Prediction on the Prices of Laboratory-Grown Diamonds based on Multiple Linear Regression Model

Jingru Xu*

Department of Business Administration, Macau University of Science and Technology, Macau, 999078, China

*Corresponding author: 1230004258@student.must.edu.mo

Abstract. Diamond as emotional investment, in recent years, the market price of it is going through rising and falling, and drastic fluctuating, more is the decline in price. The diamond industry has also been challenged, there is happening oversupply and weak demand in diamond market. The impact of lab-grown diamonds on natural diamonds has also changed the price of diamonds across the diamond industry. However, the market of diamond is still at the forefront of the jewelry industry. In order to ensure that consumers have more understanding of diamond choices, this paper adopts multiple linear regression model to forecast diamond prices, collects a large amount of data, and uses 4c criteria as the standard for diamond price prediction to explore the relationship between diamond prices and diamond measurement standards. Through the calculation, the diamond buyers will explore ways to realize diamond price, the diamond industry will be rationalized production, and balance the relationship between diamond supply and demand to provide a method.

Keywords: Diamond price; influent factors; diamond market.

1. Introduction

This paper examines the factors that influence diamond prices, and uses multiple regression linear analysis to predict their prices, thereby enhancing its professional discourse. Diamond is an alternative type of investment, known as emotional investment, and there are existing price bubbles, which affect the price of diamonds in the market [1]. Therefore, diamonds show the price fluctuations phenomenon in the market. In general, the price of natural diamonds goes through a series of price increases from the raw stone, which is caused by valued and taxed, and handled from dealers, cutters, and finally to retailers and designers [2]. And under the pressure brought by the degree of marriage and macro pressure, the sales of engagement rings have declined, and their value preservation has been criticized, and the demand for natural diamonds has decreased significantly since 2023 [3]. The saying that called “A diamond lasts forever” has been challenged.

Not only that, a new type of diamond has appeared to make the status of natural diamonds even more precarious. As a kind of diamond, lab-grown diamonds are different from moissanite. Lab-grown diamonds and natural diamonds are pure carbon crystals body, with the same physical and chemical properties, in the refractive index, transparency and other aspects of the two can be comparable [4]. Lab-grown diamonds can be understood as lab-grown crystals and are also considered part of synthetic diamonds. The development of lab-grown diamonds can produce not only high-carat, high-clarity diamonds, but also colored diamonds of different colors. Because of its excellent quality and superior price advantages, cultivated diamonds are highly sought after [5].

However, as once a highly sought after laboratory-grown diamond in the capital market, in recent years, there has been a continued downturn in the market, and some industrial terminal demand for superhard material products has weakened, laboratory-grown diamonds have suffered from supply shocks at home and abroad, and sales have declined. After experiencing a rise in market heat, capital has been involved in the field for the cultivation of high gross profit margins of diamonds, and after the continuous release of production capacity, the laboratory-grown diamond market prices have also fallen from a high level, while the laboratory-grown diamonds flooded the market, so that natural diamonds have also shrunk in price [6]. Despite the current problems facing the diamond industry, including over-optimistic expectations of diamond consumption growth and weak downstream

demand for diamonds, China still dominates the diamond jewelry consumption market [7]. At present, there are still many consumers who have the ability and enthusiasm to buy diamonds, so it is necessary to conduct research on the reliable price assessment of diamond jewelry to provide an effective means for consumers to understand the true price of diamond jewelry [8].

Because lab-grown diamonds belong to real diamonds, this paper does not distinguish lab-grown diamonds from natural diamonds in the process of diamond price prediction. In 1949 by the world's diamond monopoly, De Beers company in order to promote the diamond trade and set up the standard, often called the 4c standard, that is, according to the carat weight of the finished diamond, cut, color and clarity to grade, and according to this to set the price list, so that the diamond trade in a very standardized and orderly market, and then promote the prosperity of diamond sales [9]. These factors may have different effects on the price of diamonds, so different types of diamonds may have different criteria for measurement. For example, the cutting ratio of colored yellow diamonds is not exactly cut according to the requirements of the cutting level of colorless diamonds [10]. Therefore, the sample size should be as large as possible in order to obtain a more accurate algorithmic model for determining the value of diamonds for consumers

The above model combines the formula with specific data to describe the specific relationship between the variables, so the use of multiple linear regression analysis in this paper to predict the price of diamonds is a good idea, while citing articles to increase credibility.

