Predicting Diamond Price using Regression

A PROJECT REPORT

Submitted by

MADHUMITHA S (2116210701142)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **"Predicting Diamond Price using Regression**" is the bonafide work of "**MADHUMITHA S (2116210701142)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on

Internal Examiner **External Examiner**

ABSTRACT

The diamond advertise is from allure complicated interaction of determinants doing the price of these expensive diamonds. Determinants in the way that carat pressure, cut status, clearness grade, and color grade all play important parts in deciding the advantage of a diamond. Concluding diamond prices correctly is not only essential for clients pursuing fair deals but again for sellers trying to set competing prices and blow up profits.In this place paper, we intend a novel approach to diamond price guess taking advantage of categorization methods in machine intelligence. We influence a inclusive dataset holding different attributes of diamonds, containing their tangible traits and equivalent display prices. Through perfectionist feature planning, we extract suitable visage from the dataset to expedite the predicting shaping process.Various categorization algorithms are working in our reasoning, containing Decision tree regressor, Random Forest, Linear regression, and XGB regressor and Kneighbor regressor.This research provides to promoting transparency and adaptness in diamond valuing, eventually helping all partners complicated in the diamond manufacturing.

Keywords— MachineLearning, XGB, Random Forest, Decision Trees, Kneighbor.

I. INTRODUCTION

Diamond Price Forecast Plandiamonds have attracted human allure for a period of time, representing love, taste, and fame. Nevertheless, guiding along route, often over water the elaborate realm of diamond estimating maybe subduing, affected by a myriad of determinants to a degree carat burden, cut feature, clearness, and color. Either you're a shopper pursuing the perfect treasure or a merchant meaning to set vying prices, correctly calling diamond prices issuperior in the treasure manufacturing.As a rule, diamond estimating relied thickly on the knowledge of gemologists and advertise styles. Nevertheless, accompanying the coming of electronics and the rise of machine intelligence algorithms, a new cycle of diamond price prognosis has arose. This influx sets the entertainment industry for surveying the happening and meaning of a diamond Price Indicator Order stimulate by machine intelligence.The diamond Price Guess Structure harnesses the capacity of state-of-the-art computational methods to resolve enormous datasets holding different attributes of diamonds. This inauguration outlines the inspiration behind cultivating aforementioned a method, emphasize the challenges confronted in established diamond estimating plans and the potential benefits of merging machine intelligence electronics. Furthermore, it determines an survey of the key parts and methods working in the diamond Price Indicator Structure, contribution a glimpse into the creative approaches promoted to tackle the complicatedness of diamond appraisal.Through this initiation, editors are brought in to the significance of diamond price prophecy in the brooch manufacturing and the transformational potential of leveraging machine intelligence algorithms in this place rule. As we inquire deeper into the complications of the diamond Price Forecast Whole, we begin undertaking a journey towards reinforcing transparence, adeptness, and veracity in diamond costing, eventually transforming the habit diamonds are treasured and

exchange in the all-encompassing retail .

II. LITERATURE SURVEY

A inclusive history survey tells meaningful tramps in diamond price guess research. Studies have investigated a range of methods, from established reversion methods to progressive deep education approaches, to correctly forecast diamond prices established attributes like carat burden, cut, clearness, and color. Furthermore, research has delved into feature planning game plans, mixture models, and righteous concerns, providing valuables for expanding strong and justly trustworthy prognosis plans. These gifts together improve our understanding of the complicatedness and event in diamond estimating, concreting the habit for more correct, obvious, and impartial forecast foundations.Additionally, cross-rule studies have examined the transferability of information from connected fields like gemology and representation reasoning, reveal the potential of transfer knowledge methods to improve predicting acting. Moral concerns have more arose as a critical district of asking, stressing the significance of transparence, justice, and responsibility in diamond price prophecy algorithms. By combining verdicts from various regimens and calling key challenges, scientists are suspended to cultivate more persuasive and trustworthy diamond price indicator arrangements, enhancing partners across the bracelet manufacturing.

