Optimizing Diamond Valuation: A Machine Learning Approach to Price Prediction and What-if Analysis

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 nuanced process that blends art with science. This project aims to augment the scientific aspect of diamond valuation by constructing a data-driven regression model using a comprehensive dataset of over 200,000 diamond records. The intent is to meticulously analyze the relationships between a diamond's physical attributes and its market price, with a focus on the classical determinants such as cut, colour, clarity, and carat weight, as well as additional features that may influence valuation. A multi-faceted machine learning approach will be employed, encompassing data cleaning, feature extraction, and the application of various regression techniques, including linear models, tree-based methods, and ensemble methods like random forest and gradient boosting. The project seeks to explore the predictive power of these models, calibrating them to handle the complexity and variability inherent in diamond pricing. Alongside predictive modelling, the project will undertake a what-if analysis to simulate the effect of changes in diamond attributes on price, offering a strategic tool for scenario planning. Furthermore, we will attempt to harness optimization algorithms to identify the set of features that align with the highest valuations, aiming to provide actionable insights for maximizing returns in the diamond market. By leveraging rigorous validation techniques and acknowledging the potential constraints and pitfalls of predictive modelling, this project endeavours to produce a robust predictive tool. The ultimate goal is to offer a model that not only enhances the accuracy of diamond valuations but also provides a platform for better understanding the market dynamics at play. This endeavour acknowledges the complexity of the task at hand and the need for careful interpretation of the model's findings, aspiring to contribute meaningfully to the domain of quantitative gemology.

Keywords

Diamond Price Prediction, Machine Learning, Regression Analysis, What-if Analysis, Price Optimization, 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. Moreover, the project will explore what-if scenarios to assess the sensitivity of diamond prices to changes in their attributes. This will provide valuable insights for stakeholders, from individual consumers to large-scale retailers, by simulating the effects of market dynamics on pricing. In tandem, we will also investigate price optimization to identify the most valuable feature combinations, thereby equipping the industry with knowledge to maximize profitability and inform strategic inventory decisions. 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 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 towards 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. What-if analyses will be systematically conducted to simulate the financial impact of changes in diamond attributes, aiming to provide a robust tool for price prediction under various market conditions. Additionally, we will integrate optimization methods to identify the ideal combination of diamond features that could maximize value, offering a dual benefit of accurate appraisal and strategic market positioning. 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.

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 R. Gajjar, "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