Machine Learning Algorithms for Diamond Price Prediction

Waad Alsuraihi
King Abdulaziz University
Jeddah, Saudi Arabia
Information Systems Department

Kholoud Bawazeer
King Abdulaziz University
Jeddah, Saudi Arabia
Information Systems Department

Ekram Al-hazmi
King Abdulaziz University
Jeddah, Saudi Arabia
Information Systems Department

Hanan Alghamdi
King Abdulaziz University
Jeddah, Saudi Arabia
Information Systems Department

ABSTRACT
Precious stones like diamond are in high demand in the investment market due to their monetary rewards. Thus, it is of utmost importance to the diamond dealers to predict the accurate price. However, the prediction process is difficult due to the wide variation in the diamond stones sizes and characteristics. In this paper, several machine learning algorithms were used to help in predicting diamond price, among them Liner regression, Random forest regression, polynomial regression, Gradient descent and Neural network. After training several models, testing their accuracy and analyzing their results, it turns out that the best of them is the random forest regression.

CCS Concepts
• Information systems → Information systems → Decision support systems → Data analytics

Keywords
Machine Learning; Predictive analysis; Diamond price; Supervised; Regression; Random forest

1. INTRODUCTION
Nowadays, Machine Learning is one of the most significant driving force of Artificial Intelligence (AI), whereby computers can learn without being programmed to perform specific tasks. Machine learning algorithms learn from previous cases to produce the required results quickly and accurately. We can say that machine learning is a treasure that must be exploited to solve wide range of problems at all levels, social, economic, profession and others.

IVSP 2020, March 20–22, 2020, Singapore, Singapore

ACM ISBN 978-1-4503-7695-2/20/03…$15.00
DOI: https://doi.org/10.1145/3388818.3393715

In the diamond trading sector, buyers and investors face several difficulties in predicting diamond stones prices. This difficulty are due to the difference in the stones shapes, sizes and purity. In order to ease this problem and to aim diamond traders, this paper discusses the application of machine learning algorithms as an approach to predict diamond price through employing their features. This is done by using a dataset from Kaggle.

In this paper we used Diamonds Dataset [1], formulating the problem as a regression problem and comparing the performance of these algorithms. In this paper, we found that Random forest regression achieved the best result. All algorithms are trained using predictive supervised learning models. Predictive modelling is the task of building a model using historical data to estimate labels of new data samples [2]. As we consider predicting the price label as a numeric value, regression predictive modelling is utilized in this paper. Regression is the task of predicting a continuous output such as integer or floating pint for a data sample given its attributes [3].

This paper is structured as follows. Section II presents some of related works. Section III discusses the proposed approach, describes diamond dataset and explains the pre-processing steps. Section IV presents the results. Section V discuss these results. Section VI discusses the possible future work and concludes the paper. As we consider predicting the price label as a numeric value, regression predictive modelling is utilized in this paper. Regression is the problem of predicting a continuous output for an unseen data sample [3]. A continuous output variable is a real-value, such as a floating-point or integer value [2]. This paper is structured as follows. Section II presents some of related works. Section III discusses the proposed approach, describes diamond dataset and explains the pre-processing steps. Section IV presents the results. Section V discuss these results. Section VI discusses the possible future work and concludes the paper.

2. RELATED WORK
In this section we will talk briefly about how others have analyzed the same dataset and their results. We pick two analysis from Kaggle. In [4], the authors applied several models and reports their performance in term of R2-score, a statistical metric of how close the data are to the fitted model and the higher the R2-score, the better the model fits your data. Random Forest Regression achieved score of 0.982066, K-Neighbours Regression 0.959033, Gradient Boosting Regression 0.905833, AdaBoost Regression 0.882499, Linear Regression 0.881432, Lasso Regression

0.865866 and Ridge Regression 0.753705. Random forest algorithm gives the highest R2-score "98%"[4]. In [5], the authors trained other models including Decision Tree Regression with accuracy 100.00, Random Forest Regression with accuracy 99.50, Linear Regression with accuracy 91.87, Gradient Boosting Regression with accuracy 90.38, Lasso Regression with accuracy 90.17, AdaBoost Regression the accuracy with 85.10, Elastic Net Regression with accuracy 81.22 and Ridge Regression with accuracy 80.12. The model that gives the highest accuracy is Decision Tree Regression [5].