2. Methodology

2.1. Data Source and Description

The data set which used by the article is collected from Kaggle, and the price is in US dollar, range from $326 to $18823. All of those data are the feature of diamond, including carat, cut, color, clarity, depth (total depth percentage), table. The “x” “y” “z” refers to the length, width and depth of the diamond.

2.2. Index Selection and Description

Measuring the price of diamonds usually follows the 4c standard. They included carat, cut, color and clarity. And here are some basic introduce to those standers (Table 1).

Table 1. Basic information

4C standard	explanation
Cara	Carat refers to the weight of a diamond in metric carats
Cut	The cut grade of a diamond, includes fair, good, very good, premium, ideal
Color	It is judged from J (worst) to D (best)
clarity	Including I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best). The internal characteristics are inclusions, and external characteristics are blemishes

Carat (0.2--5.01): One carat equals 1/5 gram and is subdivided into 100 points. Carat weight is the most objective grade of the 4Cs.

Cut: The diamond graders evaluate the diamond cutting level according to the cutting ability of the diamond cutter. The more accurate the cut, the more precious it is.

Color: The color of gem-quality diamonds occurs in many hues. In the range from colorless to light yellow or light brown. Colorless diamonds are the rarest. Other natural colors (blue, red, pink for example) are known as "fancy,” and their color grading is different than from white colorless diamonds.

Clarity: Although the defects in many diamonds are so small that they need to be magnified to see them, diamonds that are truly free of inclusions or defects are rare, and they are often very scarce.

Depth: It refers the total depth percentage which equals to z / mean (x, y) = 2 * z / (x + y). It is the width of the top of the diamond relative to the widest point that allows the diamond to shine brightly, as Figure 1 shows.

