Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/45/20230274

Study of Factors in the Price of Diamonds

Yufei Li[1,a,*]

[1]Harbin No.9 High School, Harbin Heilongjiang 150023, China
a. 1811031103@mail.sit.edu.cn
**corresponding author*

Abstract: In recent years, the price of diamonds has fluctuated significantly. This article mainly studies the factors that affect the price. This article mainly uses linear regression and random forest to calculate the coefficients between prices and influencing factors, and determines the degree of impact between them by comparing their sizes. In addition, the histogram analyzes whether there is a relationship between cutting clarity color and price. However, the relationship between cutting and diamond prices is not so strong. Finally, it is concluded that the weight of the price table, as well as the length and width of the table have a significant impact on the diamond price. When purchasing diamonds, the conclusions of this article can be used to determine whether the characteristics of various aspects of the current purchase price are directly proportional to the price they sell to ensure that consumers do not enter this diamond scam.

Keywords: random forest, linear regression, diamonds prices, box plot

1. Introduction

In recent years, consumer demand for diamonds has risen sharply, especially among women aged 21 to 39. Even some jewelry brands provide other services to increase the consumption of diamonds by the younger generation. After their new CEO took office, Tiffany, a jewelry brand in New York, opened online celebrity cafes and invited traffic through flagship stores. The way stars shoot advertisements narrowed the gap with young consumers. At present, there are 30 countries with diamond resources, and the annual output is about 100 million carats. The top five countries with the highest output are Australia, Itzhar, Botswana, Russia, South Africa, and their diamond output accounts for about 90% of the world's total diamond output.

Nowadays, artificial diamonds have sprung up, and China's production of artificial diamonds accounts for over 90% of the global total, with the production of artificial diamonds in Henan Province accounting for 80%. The reason why artificial diamonds are becoming more and more popular is that the industries and fields it involves as well as the consumer market are expanding [1].

But In fact, artificial diamonds do not have a significant impact on natural diamonds because people believe that there is no way to compare the value of natural diamonds with that of artificial diamonds. Moreover, the value of artificial diamonds in recent years is also very low. In addition, with the rapid development of science and technology, there are many technologies available, and I can help people distinguish the difference between artificial diamonds and natural diamonds. People are not deceived into consuming artificial diamonds because they cannot distinguish them. The project involves using diamonds and verifying samples based on a unified diamond detector standard

DOI: 10.54254/2754-1169/45/20230274

and conducted by an independent third-party testing agency. This information provided by agency is transparent and very objective. For example, someone has previously used a gemstone microscope infrared spectrometer and Raman spectroscopy to test and study natural black single crystal diamonds. Raman spectroscopy is the fastest and most effective identification method for determining the natural nature of black diamonds [2].

The price of diamonds is inextricably related to their color. The rarer and more precious colored diamonds among diamonds have always been appreciated by many people. The demand for colored diamonds in the international and domestic markets is also increasing, so their value has also been rising. Some rare colors such as pink are only 50 carats worldwide [3]. Cut is also an important influencing factor. The clarity of diamonds also plays a significant role in the price of diamonds. Different clarity and cut have an impact on the price growth rate of the same color yellow diamond [4].

In addition to the above factors, consumers may not care about the length, width, depth, and weight of the diamonds provided on the certificate. Is this related to the price of the diamonds? This is also the main issue to be discussed in this article.

In the past four years, machine learning methods have been widely used. In this paper, linear regression and random forest methods are mainly used. Linear regression has been applied to various industries, including some who have applied machine learning methods, in which the logical regression support vector machine decision tree is used to build models and analyze the demand for new energy vehicles. In addition, it has also been applied to the field of medicine. Someone has discussed the impact of confounding factors on causal inference and the role of machine learning methods in this process [5, 6]. Linear regression can obtain coefficients while visualizing data, which is very convenient. Random forest can obtain the sequential arrangement of coefficients and influencing factors. This paper obtained the specific data of nearly 60000 diamonds from the website. Firstly, they were visualized using a box diagram, and then their scatter diagrams were drawn using a linear regression algorithm. Calculate the R2 coefficient and the coefficient of random forest using the formula.

The basic conclusion drawn in this article is that there is a very strong relationship between the price of diamonds and their color clarity, but the relationship with cutting is not obvious. In addition, the price coefficient with carat, length, width, and depth of diamonds are relatively large, indicating that there is also a very strong relationship between them.

The remainder of this paper is organized as follows. Section 2 describes the data, their sources, and how they were processed into the measures used for analysis. In the following Section 3, the author presents and describes the results and emphasize their place in the context of the war and economic connectedness. The last section concludes describing results and contributions

2. Methodology

The box plot is used to cut and analyze the influence of the clarity and color of diamonds on the prices. The author mainly use two python models: linear regression and random forest to compare remaining factors.