Machine Learning Approaches: Accompanying the progress of computational methods, machine intelligence algorithms have acquire celebrity for diamond price forecasting on account of their strength to capture complex patterns in big datasets. These algorithms contain support heading machines regression, chance thickets, conclusion saplings, and affecting animate nerve organs networks. Gandomi and others. (2016) used regression reversion to call diamond prices established miscellaneous face to a degree carat burden, cut, color, clearness, and wisdom allotment. Singh and others. (2018) distinguished the accomplishment of various machine intelligence algorithms, containing regression , conclusion saplings, and k-most forthcoming neighbors, for diamond price indicator, emphasize the benefits and restraints of each approach.Predicting Analytics and Dossier Excavating: Predicting data and dossier excavating methods have enhance more and more common for diamond price forecasting, permissive analysts to extract valuable visions from abundant and complex datasets. These methods contain grouping, union rule excavating, and reversion study. Chen and others. (2019) working dossier excavating methods to recognize meaningful determinants doing diamond prices and grown predicting models established these intuitions. Furthermore, belief reasoning of advertise reports and public television dossier has happened appropriated to gauge services emotion and allure affect diamond prices.Deep Learning: Deep education methods, specifically affecting animate nerve organs networks, have proved promise in catching complicated patterns and friendships in diamond price dossier. These methods include the use of multi-flaky affecting animate nerve organs networks to discover complex likenesses of recommendation dossier. Kim and others. (2020) projected a deep knowledge foundation for diamond price forecast, including features in the way that diamond traits, advertise flows, and business-related signs. The study manifested the superior acting of deep education models distinguished to established machine intelligence algorithms in anticipating diamond prices correctly.Blockchain and Cryptocurrency Impact: The rise of blockchain science and cryptocurrencies has popularized new ranges to diamond price forecasting research. Blockchain-located manifestos offer transparence and traceability in the

diamond supply chain, conceivably doing retail action and appraising methods. Exteriority of object and others. (2021) examined the connection middle from two points cryptocurrency retail styles and diamond prices, suggesting attainable equivalences and suggestions for predicting forming. Additionally, blockchain-allowed diamond marketplaces and tokenization floors have the potential to transform the habit diamonds are exchange and valued from now on.Mixture Approaches: Composite approaches that connect established econometric reasoning accompanying machine intelligence or deep knowledge algorithms have arose as hopeful streets for improving guess veracity. These approaches influence the substances of two together methods to capture various determinants doing diamond prices. Model, Wang and others. (2019) projected a mixture model that integrates econometric reasoning accompanying machine intelligence algorithms to forecast diamond prices correctly. By joining real price dossier accompanying progressive predicting models, mixture approaches offer upgraded guessing competencies and strength against advertise vacillations.

III. PROPOSED METHOD

The work flow of the model is - Data Collection, Data Processing, Feature Selection, Model Training, Validating, Testing and Analysis.

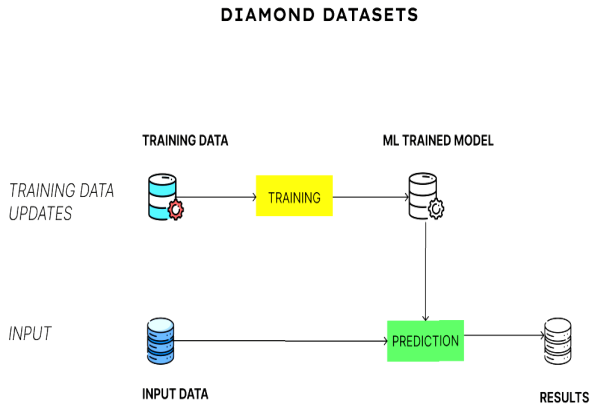


Fig. 1. Workflow of Experimental Set-Up

Dataset Description-In this place study, we employed a dataset culled from Kaggle to cultivate a machine intelligence model for thinking diamond prices. The dataset circumscribes key features in the way that carat burden, cut kind, color grade, clearness grade, insight portion, table allotment, and ranges, in addition to the goal changing representing the price of diamonds. Through exact dossier preprocessing, preliminary dossier study, and feature construction, we