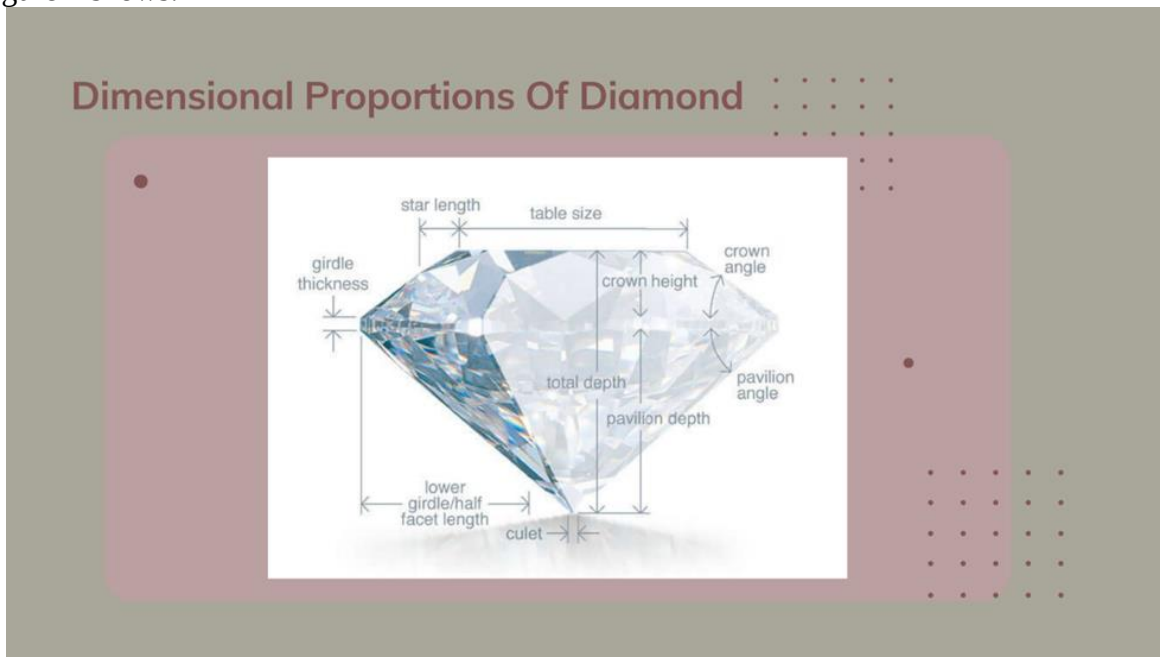

Fig. 1 Dimensional proportions of diamond [11]

2.3. Introduction to Methods

The article choses multiple linear regression model as the method to conduct analysis. It refers to a model in which multiple influencing factors are used as independent variables to explain the dependent variables. The mathematical expression of the multiple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

In multiple linear regression analysis, there is no significant difference between the i-th partial regression coefficient and 0. The null hypothesis for the significance test of multiple linear regression equations is:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 (i = 1,2, \ldots, k) \quad (2)$$

In the formula, k is the number of explanatory variables and n is the number of samples. SPSS automatically corresponds the F-value to the probability P-value, if the P-value is less than the given significance level α, refuse the original hypothesis.

First, the dependent variables and independent variables in the multiple linear regression model are determined, then the parameters are estimated, and the prediction model is tested by regression test statistics.

3. Results and Discussion

3.1. Data Pre-processing

Firstly, the relevant data set is found in Kaggle, and the relationship between the length, width and height of diamonds and the price is analyzed by using the data set, so as to accurately measure the impact of various factors on the price of diamonds. Obviously, it is impossible to reach a conclusion by measuring the price of diamonds only by one standard. Therefore, the author combines other data in the data set to analyze the relationship between carat, cut, color, clarity, depth, table and price of diamonds, and gives scatter plots and other charts. In the data, the minimum value of "x", "y" and "z"

is zero, but obviously these are wrong values and need to be excluded from the calculation. After excluding invalid data, there are 37959 non-null values in all the attributes thus no missing values.

The following figures (Figure 2, 3 and 4) show the categorical features. Since the "cut", "color", and "clarity" of a diamond are "objects", data preprocessing is required, during which they are converted into numerical variables and then fed to the algorithm.