2.1. Linear Regression

Linear regression model is a classic statistical model. Its main idea is to predict a continuous numerical variable based on known variables, while also calculating the correlation between two variables. This paper mainly uses a linear regression model to calculate the correlation between two variables to determine which factor mainly affects the price. Two authors' previous research shows that generally, such machine learning can be conducted without prior training, and can be used to find

DOI: 10.54254/2754-1169/45/20230274

the rules and patterns within the data to obtain the characteristics of the data. This is also the main feature of linear regression used in this paper.

$$\hat{y} = \hat{b}x + \hat{a} \tag{1}$$

This is the main formula used in linear regression. Y refers to the dependent variable represented by the y-axis, while x refers to the dependent variable represented by the x-axis. B refers to the coefficient of the regression line in the graph and a refers to the intercept between the regression line and the y-axis. The next two formulas show how to calculate the value in the b and a in the formula.

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2} \tag{2}$$

$$\hat{a} = \hat{y} - b\bar{x} \tag{3}$$

The $\bar{x}$ and the $\bar{y}$ in these two formulas represent the average values of the independent and dependent variables.

In the linear regression model, R^2 score is used to judge the relationship between independent variables and dependent variables. The closer the score of R^2 is, the greater the relationship between independent variables and dependent variables. The following equation is used to calculate the value of R^2.

$$R^2 = 1 - \frac{\sum_i(\hat{y}^{(i)} - y^{(i)})^2}{\sum_i\sum_i(\bar{y} - y^{(i)})^2}$$

$$=1 - \frac{(\sum_{i=1}^{m}(\hat{y}^{(i)}-y^{(i)})^2)/m}{(\sum_{i=1}^{m}(y^{(i)}-\bar{y})^2)/m}$$

$$=1 - \frac{MSE(\hat{y},y)}{Var(y)} \tag{4}$$

2.2. Random Forest

The random forest algorithm is a classifier that contains multiple decision trees and can be used to solve classification and regression problems. In this paper, the author mainly uses it to solve regression problems Someone once used machine learning methods to identify illegally operated vehicles, which is also a principle that the author mainly applies [7]. The relationship between dependent variable and independent variable of random forest judgement conforms to the following chart (Fig. 1).

DOI: 10.54254/2754-1169/45/20230274

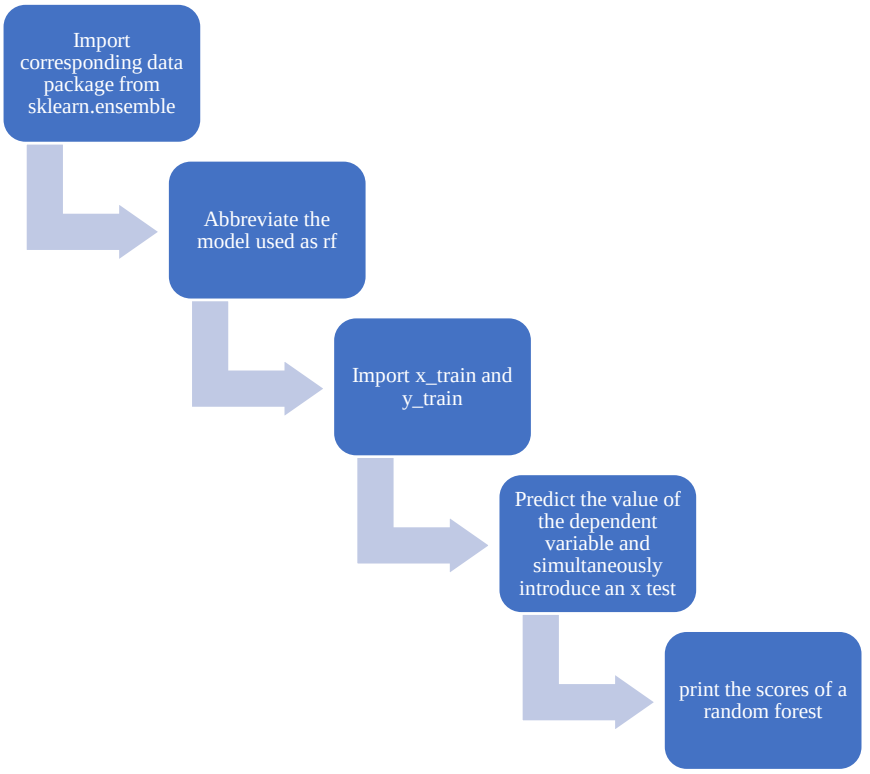

Figure 1: The process of random forest.

