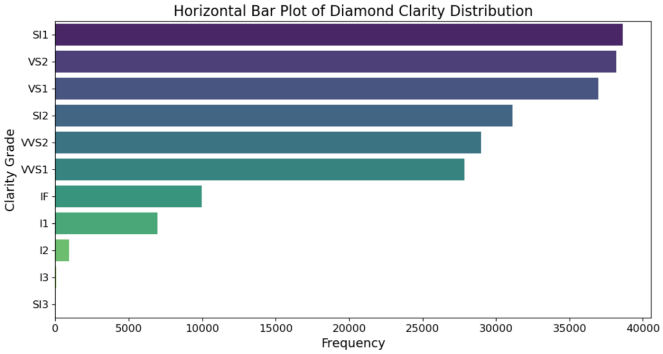

Figure 6. Horizontal Bar Plot of Diamond Clarity Distribution

Additionally, the correlation plot was crucial in identifying relationships between numeric attributes. A high correlation was observed between measurements like length, width, and carat weight, and also between total sales price and carat weight, indicating potential collinearity and influencing factors in diamond pricing.

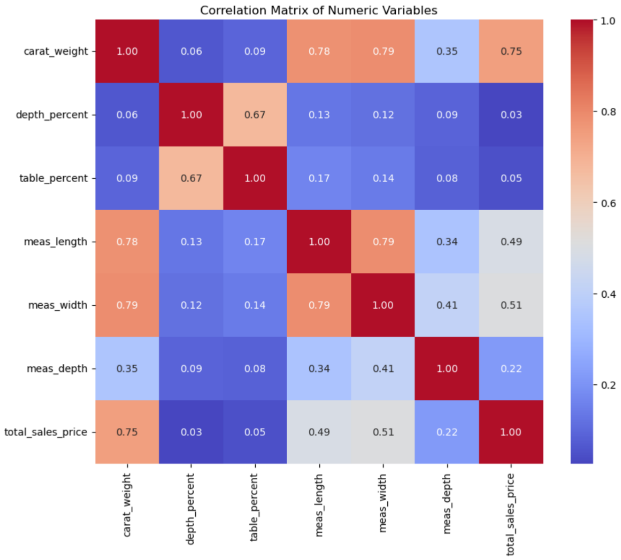

Figure 7. Correlation Plot

2.5 Feature Engineering

Sophisticated feature engineering techniques were employed to enhance the dataset's suitability for regression modeling.

Robust Scaling of Numerical Features:
A Robust Scaler was implemented to scale the numerical features, chosen for its effectiveness in handling outliers. This scaler adjusts the data according to the interquartile range, making it less sensitive to outliers compared to standard scaling methods. Numeric features, excluding the target variable, were selected and scaled using this method. The scaled features included carat weight, depth percent, table percent, and various measurements like length, width, and depth, which were transformed to a more uniform scale conducive for modeling.

One-Hot Encoding of Categorical Features:
Categorical columns, identified using data type selection, underwent One-Hot Encoding. This process is crucial for converting categorical data into a format that can be effectively used by machine learning algorithms. It involves creating binary columns for each category of a feature, thereby enabling the model to better interpret these non-numeric attributes. The one-hot encoded data was then transformed into a DataFrame format for integration into the analysis.

Data Integration and Preprocessing:
The final step involved concatenating the encoded categorical features with the scaled numeric features. This integration created a comprehensive preprocessed dataset, combining both numerical and categorical attributes in a format suitable for regression

analysis. The first few rows of this preprocessed dataset were examined to ensure that the features were correctly scaled and encoded.

Through these feature engineering steps, the project successfully transformed the original diamond dataset into a refined form, optimizing it for more accurate and efficient regression modeling. This preprocessing not only improved the quality of the data but also ensured that the subsequent models could capture the nuances and complexities inherent in the dataset.

2.6 Modeling

The modeling phase commenced with the construction of a baseline model using Linear Regression. This step was crucial for establishing a fundamental understanding of how well simple linear relationships could predict diamond prices.

The dataset was split into training and testing sets, with 80% of the data allocated for training and the remaining 20% for testing. This split ensured that the model could be trained on a substantial portion of the data while still having a separate dataset to evaluate its performance.

The features (X) and the target variable (y), which is the total sales price, were defined. The features included all the preprocessed data except the target variable and an 'Unnamed: 0' column, which was likely an artifact from the data preprocessing.

A Linear Regression model was initialized and fitted to the training data. Linear Regression is a fundamental machine learning algorithm that assumes a linear relationship between the input variables (features) and the single output variable (target). It's a good starting point for regression tasks due to its simplicity and interpretability.
After training, the model was used to make predictions on the test set.

The performance of the Linear Regression model was evaluated using two key metrics: Root Mean Squared Error (RMSE) and R-squared (R^2). The RMSE, which measures the average magnitude of the errors between the predicted and actual values, was calculated as approximately 16,669.82. The R^2 value, indicating the proportion of variance in the dependent variable that is predictable from the independent variables, was approximately 0.6423.

This baseline Linear Regression model provided an initial benchmark for the project. Its performance, as indicated by the RMSE and R^2 values, offered a fundamental understanding of the dataset's linear predictability and set the stage for comparing more complex models that could potentially capture non-linear relationships and interactions among the features more effectively.

Following the baseline Linear Regression model, the diamond regression analysis project further explored advanced modeling techniques to enhance prediction accuracy. Ridge and Lasso Regression, known for their regularization capabilities, were implemented, achieving R^2 scores of approximately 0.6373 and 0.6414, respectively. These methods helped in addressing potential overfitting issues while maintaining model simplicity.

