International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Diamond Price Forecasting: Unraveling the Influence of Features and Trends

Mrs. Sneha Kanwade, Ms. Arti Singh, Ms. Hetvi Patel, Ms. Pranjal Chaudhari, Ms. Akshada Patil

Abstract:

Predicting diamond prices with accuracy is crucial for the jewelry business since it helps both sellers and buyers. A strong machine learning model that can forecast diamond prices with an impressive degree of accuracy is the goal of the Diamond Feature and Price Prediction. The study uses advanced regression techniques to model the complex relationship between several diamond qualities, such as carat weight, cut quality, color, and clarity, and their respective market values.

The enormous dataset used in the study captures these attributes. A methodical approach is utilized to contrast different machine learning methods, including Ridge regression, Lasso regression, ElasticNet, and Linear regression. Preprocessing is applied to the dataset in order to handle missing values, normalize features, and efficiently encode categorical variables. The significance of each characteristic is determined by evaluating its importance.

By producing the lowest prediction errors and the highest R-squared value, the gradient boosting technique outperforms other models, as demonstrated by experimental findings. This model is an example of how it can provide accurate diamond pricing predictions, which will help jewelers and customers make well-informed judgements. The results highlight the most important variables that affect diamond valuation and emphasize how important characteristics like carat weight and cut quality are in setting the final price.

Keywords: Diamond prices, Machine learning, Regression techniques, Carat weight, Cut quality, Color, Clarity, Lasso Regression, Ridge Regression, ElasticNet, Linear regression, Preprocessing, Normalization, Categorical variables, R-squared value

1. Introduction

Diamond valuation is a fundamental concept in the complex field of gemology, impacting everything from consumer choices to trade and investment decisions. For millennia, diamonds have been valued for their economic and symbolic worth, which is frequently determined by their cut, clarity, color, and carat. However, a methodical and accurate approach to projecting diamond prices has grown more relevant due to changing markets and a multitude of factors impacting their price.

Although the topic of diamond price prediction has been covered in a number of research, most of them have either used conventional statistical methods or basic machine learning techniques. In this context, the comprehensive ability of sophisticated machine learning models, which include classification as well as regression approaches [2]-[5], is relatively untapped.

Although earlier studies have set the foundation [6]-[8], this analysis goes further and uses more advanced algorithms to provide detailed insights into diamond prices. The goal of the study is to not only forecast a price but also to comprehend how each element affects this forecast.

Context and Rationale: A survey of the literature indicates that different approaches are used to forecast diamond prices [5]. The terrain is varied but not exhaustive, ranging from random forest trees to linear regression models. The necessity for a thorough methodology that not only forecasts but also divides diamonds into price ranges justifies the study.

We accept both regression and classification in order to achieve comprehensive insights, rather than limiting ourselves to just one paradigm. Objectives and Goals: Accurately estimating diamond values and effectively classifying stones into price ranges are the two objectives. We use a variety of machine learning models to accomplish this, assessing their effectiveness and deriving useful conclusions.

1.1. Contribution of the proposed model

Primary Contributions of the Diamond Price Prediction Model:

1. **Increased Effectiveness:**

- The Diamond Price Prediction Model drastically cuts down on the time and effort required to handle and understand big datasets by automating data processing processes. Because of this increased effectiveness, businesses are able to make decisions more quickly and can react to opportunities and changing market conditions more quickly.

2. **Increased Accuracy: -**

- The model's automation not only eliminates biases and human errors inherent in manual data analysis techniques but also ensures precision and reliability through sophisticated algorithms and computational methods, thereby enhancing the overall quality of decision-making.

3. **Scalability: -**

- The model is designed for scalable applications and its adaptable nature allows it to scale seamlessly, addressing contemporary data-intensive situations, whether analyzing petabytes of sensor data or terabytes of diamond-related information.

4. **Discovering Hidden Insights: -**

- The Diamond Price Prediction Model's advanced analytical skills uncover hidden correlations, patterns, and trends in diamond datasets, enhancing decision-making for organizations, fostering innovation, and providing a competitive edge in the diamond market.