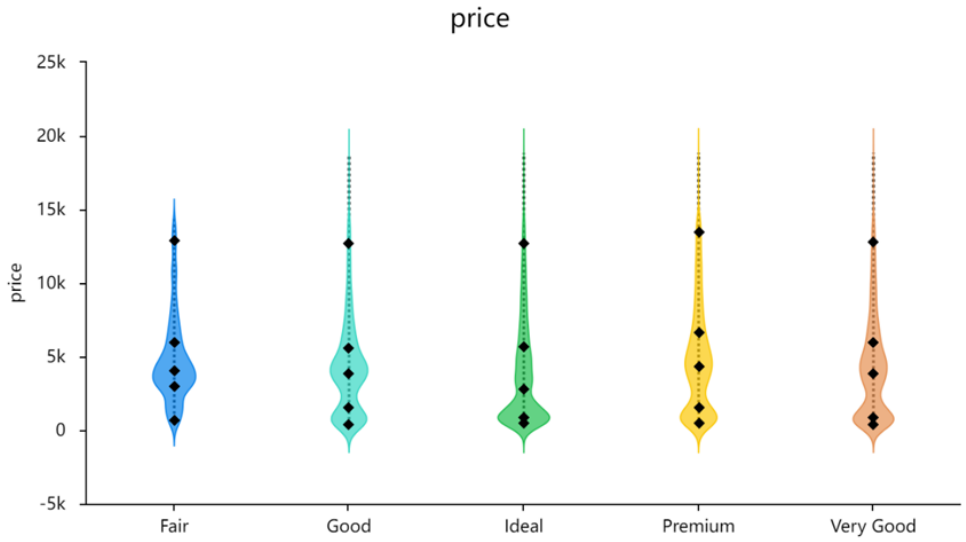

Fig. 2 Diamond cut for price

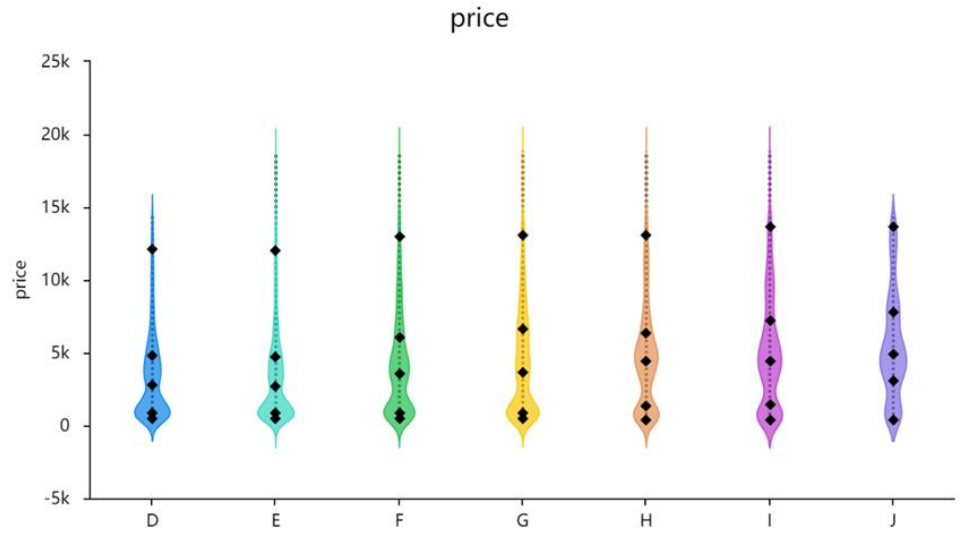

Fig. 3 Diamond colors for price

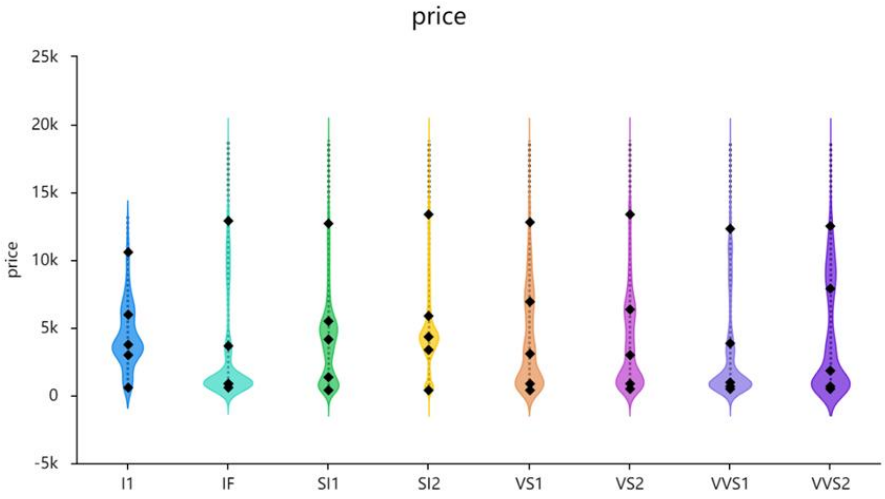

Fig. 4 Diamond clarity for price

It can be seen from observation that the highest price of the level “fair” is lower than other levels; diamonds in “D" and “E "colors are more abundant and of lower quality, while diamonds in the" J "color class are the worst quality and least abundant. Both the best and the worst clarity diamonds are the least.

After classifying the features and converting them into sequence values, we can now use the pre-processed data analysis to build a correlation matrix to make the data set cleaner before building the model and entering the algorithm.

