Research on the Diamond Price Prediction based on Linear Regression, Decision Tree and Random Forest

Zhe OuYang*

Department of Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

* Corresponding Author Email: r130026114@mail.uic.edu.cn

Abstract. Diamonds are the symbol of the pure and indestructible love and the luxury that people have always sought after. However, because people know less about diamonds, they often only rely on the introduction of salespeople and jewelers in diamond trading. Therefore, it is difficult for consumers to buy diamonds of equal value and price. To solve this problem, this paper uses Multiple Linear Regression model, Decision Tree Regression model and Random Forest Regression model to predict diamond prices based on various diamond evaluation metrics in data set, so that consumers can intuitively learn about the normal price of the evaluation metrics of selected diamonds. Through this paper, it is found that the Random Forest Regression model has the best fitting and predictive ability in diamond prediction task, which is also the most recommended model.

Keywords: Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, Diamond Price Prediction.

1. Introduction

Diamond is the hardest gemstone in the world. This kind of gemstone is generally pure and colorless, which needs to be formed in an oxygen-deficient environment under high temperature and extremely high pressure. Therefore, diamonds have the characteristics of beauty and rarity, making it be the symbol of the pure and indestructible love and the luxury that people have always sought[1]. However, in real life, diamond buyers have little understanding of diamond-related evaluation indicators such as cut, clarity, carat, color, etc., they can only obtain related information from the introduction of salespeople or jewelers. This makes consumers have difficulty in learning about the true value of diamonds due to the single source of diamond information. Hence, it is often hard for consumers to buy the diamonds worth the prices[1]. To solve this problem, this paper adopts the Multiple Linear Regression model, the Decision Tree Regression model and the Random Forest Regression model based on the data set obtained from GIA trading website to predict the price. Meanwhile, by comparing all models based on several model evaluation metrics, the model with the best performance in diamond prediction task will be recommended to diamond buyers. In this way, consumers can know the normal price of the selected diamond based on the recommended model and past data when purchasing diamonds.

2. Literature Review

Present studies have used various models to predict diamond prices based on diamond datasets. According to the study of Waad, the Linear Regression model, Random Forest model, Gradient Boosting Regressor, Polynomial Regression and Neural Network were built to do the diamond prices prediction. The study found that in the case of noisy data set, the Random Forest model has the smallest RMSE and MAE which is the best performance[2]. The study implemented by Garima, various supervised models such as Random Forest model, Decision Tree model, LASSO Regression, etc., were used to predict diamond prices. By calculating and comparing the RMSE, CV_RMSE and accuracy of each model, it was found that the Random Forest model has the smallest RMSE, CV RMSE and the largest accuracy[3]. In addition, two machine learning models: (KNN) K-Nearest

Neighbors and LASSO Regression were used by Shafilah[4]. The final results showed that compared to LASSO Regression model, the KNN model has higher accuracy and the smaller RMSE, which indicated that the KNN model has better fitting and predictive ability. Besides, Kigo used a variety of supervised machine learning models like Linear Regression model, eXtreme Gradient Boosting, Random Forest, K-Nearest Neighbors, etc. By comparing several model evaluation metrics, it was found that eXtreme Gradient Boosting was the optimal algorithm[5].

3. Data and Methods

3.1. Samples and Variables

GIA Diamond is an abbreviation for describing diamonds graded by Gemological Institute of America (GIA). Gemological Institute of America (GIA) is an independent non-profit organization whose main function is to evaluate and describe the quality of diamonds based on the 4C standards: color, clarity, cut, and carat. Gemological Institute of America is the most authoritative diamond research and evaluation organization. Therefore, in order to analyze and predict the price data of diamonds, this paper uses the data from GIA Diamond Trading Website[6] which contains the most GIA diamonds to do the analysis and prediction. This diamond data set has a total of 16703 samples. In this data set the variables: cut, clarity, color are categorical variables and the rest are continuous variables. The detail information of the data set is shown in Table 1.

