Evaluating the Predictive Proficiency of Machine Learning Algorithms: Progressive Developments in Diamond Price Forecasting

Ying Zhang
University of Leeds, Leeds, LS27FD, U.K.

Keywords: Python, Price, Prediction, Diamond.

Abstract: Distinguished for their global recognition as the most resilient mineral and enduring allure as coveted gemstones, diamonds have captivated human fascination for centuries. The popularity of diamonds extends beyond the intrinsic properties, encompassing optical brilliance and unparalleled hardness which is influenced by durability, tradition, fashion, and robust marketing strategies employed by industry producers. Despite inherent qualities, the demand for diamonds is intricately tied to perceived rarity and exclusivity. Forecasting diamond pricing presents a unique set of challenges primarily rooted in nonlinear relationships within crucial attributes like carat, cut, clarity, table, and depth. In response to the complexity, the research conducts a comprehensive comparative analysis, utilizing diverse supervised machine-learning models for precise prediction via classification and regression approaches. Meticulous evaluation of eXtreme Gradient Boosting, Random Forest, Multiple Linear Regression, k-Nearest Neighbors, and Decision Tree Regressor reveals that the eXtreme Gradient Boosting algorithm emerges as the most optimal choice, boasting an impressive R² score of 98.07% through rigorous evaluation. This research encompasses critical phases, including data preprocessing, exploratory data analysis, model training, accuracy assessment, and result interpretation. Not only sheds light on the intricacies of diamond pricing but also contributes valuable insights for leveraging advanced machine learning techniques in the realm of gemstone valuation and prediction.

1 INTRODUCTION

The Gemological Institute of America (GIA) introduced Cut, Carat, Color, and Clarity. They were providing a standardized framework for assessing and grading diamonds based on their distinct attributes in the 1940s.

The burgeoning global appetite for diamonds has precipitated an imperative for pricing paradigms characterized by both accuracy and transparency. Conventional methodologies, tethered to venerable compendia like the Rapaport Price List, grapple with the intricate challenge of assimilating and mirroring the multifarious dynamism inherent in the diamond market. The idiosyncratic attributes of diamonds manifest in diverse morphologies, dimensions, and gradations of clarity, which introduce a compounding layer of complexity in discerning their intrinsic market value.

The realm of diamond price prognostication, delving into the realm of machine learning, orchestrates a symphony of analytical prowess. This entails the meticulous training of models, leveraging historical datasets and meticulously considering variables such as carat weight, cut quality, color gamut, and clarity. These trained models, having imbibed historical intricacies, extrapolate overarching patterns to venture predictions into uncharted territories of new diamond valuations. A methodological bastion grounded in data-driven acuity, this approach finds an organic alignment with the evolving contours of the gemstone market, cherishing the imperatives of transparency, efficiency, and razor-sharp precision.

In summation, the rubric of diamond price prognostication not only dovetails with age-old valuation paradigms but also interfaces seamlessly with the kaleidoscopic shifts characterizing the contemporary market milieu. In catering to the discerning exigencies of a modern consumer cohort, this predictive discipline emerges as a linchpin, bestowing sagacious insights unto stakeholders, investors, and consumers alike. This predictive accuracy serves as a potent instrument, galvanizing investors with informed decision-making capabilities, charting the course for sagacious

Zhang, Y.
Evaluating the Predictive Proficiency of Machine Learning Algorithms: Progressive Developments in Diamond Price Forecasting.
DOI: 10.5220/0012818500004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 448-452
ISBN: 978-989-758-690-3

industry blueprints, mitigating risks for manufacturers and retailers, offering a compass for consumer choices, enlightening market analyses, fine-tuning pricing strategies for retailers, optimizing the logistics of the supply chain, ensuring the sanctity of transactions through fairness, propelling the frontiers of technological innovation, and charting the course for judicious policy formulation and regulatory oversight. In the crucible of diamond price prognostication, market dynamics are thus infused with a potent elixir that enhances market efficiency, fortifies risk management, and upholds the edifice of equitable transactions, thereby exerting a pervasive influence upon the holistic vigor and equilibrium of the diamond market.

2 LITERATURE REVIEW

Statistical Models: Exploration of statistical models within the realm of diamond price prognostication constitutes a pivotal facet of scholarly inquiries. This corpus of literature, poised at the vanguard of intellectual inquiry, embarks on an expedition into classical statistical models, an odyssey that encompasses the labyrinthine terrain of regression analysis and other intricately woven econometric techniques. The incisive examination of these models, resplendent in their mathematical complexity, unveils a panoply of methodological nuances intrinsic to the predictive tapestry of diamond valuations.