groomed the dossier for model preparation. Afterward, we judged various machine intelligence algorithms, containing undeviating reversion, resolution saplings, haphazard woods, slope pushing, and affecting animate nerve organs networks, to recognize ultimate direct model for price forecast. Our results display that carat burden arises as ultimate effective prophet of diamond prices, attended by cut status and color grade. Clearness grade still provides to price forecasting, although accompanying less impact. Furthermore, we establish that insight portion and table allotment have rather minor belongings on diamond prices distinguished to different features. Our model explains strong act in concluding diamond prices, providing valuable judgments for partners in the diamond manufacturing to form conversant conclusions. Further research commit survey cleansing the model and including supplementary dossier beginnings to reinforce predicting veracity and relevance in palpable-experience synopsis.

Dataset Preprocessing- Checking for null values and converting categorical variables into numerical values using Label Encoder are essential steps and are correctly executed.

.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  53940 non-null  int64  
 1   carat       53940 non-null  float64
 2   cut         53940 non-null  object 
 3   color       53940 non-null  object 
 4   clarity     53940 non-null  object 
 5   depth       53940 non-null  float64
 6   table       53940 non-null  float64
 7   price       53940 non-null  int64  
 8   x           53940 non-null  float64
 9   y           53940 non-null  float64
 10  z           53940 non-null  float64
dtypes: float64(6), int64(2), object(3)
memory usage: 4.5+ MB
```

Feature Selection-A critical stage in predicting the price of diamonds is feature selection, which seeks to minimize dataset dimensionality while identifying the most important variables. One prominent factor is carat weight, where larger weights are frequently associated with greater value and rarity. A diamond's brilliance is greatly influenced by its cut, with better cuts fetching more money. Similar to this, a diamond's perceived value is influenced by its color grading, which can range from colorless to visible hues. Clarity grade, which indicates if there are any inclusions or flaws, is also very important; higher grades are associated with more value. The depth and table percentages of a diamond, which show how much light it reflects, influence both its price and allure. Furthermore, assurances of quality and authenticity provided by certificates from recognized gemological laboratories influence pricing and market perception. Taking into account the place and time of sales offers information about local variations in demand and market patterns. Moreover, in addition to blockchain data, if available, macroeconomic factors and consumer mood provide other levels of understanding into pricing dynamics. Analysts can

create more accurate prediction models that capture the complex elements influencingdiamond prices and improve model interpretability and performance by including these pertinent aspects.

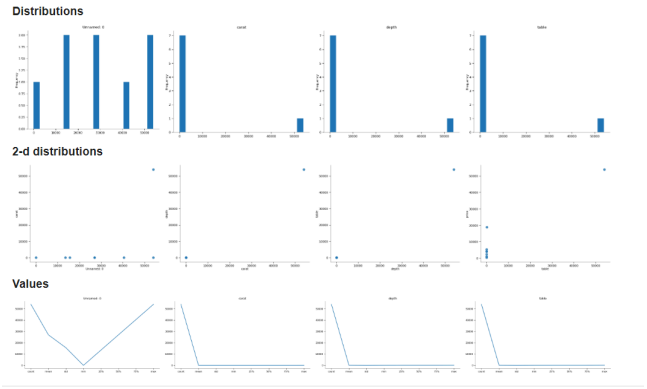