3. Data

3.1. Source of Data

The author's data comes from the Kaggle website and it shows the prices of 54000 diamonds and many of their characteristics, such as cut, width, depth and length. The Fig. 2 will illustrate all factors affecting the prices.

3.2. Factors of Prices

Figure 2: Main factors of diamonds prices.

DOI: 10.54254/2754-1169/45/20230274

The first factor from figure 2 is the weight of the diamond, which is usually clearly indicated in the certificate, and is also a very important factor in measuring the grade of the diamond. The second three factors are the length, width, and depth of the diamond, which are usually expressed in the form of a diamond image, with specific values marked thereon. In the chart, the last two factors are the color and clarity of diamonds, which are both diamonds. The measurement criteria are two of the most important among the four C, and they are a matter of great concern to everyone. Usually, when buying diamonds, people ask the seller about the color and clarity of the diamonds, which can also enable everyone to have a valuation of the diamond in their hearts.

The value of diamonds is determined by their grade, which is commonly referred to as the "Four C" standard for diamonds. In this journal, there are also specific descriptions of this standard [8].

3.3. Process and Analyze Data

In the process of data preprocessing, the paper applies the log logarithm on digital normalization in order to make the mean square error as small as possible. The below graph (Figure 3) demonstrates the trend of diamonds from April 2022 to the February 2023.

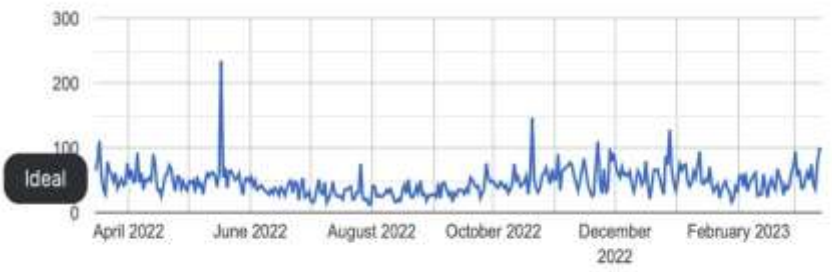

Figure 3: The prices of diamonds from 2022 to 2023.

The following scatter plots (Figure 4) and fitting lines represent a proportional relationship between diamonds prices and carat.

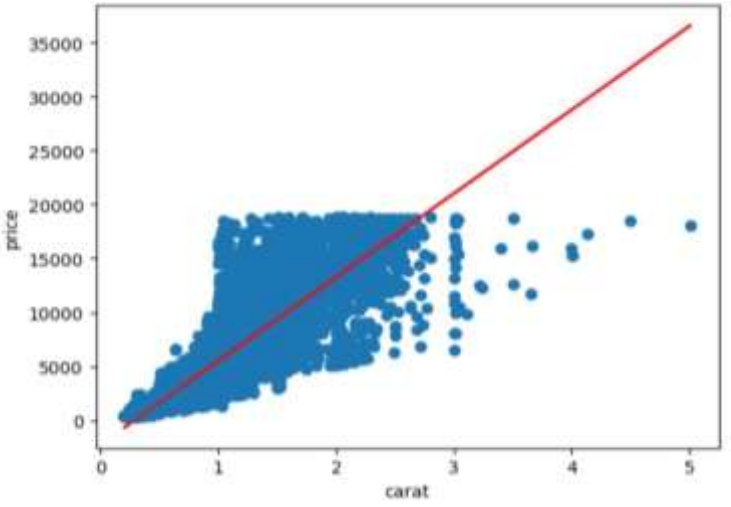

Figure 4: The scatter plot of carat and prices.

The two bar charts (Fig. 5 and Fig. 6) below illustrate that the better the color and clarity, the higher the price of a diamond will be.