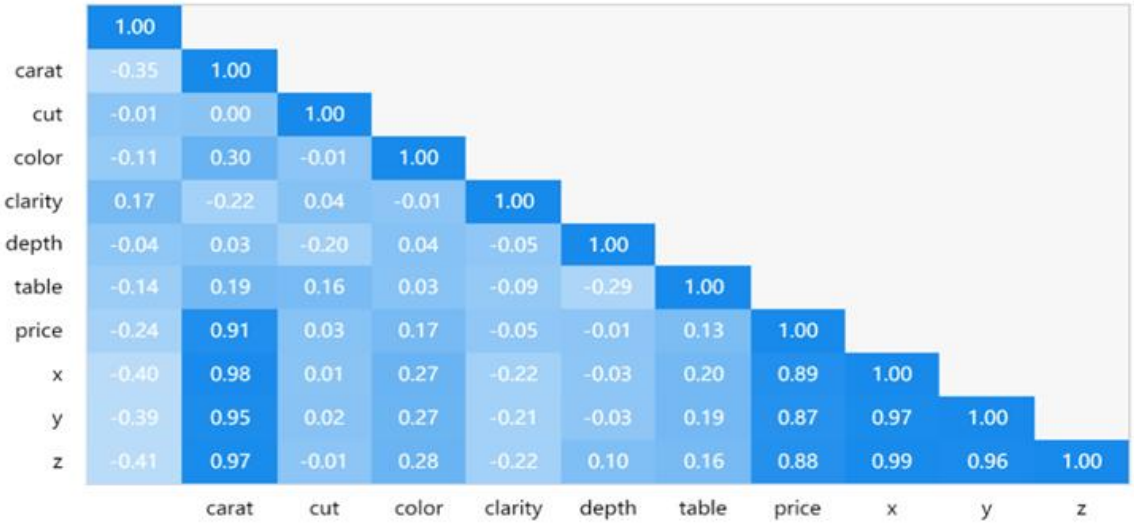

Fig. 5 Correlation matrix

It can be seen that "x "," y", "z", and “carat” are highly correlated with the variable "price ", while "cut", "clarity" and "depth" are not very correlated with the price, but they still have retention value due to fewer selected features (Figure 5).

3.2. Model Evolution

After the data visualization is completed, it is urgent to build a price prediction model. First, confirm the independent and dependent variables in the mode. Next, establish standard scalers and model pipelines for various regressor. Then fit all the models on training data, the mean value of cross-validation of training sets of all models is negative root mean square error. Select the model with the highest cross-validation score, and get the best model. After the calculation, it will obtain, r^2, adjusted r^2, mean squared error, mean absolute error and r-squared. It can be found that the model can predict the price of diamonds accurately.

Although this model can evaluate the price of an individual diamond, it cannot predict the value of the diamond on the vertical axis of time, because the diamond industry is not linked to the currency exchange rate like gold, and its future price prediction can only be judged according to the diamond market, so it is necessary to consider the future diamond price and collect the industry situation of the diamond industry at different times. Although this model can evaluate the price of an individual diamond, it cannot predict the value of the diamond on the vertical axis of time, because the diamond industry is not linked to the currency exchange rate like gold, and its future price prediction can only be judged according to the diamond market, so it is necessary to consider the future diamond price and collect the industry situation of the diamond industry at different times (Table 2).