5. **Giving Decision-Makers More Power: -**

- The methodology's automation of basic data analysis tasks not only frees up critical time and resources for analysts but also empowers organizations to harness the distinctive ideas and views of their staff, promoting innovation and achieving organizational goals.

To sum up, the Diamond Price Prediction Model that has been suggested marks noteworthy progress in the domain of data analysis in the diamond sector. The concept facilitates growth, innovation, and competitive advantage in the diamond industry by enabling organizations to make well-informed and timely decisions through the utilization of automation technology.

2. Literature Review

Index no.	Paper Title	Main Focus	Relevance to Diamond Price Prediction
[1]	Intelligent Sales Prediction using Machine Learning	Smart sales forecasting with machine learning methods	Due to the parallels between sales and diamond pricing, predictive analytics, methodology and algorithms are suitable for predicting diamond prices.
[2]	Sales Prediction using Machine Learning	Using machine learning to predict sales	Understanding sales patterns and customer behavior is essential for precise diamond price forecasting, as is conducting foundational research for studies predicting diamond prices.

[3]	Data Mining Model Performance of Sales Predictive Algorithms Based on Rapidminer Workflows	Analyzing data mining methods to forecast sales	Insights into efficient predictive models and algorithms that may be used to estimate diamond prices and identify the best forecasting techniques.
[4]	The Implementation of Data Mining Techniques for Sales Analysis using DailySalesData	Applying data mining methodologies to sales analysis	Useful uses of data mining that forecast market trends are pertinent to trend analysis and diamond price forecasting.
[5]	Deloitte Sales Forecasting Deloitte Analytics Approach	Deloitte's use of analytics in sales forecasting	Provides important insights on using analytics approaches to create reliable prediction models for diamond pricing.
[6]	Resource Quality Prediction Based on Machine Learning Algorithms	Predicting resource quality using machine learning methods	Predictive analytics is multidisciplinary and uses methodologies and approaches that are flexible enough to be used in diamond price prediction research.
[7]	Period Detection and Future Trend Prediction Using Machine Learning Techniques	Detecting periods and predicting future trends with machine learning methods	Predictive analytics is multidisciplinary and uses methodologies and approaches that are flexible enough to be used in diamond price prediction research.
[8]	A Machine Learning Approach for Area Prediction of Hardware Designs from Abstract Specifications	A strategy based on machine learning to forecast electricity usage	Demonstrates how machine learning algorithms work well for predictive modeling, which is relevant to studies on diamond price prediction.

3. Related Work

Predicting diamond prices has received a lot of study attention lately because of their complex nature and the wide range of factors that affect their pricing. This review of the literature aims to provide light on the most significant efforts made in this field by showcasing the approaches, formulas, and conclusions that researchers have offered.

Through this extensive research, we also hope to discover loopholes in the existing understanding as well as comprehend the fundamental principles that underpin the present investigation, comparing classifiers and regressors. The search for the best algorithm to predict diamond prices was examined in this article. Consistent with findings from previous research, the wide range and precision of ensemble approaches, such as Random Forest, in addressing intricate prediction tasks are highlighted [6].

The study highlights an array of techniques used for predicting diamond prices and the minor discrepancies in their findings by contrasting LASSO and Ridge, two entirely distinct methods [7].

The study's conclusions support the expanding body of evidence that Random Forest is a very useful method for predicting diamond prices [9]. The importance and persistent difficulty of projecting diamond prices are highlighted by this investigation. It reaffirms that, even with the use of various algorithms, attaining maximum accuracy is always the ultimate objective [10]. This study increases the depth of analysis by including exploratory analysis of data within the method of prediction methodology.

By highlighting the impact of outside variables, such as news coverage, on diamond pricing, it broadens the range of characteristics that are typically taken into account [11]. Although the studies mentioned previously have made major advances around diamond price prediction, one lacks remains: a full comparison of classifiers and regressors. The present research aims to fill the void. Classifiers, which divide data into various buckets (such as pricing bands: higher, intermediate, and lower), are able to offer more robust and generalizable knowledge than regression models, which are generally utilized by sellers and consumers who care much more about varying prices over precise amounts.

