Diamonds price prediction

Tuqa AbuRaddaha

Ahmad Deep

***Abstract*—Natural diamond is a mineral composed of a single element–carbon (C). It has a cubic crystalline structure. It generally occurs in the form of octahedral crystals with curved faces, with cubic crystals being rarer. Diamonds are usually colorless. And in this report we are going to make a model to predict the price of Diamonds.**

Keywords—Diamonds, linear, price (*key words)*

I. Introduction

Diamond is a form of the element carbon with its atoms arranged in a crystal structure called diamond cubic. At room temperature and pressure, another solid form of carbon known as graphite is the chemically stable form of carbon, but diamond almost never converts to it. Diamond has the highest hardness and thermal conductivity of any natural material, properties that are utilized in major industrial applications such as cutting and polishing tools. They are also the reason that diamond anvil cells can subject materials to pressures found deep in the Earth. Most natural diamonds have ages between 1 billion and 3.5 billion years. Most were formed at depths between 150 and 250 kilometers in the Earth's mantle, although a few have come from as deep as 800 kilometers. Under high pressure and temperature, carbon-containing fluids dissolved various minerals and replaced them with diamonds. A variety of machine learning regression techniques may be able to analyse the complex relationships between diamonds data features and a possible price for it.

In this study, the input will be 10 columns of diamonds features that are easily obtainable and typically used for diamonds price prediction and the output will be the price prediction for the diamond. We will run a variety of regression algorithms to try to achieve the best accuracy possible.

II. DATASET AND FEATURES

A. *Dataset and Visualization*

In this study, we collected the data from Kaggle website, it was a complete dataset [1] of 53940 rows with 10 columns. We didn’t find any missing data so we so we didn't need to replace it with any other feature. As I mentioned it has 10 features explained in the table.

TABLE I. SHORT DESCRIPTION FOR THE FEATURES

Feature	Description
Price	Price in US dollar (\$326--\$18,823)
Carat	weight of the diamond (0.2--5.01)
Cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
Color	diamond color→ from J (worst) to D (best)
Clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

B. *Visualization*

In order to understand the data deeply we used seaborn library and other useful libraries to analyze the most important plots.

