Data Science and Engineering (DSE) Record, Volume 1, issue 2.

Comparative Study of Predicting Diamond Ring Prices in Online Retail Shop

Thanapon Chaijunla[1] and Phimphaka Taninpong[2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

Thanapon_Chaijunla@cmu.ac.th

Abstract. This study aims to develop the models for predicting the retail pricing of jewelry by using data source from online retail diamond ring stores. There are 2,206 records of ring data and 187,821 records of loose diamond data. This study develops and compare a performance of three models consist of Multiple Linear Regression (MLR), Random Forest (RF), and Deep Neural Network (DNN). The evaluation metrics used for comparing algorithms are accuracy of prediction using MAE and MAPE. The results show that MAE for the ring price prediction of MLR, RF, and DNN are $688.36, $235.33, and $273.00, respectively. In addition, MAE for a diamond price prediction of MLR, RF, and DNN are $3254.03, $450.44, $445.94, respectively. The results show that RF and DNN give higher accuracy rate than MLR. However, the accuracy rate of RF and DNN are slightly different.

Keywords: Diamond ring, Loose Diamonds, Retail Price, Data Mining, Price Prediction, Deep Neural Network

1 Introduction

In e-commerce, a price of similar products is varied from brand to brand. A comparison of the retail pricing of similar product is compared through the website. Jewelry products, i.e. a diamond ring, an engagement ring, and a wedding ring, are products with high value itself. The retail prices of the jewelry product are different based on manufacturing cost and the famous of brand. In terms of manufacturing and production, the product costs consist of labor and raw material expenses, consumable manufacturing supplies, and general overhead [1]. In retail marketing, the price is varied on several factors that either predictable or unpredictable such as product brand, market cost, consumer trends, the name of designer, seasoning, the source of the material even store's location, and so on. For jewelry price prediction, basically you have to consider on the raw material and other information of the products which can indicate the value of them to a certain extent, such as a metal type, diamond weight (carat), well-defined physical properties, and production technique. However, general customers did not have good knowledge of the products, so that they will not know whether the jewelry price offered by the manufacturer is an appropriate price or not. From the aforementioned problem,

many retailers provide the good experience on the website to facilitate their customers. For example, websites have the smart filter dashboards for searching products from their properties and enable customer to interact with actual photographed of the products with multiple dimension display and high resolution.

Nowadays, huge data are available on those websites which the customers have to search for their desired jewelry. For example, if we need to find a diamond that has 1 carat of weight, a website will deliver you more than 4,000 pieces with different property and characteristic of each diamond. Moreover, the price of them will be varied in price from 2,000 dollars up to 20,000 dollars. Therefore, consumers feel overwhelmed by huge data from the website. In previous research, there are a number of studies for jewelry product price prediction. Stanislav Mamonov and Tamilla Triantoro [2] used the physical properties of a diamond to predict the diamond price by using dataset of loose diamonds scraped from an online diamond retailer. In addition, José M. Peña Marmolejos [3] implemented an analysis of diamond price prediction by implementing data mining algorithms, that using public dataset from the Kaggle repository. Recently, Yusuke Yamaura and et al. [4] study of resale pricing of the secondhand jewelry by building a multimodal model. The model uses images and attributes of the product and employs multimodal deep neural networks which are applied in computer vision.

This study aims to apply knowledge of data science for solving this business problem by developing models to predict the retail prices of ring and loose diamond by using their physical properties. In addition, the predictive result can be used as alternative information for the consumer. Moreover, this study will focus on applying data mining techniques which will enable for further extension in other business cases of jewelry industrial.

2 Data and methodology

2.1 Data

This study used two datasets: ring and loose diamond property data which are obtained from online retail diamond ring websites. There are 2,206 records of ring data and 187,821 records of loose diamond data. The quality of both datasets is perfect as there is no missing data or other features need to correct. Ring dataset consists of 7 features and loose diamond dataset 13 features. Table 1 shows all features used for developing prediction models.