Table 2. Linear regression analysis results (n=37958)

	Nonnormalized coefficient		Standardization coefficient	t	p	collinearity diagnostics	
	B	SE	Beta			VIF	Tolerability
constant	12240.201	544.296	-	22.488	0.000**	-	-
carat	8917.247	73.702	1.139	120.990	0.000**	26.317	0.038
cut	72.395	6.442	0.021	11.237	0.000**	1.056	0.947
color	-262.989	3.999	-0.128	-65.763	0.000**	1.121	0.892
clarity	333.331	3.998	0.158	83.375	0.000**	1.062	0.942
table	-93.892	3.148	-0.059	-29.823	0.000**	1.160	0.862
depth	-135.270	7.263	-0.055	-18.626	0.000**	2.568	0.389
x	-651.303	59.833	-0.204	-10.885	0.000**	104.082	0.010
y	49.595	24.949	0.016	1.988	0.047*	18.868	0.053
z	223.067	87.166	0.043	2.559	0.010*	84.811	0.012
R^2	0.872						
Adjusted R^2	0.872						
F	F (9,37948) =28789.488, p=0.000						
D-W value	1.120						

dependent variable: price. * p<0.05 ** p<0.01

From the table 2 above shows that the carat, cut, color, clarity, table, the depth, the x, y, z as independent variables, and the price as the dependent variable linear regression analysis, can be seen from the chart, the model formula is: price=12240.201 + 8917.247*carat + 72.395*cut-262.989*color + 333.331*clarity-93.892*table-135.270*depth-651.303*x + + 223.067 * 49.595 * y z, model R square value is 0.872, means the carat, cut, color, clarity, table, the depth, the x, y, z can explain 87.2% of the price change. It was found that the model passed the F test (F=28789.488, p=0.000< 0.05), which shows the carat, cut, color, clarity, table, the depth, the x, y, z in at least one affect price relationship. In conclusion, carat, cut, clarity, y and z have a significant positive influence on price. And color, table, depth and x will have a significant negative influence on price.

4. Conclusion

In this study, multiple linear regression was used to forecast the diamond price level. The model is trained and tested on the same dataset, and its performance is evaluated using relative test mean

square error (MSE) and test mean absolute error (MAE). This paper uses multiple price measures of diamonds, including "table" "depth" "cut" "colors" "clarity". The results show that the algorithm can be a more accurate tool to predict diamond prices and has good feasibility. The algorithm has a high accuracy in predicting diamond prices, providing a good measure for jewelers, designers and diamond consumers.

However, the model also has some limitations. Only the price of the diamond itself is considered horizontally during the data set collection process, and whether the diamond will depreciate due to the appearance of lab-grown diamond is not considered. In addition, the model simply measures the value of the diamond itself, not whether the overall price of the diamond in the market fluctuates.

References

[1] Marcin Potrykus. Diamond investments-Is the market free from multiple price bubbles. International Review of Financial Analysis, 2020, 83(1): 1-9.

[2] Russell Sho. What causes jewelry prices to rise. Journal of Gems and Gemmology, 2003, 13(1): 1-14.

[3] Wang Tongxu. Can the price of natural diamonds be "forever circulated". China Business Daily, 2023.

[4] Guo Sheng, Xie Bozhi. Cultivating diamonds: The importance of building a value system with brands. Journal of Gems and Gemmology, 2021, 23(6): 84-89.

[5] Li Fengfeng, Liu Yongjia, Zhang Jianhua. Development status and application research of cultivated diamond, Superhard Material Engineering, 2022, 34(4).

[6] Mo Mo, Liang Weizhang, Luo lei, Lu Zhixin. The development trend of cultivated diamond market is analyzed from the construction of international cultivated diamond brand, Journal of Gems and Gemmology, 2016, 18(6): 42-52.

[7] Zhou Qinghan. Research and implementation of diamond jewelry price estimation method based on OCR and big data technology. Jianghan University, 2023.

[8] Xu Rupeng. Grading criteria and evaluation of jewelry. Shanghai Measurement and Testing, 2006, 2.

[9] Mo Mo, Liang Weizhang, Luo lei, Lu Zhixin. The development trend of cultivated diamond market is analyzed from the construction of international cultivated diamond brand, Journal of Gems and Gemmology, 2016, 18(6): 42-52.

[10] Bao Xue, Chen Hua, Li Qi. A case study of the impact of "4C" on the price of colored yellow diamonds. Gemology and Technology, 2015.

[11] https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction, created by KANIKA KAPOOR, which used python to do research.