Predict Diamond Price with Data Mining Concepts and Techniques

Jiaxing Gao

ECE537 Final Project

1. Introduction

1.1 Objectives

The aim of this project is to find an appropriate model to predict the diamond's price based on the data from internet. There data mining algorithms will be applied to predict the diamonds, Linear Regression, Classification and Regression Tree (CART) and Neural Network, the use of these methods will be compared by the end of the report.

1.2 About Diamond

A Diamond is one of the best-known and most sought-after gemstones especially valued for their rarity and optical properties [1]. Diamond's price influenced by several factors, Carat, Color, Clarity, Cut, Shape and dimensions [2]. Those four words start with letter 'C' is the best-known diamond properties used in terms of 4C. Additional factors also included Symmetry and certification and so on.

Clarity: Refers to the absence of birthmarks or blemished on or inside the diamonds (ranking from best to worst: FL, IF, VVS1, ⋯, I2, I3) [2]

Color: Refers to the diamond how closely they approach colorlessness (ranking from best to worst: D, E, F, ⋯, Z) [2]

Carat Weight: Size of the Diamond. The carat equals to 0.2 grams, measured to the nearest hundredth of a carat [2]

Cut: Cut quality of the diamond which affect diamond's fire, sparkle and brilliance (ranking from best to worst: Ideal, Excellent, ⋯, Fair) [2]

Shape: Diamonds have variety of shapes, the most common of which is the Round, other common cuts include Oval, Emerald and so on. The shape category has no ranking system, people cannot say Round diamond is better than Emerald diamond, customer buy the diamond based on her own choice.

1.3 Data from Internet

The data is download from website of ultimatediamond.com. The dataset contains ten attributes of diamonds and totally 10457 groups of data. The first five groups of data as shown in Table 1. This data contains different types, features of Carat, Depth, Table and Price are numeric data, Color, Cut, Clarity and Symmetry are ordinal data, the Shape is nominal data and certificate is binary data.

Table 1 first 5 groups of data in the dataset

	Shape	Carat	Color	Clarity	Cut	Depth	Table	Symmetry	Cert	Price
0	Asscher	1.41	F	VVS1	Excellent	68.1	66.0	Very Good	GIA	9668
1	Asscher	1.18	I	VS1	Excellent	68.4	63.0	Excellent	GIA	4408
2	Asscher	1.83	I	VVS2	Excellent	66.6	61.0	Excellent	GIA	8220
3	Asscher	2.32	F	VVS2	Excellent	67.8	63.0	Excellent	GIA	22106
4	Asscher	1.01	I	VVS1	Excellent	70.5	67.0	Very Good	GIA	2901

2. Data Pre-processing

2.1 Data cleaning

2.1.1 Find missing values

After sorting the whole dataset, there are 50 missing data inside the dataset, and the feature which contains most missing value is Cut, it contains 15 missing values. However, comparing with the 10457 values, the 50-missing value is not crucial, so my strategy is delete the groups which contain missing values.

2.1.2 Identify Outliers

Outliers were identified though the standard – outliers normally above/lower than the 1.5x IQR. All the outliers in the numeric data are identified based on this standard, which about 200 data groups, again comparing with more than ten-thousands of data, so just drop these data. Fig 1 and Fig 2 can show the smoothness of the data after removing the outlier.