Metal Type is the metal type used for producing a band of ring. Ring width is the horizontal measurement across the widest point of a ring, which can vary depending on style. Sub diamond number is the number of sub diamond to be set on a ring. Sub Diamond total carat, or so-called carat, is the standard unit of measurement used to indicate the weight of diamonds and precious gemstones [5]. This feature is the combination weight of sub diamonds on that ring. Diamond color is an important characteristic that affects a diamond's beauty. The Gemological Institute of America (GIA) color scale is the industry standard for diamond grading [6]. The GIA diamond color grades range

from D (colorless) to Z (light yellow or brown) which the more colorless means the higher their value [7].

Table 1. All Features

Dataset	Features	Variables	Type
Ring	Metal Type	Independent	Categorical
	Ring Width	Independent	Numerical
	Sub Diamond Number	Independent	Numerical
	Sub Diamond Total Carat	Independent	Numerical
	Color Scale	Independent	Categorical
	Clarity Scale	Independent	Categorical
	Price USD	Dependent	Numerical
Loose Diamond	Shape	Independent	Categorical
	Carat	Independent	Numerical
	Cut	Independent	Categorical
	Color	Independent	Categorical
	Clarity	Independent	Categorical
	Fluorescence	Independent	Categorical
	Length Width Ratio	Independent	Numerical
	Depth Percent	Independent	Numerical
	Table Percent	Independent	Numerical
	Culet	Independent	Categorical
	Polish	Independent	Categorical
	Symmetry	Independent	Categorical
	Price USD	Dependent	Numerical

The clarity scale of diamond clarity refers to the absence of blemishes. Diamonds without these birthmarks are rare, and the rarity effect a diamond's value [8]. The GIA diamond grading scale is divided into 6 categories and 11 diamond clarity grades [9] which the less blemishes mean the higher their value. Diamond shape refers to the geometric outline and overall physical form of a diamond. Every diamond shape has its own attributes and cut specifications, which also play a large factor in the overall look of the stone. Diamond cut is the summary of a diamond's proportions evaluated using the attributes of brilliance, fire, and sparkle. While high marks of color or clarity affect a diamond, it is the cut that defines its proportions and ability to reflect light. Diamond fluorescence is the tendency of a diamond to emit a (soft) glow when exposed to ultraviolet light (UV light). The fluorescence effect is present in over 30% of diamonds and is an important consideration when buying a loose diamond. The diamond length width ratio calculates by its length divided by width. This conveys how relatively square or rectangular a fancy-shaped diamond appears when viewed from the top [6]. The depth is the distance from the table to the culet, or point, of the diamond. When discussion depth in terms of cut quality, it is described in percentages. The most ideal table percentages are between 60 and 54 percent. At this proportion, the table is large enough to allow light to enter the stone at correct angles to reflect and refract off the smaller facets below. The smallest facet of a diamond, the culet is located at the very bottom of the

stone. If the diamond ends in a point, the diamond grading report will show a value of 'None' for the culet designation. The Diamond polish and symmetry are critical components to cut quality. For maximum brilliance, every facet of a diamond should be polished after the cutting process. A symmetrical diamond will have well-balanced and properly aligned facets. If the facets are not symmetrical or not optimally shaped, they will display less sparkle [5].

This study uses the same approach in data preparation for both datasets. For numerical features, we use a standardization technique to do scaling those values, in order to reduce any bias from differentiation of magnitude, range, and units. We also use the one-hot encoding technique to convert each categorical feature to be in calculable format.

2.3 Methodology

Fig. 1 shows the study process which consists of the data gathering, data mining, and model evaluation. The data gathering is the process of data collection from websites. After the data gathering process, the data mining process proceeds by exploring data and preparing data for analysis. In the model development process, we develop the model for ring price estimation and the model for diamond price estimation separately by using supervised machine learning techniques which consist of multiple linear regression (MLR), random forest (RF), and the deep neural network (DNN). In the evaluation process, we split the data into two subsets using 80% of data for training the model and the rest of data for testing the model. Consequently, we evaluate the performance of each model to find the best one. Finally, the predictive result from both models will be combined to be a predictive price of a diamond ring. In this study, standard packages of Pandas, NumPy, Seaborn, and Matplotlib were used for reading, transforming, and assessing the datasets as well as model development.

