Diamond Price Prediction and Classification using Machine Learning Algorithms

Rajat Singh
Department of MCA,
School of Information Technology and Engineering
Vellore Institute of Technology
rajat.singh2022@vitstudent.ac.in

Hrishikesh S G
Department of MCA,
School of Information Technology and Engineering
Vellore Institute of Technology
gaikwad.hrishikesh2022@vitstudent.ac.in

Sambit Basu
Department of MCA,
School of Information Technology and Engineering
Vellore Institute of Technology
sambit.basu2022@vitstudent.ac.in

Abstract* - Diamonds are one of the most precious and valuable commodities, and their prices are influenced by several factors, including carat, cut, clarity, and color. With the help of machine learning algorithms, it is possible to predict diamond prices more accurately [2]. In this study, classification algorithms, like XgBoost, AdaBoost, Decision Tree, Gradient Boosting, Random Forest, Logistic Regression, Naïve Bayes were utilized to classify diamond prices. The performance metrics used include accuracy, precision, recall, and F1-score. This paper throws light on the use of machine learning in the "fluctuative" world diamond industry and provides insights into the importance of selecting the appropriate algorithm for accurate price classification. Overall, this study adds to the existing literature on the use of machine learning in the diamond industry and highlights the potential of this technology for improving the accuracy of diamond price prediction.

***Keywords* - Diamonds, Machine learning algorithms, Price prediction, Classification algorithms, Accuracy, Diamond industry Performance.**

INTRODUCTION

The price of diamonds is determined by various factors, namely color, clarity, cut and carat [4]. The process of diamond price determination might be complex and time-consuming, requiring a thorough understanding of the diamond grading system and market trends [1]. ML algorithms have been increasingly used in the diamond industry to automate and improve the process of diamond price classification. ML algorithms by nature are good at analyzing large data and identify patterns that are not apparent to human experts [5]. This can help diamond traders and appraisers make more accurate and informed decisions about the value of a diamond.

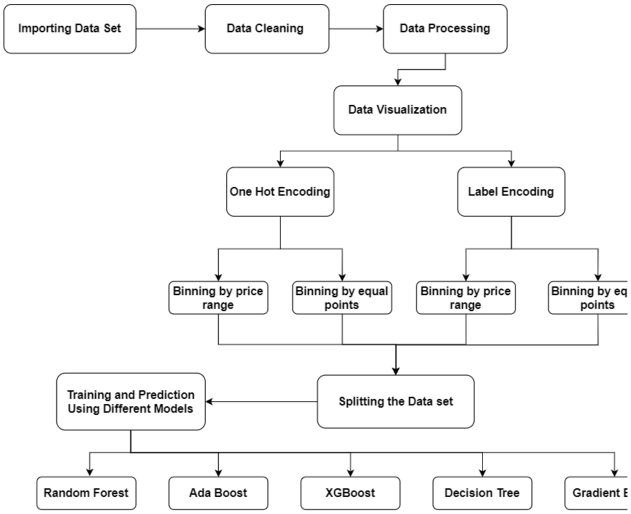
This research paper focuses on the use of machine learning algorithms for diamond price classification. Specifically, we explore the performance of four popular algorithms: XgBoost, AdaBoost, Random Forest, and Gradient Boosting. These algorithms were trained on a dataset of over 10,000 diamonds with known prices and various features such as carat weight, color, clarity, and cut.

PREVIOUS WORK AND THEIR SHORTCOMMINGS

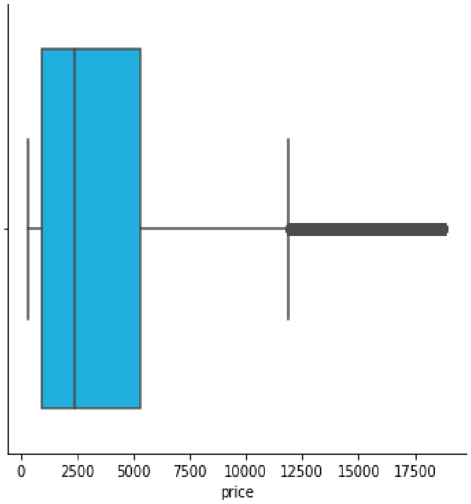
Diamond price market is one of the most fluctuating markets in the world. The price of a diamond depends on various factors such as cut, color, clarity, etc. [4] In our literature review, we found all the researchers have used the regression method to predict the diamond price based on the mentioned attributes. Now, let's look at the nature of the diamond market. If the prices fluctuate too much because of certain reasons that are not controlled by the stakeholders or the market in general, the predicted price of the regressive model would largely differ from the actual price (new price after fluctuation). Secondly, a single value predicted price doesn't give a complete overview of the diamond to the stakeholders (buyers, sellers) about the commercial value of the diamond. Thus, our proposed method of prediction is classification wherein we predict the class of diamond to which it belongs. These classes are based on prices. Therefore, the output of our model shall give the stakeholders a price range in which a specific diamond can be bought and sold.

Also, this range-based pricing technique safeguards the stakeholders from the confusion caused due to the unstable nature of the market. We reviewed many such papers, all of them use regression as the technique of prediction. We propose classification as the technique keeping in mind the above reasons. In addition, we have also used hyperparameter tuning to enhance the algorithms' performance.

