MACHINE LEARNING ALGORITHMS FOR DIAMOND PRICE PREDICTION

T. Ramaswamy [1] , S.Sanjana[2], T Meghana [3]

Associate professor ,Student, Student

Dept. of Electronics and Communications Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana

Associate Professor, Dept. of Electronics and Communications Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana

Abstract

Diamond prices have been exceedingly variable during the last century. In this research, we describe a machine learning-based strategy for predicting diamond prices in order to avoid human error. There is a lot lower danger of losing the investment with an accuracy of 98% utilizing Random Forest Regression. Linear regression, Lasso regression, Support Vector Regression, and Random Forest are all used in the proposed machine learning-based prediction model. The proposed technique predicts the value most correctly. We've also introduced a Crontab tool to automate the process, which will retrain the model to the most correct value before the diamond market opens

Keywords: Support vector regression, Random forest, Decision tree, Linear regression, Laso regression.

I. INTRODUCTION

Diamonds are one of the world's most valuable stones. It is also one of the most expensive stones, hence its price is quite erratic. The value of a diamond is determined by its structure, cut, inclusions (impurities), carats, and a variety of other factors. Diamonds are widely used in a variety of industries due to their effectiveness in cutting, polishing, and drilling. Diamonds are highly precious, thus they have been traded across borders for ages, and this trade is only growing. They are assessed and verified using the "four Cs": color, cut, clarity, and carat. Color, clarity, cut, and carat weight are all important factors to consider.

.

This metric allows individuals all around the world to acquire diamonds with a common understanding, facilitating trade and providing a fair price for what is purchased. Diamond prices are typically established for the day and are denominated in US dollars.The more obvious notion is that there is a

This machine-learning algorithm examines more than four features, resulting in a more accurate output. When a price chart is displayed on a graph, it produces a variety of formations such as pennants, wedges, flags, double bottoms, and tops. These patterns are frequently employed in currency markets and many other trading markets, such as the diamond market..

II. LITERATURE SURVEY

Background study

The price of diamonds is largely determined by the 4Cs: carat weight, cut, colour, and clarity. These factors can interact in complex ways, making it difficult to predict the value of a diamond. Traditional methods of diamond price prediction have relied on expert opinions and industry knowledge. However, with the rise of machine learning, algorithms can now be trained on large datasets of diamond prices to predict the value of new diamonds. Linear regression is one of the simplest and most widely used machine learning algorithms for diamond price prediction. It works by finding the linear relationship between the input variables (the 4Cs) and the output variable (diamond price). This algorithm can be effective if the relationship between the variables is linear, but it may not capture more complex interactions. Decision trees are another popular machine-learning algorithm used for diamond price prediction. They work by recursively splitting the input data into subsets based on the input variables, creating a tree-like structure that maps the input variables to the output variable (diamond price). Decision trees can be powerful algorithms, but they can also be prone to overfitting and may require careful tuning to produce accurate results. The Gemological Institute of America (GIA) was the first laboratory in the United States to provide contemporary diamond reports, and it is highly regarded among gemologists for its consistent, conservative grading. Diamond High Council (HRD) Antwerp-based official certification laboratory for the Belgian diamond industry.A number of for-profit gemological grading laboratories have also been established in the recent two decades.

III. METHODOLOGY

Jupyter Lab was the interactive development tool we used to write code and see data. Jupyter Lab is a versatile platform for arranging and configuring data for machine learning and data science tasks. The biggest benefit is the ability to create plugins and new components and combine them with Crontab.

It is built on top of matplotlib and tightly integrates with pandas data structures. It contains built-in capabilities that allow it to clearly visualize data with minimal programming. Sklearn is a Python-based machine-learning library. It includes techniques for classification, regression, and grouping.

.

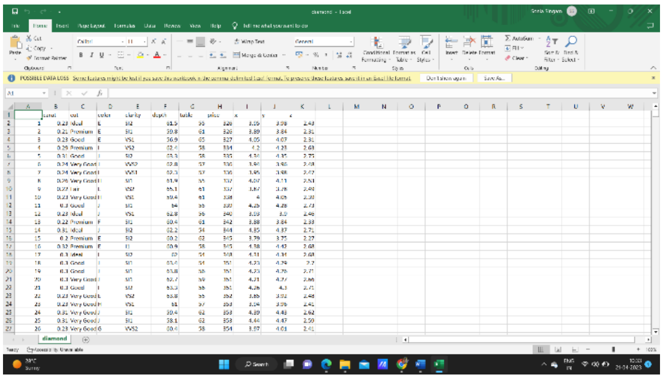
Crontab is employed to perform a regular schedule to manage a list, which is accomplished through the use of a set of scripts. Crontab, which stands for cron surface, is an assignment scheduler. In this situation, we're utilizing it each day to refresh the algorithm for predictions that are more precise. The @daily cron keyword is used, which will generate a log file each day and clear it using the cleaning up-logs bash program at 08:00 each day when the diamond's shape is created.