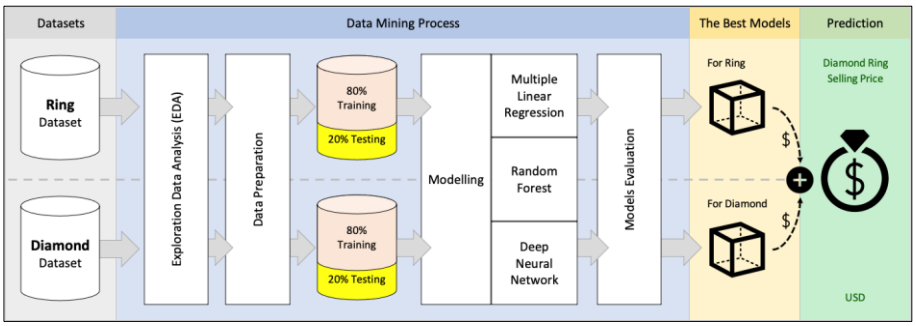

Fig. 1. Study process

3 Supervised learning Techniques

3.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is used to predict the dependent variable by two or more predictor variables. MLR is so-called a multiple regression as it extends a simple linear regression in which more than one predictor variables consider in the model. In this study, the dependent variable that we want to predict is the price of ring and loose diamond, while the predictor variables or independent variables, which are called features, are display in Table 1. Regression analysis has been used effectively to answer many questions and business cases. In addition, linear regression shows the relationship between the two variables using a straight line, therefore MLR shows a linear relationship between the dependent variable and independent variables. However, the relationship can be non-linear in which the relationship between the dependent and independent variables do not follow a straight line. By using MLR, we need to follow the assumptions: 1) There is a linear relationship between the dependent variable and the independent variables, 2) There is no multicollinearity between the independent variables, 3) The variance of the residuals is constant, and 4) the residual is normally distributed. Fig. 2 shows the scatter plots between each feature.

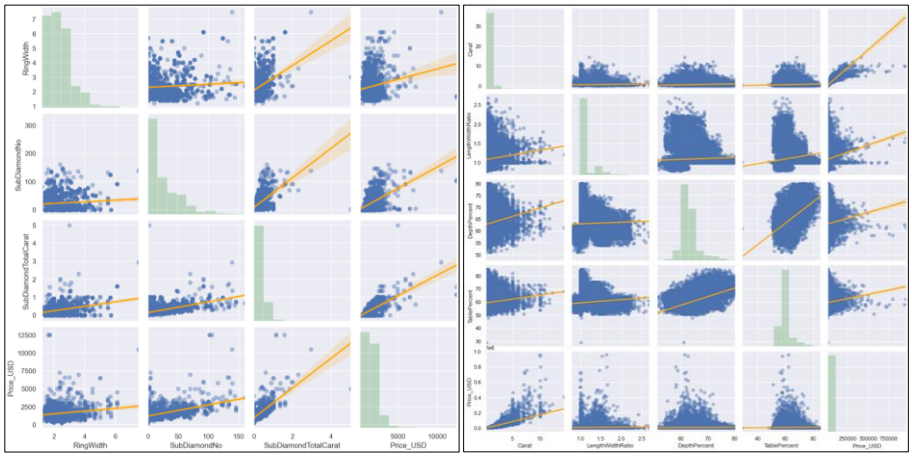

Fig. 2. The Combination Pair plot and Scatterplot to show a relationship of each feature

In this study, we apply Variance Inflation Factor (VIF) technique to verify the multicollinearity for each feature. If multicollinearity occurs, there is a variable that contributes to the variance in the dependent variable. Moreover, MLR has the assumption of homoscedasticity, therefore we apply standardized technique to manipulate the data and use histograms and boxplots to visualize and analyze the distribution of the data [10].

Multiple Linear Regression Formula [11]

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Where:

y_i is the dependent or target variable,
β_0 is the y-intercept,
β_1, β_2 and β_k are the regression coefficients that represent the change in y relative to a one-unit change in x_1, x_2 and x_k, respectively,
ε is the model's random error (residual) term.

3.2 Random Forest

Random Forest (RF), as known as decision forest, is one kind of machine learning algorithm that can be applied for a regression and classification. RF is an ensemble method as a random forest model is made up of a large number of small decision trees, called estimators, which each tree produces their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction [12] as shown in Fig. 3.

