Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/46/20230347

Prediction and Feature Importance Analysis for Diamond Price Based on Machine Learning Models

Hairong Zhang[1,a,*]

[1]*Finance, South China Normal University, Tianhe District, Guangzhou City, Guangdong Province, China*
a. zhanghairong@hnsfdx.wecom.work
**corresponding author*

Abstract: The advent of Artificial Intelligence (AI) has facilitated the prediction of diamond prices through data analysis techniques. By incorporating relevant data, various models were constructed to examine the interrelationships between different factors and subsequently forecast diamond prices, which were then subjected to rigorous verification. The findings revealed that the XGBoost model demonstrated superior performance, exhibiting a high coefficient of determination (R square) and a low Root Mean Squared Error (RMSE). Furthermore, employing the feature importance method elucidated the significance of specific factors in determining diamond prices. Notably, carat weight emerged as the most influential factor, followed by width, clarity, and color. Conversely, other factors exhibited a lesser impact on price determination. These findings provide valuable insights for stakeholders in the diamond industry, enabling them to prioritize the most influential factors when assessing and forecasting diamond prices. Future research endeavors could explore additional data sources and advanced AI techniques to further enhance the accuracy and comprehensiveness of diamond price predictions.

Keywords: diamond price prediction, machine learning, feature importance

1. Introduction

The COVID-19 pandemic has exerted a profound influence on the demand for diamonds, resulting in significant price fluctuations, surpassing a 10% variation in early 2020. Consequently, there arises a need to ascertain the intrinsic value of diamonds, independent of the perturbations arising from demand and supply dynamics. In addition, diamonds have the most transparent prices of all jewelry. Rapaport Diamond Report is the most authoritative quotation system around the world, which stipulated the 4C standard (cut/clarity/cut/color) of diamonds. This framework provides a reliable and comprehensive guideline for the evaluation and determination of diamond valuations in the market.AI is able to predict future diamond price using passed data, inserting impacted factors and building statistical models. This study builds a kind of model first, with the help of python, then to find the best one in predicting diamond price.

In recent years, machine learning has developed rapidly, and various methods have emerged, such as ARIMA, random forest, IGM-BP etc., being applied to different fields including stock price prediction, house price prediction, and medical diagnosis. There is also a lot of related work on the prediction of luxury prices. For example, about gold price prediction, there is a lot of research. In the

DOI: 10.54254/2754-1169/46/20230347

early research, considerable efforts were devoted to the prediction of gold prices, employing diverse methodologies. For instance, Wang et al. use SARIMA to predict gold futures price and get a low mean relative error [1]. He et al. uses ICEEMDAN method to decompose gold futures price, then use SSA-ELM to predict gold futures price and finally analyse the price and get a result [2]. This integrated approach achieved a better performance than independent one.The IGM-BP gold price prediction model used by Huang Qian was established by combining six international economic indicators, namely the broad effective exchange rate of the US dollar, the US CPI index, the crude oil price, the Dow Jones index, the US Federal funds rate and the US inflation index and this model can predict gold price with high accuracy [3]. In addition, Yuan used CEEMDAN-PCA-LSTM model and get a low RMSE [4]. They all make great predictions about gold futures prices. However, none of the aforementioned studies have comprehensively compared multiple methodologies within a single paper, nor have they emphasized the significance of feature importance within a given model.

This study compared 10 models such as LinearRegression, Decision Tree and Support Vector Regression using python and found the optimal model is XGBRegressor. Then this study evaluated xgb model with Adjusted, Mean Absolute Error (MAE), Mean Squared Error MSE, Root Mean Square Error (RMSE), and found R square is higher than 98%, so this model is feasible. Finally, feature importance in this model was obtained.

2. Method

2.1. Dataset Description and Preprocessing

The dataset utilized for this study originates from the Kaggle platform, specifically from the source provided [5]. It comprises a total of 53,940 data points, each associated with 10 distinct features. These features are denoted as carat, cur, clarity, color, price, depth, table size, and the dimensions x, y, and z, where x represents the length, y represents the width, and z represents the depth of the diamond in millimeters.

The primary objective of this dataset is to explore and evaluate various models for predicting diamond prices. The dataset was divided into two subsets for this purpose: a training set consisting of 80% of the data and a test set comprising the remaining 20%. The goal is to identify and select the optimal model that achieves the most accurate predictions of diamond prices.