3. PROPOSED METHOD

3.1 Dataset Description

Since ancient times, diamonds have been used as ornamental items. The value of diamonds makes it very useful for industrial applications and desirable jewelries [6]. Multiple organizations have been created for grading and certifying diamonds based on the "four Cs", which are color, cut, clarity, and carat. In this work, diamonds dataset contains the prices and other attributes of almost 54,000 diamonds [1].

The attributes are:

- Data frame [53940 instances & 10 features]
- Price in US dollars [$326 -$18,823]
- x length in mm [0 - 10.74]
- y width in mm [0 - 58.9], [See Figure 1]
- z depth in mm [0 - 31.8]
- Carat weight of the diamond [0.2-5.01]
- Cut quality of the cut [Fair, Good, Very Good, Premium, Ideal]
- Color diamond [from J (worst) - to D (best)]
- How clear the diamond is [I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)]
- Table width of top of diamond relative to widest point [43-95]. [See Figure 1].
- Depth total depth percentage = z / mean (x, y) = 2 * z / (x + y) [43 - 79]. [See Figure 1]

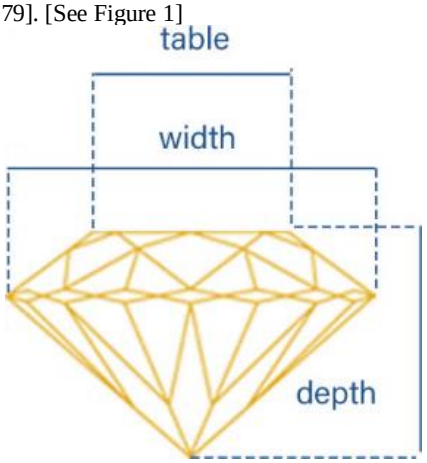

Figure 1. Diamond stone features: table, depth and width

3.2 Feature Correlation

We use correlation coefficient measures to measure the strength of the relationship between two features. A correlation coefficient is a statistical metric of the degree to which the changes of one variable can predict the change of another one. [7]. The strength of the relationship between the characteristics of the diamond data set was measured, [See **Figure 2**]

As it can be noticed, X, Y, and Z have a strong relation with price. Carat Also has a significant relation with price.