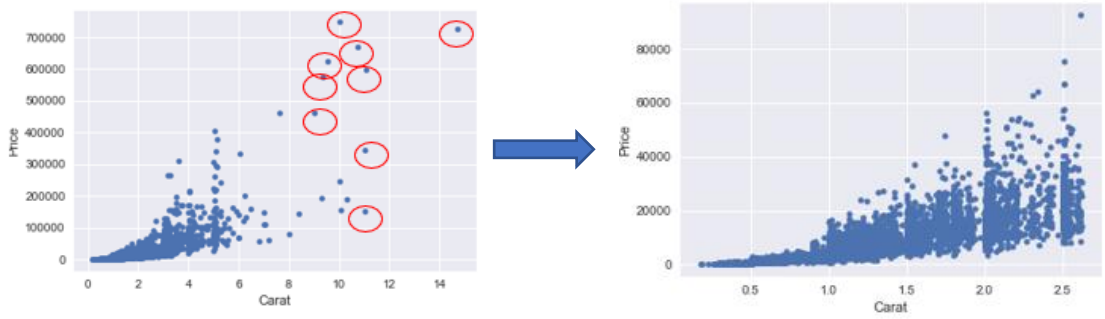


Fig 1 Carat vs Price before identifying outliers

Fig 1 After identifying outliers

2.2 Data Transformation

2.2.1 Data Mapping

Ordinal data has a meaningful order (ranking), to build the model, it needs to convert the spring data to numeric data. And the bigger the number is, the higher level of the data is. For example, the Clarity category has 11 levels, the I3 is the worst quality, and the FL is the best quality, so mapping the data from 0 to 10 to represent the Clarity quality of the diamond from bad to good. Table 2 shows the dataset after mapping.

Table 2 Dataset after mapping

Carat	Color	Clarity	Cut	Depth	Table	Symmetry	Cert
1.41	8	8	3	68.1	66.0	2	1
1.18	5	6	3	68.4	63.0	3	1
1.83	5	7	3	66.6	61.0	3	1
2.32	8	7	3	67.8	63.0	3	1
1.01	5	8	3	70.5	67.0	2	1

2.2.2 Data Transformation

From the Table 1, we can find that the majority of Carat data is around 1 or 2, and the highest before data cleaning is 15, however the Depth and Table data is around 68 or 70. They are not in the same scale, which will cause the high error especially when modeling with Neural Network and Linear Regression [4]. Therefore, Z-Score method applied here to help transfer the original data to scaled data. The Z-Score method function is $v' = \frac{v - mean_A}{Stand\_dev_A}$. Table 3 shows the before and after Z-Score transformation.

Table 3 Z-Score Transformation

	Carat (first 5)	**Depth (first 5)**	**Table (first 5)**
Before transformation	[1.14, 1.18, 1.83, 2.32, 1.01]	[68.1, 68.4, 66.6, 67.8, 70.5]	[66.0, 63.0, 61.0, 63.0, 67.0]
After transformation	[-0.033, -0.296, 0.447, 1.008, -0.4907]	[0.851, 0.913, 0.540, 0.789,1.348]	[0.88, 0.313, -0.066, 0.313, 1.07]

2.3 Split Dataset into Training and Testing Part

In order to evaluate the models' result, 65% of the original dataset randomly selected as training sample, the remaining observations being used as test sample. The predictor variable (X) is Carat, Depth, Table, Color, Cut, Clarity, Symmetry, Certificate and Shape. The response of the model (Y) will be price.

3. Modeling

Since there are complex relationships between the chosen predictors and the target price variable and consider the diverse measurement level associated with the predictors, three possible modeling methods are chosen to model the data, Linear Regression, Classification and Regression Trees (CART) and Neural Networks. They will be introduced as follow

3.1 Multiple Linear Regression Model

The linear regression method has an assumption that it assumes the predictor and the response has a linear relationship and this model sensitive to outliers [5] so the data pre-processing is important for this model. Generally, the equation of multiple linear equation as follow:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_p x_p + \epsilon$$

- Where θ_0, θ_1, …., θ_p represent the regression coefficients, which increase in *Y* associate with one or more independent variables (predictors) x_1, x_2 ,….,x_n is held constant. ϵ is the statistical error

If we want to use the multiple linear regression model, the data has to meet four assumptions of linear regression, for this project, the price data is highly left-skewed, so to meet the second assumption of the Linear Regression – data should not be non-linear we have to perform log-

transformation to ensure the data is linear [5]. From the Fig 2 we can see that the transformed data is much better than the previous skewed data, not perfect but better.