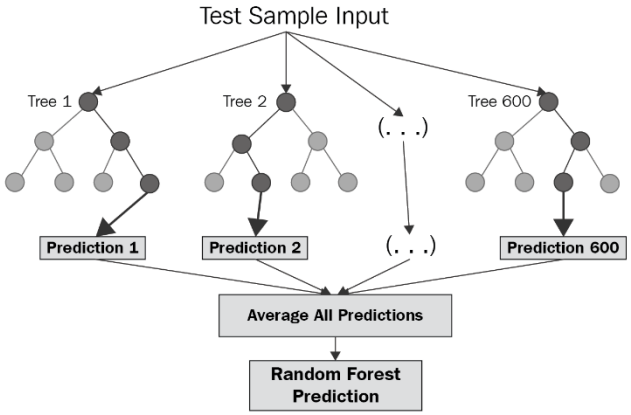

Fig. 3. Random Forest Structure [13]

In this study, we use RandomForestRegression which is a standard package of Scikit-learn to develop the Random Forest model. However, there are many hyperparameters of random forest that have to be tuned such as a number of trees in the forest, maximum depth of the tree, the minimum number of samples required to split and internal node, and so on. Number of trees is one of the important parameters that need to be tuned, and the proper number of trees is 150 tress which is evaluated by considering Mean Absolute Error as shown in Fig. 4.

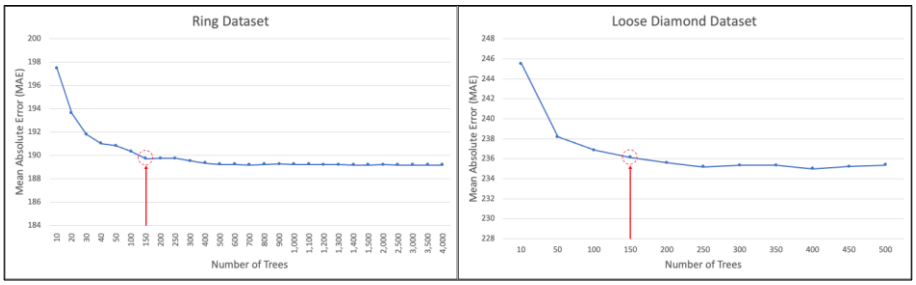

Fig. 4. Fine-tuning Hyperparameter (n_estimators: a number of trees)

3.3 Deep Neural Network

Deep Neural Network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layer as shown in Fig.5. The DNN turn the input layer into the output layer by finding the correct mathematical manipulation. The probability of each output is computed through the layer in the network. DNN is typically feedforward network in which data flows from the input layer to the output layer without looping back. At the first stage, the DNN creates a map of virtual neurons and assigns the initial weights using the random value to connect between them. Weights and the values of inputs are multiplied and return an output which is the value between 0 and 1. In the iteration, an algorithm will adjust the weights in order to recognize a particular pattern. Therefore, the algorithm will be iterated, and terminated when it determines the correct mathematical manipulation to fully process the data [14].

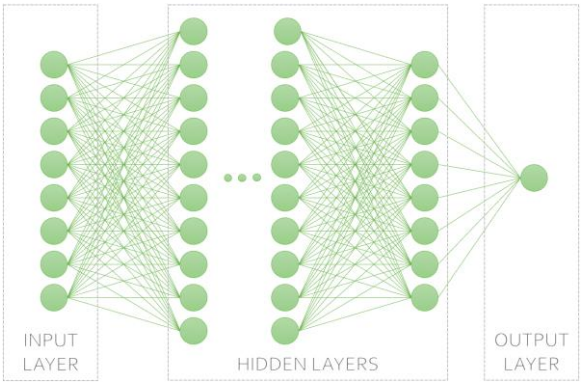

Fig. 5. Deep Neural Network (DNN) Structure

In order to create Deep Neural Network (DNN) model, we need to design a structure of the neural network such as a number of nodes in each layer and number of layers. Moreover, we have to set the hyperparameters for the network as well as the activation function, learning rate, batch size, a number of epochs, and etc. In the tuning stage, we use the technique as known as Grid Search Cross Validation by using the standard package of Scikit-learn called GridSearchCV, to tune the hyperparameters. After that

we use another package called MLPRegressor to construct the models by using the parameters from tuning stage.

