Automated Diamond Price Prediction Using Machine Learning

Abirami R and Agniswar P

SRM University AP

ABSTRACT

The price of diamonds has been extremely volatile over the last century. Investing in diamonds has been extremely fruitful only for some, and for the others, it seems like a gamble. In the current paper, we present a machine learning-based method to predict the price of diamonds to prevent human error. With an accuracy of98% using Random Forest Regression, there are much lesser chances of losing the investment. The proposed machine learning-based prediction model uses Linear regression, Lasso Regression, Support Vector Regression, and Random Forest. The proposed method gives the most accurately predicted value. We have also added a feature to automate the process using Crontab, this would retrain the model before the diamond market opens to the most accurate value.

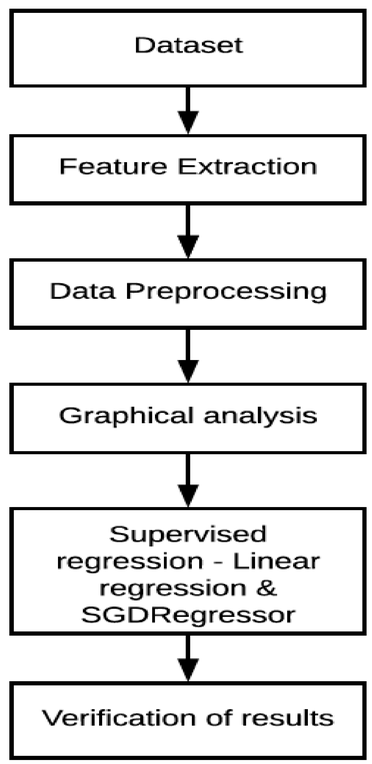
INTRODUCTION

Diamonds are one of the most valued gems in theworld. It is also one of the most expensive gems andhence has an extremely volatile price. The value of diamonds depends upon their structure, cut, inclusions (impurity), carats, and many other features. The usesof diamonds are many, such as in industries, as theyare effective in cutting, polishing, and drilling. Since diamonds are extremely valuable, they havebeen traded across different countries for centuries nowand this trade only increases with time. They are graded and certified based on the "four Cs", which are color, cut, clarity, and carat. Color, Clarity, Cut, and Carat weight. These are the only metrics that are beingused to the quality of diamonds and sets the price of thediamond. This metric allows uniform understanding forpeople across the world to buy diamonds, which allowsease of trade and value for what is purchased. Diamond prices are usually set for the day are traded in US Dollars. To better predict the price of diamonds, the Kaggle diamond dataset is used and a scatterplot of metrics such as carats, price, and the coloris used to understand the nature of their relationships. The more obvious thought is that there is a strong relationship between carat and price, but it is observed that this trend does not seem to hold true anymore, and thus leading to higher volatility in the price of diamonds. This machine learning model analyses more than 4 features and can thus produce a more accurate result. When the price chart is plotted in a graph, it leads to various formations such as pennants, wedge, flags, double bottoms and tops. These formations often used in the currency markets,as well as many other trading markets, like the diamond market. The software used here are Jupyter Notebook (anaconda navigator), NumPy which is an array processingpackage, Pandas which is a data manipulation tool, Scikit learn which is a Machine Learning library, Matplotlib and Seaborn which are Data Visualization tools and Crontab which is a scheduling tool

METHODOLOGY

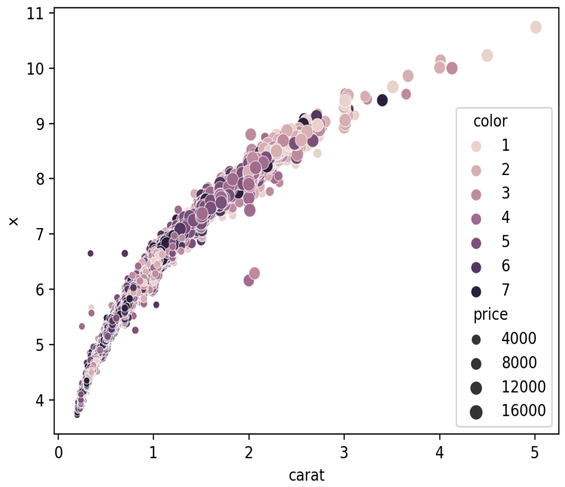
The tools used here are Jupyter lab, Matplotlib, Numpy, Seaborn and Sklearn. The interactive development platform we used Jupyter lab to write codeand visualize data. Jupyter lab is a flexible environmentto arrange and configure data suitable for machinelearning and data science projects. The main advantageis to write plugins and new components and integratethem with Crontab. Matplotlib is a python library used to visualize and animation to analyze the given data. Numpy is a python library used to work with arrays to work faster than lists. Pandas is a powerful tool that is used for data analysis. It is also a flexible manipulation tool built on Python for data cleaning, merging, reshaping, selecting, and wrangling. Seaborn is a library built on matplotlib and is used tomake statistical graphs and high-level visualization. Itbuilds on top of matplotlib and integrates closely withpandas data structures. It has inbuilt features that help visualize data clearly with minimum code. Sklearn is a machine learning library built using Python. It has various features like classification, regression, and clustering algorithms.