Fig 2 Data Before and after log-transformation

For this project, the features of the diamonds are the predictors, and the price is the response, all the predictors multiply by their coefficients and sum up together is equal its price. Therefore, to build this model is to find all the $\theta_0, \theta_1, \cdots, \theta_p$ of this equation. The method is Least Square method. The first step is put all the data in an equation format, then extract all the coefficients to build the coefficient matrix **X*b = y.** The coefficient of each feature is equal to the solution of $\hat{\beta} = (X^T X)^{-1} X^T y$.

3.2 Classification and Regression Tree (CART)

3.2.1 Crate CART model

When trying to explain the variance of the unit price of diamond, these algorithms will attempt to progressively divide the original sample of observations into descendent nodes, in order to decrease price diversity. The main different between CART and other decision tree algorithm is the CART is a binary tree, each root node represents a single input and split two points on that variable [6]. Selecting input variables and split points on those variables until a tree is constructed is the crucial step to create a CART model. Greedy algorithm will be used to find out the proper split point. For this project, the cost function $J = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$ will be used as the criteria to find out the best split point (which has the lowest cost value).

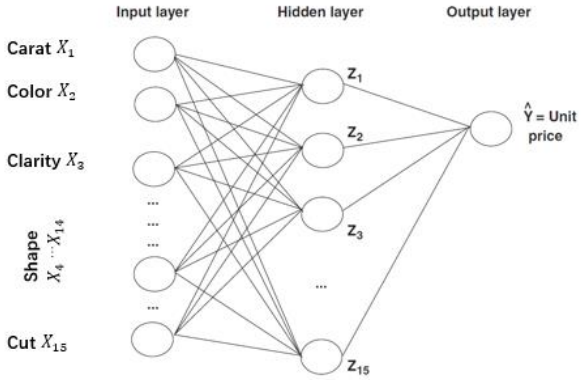
Although the response is better and easier to use regression tree part, the classification part still can be used if transform the continuous data (diamond price) to categorical data through binning method. For this project, I divided the price data into five equal intervals, [0, 9960), [9960, 19920), …, [39840, 49650). Then we can use the Gini index method ($Gini = 1 - \sum p_j^2$)to find the split point of each feature and using the Gini importance to find feature importance.

3.2.2 Prevent Over-fitting

The complexity of a decision tree is defined as the number of splits in the tree. Simpler trees are preferred. They are easy to understand, and they are less likely to overfit the data [6]. In this project, the max depth of the tree can grow is the variable that I can control to prevent overfitting the data. The method is iterating the max depth from 1 to 30 and find out the best validation score for this model (16 in this case).

3.3 Neural Network

The third model can be used to predict the diamond price is the Neural Network, this kind of models may be competitive for prediction purpose, although they yield result which are more difficult to interpret than the logical type models like regression trees [7]. The typical neural network contains three kinds of layers, hidden layer ($Z_1, Z_2, \dots, Z_M$), input layer ($X_1, X_2, \dots, X_M$) and output layer ($\hat{Y}$). The derived features are given by $Z_m = \sigma(\alpha_m^T x)$ and the output units yielded by $\hat{Y} = \beta^T z$. The neural network's α and β weights are associated with links between input and hidden layer and hidden and output layer, respectively. To minimize a total sum of squared error function: $E(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} E_i$. The generic approach to minimization is by gradient descent, called backpropagation in this setting. The figure below shows the layout of the neural network for this project.



Layout of the Neural Network

4. Validation

To judge the accuracy and the consistency of results, reliable estimates of the models' predictive capacity should be delivered. Therefore, two validation methods will apply to all of these three models.

4.1 K-fold cross-validation.

In this method, the original sample will be randomly partitioned into K equal sized subsamples. Of the Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated *k* times (the *folds*), with each of the *k* subsamples used exactly once as the validation data. The *k* results from the folds can then be averaged to produce a single estimation [8].