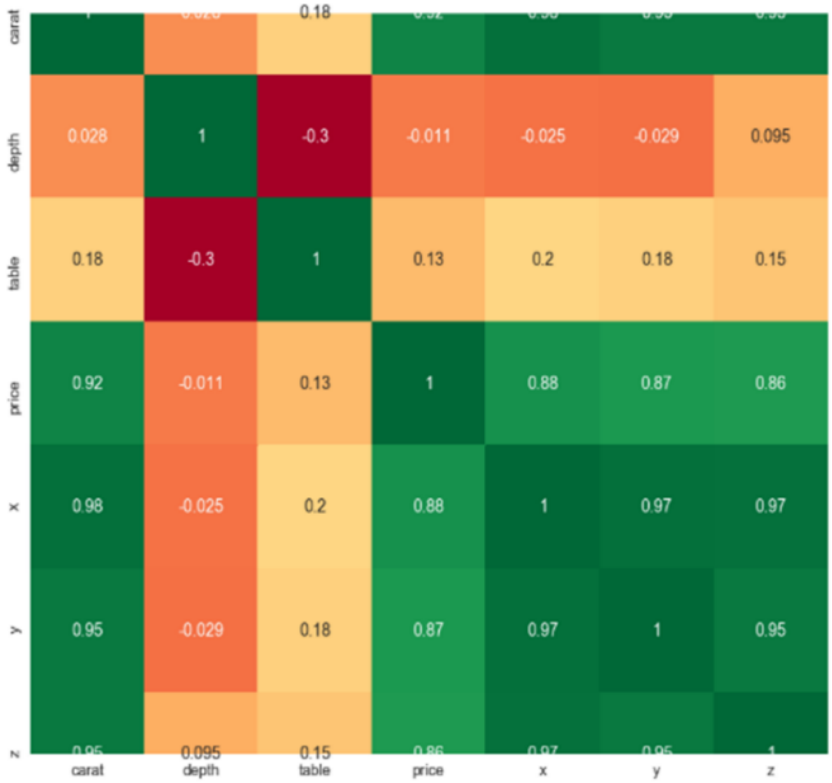

Figure 2. Correlation coefficients of the diamond dataset features

3.3 Pre-processing

Pre-processing means apply a transformation to a dataset before feeding it to the algorithm. Data Pre-processing means performing some cleansing techniques to convert the raw data into a clean data set [8], in order to achieve better results when training the machine learning models. For example, some algorithms do not support null values or categorical data. In this work, the dataset has no null values (or Missing values), but there are three categorical features that need transformation. Thus, before manipulating the dataset, we convert the nominal values (object type) in color, cut and clarity attributes, to numeric values by using one hot encoding. We split the data set to two part – the first part is training set (80%) to create the model and the another is test set (20%) use it to validate the model.

3.4 Models Descriptions

In this paper, we trained several models to predict the price label. Following are brief descriptions about these models:

3.4.1 Linear Regression

Linear Regression is a supervised machine learning algorithm based [9]. Linear model makes its prediction by the following equation:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Where:

- x: input from training data.
- y: label
- θ_0: intercept
- $\theta_{1,2,n}$: coefficient of x.

3.4.2 Random forest regression

Random forest regression is a supervised learning algorithm which uses ensemble learning method for both classification and regression problems. It is kind of Ensemble learning which means it takes the result of prediction from group of models [11]. Random forest contains several decision trees, which are based on the following Gini impurity equation. [10]

$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

Where

- pi,k is "the ratio of class k instances among the training instances in the ith node" [10].

3.4.3 Gradient Boosting Regressor

The seconds model is Gradient Boosting Regressor. It is a machine learning algorithm for regression and classification problems, which is based on an ensemble of weak learners, typically decision trees[12]. This model it will be learned by taking weighted sum of M base learners (additive modelling).

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

Where

- $F\_m(x)$: "Model obtained by adding m (=1, 2,…, m) base learners"[13].
- $h\_m(x)$: "represents the direction in which the loss function decreases w.r.t. $F\_m\text{-}1(x)$" [13].
- γ: "corresponds to the hyperparameter α in terms of the utility" [13].

3.4.4 Polynomial regression

Polynomial regression is a regression analysis model in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x [14]. To generate polynomial regression equation from linear regression equation, we only add powers of the original features as new features [10].

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2$$

3.4.5 Neural network

Neural network is a machine learning model that make use of an architecture inspired by the neurons in the brain [15]. It uses Perceptron learning rule to make the prediction. The rule is:

$$w_{i,j}^{(\text{next step})} = w_{i,j} + \eta(y_j - \hat{y}_j)x_i$$

Where

- wi,j is the weight of the ith input and the jth output neuron [10].
- xi is the ith input of the current instance [10].
- yj is the output of the jth output neuron for the current instance [10].
- yj is the target output of the jth output neuron for the current instance [10].
- η is the learning rate [10].

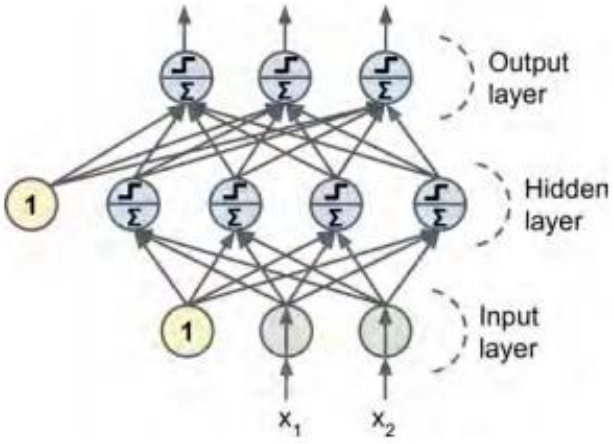

Figure 3. Layers of deep neural networks (simple model) [10]

Deep learning is neural networks composed of several layers [see figure 3] [10] . The neural networks have three types of layers are input, hidden and output. Each neuron is connected to the next layer [10].

Neural networks have many parameters that we can change their value to fit the data, but in this paper, we focused on the layers and their neurons (units). We tried to change the number of layers and

the neurons (units), because they are among the factors affecting the accuracy of the prediction [10].