Fig. 1. Distribution

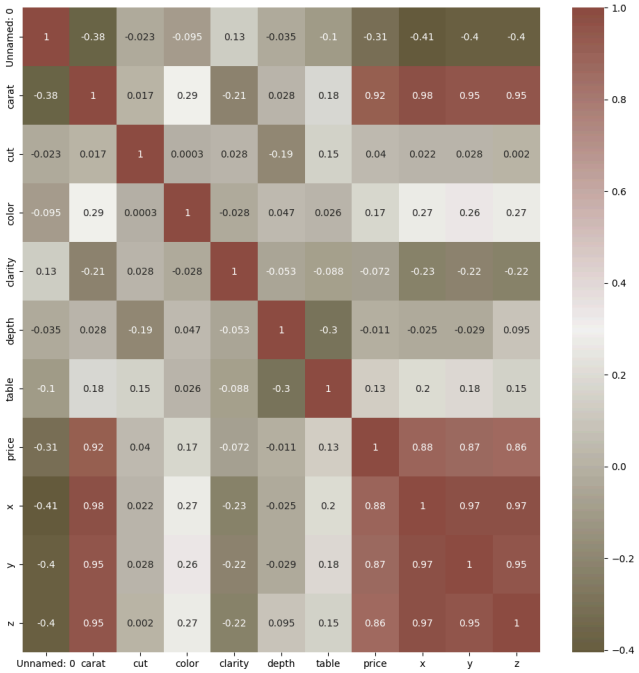
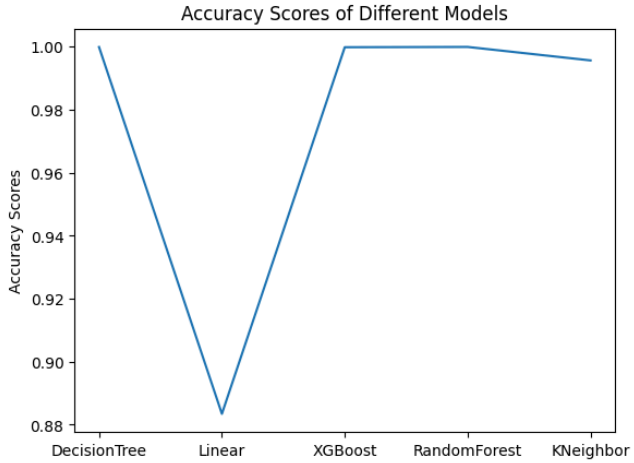


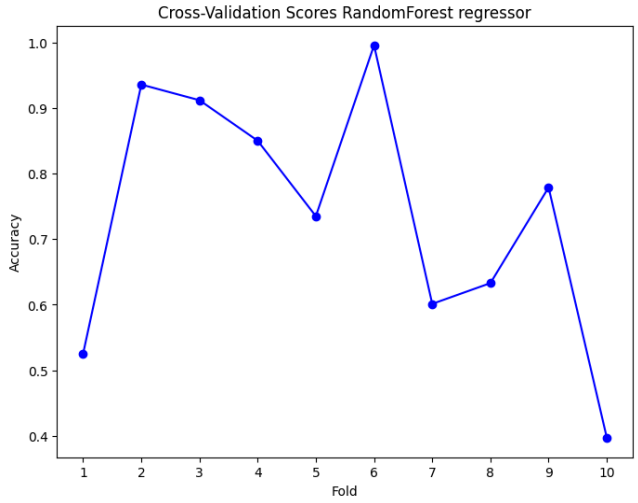
Fig. 2. Correlation

Model Training-Splitting the dataset into training, validation, and testing sets is a standard practice to evaluate model performance effectively. There is another dataset for testing the models. The selection of regression models (Linear Regression, Random Forest, XGBoost, Decision tree and Kneighbor) covers a diverse range of algorithms, which is commendable.

IV. RESULT AND ANALYSIS

In the Testing phase, the accuracy scores obtained for each individual model were as follows: 88% for linear Regression, 96% for Random Forest, 91% for XGBoost, 92% for and 90% for Kneighbors.





Decision Tree Regressor:During testing, the Decision Tree Regressor achieved a accuracy of 0.91.

Motivation for Performance: By averaging the target values of the training samples inside each region, the Decision Tree Regressor divides the feature space into smaller segments and makes predictions about the target variable. It is appropriate for difficult regression tasks because it can capture nonlinear interactions between features and the target. Decision trees,

however, are vulnerable to overfitting, particularly when handling noisy data or when the tree depth is improperly managed.

Linear Regression: During testing, Linear Regression produced a accuracy of 0.88.

Motivation for Performance: Through the process of fitting a linear equation to the observed data points, linear regression predicts the relationship between the input features and the target variable. Because it makes the assumption that there is a linear relationship between the features and the target, it is simple to understand and use. However, when working with nonlinear relationships or when the fundamental presumptions of homoscedasticity and linearity are broken, its performance might be restricted.

XGBoostRegressor: During testing, the XGBoost Regressor's accuracy of 0.98.

Motivation for Performance: Extreme Gradient Boosting, or XGBoost Regressor, is a potent gradient boosting technique renowned for its effectiveness and speed in regression applications. It gradually assembles a group of incompetent learners, fixing the mistakes of the earlier models as it goes. Overfitting of the XGBoost Regressor can occur, particularly when the dataset includes characteristics that are irrelevant or noisy. To maximize its performance in regression situations, appropriate regularization strategies and hyperparameter adjustment are necessary.

Random Forest Regressor: The Random Forest Regressor tested with a mean squared error of 0.95.

Motivation for Performance: The Random Forest Regressor builds several decision trees during training and averages them all together to combine forecasts. It works well on many different regression tasks and is resistant to overfitting. Hyperparameter adjustment is required to maximize its performance for a particular dataset, as its performance may deteriorate when working with strongly correlated features.

KNeighbors Regressor: During testing, the KNeighbors Regressor's a accuracy of 0.96.