ARCHITECTURE DAIGRAM



[Fig. 3]



[Fig. 4]

RESULTS AND DISCUSSION

During the Exploratory data analysis of the dataset, we found that most of the diamonds listed are in the price range of 800 to 5000 dollars, as shown in Fig. 4, whereas the entire price class for the data set ranged from 100 to 18000 dollars. So, there is a heavy portion of the data belonging to a specific range. This finding helped identify that the binning of data based on equal price range and binning of data based on equal number of points(frequency) in a bin shall give us 2 different results. As mentioned in the flowchart (Fig. 3) above, One Hot Encoding [6] has been used on the dataset and then performed binning based on equal price range and then by equal number of data points. Similarly, we Label encoded the data and then classified(binned) them by equal price ranges or again by equal number of data points. As a result, 4 different combinations of data were found that were fed into our machine learning algorithms for training. Below is the table that has best performing algorithms and their accuracy.

Random forest gives accuracy of 87.425% and XgBoost classifier gives 87.073%

Various classification algorithms:

Algorithms	**Best Score**
KNN	79.274%
Gaussian Naive Bayes	70.919%
Decision Tree Classifier	84.588%
XGB Classifier	87.073%
Random Forest Classifier	87.425%
Gradient Boosting Classifier	84.041%
AdaBoost Classifier	60.571%

Support Vector Machine	69.389%
Logistic Regression	70.632%

TABLE I.

The hyperparameter findings are not satisfactory as shown in the below table. It showed an improvement, but that improvement is not worth the time taken to train the hyperparameters. So, in the final model building, the hyperparameter tuned models shall be discarded. Reasons for these unsatisfactory results may be the varied nature of the dataset. Hyperparameter tuning works well for data which has linear correlation throughout, so in such datasets, one changed hyperparameter can work well for all tuples in that data. For the current dataset, a single parameter adjustment could not be found as it is simply not possible, i.e., there is no such set of same parameters that can accommodate for all the trend changes in the dataset. Therefore, the hyperparameter tuning extension is not worth doing.

Various classification algorithms with hyperparameter tuning:

Algorithms	**Best Score**
Decision Tree Classifier	84.618%
Random Forest Classifier	72.952%
XgBoost Classifier	46.726%

TABLE II.

CONCLUSION

Random Forest Classifier is the best performing model on this dataset, with accuracy: 87.425%. As mentioned above, regression analysis could have achieved higher MSE, but in the real world, a range of price would be more beneficial to the stakeholders rather than an exact price. Tuning the hyperparameters in some models is not improving the performance as such, this is a major finding. This model now can be used in APIs to predict the price range of a newly found diamond in the current market. There are places of improvement in this paper, like in the hyperparameter tuning. The hyperparameters can be better tuned to achieve a higher result. Also, we are training and testing the dataset only through Hold-out cross validation method. The other types of cross validation have not been explored in this paper.

REFERENCES

[1] : Sharma, G., Tripathi, V., Mahajan, M., & Srivastava, A. K. (2021, January). Comparative analysis of supervised models for diamond price prediction. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 1019-1022). IEEE.

[2] : Alsuraihi, W., Al-hazmi, E., Bawazeer, K., & AlGhamdi, H. (2020, March). Machine learning algorithms for diamond price prediction. In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing (pp. 150-154).

[3] : Chaijunla, T., & Taninpong, P. Comparative Study of Predicting Diamond Ring Prices in Online Retail Shop.

[4] : Basysyar, F. M., & Dwilestari, G. COMPARISON OF MACHINE LEARNING ALGORITHMS FOR PREDICTING DIAMOND PRICES BASED ON EXPLORATORY DATA ANALYSIS.

[5] : Mamonov, S., & Triantoro, T. (2018). Subjectivity of diamond prices in online retail: insights from a data mining study. Journal of theoretical and applied electronic commerce research, 13(2), 15-28.

[6] : Yong, Z., Jianyang, L., Hui, L., & Xuehui, G. (2018, June). Fatigue driving detection with modified ada-boost and fuzzy algorithm. In 2018 Chinese Control And Decision Conference (CCDC) (pp. 5971-5974). IEEE.

[7] : Dutta, J., Kim, Y. W., & Dominic, D. (2020, November). Comparison of gradient boosting and extreme boosting ensemble methods for webpage classification. In 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 77-82). IEEE.

[8] : Chen, H., Ai, H., Yang, Z., Yang, W., Ye, Z., & Dong, D. (2020, September). An Improved XGBoost Model Based on Spark for Credit Card Fraud Prediction. In 2020 IEEE 5th International Symposium on Smart and Wireless Systems within the Conferences on Intelligent Data Acquisition and

Advanced Computing Systems (IDAACS-SWS) (pp. 1-6). IEEE.

[9] : Bulbul, H. I., & Unsal, Ö. (2011, December). Comparison of classification techniques used in machine learning as applied on vocational guidance data. In 2011 10th International Conference on Machine Learning and Applications and Workshops (Vol. 2, pp. 298-301). IEEE.

[10] : Yue, Y., & Yang, Y. (2020, February). Improved Ada Boost Classifier for Sports Scene Detection in Videos: from Data Extraction to Image Understanding. In 2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 1-4). IEEE.