2.1.1. Preprocessing

Firstly, the pairplots are made to observe the relationship of the factors. However, there is some data disturbing us to find a broad view of the relationship between factors. These isolated data are named “outliers”. The data that “depth” is more than 75 or less than 45 is useless, so that for table size is more than 80 or less than 40; y is greater than 30; z is greater than 30 and z is greater than 2. These data are outliers in this data set, they need to be moved out.

For non-statistics factors, violin matrices are made to show the relationships shown in Figure 1. From the violin matrix for cut, it shows that most ideal diamonds have low prices while fair diamonds have higher prices concentrated in around 3000. The better the cut, the higher the price is. From the chart for clarity, SI1, SI2, I1have highest prices, and for VS1 VS2 WS1 WS2 and IF, prices mainly concentrate on around 1000. However, for color, the prices don’t show munch differences.

DOI: 10.54254/2754-1169/46/20230347

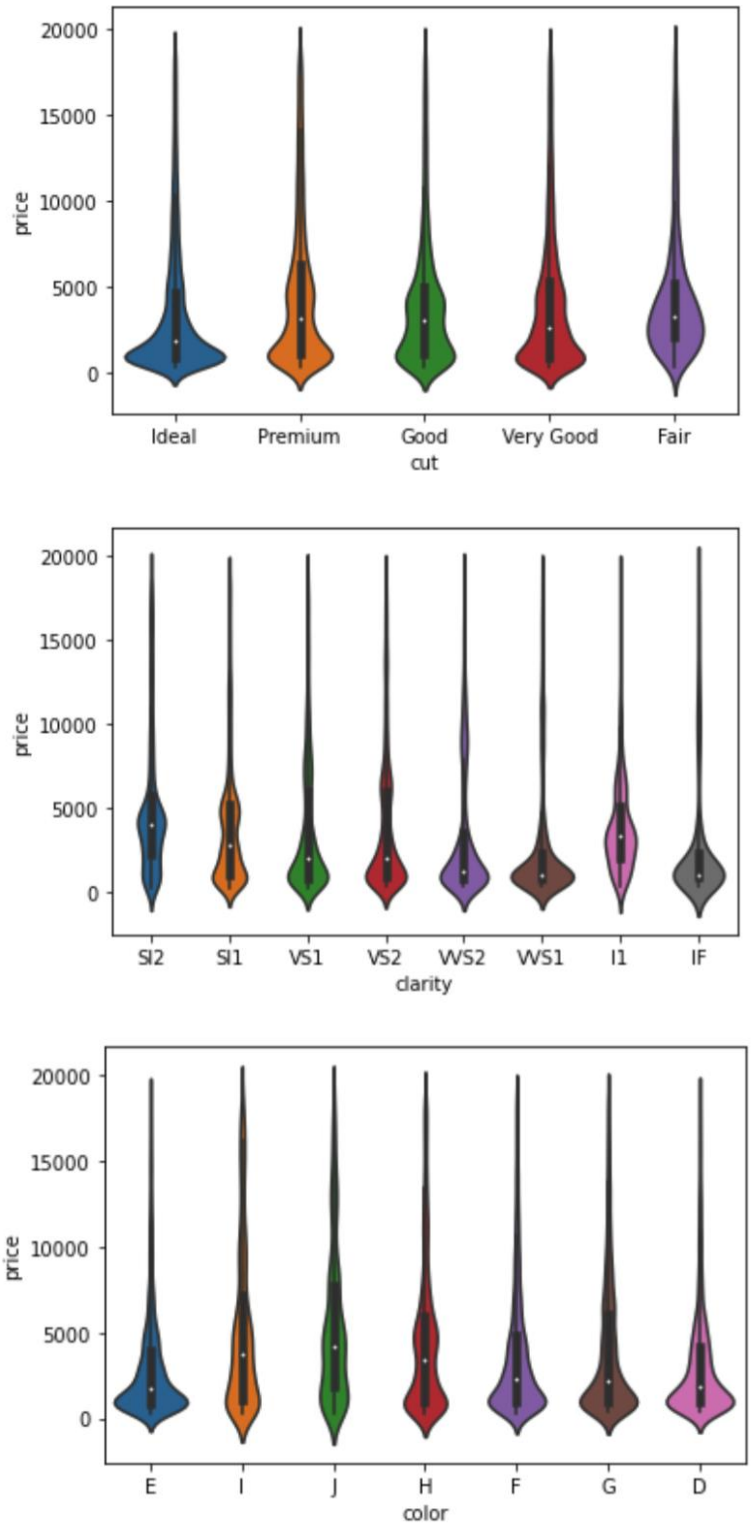

Figure 1: The relationship of the factors (Photo/Picture credit: Original).