Feature Importance: Studies that traverse the expanse of feature importance within the precincts of diamond price prediction exemplify another facet of erudite discourse. This corpus of scholarly exploration contemplates the salience of diverse features in the predictive matrix, casting an eloquent spotlight upon the integral role played by each facet of the renowned 4Cs. Moreover, the discourse unfurls into the realm of speculative contemplation, entertaining the prospect of introducing additional variables into the predictive equation

Time Series Analysis: Temporal considerations, woven into the fabric of diamond prices like an intricate tapestry, beckon the scholarly gaze toward the frontier of time series analysis. This intellectual endeavor, steeped in analytical sagacity, endeavors to decipher the cryptic language of temporal dynamics inherent in diamond valuations.

Ensemble Methods: The intellectual crucible expands further into the province of ensemble methods, where the alchemy of knowledge metamorphoses into predictive prowess. In this domain, the erudite fraternity contemplates the efficacies of illustrious ensemble methods, including the arboreal complexity of Random Forests and the orchestrated ascent of Gradient Boosting.

Evaluation Metrics: The compendium of literature, marked by its incisive scrutiny, engages in an erudite discussion on the myriad metrics adorning the evaluative tapestry. From the pragmatic expanse of Mean Absolute Error to the geometric profundities of Root Mean Squared Error and the metric symphony of R-squared, the scholarly discourse undertakes a comprehensive survey, offering a lexicon that encapsulates the multifaceted dimensions of predictive model performance assessment.

3 PROBLEM STATEMENT

3.1 Dataset & Dimensional Proportions of Diamond

According to the index of cut, carat, color, and clarity as shown in Figure 1, Table 1.

Table 1: Dataset overview (from Kaggle) (Mirzaei).

Diamond weight in carat	carat	0.2-5.01	Dimensions(mm)
Diamond cutting precision	cutting	Fair, Good, Very Good, Ideal	X length (0-10.74)
Diamond colour	color	From J to D	Y width (0-58.9)
A measure of diamond clarity	clarity	(I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF)	Z depth (0-31.8)

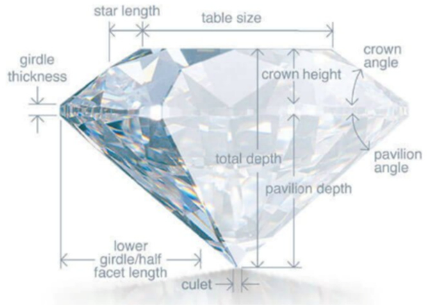


Figure 1: Dimensional proportions of diamond (from Kaggle) (Karnika Kapoor).

3.2 Objectives of the Study & Research Question

The objective is to evaluate the forecasting efficacy of Python models in predicting diamond prices.

The research question seeks to understand the predictive performance of these models in the complex domain of diamond pricing.

4 MATHEMATICAL METHODOLOGY AND EQUATIONS

4.1 Multiple Linear Regression (MLR)

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable (Hayes).

$$y_{\mathrm{i}} = \beta_0 + \beta_1 x_{\mathrm{i1}} + \beta_2 x_{\mathrm{i2}} + \ldots + \beta_{\mathrm{p}} x_{\mathrm{ip}} + \grave{o} \qquad (1)$$

4.2 Extreme Gradient Boosting (XGBoost)

The objective function J, encompassing both training loss and regularization is a pivotal component in various tasks like regression, classification, and ranking, wherein the paramount goal is to optimize the parameters (denoted as θ) for optimal alignment with training data (X_i) and corresponding labels (Y_i). Training the model entails the meticulous definition of this objective function, serving as a yardstick to gauge the model's efficacy in fitting the training data (T. Chen, 2014).

4.3 K-Nearest Neighbors (KNN)

It is a non-parametric and instance-based method that makes predictions based on the similarity of input data points (Larose & Larose, 2014).

4.4 Random Foreast (RFs)

A random forest is an ensemble learning technique in machine learning, specifically designed for both classification and regression tasks. It operates by constructing a multitude of decision trees during the training phase and outputs the average (for regression problems) or the mode (for classification problems) of the individual trees' predictions (Rigatti, 2017).

4.5 Means Squared Error (MSE)

Average squared difference between the estimated values and the actual value (Error, 2010):

$$MSE = \frac{\Sigma\left(y_i - p_i\right)^2}{n} \qquad (2)$$

where y_i is the ith observed value, p_i is the corresponding predicted value for y_i, and n is the number of observations (Error, 2010).

4.6 Root Mean Squared Error (RMSE)

RMSE calculates the average difference between observed outcomes and the model's predictions. Lower RMSE values indicate superior predictive performance. f = forecasts(expected values or unknown results), o = observed values, (known results), (zf i −zoi) 2 = differences, squared and N = sample size (Chai & Draxler, 2014).

$$RMSE_{fo} = \left[\sum_{i=1}^{N}\left(z_{fi} - z_{oi}\right)^2 / N\right]^{\frac{1}{2}} \qquad (3)$$

4.7 Mean Absolute Error (Mae)

MAE is a robust and intuitive metric for model accuracy, calculating the total error by summing magnitudes and dividing by n. Lower MAE values indicate superior model performance and enhanced prediction accuracy (Hodson, 2022).