4 Results

In this section, the exploration data analysis is processed for ring dataset. Table 2 shows that ring width is approximately 2.36 cm. Table 2 also shows that the average of sub diamond number is 24.04 pieces and minimum is zero which means there is no sub diamond on that ring. In addition, the average of carat is approximately 0.29 carat and the minimum value is zero which means that there is no sub diamond on that ring. Fig. 6 shows the frequency distribution of metal type and platinum is the most popular metal type. Fig. 7 shows that F/G is the most popular color scale to be used for sub diamond. Fig. 8 shows that SI1: Slightly Included 1 is the most popular clarity scale to be used for sub diamond.

Table 2. Descriptive Statistics for each numerical feature of ring dataset

Features	count	mean	std	min	25%	50%	75%	max
Ring Width	2,206	2.36	0.78	1.2	1.8	2.2	2.6	7.5
Sub Dia. Number	2,206	24.04	28.17	0	2	16	38.75	160
Sub Dia. Total Carat	2,206	0.29	0.30	0	0.05	0.24	0.44	5
Price USD	2,206	1,575.43	903.49	180	1,050	1,450	1,890	12,500

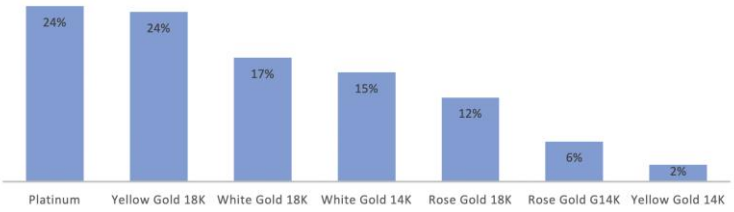

Fig. 6. Frequency distribution of metal type

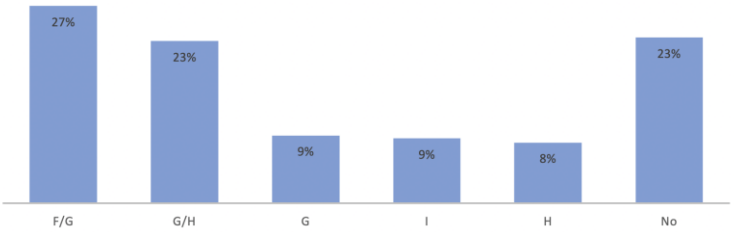

Fig. 7. Frequency distribution of color scale

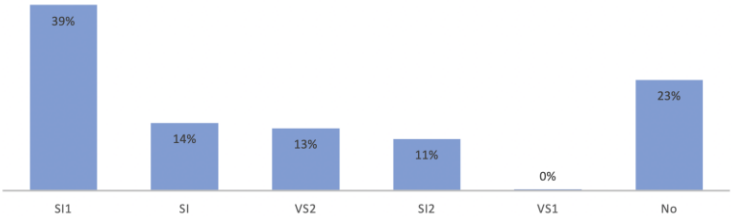

Fig. 8. Frequency distribution of clarity scale

Note: "No" is mean these is not any sub diamond on that ring.

The exploration data analysis is also the process for loose diamond dataset. Table 3 shows the average diamond weight is approximately 0.68 carat with the minimum weight is 0.32 carat and maximum weight is 14.38 carat. Fig. 9 shows the frequency distribution of diamond shape and round is the mostly shape in this dataset. The grade of diamond cutting for this dataset is in the high-quality level as shows in Fig. 10. In Fig. 11 shows that E and D are the most diamond color scales in this dataset. In addition, VS1: Very Slightly Included 1 and VS2: Very Slightly Included 2 are the most diamond clarity grades in this dataset as shows in Fig. 12. Approximately 70% of all diamond are no fluorescence and most of them no culet as shows in Fig. 13 and Fig. 14. The frequency distribution of polishing and symmetry qualities will reflect and correspond to the cutting quality that show in Fig. 15 and 16.