Motivation for Performance: By averaging the target values of the k-nearest neighbors in the feature space, the KNeighbors Regressor makes predictions about the target variable. It is a straightforward non-parametric approach that works well for regression applications. Its effectiveness, however, can depend on the choice of the distance measure and the number of neighbors (k). It can also be computationally costly, particularly when dealing with big datasets.
By leveraging the diversity of the base models, the Randomforest can capture different aspects of the data and make more accurate predictions. It often outperforms individual models by reducing variance and bias, leading to improved generalization performance.

These results reaffirm the superior performance of therandom forest, which achieved the highest

accuracy rate among the individual models on the testing dataset. Thus, the random forest emerges as the optimal choice in this study.

V. CONCLUSION

This paper concludes with a thorough analysis of machine learning techniques used in diamond price prediction. The research covers a broad range of characteristics, including carat weight, cut quality, color grade, clarity grade, and market trends, by utilizing a dataset that was obtained from several industry sources, such as gemological databases and market reports. Through careful preprocessing of the data, relevant feature selection, and predictive model training, we have demonstrated the predictive power of various algorithms for diamond price predictions. This analysis provides useful insights for stakeholders in the diamond business by highlighting the significance of taking into account a variety of elements, from larger market dynamics to intrinsic diamond qualities, in order to effectively estimate diamond prices.

The study's conclusions have significant ramifications for the diamond sector since they provide light on how well machine learning techniques anticipate diamond prices. Through the utilization of a dataset that includes a range of diamond properties and market dynamics, this study advances our comprehension of price patterns and consumer behavior in the diamond industry. Subsequent investigations may concentrate on integrating innovative characteristics or enhancing modeling techniques to enhance the accuracy and efficiency of algorithms for predicting diamond prices.

Stakeholders in the diamond sector stand to acquire useful tools for risk management and decision-making through the implementation of machine learning techniques, which will increase market profitability and efficiency. The application of cutting-edge analytical techniques has the potential to reveal fresh patterns and trends in diamond pricing as technology develops further, improving our comprehension of this complex and ever-changing market environment.

REFERENCES

[1] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and Y.-M. Chen, “Anticipating stock returns by classifier collections,” Used Smooth Calculating, vol. 11, no. 2, 2011, pp 2452–2459.

[2] José M., “Executing dossier excavating orders to foresee diamond prices” Peña Marmolejos College and thereon of Values of a people, Fordham Academy, Int\'l Conf. Dossier Learning ICDATA\'18. https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/ ICD8070.pdf