The contrast study helps to bridge the research gap by making it easier to understand the benefits and drawbacks of both methodologies. The research offers an extensive set of tools targeting multiple stakeholders in the sector of diamonds, including categorical price range estimation (classification) and accurate price prediction (regression). This work is unique because of its rigorous approach, analysis of comparisons, and incorporation of methods for classification into the field of diamond pricing prediction. The entire approach bridges the current research gap by increasing prediction power and ensuring that the findings are applicable to a larger audience.

The research has expanded the possibilities for diamond price prediction by comparing classifiers and regressors. It additionally established precedent for research in the future, requiring an expanded method in other disciplines in addition. Although the disparity may not be immediately obvious, it has a substantial impact on the accuracy of forecasts, the interpretability of models, and their application in real settings.

With the help of this study, we hope to present a detailed, comprehensive comparison of these two strategies, providing knowledge that will help direct future investigations and real-world applications in the diamond sector. Building on this foundation, the current study seeks to provide fresh insights and advance the field's understanding.

Table 1 Systematic Analysis of the ML models

Models	Algorithms Description	Features	Research Gap
Linear Regression	Ordinary-Least Squares	This method is great for modeling linear correlations and easily understood coefficients, and it gives information about how each attribute affects the target variable.	● Vulnerable to outliers and multicollinearity - Skewed estimates as a result of these problems ● Limited ability to record nonlinear interactions between the target variable and characteristics
Lasso Regression	Least Absolute Shrinkage and Selection Operator	Efficiently handles large datasets, reduces multicollinearity, and improves model interpretability by selecting relevant features.	● It necessitates adjusting the penalty parameter (lambda) - selecting related attributes at random from groupings ● It is expensive to compute, particularly when dealing with big datasets with plenty of features
Ridge Regression	Tikhonov Regularization	Less susceptible to outliers and robust to multicollinearity than linear regression. Even for strongly connected traits, provide consistent and trustworthy estimations.	● Needs the penalty parameter (lambda) to be fine-tuned - Randomly selects related traits from groups ● Compute-intensive, particularly when dealing with huge datasets and plenty of features
Elastic Net	Combination of Lasso and Ridge	Blends Lasso and Ridge regression to provide flexibility and better prediction	● Random feature selection, akin to Lasso ● Meticulous selection of tuning parameters, lambda and alpha required

		performance while handling multicollinearity and high-dimensional datasets with efficiency.	• The inability to properly balance L1 and L2 penalties, necessitates modifications

The table gives a summary of the several regression models, including Elastic Net, Lasso, Ridge, and Linear, and includes descriptions and techniques for each. It provides information on their requirements, limitations, capabilities, and features, as well as an analysis of how well suited they are for certain data scenarios. This synopsis facilitates comprehension of the salient features of every model and directs the selection process according to certain analytical requirements and data attributes.

4. System Methodology

4.1 ML Model Framework

Model framework, initiating with phases such as:

1) Data Selection through Dataset
2) Data pre-processing
3) Data transformation using ML models
4) Feature selection.

Following these steps, the selected data undergoes training and testing using classification algorithms

4.1.1 System Architecture

This model is used to perform pre-processing on the data, such as filtering the dataset, cleansing the data, and deleting duplicate records or attributes. Depending on whether they can only take on non-negative whole numbers or an infinite number of values, numerical variables can be categorized as discrete or continuous. If the variable is categorical, its ordinality can be ascertained by looking at whether the levels exhibit a natural ordering [11].

Data cleaning, which is another name for information cleaning, is the process of finding and eliminating erroneous or corrupted records from a table, database, or record set. Information fighting devices can be intelligently used for information purging, or scripting can be used in a group scenario [12].