Table 3. Descriptive Statistics for each numerical feature of diamond dataset

Features	count	mean	Std	min	25%	50%	75%	max
Carat	187,821	0.68	0.63	0.32	0.32	0.5	0.8	14.38
L/W Ratio	187,821	1.09	0.20	0.78	1.01	1.01	1.01	2.68
Depth %	187,821	63.02	3.35	51	62.4	61.4	63.3	80
Table %	187,821	59.44	4.03	29	57	58	61	85
Price USD	187,821	4,263.04	14,117.55	178	741	1,250	3,040.25	960,152

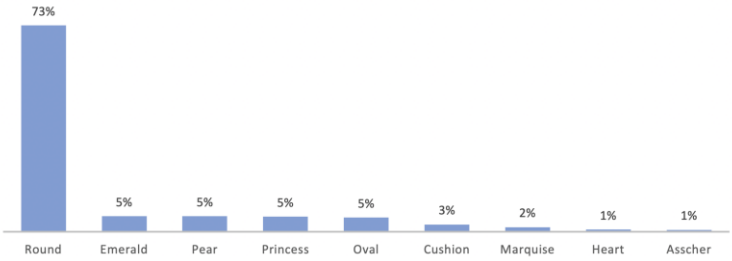

Fig. 9. Frequency distribution of diamond shape

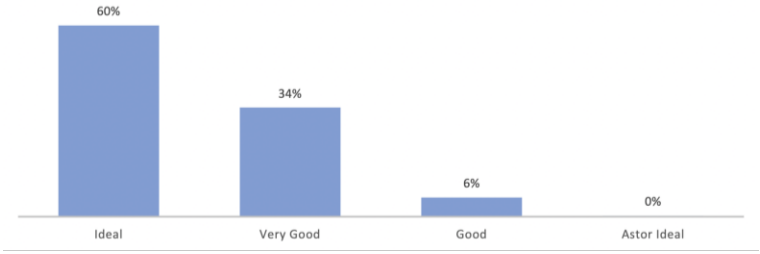

Fig. 10. Frequency distribution of diamond cut

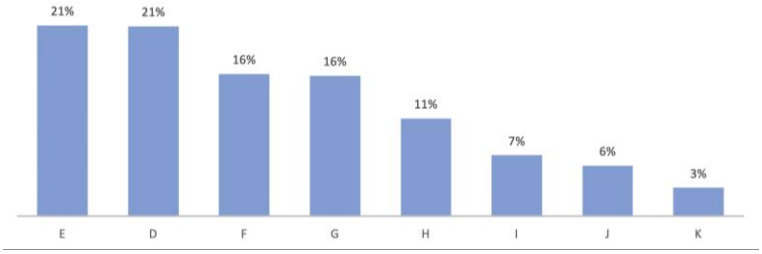

Fig. 11. Frequency distribution of diamond color

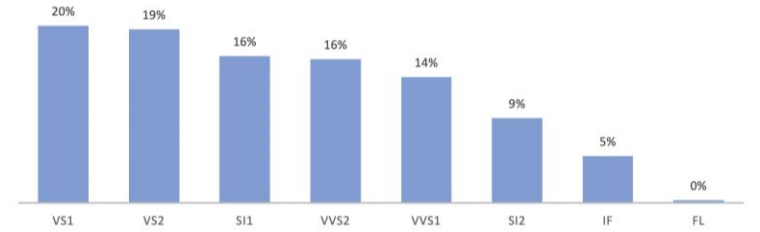

Fig. 12. Frequency distribution of diamond clarity

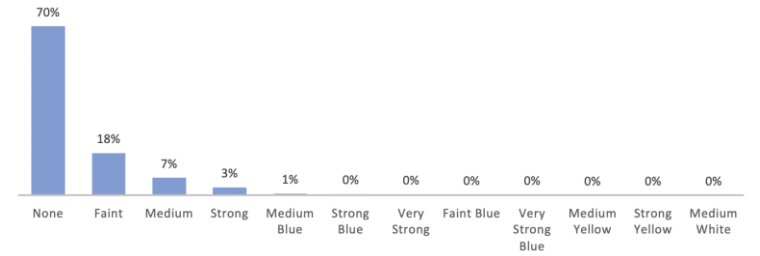

Fig. 13. Frequency distribution of fluorescence

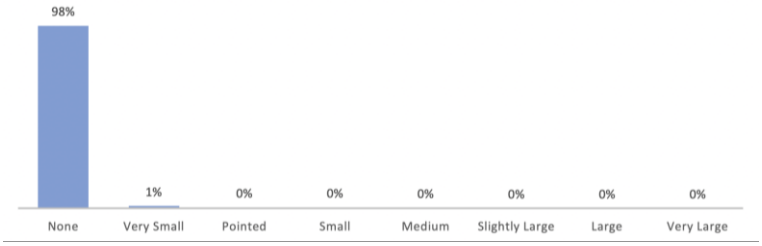

Fig. 14. Frequency distribution of culet

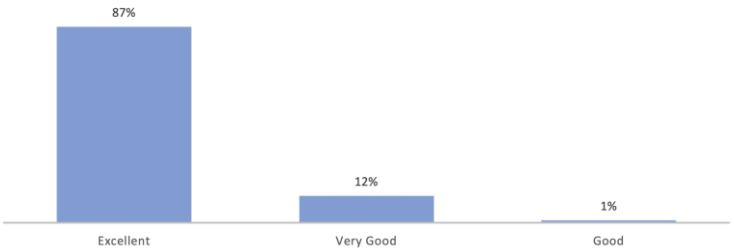

Fig. 15. Frequency distribution of polish

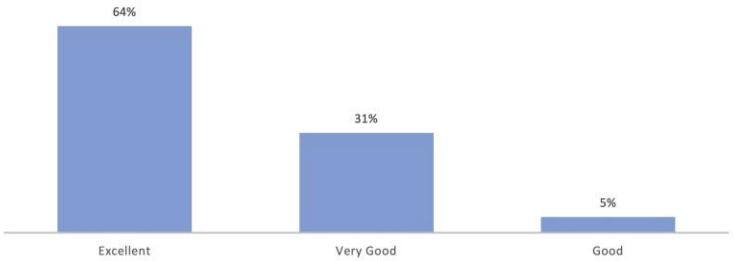