4. PERFORMANCE AND VALIDATION METRICS

To check the rate of error we used MSE and RMSE. The MSE is an estimator measures the average of error squares i.e. the average squared difference between the estimated values and true value [16]. The RMSE is a quadratic scoring which measures the average volume of the error. It is the difference between prediction and the corresponding actual values. RMSE gives a relatively high weight to large errors, thus, it is preferred when large errors are particularly undesirable [17]. We also use cross validation to estimate the generalization of the models on unseen data [18].

5. RESULTS AND DISCUSSIONS

All the logarithms (or models) that mentioned earlier in Models descriptions are trained by Diamonds dataset.

The results were as follows, when we applied the liner regression, we did some sort of testing, but the result was there under fitting because one of its result of predict was 4784 and the actual is 16231... etc. So, there is high error between the Actual and predicted values.

In addition, when we use the cost function such as RMSE and MAE and their results showed that there was a high error.

For Gradient Boosting Regressor, there was low error between the Actual and predicted values compared with other models.

After that, we use Polynomial regression with two different degrees, when the degree equals 1 and the next time when it equals 2. When the degree is 1, there is high error between the Actual and predicted values compared to other models. We notice underfitting in the model. So, we need to increase the complexity by increase the degree of polynomial equation.

We made the model with a degree equals 2 and test the result. We found there is low error between the Actual and predicted values compared with other models. Also, it near to be overfitting. So, we should not increase the degree or use regularizing method.

About the Neural Network, in the first case, we used three layers the first one for the input and has 26 inputs (features). The second layer has 26 units (neurons). The last one for the output which is the labeled feature (price). As shown in Figure 4 there is a good alignment between predicted values and actual values,

Then, we tested using cost functions and the result of the error was high. The actual values are [16231, 4540, 5729, ..., 1014, 2871, 6320] and the predicted results are [12755.024, 5089.3423, 6623.827, ..., 1501.631, 2848.1982, 7164.5537]. So, there is also underfitting in this model.

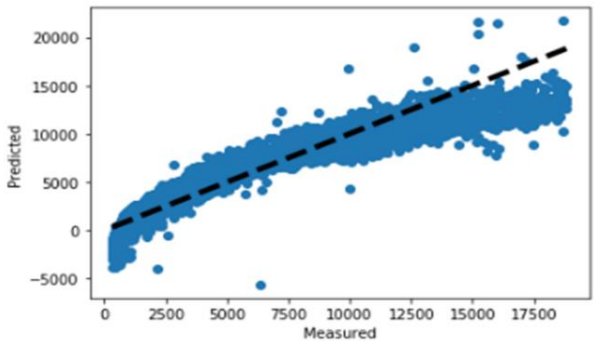

Figure 4 The scatter plot for Neural network results

In the second case, we used four layers, the first one for the input and has 26 inputs (features). The second and third layers have 10 units (neurons). The last one for the output which is the labeled feature (price). There was not much difference from the first case.

In the third case, we used six layers, the first one for the input and has 26 inputs (features). The second layer has 10 units (neurons). The third layer has 40 units (neurons). The fourth layer has 30 units (neurons).). The fifth layer has 10 units (neurons).). The last layer has 10 units (neurons). The last one for the output which is the labeled feature (price). Also, this case has a result very close to the second case.

While the *Forest regression* has the lowest error between the Actual and predicted values compared with other models. We did some sort of testing; the actual result is [15128.7 4681.5 5731.7 ... 1003.4 2874.6 6300.7] and the predation results were [16231, 4540, 5729, ... , 1014, 28716320]. There is very little difference. This makes the random forest model the best model for predict the price values, compared with the other models.

Table 1 summarizes the results of all trained algorithms.

Table 1. A comparison of accuracy results for all methods trained with a diamond dataset

Model	Parameters	MAE	RMSE
Liner regression	-	742	1128.5
Forest regression	-	112.93	241.97
Gradient Boosting Regressor	-	938	1406
Polynomial regression	Degree =1	742.256	1128.569
	Degree =2	401.497	672.398
Neural Network	Layer 1= 26 input Layer 2 = 26 unit Layer 3 = 1 output	3103 until it become 992 at epochs=100	-
	Layer 1= 26 input Layer 2 = 10 unit Layer 3 = 10 unit Layer 4 = 1 output	2803.8015 it become 764.1069 at epochs=100	-
	Layer 1= 26 input Layer 2 = 10 unit Layer 3 = 40 unit Layer 4 = 20 unit Layer 5 = 10 unit Layer 6 = 1 output	1921.5269 until it become 769.9570 at epochs=100	-

As we compared the results of the logarithms that we trained with the logarithms that were trained by the authors that we mentioned them previously in the Related work section. Table 2summarizes the results.