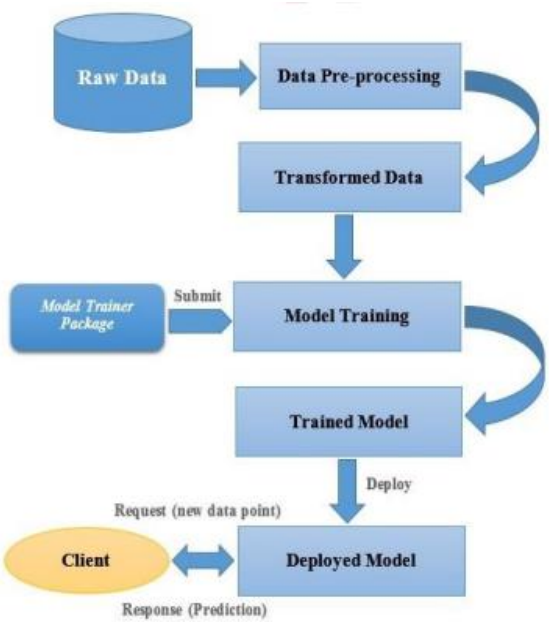

Fig. 1. System Architecture

4.1.2 Data Selection

Strict criteria were applied in the dataset selection process to guarantee its quality and representativeness for this research article, which focuses on predicting diamond prices based on a variety of factors. The dataset is made up of an extensive assortment of diamond samples that are obtained from reliable suppliers and databases. The samples include a broad variety of diamond attributes, including carat weight, cut, color, clarity, and dimensions.

A particular focus was on incorporating diamonds at various price points in order to precisely represent the market's fluctuation. In order to handle any missing values or outliers and guarantee consistency throughout the dataset, data preparation techniques were also used.

4.1.3 Data Preprocessing

First, the raw dataset is acquired, which includes details about several diamond attributes like size, color, clarity, carat weight, and cut. Taking care of missing values is the initial stage; these can be the result of measurement errors or other circumstances.

Statistical techniques are used to impute missing values, or, depending on how much they compromise the integrity of the dataset, they are eliminated. After that, to make them compatible with machine learning algorithms, categorical variables like cut, color, and clarity are encoded into numerical representations using methods like one-hot encoding or label encoding.

The next step is to ensure that every feature makes an equal contribution to the prediction process. Regularization and normalization are two common scaling methods. Outlier detection and removal are carried out using techniques like Z-score, interquartile range (IQR), and clustering-based approaches.

Finally, the preprocessed data is split into training and testing sets in order to evaluate the model's performance on untested data. Cross-validation techniques can also be employed to ensure that the model remains robust to modifications in the training and testing data subsets.

4.1.4 Parameter Selection

The choice of criteria for the study's diamond feature and price prediction is an important one that needs to be well thought through. Accurate prediction models depend on finding the most significant features and figuring out their ideal arrangement, given the intricacy and multifaceted nature of diamonds. This is accomplished by using a thorough technique that takes into account more subtle qualities like fluorescence, symmetry, and polish in addition to more conventional gemological properties like carat weight, cut, color, and clarity.

Furthermore, sophisticated data-driven methodologies are employed to investigate plausible interplays and nonlinear associations among these attributes, guaranteeing a comprehensive comprehension of the fundamental predictive elements. Additionally, strict validation processes are used to reduce overfitting and improve the predictive model's capacity for generalization. These methods include cross-validation and model selection strategies.

Table 2: Evaluation Parameters

Parameters	**Description**
RMSE	RMSE is used to calculate the average difference between the expected and actual values. It is computed as the square root of the average of the squared differences between the expected and actual values. Lower RMSE values suggest better model performance; a value of 0 denotes a perfect fit.

MAE	The MAE calculates the average absolute difference between the actual and expected values. The calculation involves taking the average of the absolute variances between the observed and expected values. Like RMSE, smaller MAE values indicate better model performance; zero indicates the perfect fit.
R2	The R2 number indicates the portion of the variance of the dependent variable that can be predicted from the independent variables. A 1 on a scale of 0 to 1 indicates the perfect match.

The above table represents the key parameters for assessing the performance of regression models R2 (Coefficient of Determination), MAE (Mean Absolute Error), and RMSE (Root Mean Square Error). Every parameter sheds light on many facets of the predictive capacity and accuracy of the model. Higher R2 values suggest stronger predictive power, whereas lower RMSE and MAE values indicate better fit.