1. Dataset

Machine learning databases are used to train neural network methods. A dataset is an example of how machine learning can aid in forecasting, with labels representing the achievement or defeat of a certain hypothesis. The best approach to get started with machine learning is to use libraries such as Scikit-learn or Tensor flow, which enable one to execute the majority of tasks while having to write codes.

Machine learning techniques are classified into three categories: supervised (learning from examples), unsupervised (learning through grouping), and reinforced learning (rewards). The practice of teaching a computer to recognize similarities in information is known as supervised learning. Algorithms for supervised learning are used in the following methods: random forest, nearest neighbors, the weak rule for big values, the ray-tracing method, and the SVM method.

Machine learning datasets come in an array of formats and can be obtained from a range of sources. The three most popular types of machine learning datasets are data are written, picture information, and information from sensors. A database is a collection of records that may be utilized to predict collection of records that may be utilized for predicting potential events or outcomes based on historical data. When using methods of machine learning, facts are often labeled so that the system understands what outcome to forecast or categorize as an abnormality. As an illustration, in the event that you had to forecast whether or not a customer would churn, you could label the dataset "churned" versus "not churned" so that the method of machine learning could gain insight from previous data. Machine learning datasets can be generated from either.



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41

Fig:1 Dataset with different specifications.

2. Feature extraction

If an amount of features grows comparable (or even greater!) than the number of observations recorded in an information set, a Machine Learning model is prone to overfit. To avert this sort of difficulty, major normalization or methods for reducing dimensionality (Feature Extraction) must be used. A dataset's dimensional in the context of machine learning is equal to the number of variables used to represent it.

The feature extraction process aims to minimize the number ofof the feature extraction process is to minimize the number of features in a dataset by producing fresh ones using current ones (and subsequently deleting the initial ones). The new reduced set of features should then be able to summarise the majority of the data contained in the initial assortment of attributes. A condensed version of the original characteristics can thus be generated by combining the original set.

Choosing Features is another method for reducing the number of dataset features of characteristics in an information set. Feature Selection differs from Feature Extraction because the choice of feature the significance of the present characteristics in the source data and removes less significant ones (no new features are created).

3. Data Preprocessing

Data preprocessing is the procedure of arranging unprocessed information for use with a machine learning model. It is the initial and most important stage in developing a machine-learning model. While developing a machine learning project, we do not always come across clean and prepared data. And, before performing any actions on statistics, it must be cleaned and prepared. Therefore we employ an information preparation activity for this.

A. Acquiring the dataset

The initial resource that we needed to develop a machine-learning for an available dataset was because a machine-learning model is entirely dependent on data. A database is the collection of data for a certain problem in the right format.

Datasets can be of various forms for multiple uses. For example, if we want to construct a machine learning model for business reasons, the collection of data is going to be distinct from the information necessary for a liver illness. As a result, every data set is distinct from the others. We normally save the information in question as a CSV file before using it in the source code. But there may be situations when we must use an HTML or xlsx file.

B. Importing Libraries

To do data preparation with Python, we must first import several established libraries for Python. These libraries are used to carry out specific tasks. For preprocessing information, we are going to use several particular archives, which include:

Numpy: Numpy The Python library is used in programming to include any form of mathematical computation. It is Python's essential package for computation in science. It also allows for the addition of big, multi-dimensional arrays and matrices.

C. To import datasets

We must now integrate the collection of data that we acquired for our machine learning research. However, before importing a dataset, we must make the present folder a functional directory.

D. Handling the data that is missing

The following stage in data preparation is to deal with missing data in the datasets. If the information we are using contains certain data that isn't present, it may pose a significant challenge to our machine-learning model. As a result, handling the values that are absent in the information set is required.

E. Encoding the categorical data

Classified information is data that falls into certain categories, including both of the categorical variables in our dataset, State and Bought. Because artificial intelligence models are entirely based on mathematics and numbers, having a variable with a category in our data set may cause problems while developing the algorithm. As a result, these category data must be encoded into integer.

F. Testing and training

The division of a training set and an evaluation set during the process of machine learning data preparation. This is an important stage in data preparation since it allows us to improve the performance of our machine-learning model.