Table 1. Description of values and definitions of all variables in diamond data set

Variables	Definitions	Values
Price (RMB)	Price in RMB	1353.31 - 976295
Carat weight	Weight of the diamond	0.3 - 81
Color	An indicator used to measure the apparent size of a diamond	E, M, K, L, D, J, I, H, G, F
Cut	Quality of the cut, which is an important indicator for the evaluation of the valuable of diamond	Fair, Excellent, Good, Very Good
Clarity	The degree to which it is free of inclusions	SI2, SI1, VS1, VS2, VVS1, VVS1, I1, IF, FL
Depth	The depth of diamond	50.0928 – 80.0
Table	Width of top of diamond relative to widest point	32.5 – 78.0

3.2. Data Preprocessing

3.2.1 Handling outliers

The box-plot is first used to do the detection for all data of every continuous variable. The results are shown in figure 1- figure 3. In these three figures, the outliers of the corresponding variable are shown by red points, and the box are shown by a black box. Based on the detection of the above box plots, it is found that there are a total of 1145 samples are outliers which should be deleted.

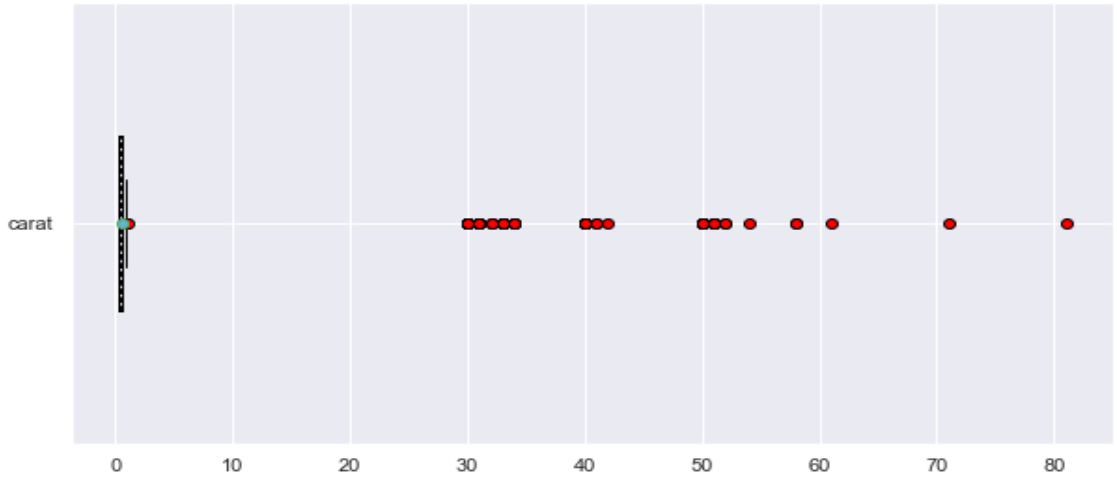

Fig. 1 Box-plot of carat

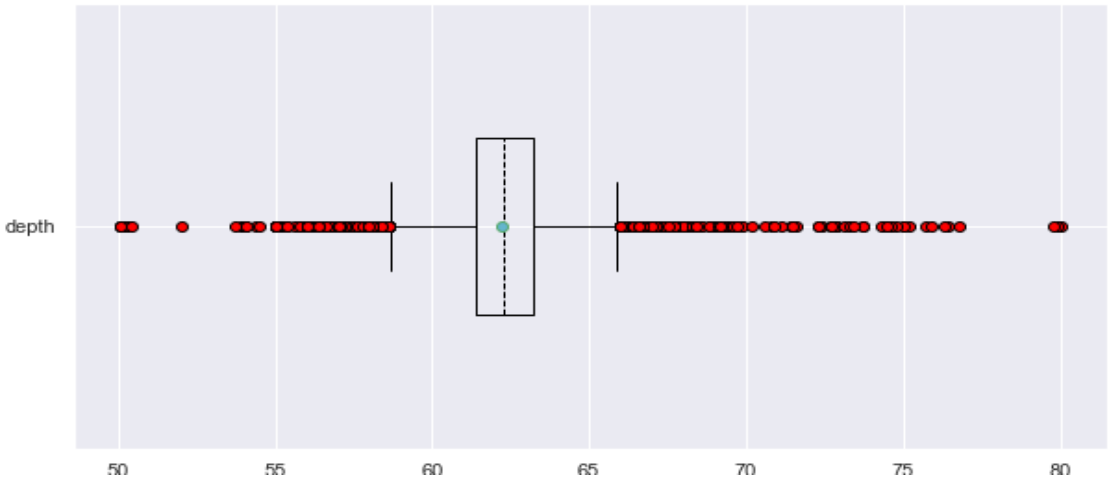

Fig. 2 Box-plot of depth

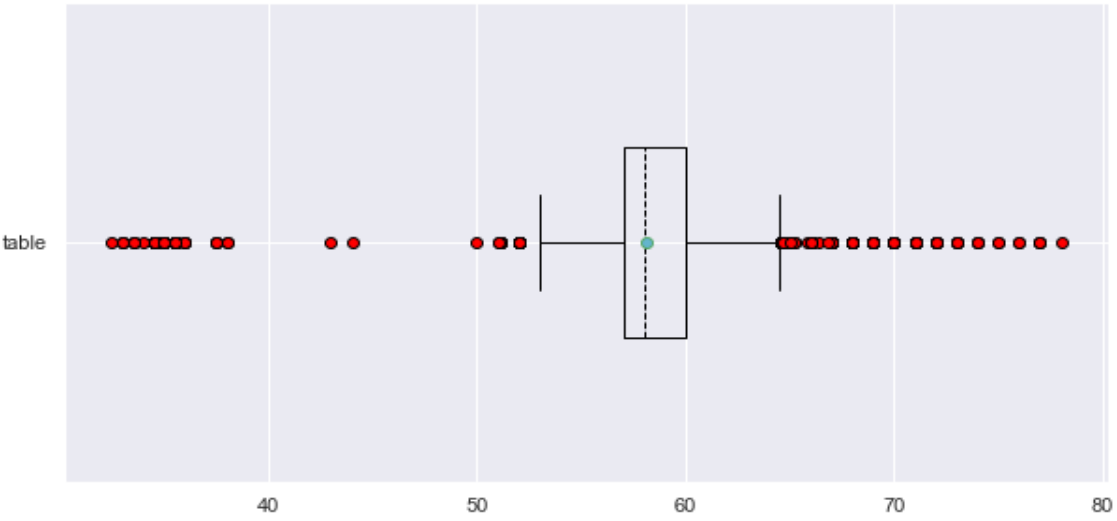

Fig. 3 Box-plot of table

3.2.2 Transformation for Multiple Linear Regression model

The Box-Cox transformation can make the distribution of data closer to the normal distribution by calculating the parameters lambda that controls the degree and the way of transformation. Therefore, in order to satisfy the assumptions of the Linear Regression model and let the parameters estimated by least squares method reach the optimal value[7], this paper chooses to use Box-Cox transformation to perform appropriate transformations on the data.