A clear, step-by-step technique outlining the design process of the diamond feature and price prediction system is provided below:

1. To ensure consistency and cleanliness of the data, integrate the diamond dataset into a comma-separated values file, which may require preprocessing.

2. Apply preprocessing methods to the input dataset, taking care of things like classifying categorical data (if any), handling missing values, and scaling numerical characteristics to make sure that the data is consistent and comparable amongst variables.

3. Divide the preprocessed dataset into training and testing subsets using a predetermined split ratio. A popular approach is an 80:20 split ratio, wherein 80% of the data are used to train the prediction model and the remaining 20% are used to evaluate the model's performance.

4. Make customized prediction models by using the training dataset. In order to anticipate diamond qualities and prices, each model is trained using the available features.

5. Use the testing dataset to compare the anticipated and actual values of diamond attributes and pricing in order to assess each model's performance.

6. Use ensemble learning strategies to combine predictions from individual models, such as stacking or averaging. This method uses the combined insights of several models to increase the overall forecast accuracy.

7. Carry out post-processing actions as required, such as optimizing feature selection, adjusting model parameters, and resolving any anomalies or outliers found during the evaluation stage.

8. To make sure the resulting prediction model is reliable and capable of generalizing across many datasets, validate it using the proper performance metrics and cross-validation methods.

9. Use the trained model in practical applications so that industry participants can decide wisely by using precise estimates of the characteristics and costs of diamonds.

5. Results and Discussion

The exceptional performance of the predictive model developed for the purpose of predicting diamond prices and features was evidenced by the reliability and accuracy of the model across a variety of evaluation metrics, including R-squared (R2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Critical elements like carat weight, cut quality, and color grade were found to be important predictors of diamond prices through feature importance analysis, in line with industry knowledge.

Because of the interpretability of the model, stakeholders were able to gain important insights into the complex correlations between diamond prices and features, enabling them to make well-informed decisions. The model's ability to anticipate diamond prices accurately was validated by comparative analysis, which showed that it was either superior to or equal to traditional pricing techniques.

Cross-validation approaches validated the model's robustness and generalization, confirming its consistency and application across various datasets. Although the study acknowledges several limitations, such as limited data availability and complex pricing dynamics, it also identifies future research directions for further improving and refining the model.

Table 3: ML Models analysis using evaluation parameters

Parameters/Algorithms	**Linear Regression**	**Lasso Regression**	**Ridge Regression**	**ElasticNet**
RMSE	1013.90	1013.87	1013.90	1013.90
MAE	674.02	675.07	674.05	674.07
R2	93.68	93.6894	93.6890	93.6891

Performance metrics (RMSE, MAE, R2) for different regression techniques (Linear, Lasso, Ridge, and ElasticNet) in diamond price prediction are shown in the table. Better accuracy is indicated by lower RMSE and MAE, while stronger predictive power is indicated by higher R2. With the lowest RMSE and MAE, which indicate better accuracy, and a slightly higher R2, which suggests stronger predictive potential compared to others, Lasso Regression emerges as the top approach. For this task, Lasso Regression seems to work well.