Table 2. A comparison of the methods used in this paper with methods used in other related works

	Proposed models		Work in [5]		Work in [4]	
Models	**RMSE**	**MAE**	**RMSE**	**RMSE Mean**	**RMSE**	**MAE**
Linear regression	1128.57	742.26	1142.27	1126.90	1382.53	926.72
Gradient boosting regressor	1406.26	938.02	1242.20	1235.61	1232.08	720.72
Random forest	241.98	112.94	282.70	577.40	559.57	283.57

6. CONCLUSION AND FUTURE WORK

This paper demonstrates that machine learning can contribute to the economic field and help traders solve their problems, such as forecasting prices. In this paper, we applied several models and conducted accuracy tests on them. We saw that the random forest has the best result despite the noise in this dataset. Therefore, we recommend using a random forest or make ensemble models. Ensemble learning helps improve machine learning results and solve the problems in (bagging), bias (boosting) by combining several models [19]. In future, we aim to enhance our results through building combination of ensemble models.

7. REFERENCES

[1] Diamonds | Kaggle. (n.d.). Retrieved 5 December 2019, from https://www.kaggle.com/shivam2503/diamonds

[2] Difference Between Supervised, Unsupervised, & Reinforcement Learning | NVIDIA Blog. (n.d.). Retrieved 21 October 2019, from https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

[3] Brownlee, J. (2017, December 10). Difference Between Classification and Regression in Machine Learning. Retrieved 21 October 2019, from Machine Learning Mastery website: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/

[4] Diamonds In-Depth Analysis. (n.d.). Retrieved 5 December 2019, from https://kaggle.com/fuzzywizard/diamonds-in-depth-analysis

[5] Diamond Price Modelling. (n.d.). Retrieved 5 December 2019, from https://kaggle.com/tobby1177/diamond-price-modelling

[6] Diamond (gemstone). (2019). In Wikipedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Diamond_(gemstone)&oldid=918934528

[7] What is correlation coefficient? - Definition from WhatIs.com. (n.d.). Retrieved 5 December 2019, from WhatIs.com website: https://whatis.techtarget.com/definition/correlation-coefficient

[8] Data Preprocessing for Machine learning in Python. (2017, October 29). Retrieved 5 December 2019, from GeeksforGeeks website: https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/

[9] ML | Linear Regression. (2018, September 13). Retrieved 5 December 2019, from GeeksforGeeks website: https://www.geeksforgeeks.org/ml-linear-regression/

[10] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

[11] Chakure, A. (2019, July 7). Random Forest and its Implementation. Retrieved 5 December 2019, from Medium website: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f

[12] Grover, P. (2019, August 1). Gradient Boosting from scratch. Retrieved 5 December 2019, from Medium website: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

[13] Mahto, K. K. (2019, February 25). Demystifying Maths of Gradient Boosting. Retrieved 15 December 2019, from Medium website: https://towardsdatascience.com/demystifying-maths-of-gradient-boosting-bd5715e82b7c

[14] Polynomial regression. (2019). In Wikipedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Polynomial_regression&oldid=928707881

[15] Simpson, M. (n.d.). Machine Learning Algorithms: What is a Neural Network? Retrieved 5 December 2019, from https://www.verypossible.com/blog/machine-learning-algorithms-what-is-a-neural-network

[16] Python | Mean Squared Error—GeeksforGeeks. (n.d.). Retrieved 21 October 2019, from https://www.geeksforgeeks.org/python-mean-squared-error/

[17] Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). (n.d.). Retrieved 21 October 2019, from http://www.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.htm

[18] Brownlee, J. (2018, May 22). A Gentle Introduction to k-fold Cross-Validation. Retrieved 21 October 2019, from Machine Learning Mastery website: https://machinelearningmastery.com/k-fold-cross-validation/

[19] Smolyakov, V. (2019, March 7). Ensemble Learning to Improve Machine Learning Results. Retrieved 5 December 2019, from Medium website: https://blog.statsbot.co/ensemble-learning-d1dcd548e936