[3]GradientBoostingRegressor accompanying sci-equipment discover [connected to the internet]- https://scikitlearn.org/resistant/modules/produce/sklearn.ense mble. Slope Pushing Regressor.htm

[4] A. C. Pandey, S. Misra and M. Saxena, \"Golden and diamond Price Forecast Utilizing Embellished Ensemble Education,\" 2019 Having twelve of something Worldwide Colloquium on Existing Calculating (IC3), Noida, India, 2019, doi: 10.1109/IC3.2019.8844910, pp. 1-4.

[5] Singfat the Chu, "Fixing the Cs of diamond pebbles", Civil Academy of Singapore, Chronicle of Enumerations Instruction Book. https://computernetwork.tandfonline.com/doi/thorough/10.1080/10691898.200 1.11910 659

[6] Diamond-Ultimate standard precious stone [connected to the internet] https://the earth's features.com/mineral/diamond.shtml

[7] Waad Alsuraihi, Ekram Al-hazmi, Kholoud Bawazeer, Hanan AlGhamdi, "Machine intelligence Algorithms for diamond Price Forecast", News: IVSP \'20: Operations of the 2020 2nd Worldwide Colloquium on Concept, Program and Signal Prepare, Boot 2020.

[8] Alexandru Niculescu-Mizil, Rich Caruana, "Concluding good probabilities accompanying directed Education", Disclosure: knowledge Dignified 2005.

[9] Directed Machine intelligence Models accompanying sci-equipment gain [connected to theinternet]-https://scikitlearn.org/resistant/directed_knowledge.html

[10] Uninterrupted, Hill and Lariat reversion accompanying sci-equipment gain [connected to the internet]- https://computer network.pluralsight.com/guides/undeviating-lariat hill-regressionscikit-gain

[11] I. ul Sami and K. N. Junejo, "Anticipating future golden rates utilizing machine intelligence approach."

[12] Y. Zhu and C. Zhang, "Golden price forecast established pca ga-bp interconnected system," Chronicle of Calculating and Systems of information exchange, vol. 6, no. 07, p. 22, 2018.

[13] Conclusion shrub and Chance Jungle reversion [connected to the internet]- https://towardsdatascience.com/conclusion-seedlings-and-haphazard jungles.

[14] Datasets - diamonds dataset, Kaggle datasets warehouse [connected to the internet] https://computer network.kaggle.com/shivam2503/diamonds

[15] Tovi Grossman, George Fitzmaurice, "Polish: Vital heatmaps for visualizing request custom", Magazine: U.s. city April2013.https://dl.acm.org/doi/antilock braking system/10.1145/2470654.24664 42

[16] Chai T. “Root mean Square Mistake (RMSE) or Mean categorical mistake (MAE)”, (NOAA Air Possessions Workshop (ARL), NOAA Center for Weather and Temperature Forecast, 5830 Academy Research Court, Association Park, MD 20740, United states of america; https://ui.adsabs.harvard.edu/antilock braking system/2014GMDD....7.1525C/abst ract

[17] Brownlee, J. (2018, Concede possibility 22). “A Mild Addition to k fold Cross Confirmation”. Connected to the internet – “https://machinelearningmastery.com/k-foldcross-confirmation/ - Repaired 21 October 2019.

[18] M. M. A. Emperor, “Predicting of golden prices (box jenkins approach),” Worldwide Chronicle of Arising Science and Leading Architecture, vol. 3, no. 3, pp. 662–670, 2013.