Python is used to calculate the parameters required for each continuous variable in the Box-Cox transformation. All lambdas and their corresponding variables are shown in the Table 2. The lambda of price is 0.0237, the lambda of carat is -1.0044, the lambda of depth is 4.6974, the lambda of table is -4.1735.

Table 2. Lambda of all continuous variables in diamond data set

Variables	Lambda
Price	0.0237
Carat	-1.0044
Depth	4.6974
Table	-4.1735

After obtaining lambda for every continuous variable, the transformation for each variable is performed based on the following equation.

$$y_i^{\lambda} = X\theta + \varepsilon = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}; & \lambda \neq 0 \\ log y_i\,; & \lambda = 0 \end{cases} \quad (1)$$

Where $\boldsymbol{y}^{(\lambda)} = (y_1{}^{(\lambda)}, y_2{}^{(\lambda)}, \ldots, y_n{}^{(\lambda)})'$; X is a matrix of known data of each feature; $\boldsymbol{\theta}$ is a vector of unknown parameters; $\boldsymbol{\varepsilon}$ is a vector of error which is the difference between observed value and predicted value

3.2.3 The splitting of data set

Splitting data set into training set and testing set is useful to train the model during the process of model building and evaluate the performance of model[8]. Therefore, in this paper, the diamond data set is split into training set(70%) and testing set(30%).