Fig. 16. Frequency distribution of symmetry

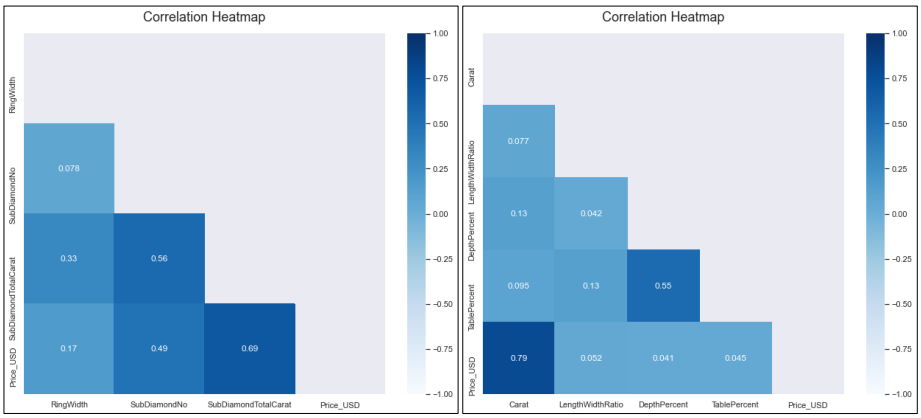

Fig. 17. Correlation Heatmap Charts of each numerical feature from both datasets

Fig. 17 shows the correlation heatmaps of each numerical feature. The left correlation heatmap shows that diamond weight and number of diamond piece features have the strong positive correlation coefficient with Price_USD feature. That mean the increment or decrement of values from both features will influence to the ring price. In addition, the right correlation heatmap shows that the weight of diamond (Carat) also delivers the highest positive correlation coefficient score with its price (Price_USD) as well. In this study, we developed three models to compare the performance of each model which consist of multiple linear regression, random forest, and the deep neural network. Moreover, we evaluate the accuracy of each model by using mean absolute error (MAE) and mean absolute percent error (MAPE). The experiment results of this study are shown in Table 4.