The simulation is going to struggle with comprehending the relationships among the representations. If we train our model very well and its training accuracy is likewise quite good, but then we give it a fresh dataset, the model's accuracy will suffer. As a result, we always strive to create a machine-learning model that outperforms the set that was used for training.

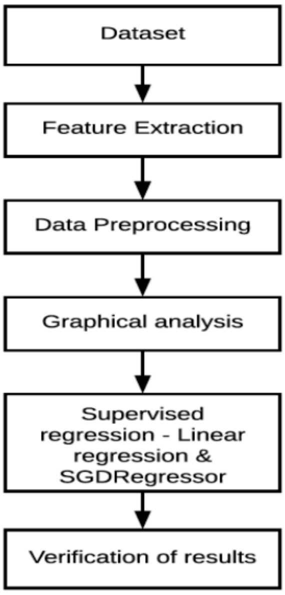


Fig:2 Data pre-processing flow diagram.

4. Graphical Analysis

The graphical analysis generates images of the information, helping in understanding the patterns and relationships between parameters associated with the process. The use of graphics is frequently used as the starting point for any problem-solving strategy.

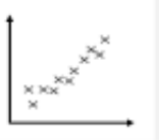


Fig:3 Shows the sample graphical analysis.

5. Supervised regression

Data analysts employ a wide range of statistical methods to uncover trends in large amounts of data that lead to significant conclusions. These numerous algorithms can be divided into two categories depending on how they "learn" from information to create predictions: supervised and unsupervised learning.

6. Verify the results

Hence this is the final step after the sequence of operations performed on the model. The result must show the appropriate and significant outcome.

IV. ALGORITHM

There are several machine learning algorithms that can be used for diamond price prediction, including linear regression, decision trees, random forests, support vector machines, and neural networks.

It involves fitting a linear equation to the data and using this equation to predict the value of new diamonds based on their 4Cs. Linear regression is easy to understand and implement, but may not capture complex relationships between the 4Cs and diamond prices.

Decision trees are another popular algorithm for diamond price prediction. Decision trees involve splitting the data into smaller subsets based on different criteria, such as carat weight or colour. The algorithm then uses these subsets to make predictions about the value of new diamonds. Decision trees are easy to interpret and can capture complex relationships between the 4Cs and diamond prices, but may be prone to overfitting.

Random forest is a variation of decision trees that can improve their accuracy by combining the predictions of multiple trees. Random forest involves creating many decision trees using different subsets of the data and combining their predictions. This approach can reduce the risk of overfitting and improve the accuracy of the model.

Support vector machines (SVMs) are another algorithm that can be used for diamond price prediction. The algorithm then uses this hyperplane to predict the value of new diamonds. SVMs can be very accurate but may be sensitive to the choice of the kernel function.

Finally, neural networks are powerful and flexible algorithms that can be used for diamond price prediction. Neural networks involve creating a network of interconnected nodes that can learn complex relationships between the 4Cs and diamond prices. They are particularly effective for capturing nonlinear relationships and can be trained using large amounts of data. However, neural networks can be complex to implement and may be prone to overfitting.

The algorithm formula for diamond price prediction with machine learning will depend on the specific algorithm used. However, as an example, I can provide the formula for a simple linear regression model:

Price = b0 + b1 x Carat + b2 x Cut + b3 x Colour + b4 x Clarity + E

where:

Price is the predicted diamond price
Carat, Cut, Colour, and Clarity are the four C's of the diamond
b0, b1, b2, b3, and b4 are the coefficients of the linear regression model
E is the error term
The coefficients (b0, b1, b2, b3, and b4) are estimated during the training phase of the model, using a dataset of diamonds with known prices and 4C values. Once the coefficients are estimated, the model can be used to predict the price of new diamonds based on their 4C values.

V. ARCHITECTURE

The architecture of a machine learning algorithm for diamond price prediction typically involves several steps:

•Data collection: Collecting data about diamonds, including their carat weight, color, cut, clarity, and other feature that affect their value.

•Data pre-processing: Cleaning the data, handling missing values, and transforming the data into a format suitable for analysis.

•Feature engineering: Selecting the most important features that are relevant for predicting diamond prices and creating new features that capture complex relationships between different features.

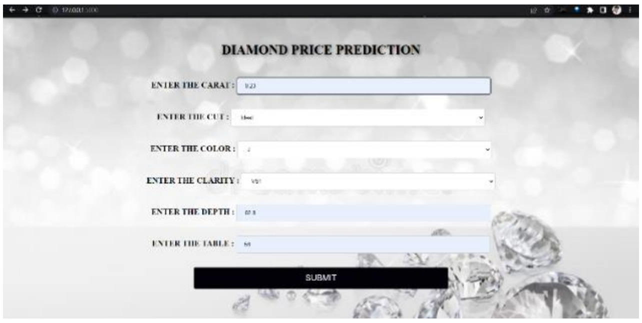
• Model selection: Choosing an appropriate machine learning algorithm, such as linear regression, decision trees, random forests, or neural networks, that can effectively learn the relationships between the features and the diamond prices.