3.2.4 Transformation for categorical variables

In order to build the Multiple Linear Regression model, Decision Tree model and Random Forest model, python is used to transform all categorical variables which are cut, clarity and color to dummy variables. After finishing the transformation, the new diamond data set have 26 variables in total.

3.3. Correlation Analysis

The correlation matrix is composed of the correlation coefficient between the columns of the matrix, and the heat map can give different shades and colors based on different correlation coefficients, which can also useful to show the strength of the correlation between different variables. Therefore, to analyze the data more reasonably, the correlation matrix of the four continuous variables which are price, carat, depth, table is plotted before the models are built.

According to the correlation matrix in figure 4, the correlation coefficient between each feature is low indicating that the correlation between each independent variable is weak. In addition, for the correlation coefficient between the features and the dependent variable price, the correlation coefficient value of carat against price is 0.89 which is the largest correlation coefficient. Hence, the feature carat has the strongest correlation with price, which implies that the carat may be significant feature in the process diamond prediction.

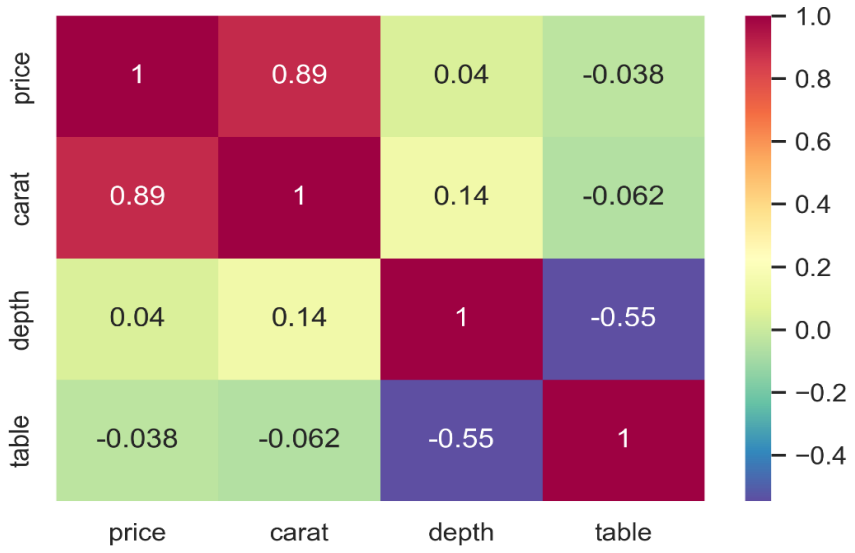

Fig. 4 Correlation Matrix of features in diamond data set

3.4. Methodology

This paper builds three prediction models which are Multiple Linear Regression Model, Decision Tree model and Random Forest model so as to do the prediction of diamond price.

3.4.1 Multiple Linear Regression model

Linear Regression model is mainly used to study the linear relationship between independent variables and dependent variables in a data set. For those Linear Regression models that based on the data sets with multiple dependent variables or independent variables are called Multiple Linear Regression model[9].

In the process of building Multiple Linear Regression model, the least square method is used to do the parameter estimation and the forward selection is used to do the features selection.

The Multiple Linear Regression model is built based on the following equation

$$Y = X\beta + \varepsilon \tag{2}$$

Where Y is a column vector of diamond prices; X is a matrix of known data of each feature; β is a column vector of unknown parameters; ε is a column vector of error values

The parameters are estimated by following equation

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3}$$

Where Y is a vector of diamond prices; X is a matrix of known data of each feature; $\hat{\beta}$ is a column vector of estimators of the unknown parameter vector β

The Multiple Linear Regression model predicts the diamond prices by the following equation

$$\hat{Y} = X\hat{\beta} \tag{4}$$

Where $\hat{Y}$ is the column vector of the predicted values of Y; X is a matrix of known data of each feature; $\hat{\beta}$ is a column vector of estimators of the unknown parameter vector β.