2.1.2. Correlation Matrix

In order to better and clearly show the relationship among these factors, correlation matrix is made to illustrate the correlation expressed in statistics shown in Table 1. The highest relationship is measured

DOI: 10.54254/2754-1169/46/20230347

by 1/-1, while the lowest is 0. It shows that carat, x, y, and z have high linear relationship (>90%) with the price. However, the other factors have very low linear relationship (<20%).

Table 1: Correlation matrix for relationship between variables.

	carat	cut	color	clarity	depth	table	x	y	z
carat	1.0	0.018134	0.291480	-0.213628	0.028503	0.185768	0.978713	0.977827	0.977947
cut	0.018134	1.0	0.000177	0.028091	-0.196646	0.175117	0.022343	0.028198	0.001088
color	0.291480	0.000177	1.0	-0.027710	0.049006	0.027511	0.270522	0.270314	0.275133
clarity	-0.213628	0.028091	-0.027710	1.000000	-0.053462	-0.089573	-0.225678	-0.222734	-0.229760
depth	0.028503	-0.196646	0.048006	-0.053462	1.000000	-0.298641	-0.024556	-0.027702	0.097157
table	0.185768	0.175117	0.027511	-0.089573	-0.298641	1.000000	0.199350	0.193502	0.159510
price	0.940755	0.040021	0.178233	-0.083677	-0.005483	0.142084	0.925321	0.926832	0.921269
x	0.978713	0.022343	0.270522	-0.225678	-0.024556	0.199350	1.000000	0.998656	0.991709
y	0.977827	0.028198	0.270314	-0.222734	-0.027702	0.193502	0.998656	1.000000	0.991372
z	0.977947	0.001088	0.275133	-0.229760	0.097157	0.159510	0.991709	0.991372	1.000000

2.2. Machine Learning Model

In this study, several machine learning models called LinearRegression, Lasso, DecisionTree, RandomForest, KNeighbors, XGBRegressor, ElasticNet, RidgeCV, GradientBoostingRegressor and Support Vector Regression were utilized. Their performance have been demonstrated in many studies [6-10]. More detailed information about these algorithms can be found as follows: 1) Linear regression is a regression analysis which uses a least square function to model relationships between one or more independent and dependent variables. 2) Lasso is a method with the idea that when optimizing the objective function, not only the fitting degree of regression coefficient should be considered, and also consider the absolute value of regression coefficient, so as to achieve feature selection. 3) Decision tree is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve. 4) Random forest is a method that fits a number of dependent decision tree classifiers and uses averaging to improve the predictive accuracy and control overfitting. 5) KNeighbors stands that the sample belongs to a class when the majority of the k nearest samples belong to the class. 6) XGBRegressor is a model structure in the Xgboost library for regression problems. It is an integreated learning method based on decision tree, which can deal with high dimensional and sparse data. 7) ElasticNet combines the regularitization method of Lasso regression and ridge regression and controls the size of penalty term by two parameters seta rou. 8) RidgeCV is repeated CV for ridge regression description performs repeated cross validation to evaluate the result of Ridge regression where the optimal Ridge parameter lambda was chosen on a fast evaluation scheme. 9) GradientBoostingRegressor is gradient enhancement of regression. In GB,

DOI: 10.54254/2754-1169/46/20230347

the additive model is established in the way of forward stage. 10) SVR is a supervised learning algorithm used to predict discrete values, to find the best fit line.

3. Results and Discussion

3.1. The Performance of Different Models

The best model to predict the diamond price is XGBRegressor, because it has the lowest root mean squared error (323.1735666) which means that the model can predict the price with smallest errors. In addition, the R square for xgbregressor is 98.84% shown in Table 2, which means that imitative effect is great.

Table 2: The performance of different models.