4.2 R-square validation

R-square aim to find the proportion of the variance in the dependent variable that is predictable from the independent variables. It represents how much of the data can be explained by the model [3].$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$

5. Modeling Result

5.1 Model Performance

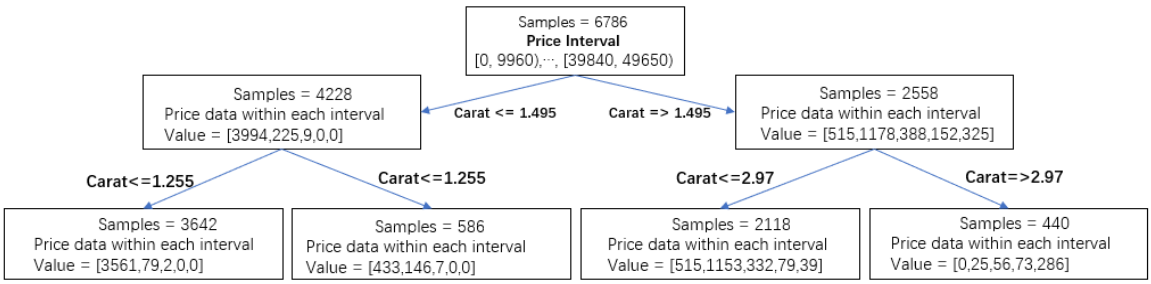
The table 4 shows the modeling result of three algorithms. The neural network gets the highest accuracy, the CART is second. Linear Regression is not that accurate as other two methods; however, it is the fast model to generate the result.

Table 4 Validation Result of Three Models

	Linear Regression	CART	Neural Network
Cross-Validation Score	78%	89%	96%
R-Square Score	80%	90%	97%
Model Time	1ms	2s	30s
Interpretability	Normal	Easy	Hard

5.2 Result Interpretability

Although CART tree returns a lower than the neural network precision, they may still be considered well fitted models and they may add some useful insights concerning diamond pricing. For example, in the generated tree (as shown in following figure), the first and second branching predictor is Carat, the higher the carat is the higher price the diamond will have. Also, from this diagram we can find the bigger size diamond is rare than the small size diamond and if the carat more than 2.97 ct, its value at least ten thousand dollars.



Secondly, the variable importance in CART can be calculated through find the sum of the decrease in impurity for each of the surrogate variables at each node [9]. Then we can find top five import features to contribute to the price of the diamond, as shown in table 5. From this table, we can find the carat is the most important feature to identify a diamond's price. And round shape diamond is normally higher price than the other shape if they have same other features.

Table 5 Top Five Importance Features Contribute to Diamond Price

Feature	Importance	Feature	Importance
Carat	0.8161	Depth	0.0248
Color	0.0852	Shape_Round	0.0191
Clarity	0.0381	Table	0.0071

6. Conclusion

In this project, Linear Regression, CART and Neural Network are used to model diamond unit prices, based on a set of diamonds characteristics: 4Cs, Shape, Depth, Table and Certification. While the neural network model achieves better precision, the CART are able to add some insights concerning the specific roles of predictors in pricing. For Linear Regression model, it has advantage in fast and easy modeling, however, it has big error when deal with upper-end priced diamond, which is non-linear data.

Finally, some further research may should be done,

The dataset seize is still too small, only have then thousand data in the current dataset, also, there are more features like fluorescence, should be added in the dataset to get more accurate prediction.

7. References

[1] Diamond, From Wikipedia https://en.wikipedia.org/wiki/Diamond_(gemstone)

[2] Liddicoat, R.T. (ed.) 1993. The GIA Diamond Dictionary (3rd edition), Gemological Institute of America, Santa Monica.

[3] A valuation model for cut diamond, Margarida G.M.S. Cardso, Luis Chamble, International Transactions in Operational Research, March 2005

[4] Smith, K., Gupta, J.N.D., 2000. Neural networks in business: techniques and applications for the operations researcher. Computers & Operations Research, 27, 1023–1044

[5] Dr. James Lani, Statistic Solutions, http://www.statisticssolutions.com/assumptions-of-linear-regression/

[6] Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth Inc., California.

[7] Mitchell, T., 1997. Machine Learning. McGraw-Hill, New York.

[8] Andrew Ng Coursera.org "Machine Learning Online Course" Stanford University

[9] Noel O'Boyle, "Supervised Classification" https://www.redbrick.dcu.ie/~noel/R_classification.html