Finally, In the process of building a Multiple Linear Regression model, the forward selection method usually is used to select the optimal features according to several model evaluation criteria. The method begins with the selected model with only the dependent variable. In each round of selection, the algorithm will take out the variables which are not selected into the selected model one by one and form a new model with all the variables in the selected model. After calculating and

comparing the model evaluation criteria of all new models, the new features in the optimal model in this round of selection will be added to the selected model. The algorithm will not stop until all features have been added to the selected model[10].

To select a model with outstanding fitting and predictive ability, and without overfitting or underfitting, this paper sets the evaluation criteria for feature selection as follows: coefficient of determination (R squared), adjusted coefficient of determination (Adjusted R squared), mean absolute error of Cross Validation (CV_MAE), Mallow's Cp.

3.4.2 Decision Tree Regression

A decision tree is a supervised machine learning model that deals with classification or regression problems by finding and generating optimal split points and forming a tree structure. Because diamond price is a continuous variable, this paper uses the least squares regression tree algorithm in decision tree to make prediction about diamond prices[11].

There are 4 steps in generating a least square regression tree.

Step 1 is to find the optimal split point. The feature space which is a space includes all features in diamond data set needs to be divided into several cells. After that, the mean value of all observations in the subregion that produced by all split points in each cell is calculated. The equation is shown as,

$$\hat{c_m} = \frac{1}{N_m} \sum_{x_i \epsilon R_m(j,s)} y_i \ (m = 1,2) \tag{5}$$

Where $\hat{c_m}$ is the mean value of the observations in subregion; m is the subregion of the current cell; j is the split variable; s is the split point; N_m is the number of diamond samples in the current subregion

Step 2 is to select the optimal split variable and split point. To accomplish this, the square difference of two subregion is used as loss function which is calculated by following equation

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right] \tag{6}$$

Where c_1, c_2 are the mean values of the observations in two subregions; j is the split variable; s is the split point; N_m is the number of diamond samples in the current sub region

Step 3 is to select the optimal split variable and split point with the minimum loss value by comparing the loss value of each split point.

Finally, repeat the above steps so as to find all optimal split points, and the tree generation will not stop until the conditions are satisfied.

3.4.3 Random Forest Regression

The Random Forest Regression model is an ensemble machine learning method based on the Decision Tree Regression model, which can also be used to solve regression problems. This model aggregates a large number of decision tree during the generation process, and integrates the output of each decision tree so as to produce the final output of the model[12].

For the generation of Random Forest Regression model, the data will be split into several sub data set by sampling with replacement, and the steps to generate decision trees for all sub data set are as the same as those steps in Decision Tree Regression model generation.

3.4.4 Evaluation metrics

To select the model with the best performance in diamond prediction task, the root mean square error (RMSE) and Mean absolute error (MAE) are chosen to evaluate the predictive ability of models[13]. In addition, the Adjusted R Square (adjust R square) is also chosen as an evaluation criterion to evaluate fitting ability of the models with different number of features.

4. Results and Discussion

4.1. Result of Multiple Linear Regression Model

In order to make a preliminary judgment on the predictive ability of the Multiple Linear Regression model, the line plot is used. To make it easier to observe, only the first five hundred observations and their corresponding predicted values are used to plot the line plot. In the line plot, the observations and the corresponding predicted values which are predicted by Multiple Linear Regression model are plotted as line respectively. The result is shown in figure 5. The red line in figure 5 represents the data of the testing set which is used to do the prediction, and the blue line represents the predicted data obtained based on testing data.

According to the Line chart in figure 5, it is found that the blue line coincides with the red line in a small number, and most of the blue line coincides poorly with the red line. Therefore, the predictive ability of this model may not good.