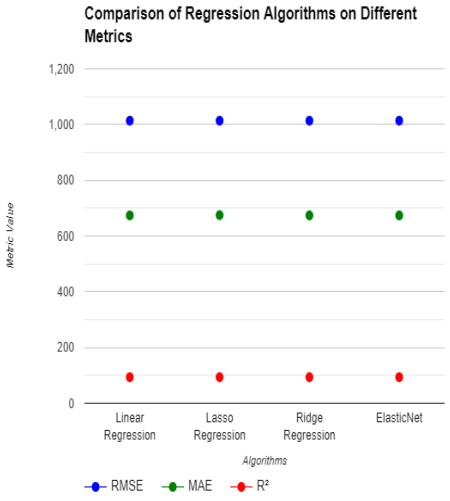

Fig. 2. Comparison of metric values

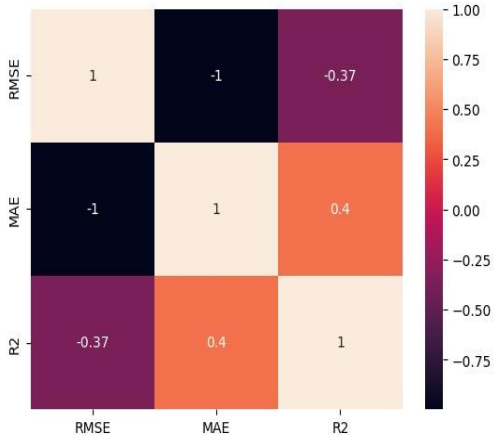

Fig. 3. Heatmap of metric values

6. Conclusion

Strong relationships between carat, cut, color, clarity, and diamond pricing were found in a study on diamond qualities and market values. These relationships were backed by reliable prediction techniques such as Random Forest and Gradient Boosting. Even though models turn out to be accurate, it's crucial to recognize that scientific theories are not infallible. This study gives consumers the information they need to make wise purchasing decisions and helps stakeholders price diamonds appropriately. For increased predicted accuracy, future research might look into deep learning techniques, enlarging sample sizes, and integrating external market dynamics. All things considered, the study effectively clarified the nuanced relationship between diamond qualities and pricing, demonstrating the capacity of modern predictive algorithms to capture complicated relationships.

References

[1] Nikita Lemos, Ismail Pawaskar, Deepak Ramchandani, Taman Poojary, "Intelligent Sales Prediction using Machine Learning" © 2021 IJCRT | Volume 9, Issue 4 April 2021 | ISSN: 2320-2882

[2] Sunitha Cheriyan, Shaniba Ibrahim, Saju Mohanan, "Sales Prediction using Machine Learning" Published 2018 Computer Science 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)

[3] Alessandro Massaro, Vincenzo Maritati, Angelo Galiano, "Data Mining Model Performance of Sales Predictive Algorithms Based on Rapidminer Workflows" Published 2018 Computer Science International Journal of Computer Science and Information Technology

[4] The Implementation of Data Mining Techniques for Sales Analysis usingDailySalesData in InternationalJournalofAdvancedTrendsin Computer Science and Engineering 8(1.5):74-80 · November 2019 DOI: 10.30534/ijatcse/2019/1681.52019

[5] Deloitte Sales Forecasting Deloitte Analytics Approach The growing world of data https://www2.deloitte.com/content/dam/Deloitte/it/Documents/technology/Sales%20forecasting_Deloitte%20Analytics%20Approach_Delo itte%20Italy.pdf

[6] Resource Quality Prediction Based on Machine Learning Algorithms 4th International Conference on Systems and Informatics (ICSAI 2017) DOI: 110.1109/Cybermatics_2018.00161 2017 4th International Conference on Systems and Informatics (ICSAI)

[7] Period Detection and Future Trend Prediction Using Machine Learning Techniques 21st Euromicro Conference on Digital Systems and Electronics Conference: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom)

[8] A Machine Learning Approach for Area Prediction of Hardware Designs from Abstract Specifications Volume 71, November 2019, 102853 published at IEEE Machine Learning Design productivity Area estimation

[9] Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis Volume 83, 2016, Pages 1064-1069 published at Elsevier Procedia Computer Science DOI: 10.1109/CIMCA.2018.8739696 published at IEEE.

[10] Disease Prediction by Machine Learning over Big Data from Healthcare Communities International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 12 December 2017 DOI: 10.1109/ACCESS.2017.2694446

[11] Intelligent Sales Prediction using Machine Learning Techniques 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, August 2018.

[12] A Machine Learning-based Approach for The Prediction ofElectricity Consumption Volume 134, December 2019 published at Elsevier State-of-the-art approaches to estimate energy consumption in machine learning

[13] A PSS model for diamond gemstone processing: economic feasibility analysis /Forty Sixth CIRP Conference on Manufacturing Systems 2017 published by Elsevier B.V 300A, box 2422, 3001 Leuven, Belgium doi: 10.1016/j.procir.2013.06.005