DOI: 10.54254/2754-1169/45/20230274

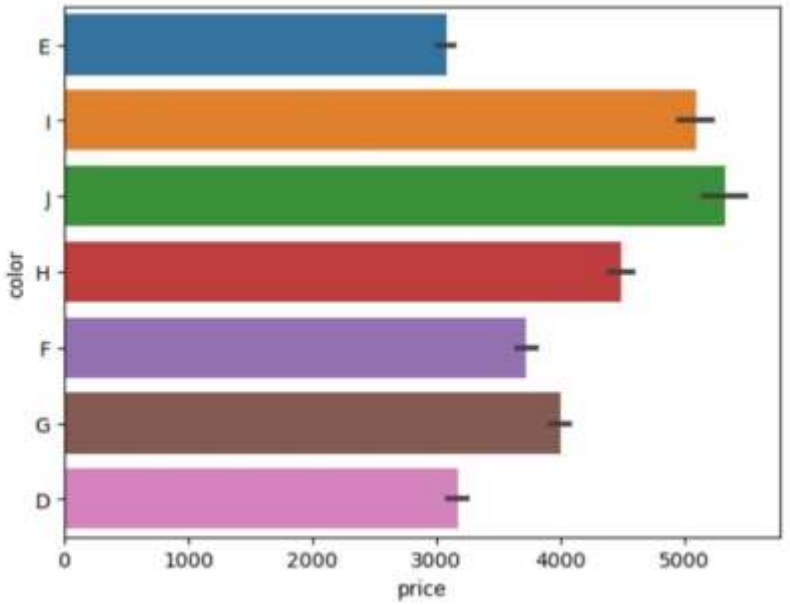

Figure 5: The cut and corresponding prices.

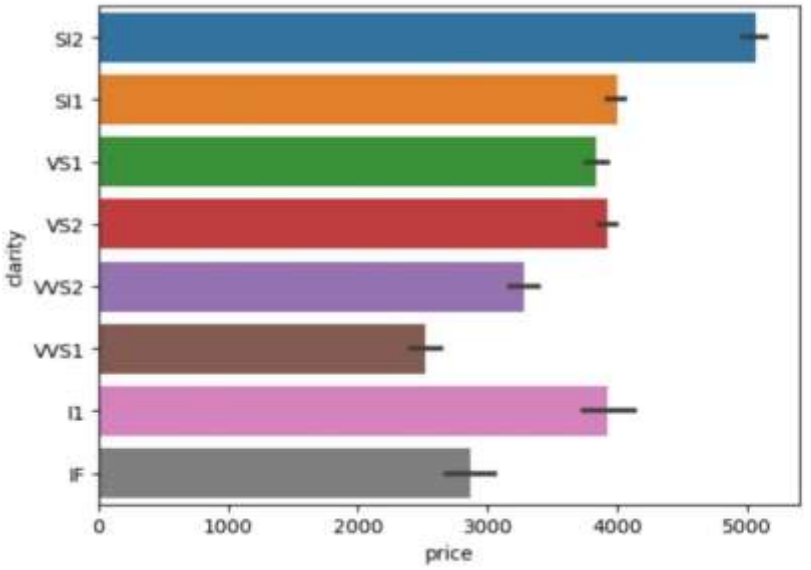

Figure 6: The clarity of diamonds and corresponding prices.

4. Results

4.1. The Comparison of Scores Between Two Models

To summarize, the three figures above briefly demonstrate some main factors that affect prices and the possible relationship between them and prices

The paper uses the table 1 to present the specific values between the prices obtained through linear regression and random forest and some influencing factors mentioned above.

DOI: 10.54254/2754-1169/45/20230274

Table 1: The scores among diamonds prices and factors.

Linear regression	Random forest
Carat: 0.84933	Carat: 0.8681
Table: 0.01615	Table: 0.01777
Length: 0.78222	Length: 0.86697
Width: 0.74894	Width: 0.87262
Depth: 0.74175	Depth: 0.85411

From the results shown in the table, it can be seen that the coefficients between the carat, length, depth, and width of diamonds and their prices are relatively large, indicating that they play a significant role in the increase and decrease of prices. On the other hand, the score between the table and the price is only close to0.02, which proves that the relationship between them is very small, and it may not be able to make a significant impact on the price change. The author used mean square error to test the accuracy of the two models, as the amount of data is relatively large, so the accuracy is very high. The accuracy of both models is 15657364. This accuracy indicates that both models are very accurate and the result is also credible.