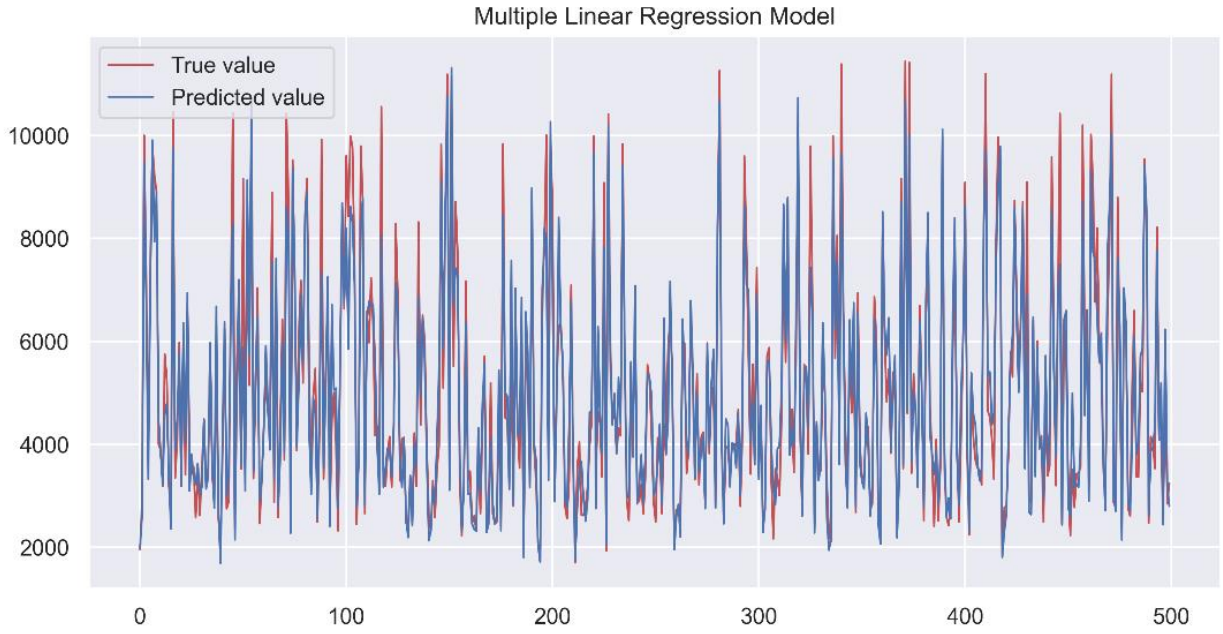

Fig. 5 Line chart of Multiple Linear Regression model

4.2. Result of Decision Tree Regression Model

The Line chart which includes the polylines of the first five hundred testing data and predicted data is plotted to do the preliminary judgement of the predictive ability of Decision Tree Regression model. The result is shown in figure 6.

Based on figure 6, it is easy to find that the number of the coincident between blue lines and red lines is higher than the number Multiple Linear Regression model, which indicates that the predicted ability of this model may be better than Multiple Linear Regression model.