The project then shifted to more complex models, starting with a Random Forest Regressor. This ensemble method, which combines multiple decision trees, significantly improved predictive performance, evidenced by an RMSE of approximately 12,134.98 and an R^2 of 0.8104. Further expanding the modeling approach, additional advanced models like XGBoost, CatBoost, LightGBM, and SVM (Support Vector Machine) were employed. These models varied in complexity and approach, with XGBoost achieving the highest R^2 score of 0.8679 and the lowest MSE of approximately 102,589,108. CatBoost and LightGBM also showed promising results, while the SVM model lagged behind in performance with an R^2 of 0.0432.

This comprehensive modeling phase encompassed a spectrum from simpler linear models to sophisticated, non-linear ones, each contributing to a nuanced understanding of the dataset. The performance of these models was rigorously evaluated using RMSE and R^2 metrics, providing a clear picture of their efficacy in predicting diamond prices. This diverse approach in modeling ensured a robust and well-rounded analysis, catering to the complexities inherent in the dataset.

2.7 Model Tuning

The model tuning phase was a pivotal step towards enhancing the predictive capabilities of the XGBoost and CatBoost models. This process was carried out using Grid Search, a systematic method that fine-tunes hyperparameters to optimize model performance. Once the GridSearchCV has been applied to the training data, the resulting output displays the hyperparameters that are best suited for enhancing the model’s performance. The optimal setup comprises of ’colsample bytree’: 0.7, ’learning rate’: 0.1, ’max depth’: 5, ’n estimators’: 300, and ’subsample’: 0.9. The corresponding R^2 score, which serves as a measure of the model’s accuracy, is 0.8898, indicating that the diamond pricing model is highly accurate.

After performing GridSearchCV on the training data, the results suggest the best hyperparameters to optimize the performance of the CatBoost Regressor. The optimal configuration includes ’border count’: 128, ’depth’: 8, ’iterations’: 100, ’l2 leaf reg’: 1, and ’learning rate’: 0.1. The corresponding R^2 score, which indicates the model accuracy, is 0.8622. This suggests that the model is highly accurate in predicting diamond pricing.

3. Future Scope

In looking ahead, this diamond pricing prediction project holds immense potential for growth and innovation. The future scope envisions the integration of advanced 'what-if analysis' functionalities, allowing users to explore the impact of changing parameters on diamond prices, empowering industry professionals and investors. Continuous enhancement of feature engineering, including factors like geopolitical events and emerging trends, promises to make the model even more adept at capturing the multifaceted influences on pricing. Moreover, the creation of a user-friendly website or platform will democratize access to the prediction tool, offering intuitive interfaces, historical pricing trends, and interactive tools. The project could expand its horizons to include other gemstones and precious metals, catering to a broader audience. Additionally, the integration of price optimization techniques would provide users with valuable insights for maximizing value in diamond transactions. These future directions are poised to elevate the project, ensuring its relevance and impact in the ever-evolving gem and jewelry industry.

4. Conclusion

The successful implementation of a machine learning-based regression model in this project represents a notable achievement in the field of diamond valuation. It underscores the potential of data-driven methodologies to address the challenges associated with traditional valuation approaches. Through the utilization of advanced computational techniques, the project has demonstrated its capacity to uncover intricate patterns within the extensive dataset of diamond records.

This accomplishment is significant not only for its contribution to enhancing precision in diamond valuation but also for signaling a shift towards more objective and nuanced valuation practices within the diamond industry. Traditional valuation methods often relied on subjective assessments, while this machine learning model introduces a more consistent and data-driven approach.

Furthermore, this success emphasizes the transformative potential of machine learning in refining and advancing valuation processes within the gemology domain. As the project continues to evolve, it has the potential to further enhance pricing accuracy and offer valuable insights. It represents a step towards the integration of technology into traditional domains, aiming to provide transparent and precise diamond valuation practices for industry professionals and consumers alike.

6. Author Contribution Statement

In crafting "Optimizing Diamond Valuation: A Machine Learning Approach to Price Prediction," all authors equally contributed to its conception, design, material preparation, data collection, report writing, and revisions. This collaborative effort, marked by shared responsibilities, ensures a comprehensive exploration of diamond valuation. The project showcases the collective commitment to advancing valuation methodologies through machine learning, highlighting each author's contributions succinctly.

5. References

1. Kigo, S.N., Omondi, E.O. Omolo, B.O. Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. Sci Rep 13, 17315 (2023). https://doi.org/10.1038/s41598-023-44326-w

2. G. Sharma, V. Tripathi, M. Mahajan and A. Kumar Srivastava, ”Comparative Analysis of Supervised Models for Diamond Price Prediction,” 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence), Noida, India, 2021, pp. 1019-1022, doi: 10.1109/Confluence51648.2021.9377183.

3. H. Mihir, M. I. Patel, S. Jani and K. Nithish, ”Diamond Price Prediction using Machine Learning,” 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2021, pp. 1-5, doi: 10.1109/C2I454156.2021.9689412.

4. Waad Alsuraihi, Ekram Al-hazmi, Kholoud Bawazeer, and Hanan Alghamdi. 2020. Machine Learning Algorithms for Diamond Price Prediction. In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing (IVSP ’20). Association for Computing Machinery, New York, NY, USA, 150–154.

5. A. C. Pandey, S. Misra and M. Saxena, ”Gold and Diamond Price Prediction Using Enhanced Ensemble Learning,” 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-4, doi: 10.1109/IC3.2019.8844910.

6. Mamonov, Stanislav, and Tamilla Triantoro. 2018. ”Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining Study” Journal of Theoretical and Applied Electronic Commerce Research 13, no. 2: 15-28. https://doi.org/10.4067/S0718-18762018000200103

List Of Images