•Model evaluation: Testing the model on a separate set of data to assess its performance and identify areas for improvement.

Model deployment: Deploying the trained model in a production environment where it can be used to predict diamond prices in real-time

VI. RESULTS

Linear Regression, Lasso Regression, Support Vector Machine, and Random Forest are the four kinds of algorithms employed here. The findings are listed right now, along with the precision of their scores. According to the surface, Random Forest produces the most accurate results because it employs several trees, with an accuracy rate of 98%. In addition, the lasso regression method produces results that are nearly exactly the same as regression using linear models. Support Vector Regression, also known as SVR, because 0 indicates there is no mistake and 1 indicates an elevated level of inaccuracy.


DIAMOND PRICE PREDICTION

ENTER THE CARAT : 0.23

ENTER THE CUT : Ideal

ENTER THE COLOR : [illegible]

ENTER THE CLARITY : VS1

ENTER THE DEPTH : 62.8

ENTER THE TABLE : 56

Fig:4 Input data in the result.

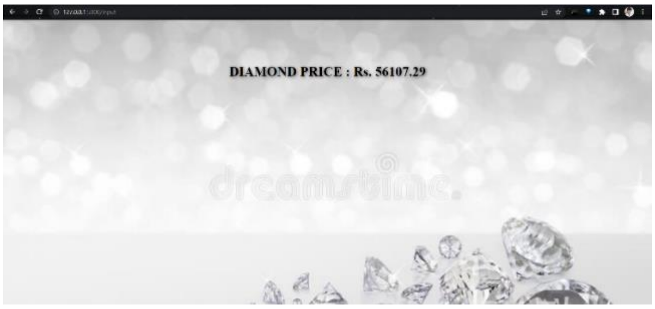

DIAMOND PRICE : Rs. 56107.29

Fig:5 Output (the price of the diamond)

VII. APPLICATIONS

Elegant, unique artifacts always piqued people's interest. Diamonds were prized as gemstones since ancient times, revered as a result of their splendor, and are today regarded as the greatest extravagance in jewels. Diamonds, on the other hand, are prized for much more than just their dazzling brilliance. Their distinct attributes in appearance elevate them above all other gems in value.

Some Applications:

i. We've all heard of De Beers' renowned motto, "A diamond is eternal," which originally appeared in the year 1947, and the brilliant color of a diamond placed in an engagement ring, studs, or similar exquisite Schmuck. Diamonds possess a long history in the marital industry, and they are the typical stone in engagement and wedding rings, expressing everlasting affection and dedication.

ii. Diamonds' outstanding durability and distinctive characteristics render them useful for a wide range of industrial applications. The majority of diamonds extracted deficiency of the grade required for the creation of jewels and 80% of all rough diamonds are used in industrial applications.

iii. Diamonds may have medicinal advantages. According to a scientific study, miniature diamonds (small diamond particles) may be an indicator of the efficacy of cancer medicine once supplied to clients, enabling clinicians to track the growth of cells. Experts are also investigating the use of diamonds to assist those with impaired vision, as well as diamonds as a possible material for bionic eyes and eye implants.
Much dental equipment has diamond edges to assist surgeons to drill more efficiently while avoiding damaging equipment.

VIII. CONCLUSIONS

When investing, knowing how to forecast diamond prices is critical. It lessens reliance on individuals and rumors. This work contains an in-depth evaluation along with studies estimating diamond pricing using criteria more compared to the renowned 4Cs.

IX. REFERENCES

[1]anwala, S. (May 16, 2021). Medium.https://medium.com/@sp7091/regressionapproaches-to-predict-diamond-price-258478a485Cj

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

[3] A. Khawaja, and M. Ahmad, "Predicting the price of diamond using machine learning techniques," in 2019 International Conference on Frontiers of Information Technology (FIT), 2019, pp. 271-276. (https://ieeexplore.ieee.org/document/9037136)

[4] H. D. Durrani, M. I. Khan, and M. J. Hussain, "Predicting diamond prices using machine learning algorithms: A comparative analysis," in 2020 International Conference on Computer, Communication, and Signal Processing (ICCCSP), 2020, pp. 1-5. (https://ieeexplore.ieee.org/document/9172045)