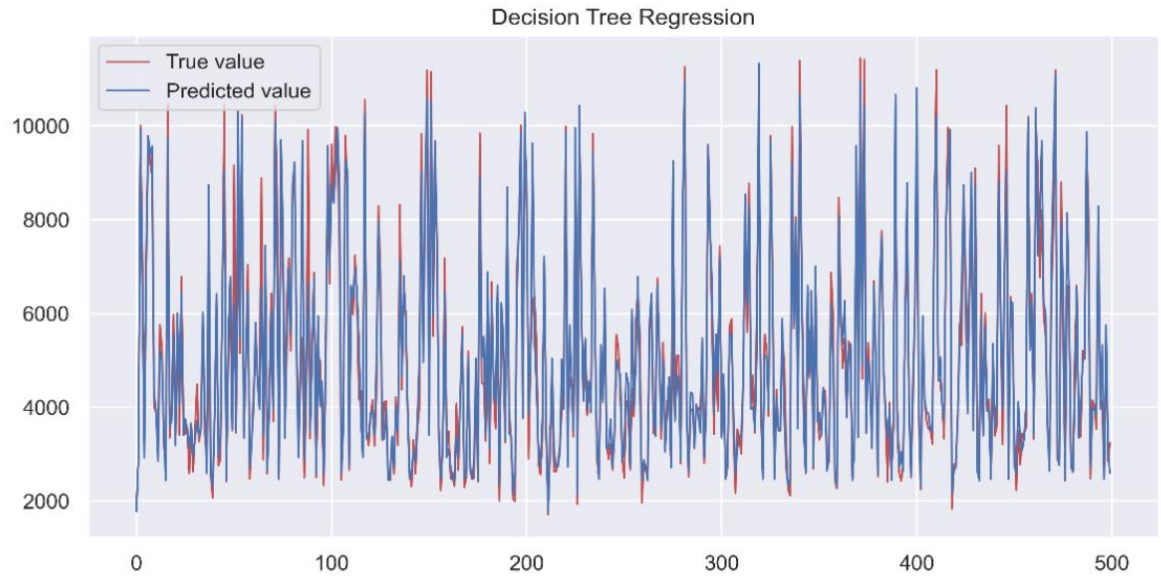

Fig. 6 Line chart of Decision Tree Regression model

4.3. Result of Random Forest Regression Model

The Line chart which includes the polylines of the first five hundred testing data and predicted data is also plotted so as to do the preliminary judgement of the predictive ability of Random Forest Regression model.

By observing the figure 7, it is found that the degree and the number of coincident between blue lines and red lines are similar to the graph in Decision Tree Regression model, which implies that the predictive ability of this model may be close to the Decision Tree Regression model.

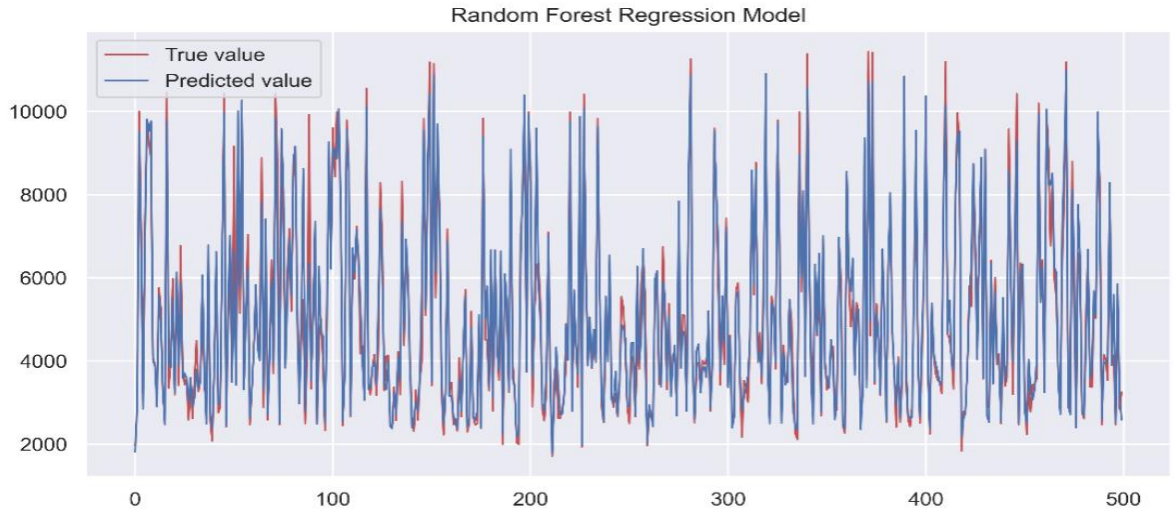

Fig. 7 Line plot of Random Forest Regression Model

4.4. Further Evaluation

This paper further evaluates the three models, the evaluation metrics: RMSE, MAE, Adjusted R square are used. The values of all evaluation criteria based on the three models are shown in table 3.

Table 3. Values of all evaluation metrics of three models

Models	MAE	RMSE	Adjusted R-square
Multiple Linear Regression	485.281446	705.486882	0.929392
Decision Tree Regression	420.774653	650.290445	0.932586
Random Forests Regression	340.754894	557.970444	0.953835

According to Table 3, it can be found that,

For adjusted R square, the Random Forest Regression model has the largest value of adjusted r square. In addition, the adjusted r square of Decision Tree Regression model is slightly larger than the Multiple Linear Regression model.

For MAE and RMSE, among three models, the Random Forest Regression model has the smallest value of these two criteria while the Multiple Linear Regression model has the largest value.