Table 4. Model Performance Results

Datasets:	**Ring**		**Loose Diamond**	
Average Price:	$ 1,604.25		$ 4,317.84	
Models	MAE	MAPE	MAE	MAPE
MLR	$ 688.36	57.54%	$ 3,254.03	243.36%
RF	$ 235.33	17.60%	$ 450.44	7.08%
DNN	$ 273.00	20.21%	$ 445.94	11.39%

The results in Table 4 show that Random Forest model provides the lowest MAE and MAPE for ring price estimation, and MAE is 235.33 dollars and MAPE is 17.60%. In addition, DNN provides the lowest MAE for loose diamond price estimation and the value of MAE is 445.94 dollars. However, RF provides the lowest MAPE for loose diamond price estimation and the value of MAPE is 7.08%. In summary, the model that gives the lowest MAE for ring dataset is Random Forest (RF) and the model that provides the lowest MAE for loose diamond dataset is Deep Neural Network (DNN). Moreover, RF gives the lowest MAPE for both datasets.

5 Discussion and Conclusion

In this work, we developed three models to predict retail pricing of a diamond ring by using their physical properties. This study uses two datasets: Ring and Loose Diamond dataset. The results also show that the proper model for this study is the Random Forest model which gives the mean absolute percentage error of 17.60 % for ring dataset and 7.08 % for diamond dataset. Our proposed model for diamond price estimation gives the opposite result to Stanislav Mamonov and Tamilla Triantoro [2] who develop a model to predict diamond prices in online retail with the range of data between 0.2 to 2.5 carat diamonds. The best model performance from their studied is Artificial Neural Network follow by MLR and RF. Future work will improve the performance of the model, there are two possible options which will not be covered in this study. The first option is collecting more data in order to have more information for training the model. Another option is separated development each model for each range or group of data.

References

1. Production Costs, https://www.investopedia.com/terms/p/production-cost.asp, last accessed 2021/02/03.
2. Stanislav Mamonov, Tamilla Triantoro. Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining Study. Journal of Theoretical and Applied Electronic Commerce Research. 2018; 13(2):15-28.
3. José M. Peña Marmolejos. Implementing Data Mining Methods to Predict Diamond Prices. 14th International Conference on Data Science ICDATA 2018, 112-116, Las Vegas, NV USA (2018).
4. Yusuke Yamaura, Nobuya Keneaki, Yukihiro Tsuboshita. The Resale Price Prediction of Secondhand Jewelry Items Using a Multi-modal Deep Model with Iterative Co-Attention. CoRR. 2019; abs/1907.00661.
5. Diamond education, https://www.brilliance.com/education/diamonds, last accessed 2021/02/12.
6. Diamond education, https://www.bluenile.com/education/diamonds, last accessed 2021/02/12.
7. GIA 4Cs Color, https://www.gia.edu/gia-about/4cs-color, last accessed 2021/03/01.
8. GIA 4Cs Clarity, https://www.gia.edu/gia-about/4cs-clarity, last accessed 2021/03/01.
9. Diamond Clarity, https://www.bluenile.com/education/diamonds/clarity, last accessed 2021/03/01.
10. What is a Multiple Linear Regression, https://corporatefinanceinstitute.com/resources/knowledge/other/multiple-linear-regression/, last accessed 2021/03/01.
11. ดร.สุทิน ชนะบุญ, สถิติและการวิเคราะห์ข้อมูลในงานวิจัยเบื้องต้น. บทที่ 6 การวิเคราะห์ข้อมูลเชิงอนุมาน. สำนักงานสาธารณสุขจังหวัดของแก่น, ขอนแก่น (2560)
12. What is a Random Forest, https://deepai.org/machine-learning-glossary-and-terms/random-forest/, last accessed 2021/03/01.
13. Random Forest Regression, https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f, last accessed 2021/03/01.
14. Deep neural network (DNN) is an artificial neural network (ANN), https://bangaloreai.com/blog/meet-the-bitcoin-cash-hyper-mini-sprint-car/, last accessed 2021/03/01.