Average of all absolute errors between paired observations expressing the same phenomenon (Hodson, 2022):

$$\mathrm{MAE} = \left[n^{-1}\sum_{i=1}^{n}\left|e_i\right|\right] \qquad (4)$$

4.8 R Squared (R^2)

R-squared (R^2), is a statistical measure that assesses the proportion of variability in the dependent variable (target) explained by the independent variables (features) in a regression model. It provides insights into the goodness of fit of the model to the observed data (Kigo et al, 2023).

4.9 Adjusted R^2

R^2 that adjusts for the number of predictors in a regression model. While R-squared measures the proportion of variability in the dependent variable explained by the independent variables, the adjusted R-squared penalizes models with unnecessary variables, providing a more reliable measure of goodness of fit (Kigo et al, 2023).

4.10 the Results from the Models

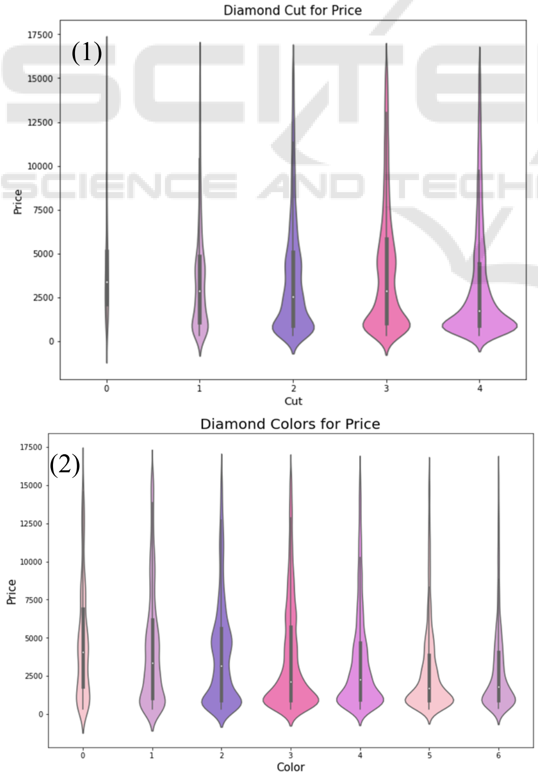
LinearRegression: -1333.321776;
DecisionTree: -759.075442;
RandomForest: -550.501632;
KNeighbors: -828.945611;
XGBRegressor: -558.378097.

5 SIMULATED DATA ANALYSIS

As Figure 2(1), "Ideal" diamond cuts are the most number while the "Fair" is the least—more diamonds of all of such cuts for the lower price category.

As Figure 2(2), "J" color diamonds which are the worst are most rare however, "H" and "G" are in number even though they're of inferior quality as well.

As the Figure 2(3), Diamonds of "IF" clarity which is best as well as "I1" which is worst are very rare and the rest are mostly in-between clarities.



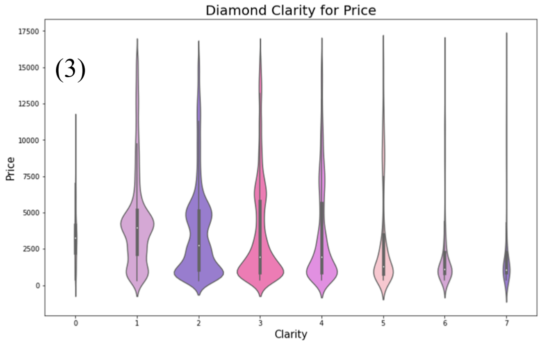
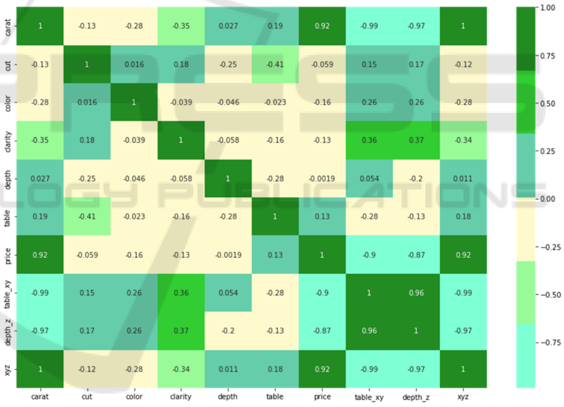


Figure 2: List of categorical variables. (1. Diamond Cut for Price; 2. Diamond Colors for Price; 3. Diamond Clarity for Price), (Picture credit: Original).