Above all, the Random Forest Regression model has the strongest fitting ability and predictive ability. However, although the rest two models have closer fitting ability based on their adjusted r square, the predictive ability of the Decision Tree Regression model is better than the Multiple Linear Regression model.

4.5. Discussion

Among three models, the result of the Random Forest model is obtained by considering multiple sub-decision trees, which can make it well avoid the influence of individual decision tree outliers and lead to inaccurate overall prediction results. In addition, the Decision Tree model is built with all features and samples, while the Random Forest model divides all samples by replacement sampling and builds decision trees for some samples, so that each decision tree in the model uses fewer features and data[14]. Therefore, the decision tree model is more prone to overfitting problems, which makes the model perform poorly in the test set.

For Multiple Linear Regression model, the model has several assumptions about the data set and individual variables, such as the need for a linear relationship between the independent and dependent variables. Hence, it is difficult for the model to capture nonlinear relationship. However, the data in real life is often impossible to meet all the assumptions of this model, which makes it difficult for the model to fit and predict the data well in the prediction task in real life.

5. Conclusion

This paper uses the diamond dataset to build a Multiple Linear Regression model, a Decision Tree Regression model and a Random Forest Regression model, and predicts the diamond prices based on the features in the dataset. By calculating and comparing the value of MAE, RMSE and adjusted r-square of these three models, it is found that the Random Forest Regression model is the best performing model in the diamond prediction task which is also the recommended model.

Through the research, diamond buyers and investors can more intuitively understand the normal price of selected diamonds, which can help them make more rational and wise diamond buying or investment decisions. Besides, by using the recommended model to predict the diamond prices based on various evaluation indicators, it is conducive to providing pricing guidance for diamond sellers, promoting more fair and reasonable diamond trading prices, and reducing unfavorable transactions caused by information asymmetry in diamond trading.

However, in addition to the evaluation criteria of the diamond itself, the price of a diamond can also be affected by various factors such as the rent of the diamond shop, the economic situation of the trading area or the supply and demand relationship of the diamond market. Therefore, in the future diamond price prediction task, more features can be added to the diamond data set so as to make it much closer to the realistic diamond trading scenarios. Meanwhile, the Random Forest model can be optimize based on the more comprehensive diamond data set so as to improve its predictive ability, or more machine learning models can be tried to find a better performance model in diamond prediction task.

References

[1] Matlins A. Jewelry & Gems. The Buying Guide: How to Buy Diamonds, Pearls, Precious and Other Popular Gems with Confidence and Knowledge. Springer Science & Business Media, 2012.

[2] Alsuraihi W., Al-Hazmi E., Bawazeer K., et al. Machine learning algorithms for diamond price prediction. In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing, 2020, 150-154.

[3] Sharma G., Tripathi V., Mahajan M., et al. Comparative analysis of supervised models for diamond price prediction. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, 1019-1022.

[4] Fitriani S. A., Astuti Y., Wulandari I. R. Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction. In 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), 2022, 135-139.

[5] Kigo S. N. Assessing predictive performance of supervised machine learning algorithms: an alternative model for diamond pricing. Strathmore University, 2022.

[6] White Diamond Search. GIA Diamond Trading Website, http://gia.airland.vip/newManager/index.html

[7] Sakia R. M. The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society Series D: The Statistician, 1992, 41(2): 169-178.

[8] Picard R. R., Berk K. N. Data splitting. The American Statistician, 1990, 44(2): 140-147.

[9] Linear regression. Wikipedia, the free encyclopedia, 2023, https://en.wikipedia.org/wiki/Linear_regression

[10] Blanchet F. G., Legendre P., Borcard D. Forward selection of explanatory variables. Ecology, 2008, 89(9): 2623-2632.

[11] Li Hang. Statistical Learning Methods. Tsinghua University Press, 2019, 55-73.

[12] Random forest. Wikipedia, the free encyclopedia, 2023, https://en.wikipedia.org/wiki/Random_forest

[13] Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). 2019, https://resources.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.htm

[14] Fratello M., Tagliaferri R., Decision trees and random forests. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 2018.