Crontab is used to run a regular schedule to manage a list and this is done with the help of a list of commands. Crontab, short for cron table is a job scheduler to execute tasks. In this case, we use it to retrain the model for more accurate predictions every day. @daily cron keyword is used, this will create a log file everyday and purge using cleanup-logs shell script at 08:00 every day, before the diamond .



RESULTS

The regression scatters plot in Fig. 2 explains the relationship between color and price with respect to carat and x. Color has been put into seven categories where 1 is the best and 7 is the worst color. Price and carat are linearly proportional, since if the number of carats increases then the price also increases and vice versa.



The four algorithms used here are Linear Regression, Lasso Regression, Support Vector Machine, and Random Forest. The results are tabulated below with their respective accuracy scores. From the table, we can see that Random Forest has the most accurate results since it makes use of multiple trees, with 98% accuracy. Secondly, lasso regression has an accurate result almost equal to linear regression. Support Vector Regression(SVR) has the worst result of 68% accuracy. The shows the r2 score is 90% which means the score is good and the prediction is very accurate. The mean Square error is 0.151 which means that we have a very small margin of error since 0 is no error and1 is a high margin of error. Mean Absolute Error is around $805 which is a very small error since the values of diamonds are in hundreds and thousands of dollars. Explained Variance Score is 0.905 which is almost 1, which is the best score.

Regression Model	Accuracy score
Linear Regression	0.9053314459955231
Lasso Regression	0.905374119451812
Epsilon-Support Vector Regression	0.6884566621516661
Random Forest	0.9820887109010709

CONCLUSION

The prediction of diamond price is extremely important to know while investing. It reduces dependency on people and hearsay. In this paper, a detailed study and research has been made on how to make a better prediction model to determine diamond price, based on factors other than well known 4Cs.

RECOMMENDATIONS

[1]anwala, S. (2020, May 14). Regression-based machinelearning approaches for diamond price prediction! Medium.https://medium.com/@sp7091/regression-approaches-to-predict-diamond-price-258478a485c9

[2] 4Cs of Diamond Quality by GIA — Learn about DiamondBuying — What are the Diamond 4Cs. (2019, August 22).GIA 4Cs. https://4cs.gia.edu/en-us/4cs-diamond-quality/

[3] G. Sharma, V. Tripathi, M. Mahajan and A. Kumar Srivastava,"Comparative Analysis of Supervised Models for DiamondPrice Prediction," 2021 11th International Conference onCloud Computing, Data Science Engineering (Confluence),Noida, India, 2021, pp. 1019-1022, doi: 10.1109/Conflu-ence51648.2021.9377183.

[4] 4Cs of Diamond Quality by GIA — Learn about DiamondBuying — What are the Diamond 4Cs. (2019, August 22).GIA 4Cs. https://4cs.gia.edu/en-us/4cs-diamond-quality/

[5] S. Ray, "A Quick Review of Machine Learning Algo-rithms," 2019 International Conference on Machine Learn-ing, Big Data, Cloud and Parallel Computing (COMITCon),Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMIT-Con.2019.8862451

[6] I. Kumar, K. Dogra, C. Utreja and P. Yadav, "A ComparativeStudy of Supervised Machine Learning Algorithms for StockMarket Trend Prediction," 2018 Second International Confer-ence on Inventive Communication and Computational Tech-nologies (ICICCT), 2018, pp. 1003-1007, doi: 10.1109/ICI-CCT.2018.8473214.

[7] LinearRegression (n.d.).LinearRegression.http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm[11] Lasso Regression: Simple Definition. (2020, September16). Statistics How To. https://www.statisticshowto.com/lasso-regression/

ACKNOWLEDGEMENTS

The authors would like to thank SRM University APfor this unique opportunity, to help conduct research onthis topic and present a fruitful solution.