In Figure 3, "x", "y" and "z" show a high correlation to the target "price", "depth", "cut" and "clarify" show a low correlation (<|0.1|), dropping though due to presence of only a few selected features.

Cross Validation: using the negative root mean squared error: The higher the score the better the model.



	carat	cut	color	clarity	depth	table	price	table_xy	depth_z	xyz
carat	1	-0.13	-0.28	-0.35	0.027	0.19	0.92	-0.99	-0.97	1
cut	-0.13	1	0.016	0.18	-0.25	-0.41	-0.059	0.15	0.17	-0.12
color	-0.28	0.016	1	-0.039	-0.046	-0.023	-0.16	0.26	0.26	-0.28
clarity	-0.35	0.18	-0.039	1	-0.058	-0.16	-0.13	0.36	0.37	-0.34
depth	0.027	-0.25	-0.046	-0.058	1	-0.28	-0.0019	0.054	-0.2	0.011
table	0.19	-0.41	-0.023	-0.16	-0.28	1	0.13	-0.28	-0.13	0.18
price	0.92	-0.059	-0.16	-0.13	-0.0019	0.13	1	-0.9	-0.87	0.92
table_xy	-0.99	0.15	0.26	0.36	0.054	-0.28	-0.9	1	0.96	-0.99
depth_z	-0.97	0.17	0.26	0.37	-0.2	-0.13	-0.87	0.96	1	-0.97
xyz	1	-0.12	-0.28	-0.34	0.011	0.18	0.92	-0.99	-0.97	1

Figure 3: Correlation matrix (Picture credit: Original).

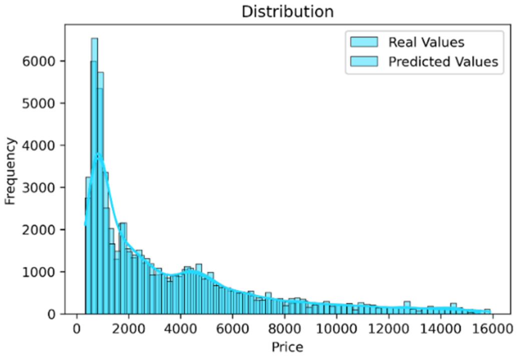


Figure 4: Comparison of Actual Price and Predicted Price (Picture credit: Original).

Figure 4 illustrates that XGBoost yielded optimal results in predicting diamond prices through regression, establishing itself as the preeminent tool for proficient diamond price prediction.

6 RESULTS

The model comparison results are shown in Table 2.

Table 2. Comparison of models.

	R^2
XGB Pipeline	0.9806573031144777
Random Forest	0.9805528549055725
Linear Regression	0.9176167819635159
KNN Regression	0.9256053449796051
Decision Tree Regressor	0.9787072279719632

7 CONCLUSION

The findings highlight that the eXtreme Gradient Boosting (XGBoost) algorithm emerges as the top performer, excelling not only in diamond classification based on cut but also in regression-based diamond price prediction. The robustness and accuracy demonstrated by XGBoost position it as the preferred tool for predicting diamond prices, showcasing its efficacy across different methodologies.

Establish an Online Interactive Space for Transparent Pricing:

The study proposes the creation of an online interactive platform, where diamond attributes can be inputted. The model would then generate the most accurate cut category, a key determinant of diamond prices. By providing justifiable price estimates based on transparent and data-driven criteria, such a platform could contribute to eliminating information asymmetry in the diamond market. This recommendation aims to address issues of price obfuscation by various diamond retailers, promoting transparency and empowering consumers with accurate pricing information.

In conclusion, this study not only contributes insights into effective algorithms for diamond price prediction but also provides practical recommendations for further research and policy considerations to enhance transparency in the diamond market. The prominence of XGBoost in both classification and regression scenarios suggests its potential applicability and reliability in real-world diamond pricing scenarios.

REFERENCES

Mirzaei, Diamond Price Prediction https://www.kaggle.com/code/amirhosseinmirzaie/diamond-price-prediction/notebook

Karnika Kapoor, Diamond Price Prediction, https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction

Hayes, Multiple Linear Regression (MLR) Definition, Formula, and Example, https://www.investopedia.com/terms/m/mlr.asp

T. Chen, University of Washington Computer Science, 22(115), 14-40 (2014)

D. T. Larose, C. D. Larose, k-nearest neighbor algorithm (2014).

S. J. Rigatti, Journal of Insurance Medicine, 47(1), 31-39 (2017)

M. S. Error, MA: Springer US, 653-653 (2010)

T. Chai, R. R. Draxler, Geoscientific model development, 7(3), 1247-1250 (2014)

T. O. Hodson, Geoscientific Model Development, 15(14), 5481-5487 (2022)

S. N. Kigo, E. O. Omondi, B. O. Omolo, Scientific Reports, 13(1), 17315 (2023)