Models	MSE
LinearRegression	879.541973
Lasso	880.825646
DecisionTree	447.394141
RandomForest	326.890693
KNeighbors	503.481810
XGBRegressor	323.173566
ElasticNet	1040.324532
RidgeCV	879.506014
GradientBoostingRegressor	395.257697
SVR	1727.052697

3.2. Feature Importance of Models

From Figure 2, “carat” is the most important factor in this model, taking up over 70% of feature importance. This, “y” is the second important factor in the model, and it took up nearly 20% of feature importance. The residual importance mainly consists of clarity, color and “x”.

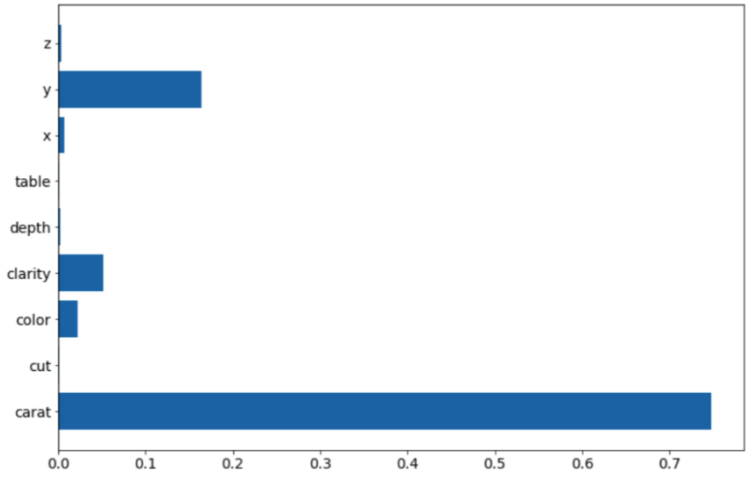

Figure 2: The feature importance from the model (Photo/Picture credit: Original).

The model shows that price is most highly related to “carat”. It has a high coefficient with the price. Then, “y”, clarity and color are also important considerations. However, cut, table size and depth have low relationship with price. They contribute little to the price estimation.

DOI: 10.54254/2754-1169/46/20230347

4. Conclusion

This research employs a comprehensive analysis to evaluate various models for predicting diamond prices, highlighting the XGBoost model as the most effective. The model consistently exhibits the lowest RMSE and a high coefficient of determination (R square) exceeding 98%. Moreover, both the Random Square and Gradient Boosting Regressor models demonstrate favorable performance, characterized by low RMSE values. Notably, the analysis identifies carat as the most significant feature in diamond price prediction, exerting a substantial influence on the forecasted prices. Furthermore, width, clarity, and color also emerge as important factors contributing to accurate price predictions. These findings contribute valuable insights for stakeholders in the diamond industry, enabling them to prioritize essential features when analyzing and forecasting diamond prices. Future research could explore advanced modeling techniques and additional factors to further enhance the precision and comprehensiveness of diamond price predictions.

References

[1] *Wang, P., et al. (2022) Gold Price prediction based on SARIMA model (in Chinese). Electronic Technology and Software Engineering, 233-237.*

[2] *He, L. (2023) Gold Futures price prediction based on ICEEMDAN-SE-SSA-ELM algorithm (in Chinese). Journal of lanzhou liberal arts college (natural science edition), 35-39. DOI: 10.13804 / j.carol carroll nki. 2095-6991.2023.01.013.*

[3] *H, Q., et al. (2022) Research on gold price prediction based on IGM-BP model (in Chinese). China Price, 87-89.*

[4] *W, Jin., (2022) Research on Gold Futures price prediction based on CEEMDAN-GRU (in Chinese). Beijing jiaotong university. DOI: 10.26944 /, dc nki. Gbfju. 2022.001336.*

[5] *Kaggle (2021) Retrieved from https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction*

[6] *Qiu, Y., Chen, P., Lin, Z., et al. (2020) Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 1: 1023-1027.*

[7] *Qi, Y. (2012) Random forest for bioinformatics, Ensemble machine learning: Methods and applications. Boston, MA: Springer US, 307-323.*

[8] *Speiser, J. L., Miller, M. E., Tooze, J., et al. (2019) A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134: 93-101.*

[9] *DeMaris, A. (1995) A tutorial in logistic regression. Journal of Marriage and the Family, 956-968.*

[10] *Hilbe, J. M. (2009) Logistic regression models. CRC press.*