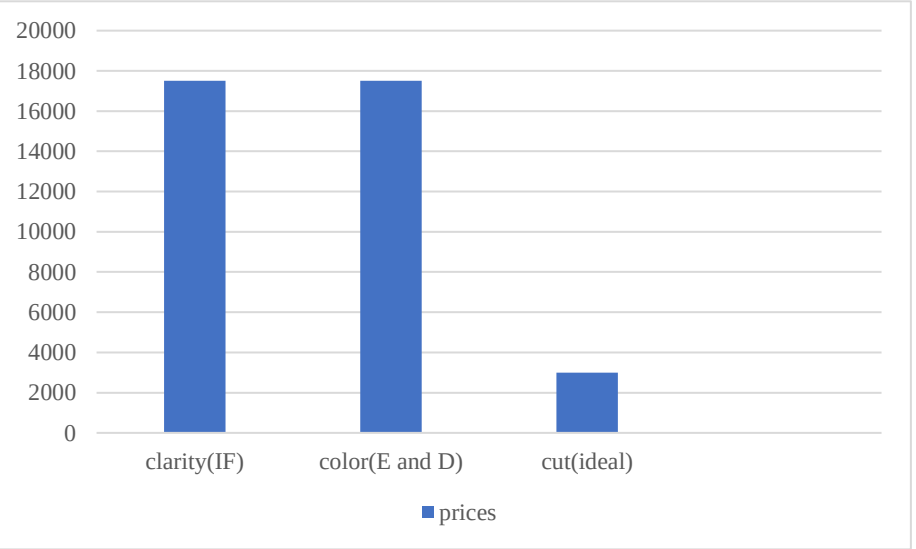

Figure 7: Color sharpness and their maximum price.

The bar chart (Fig. 7) above shows the relationship between cleanliness, color, and cutting price obtained by the author from the box diagram. The statistics for clarity, color, and cut are the highest level. The price of the highest grade diamond corresponding to clarity and color is also the highest, around 17500 which is the highest prices of diamonds in the data of writer. However, although the cut grade is also the highest, the price is only 3000, so it can be seen that the relationship between clarity and color and price is stronger than that between cut. Wang said “the color making factors of color diamonds are very complex and valuable. Color diamonds are very scarce, possibly reaching 1 in a million.” [9].

In addition, Lv and Wang mentioned that the clarity of diamonds is also a major criterion for measuring diamonds and purer the diamond, the higher the price [10]. These undoubtedly prove that the price of a diamond is closely related to its clarity and color.

DOI: 10.54254/2754-1169/45/20230274

4.2. Comparison of Two Models

The linear regression model is better in this paper because it can contain a large amount of data and can be imported from the sky database, which is very easy to operate and easy to use. In addition, the confidence regression model can also provide some scatter plots and the coefficient of the fitting line, which can also verify whether the R^2 score is very accurate, and whether there is a positive correlation or a negative correlation between the two factors. He and Duan mentioned that the linear regression can help people save time, obtain data, and are currently very popular machine learning methods and practical tools, which contain a large number of data sets and are easy to use. Great for everyone [11].

5. Conclusion

This article mainly studies the influencing factors of diamonds. Linear regression is used to visualize the data in a scatter plot and obtain the coefficients between the diamond price and the influencing factors. At the same time, the random forest method is used to calculate and compare whether the coefficients between the two methods differ greatly. In addition, the color clarity of diamonds can indeed have a significant impact on the diamond price, but in addition, the length, width, and carat number of diamonds also affect the prices.

The surface of a diamond can also have a significant impact on the price but it may be that the linear regression method and the random forest method are not suitable for his measurement, so the data obtained may be biased. It is hoped that logical regression can be used for testing in the next experiment.

References

[1] *Xia, XQ., Yang Z.P.: Henan has become the largest production base for artificial diamonds, http://www.cnli.net, 2021/11/10, last accessed 2023/04/2*
[2] *Qian, W.J., Feng, Z.R., Lu, X.Y.: Identification characteristics of natural black single crystal diamonds. Shanghai Metrology Testing 2, 8-10 (2021).*
[3] *He, Z.M.: On the color origin of pink diamonds. Art Appreciation Journal 12, 162-163 (2017).*
[4] *The magnificent 21.37 carat natural colored Swallow Diamond yellow VS2 clarity diamond with diamond ring has a transaction price of 125000 yuan, Art Appreciation Journal 34, 20-20 (2018).*
[5] *Xie, Y.F., Li, F., Cheng, D.: Analysis of new energy vehicle demand based on different machine learning methods. Modern Industrial Economy and Information Technology 2, 101-102 (2022).*
[6] *Lan, Y.S., Zheng, S., Li. J.: The application of machine learning methods in causal inference with confounding and factor control. Journal of Medical Informatics 11, 20-33 (2022).*
[7] *Huang, L.: Research on urban illegal operating vehicle identification based on stochastic forest model. Journal 12, 1-5 (2022).*
[8] *The Four C Standards for Diamonds. Chinese Baoyu 1, 106-107 (2017).*
[9] *Wang, W.Q.: Selection of color diamonds. Journal of Quality and Standardization 5, 31-34 (2020).*
[10] *Lv, L.F., Wang, Y.F.: Gem evaluation criteria. The Consumer Guide Journal 2, 37-49 (2018).*
[11] *He, X.N., Duan, F.H.: A linear regression case based on python. Microcomputer Applications Journal 11, 35-37 (2022).*