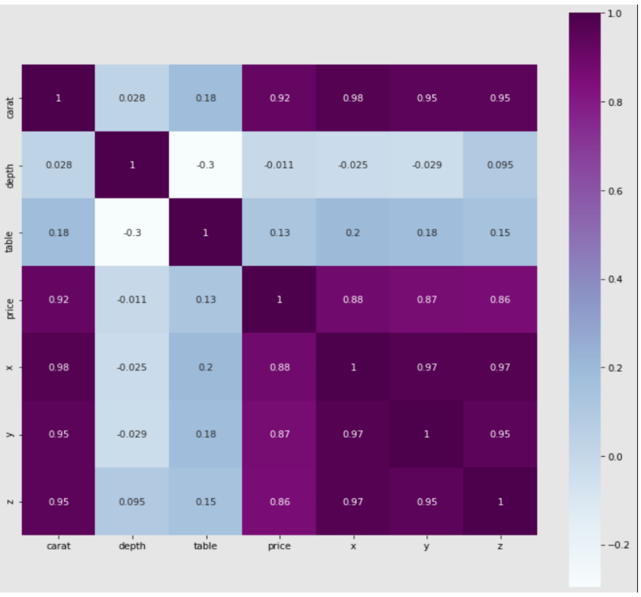


Fig 1. The correlation between the different features

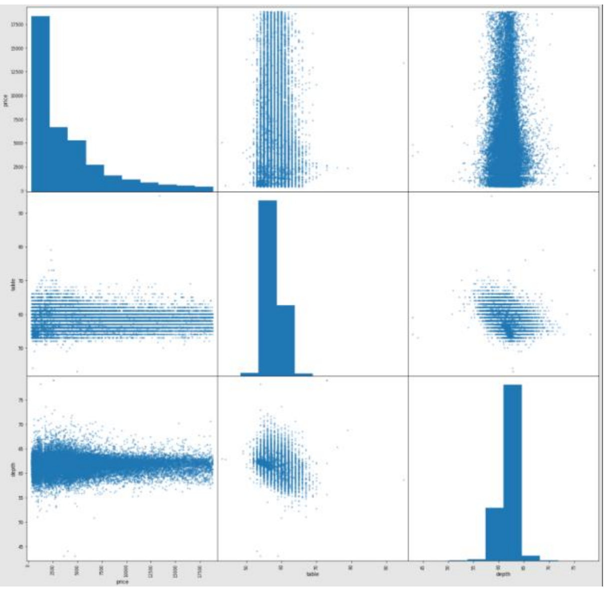


Fig 2. Comarison between Price, table, and depth

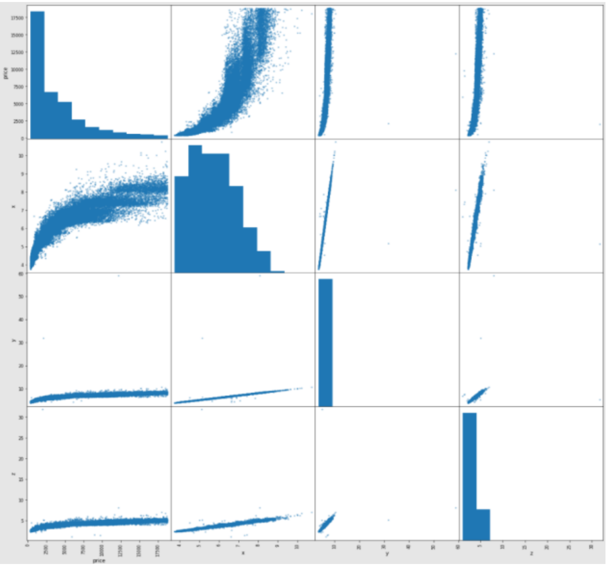


Fig 3. Comarison between Price, x, y, and z

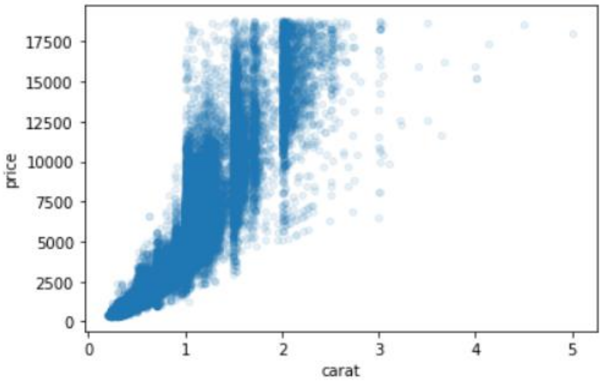
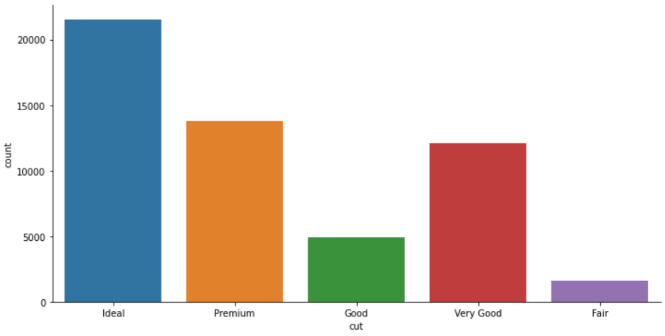
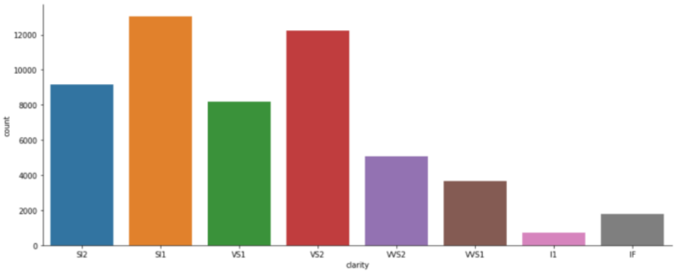


Fig 4. Scatter matrix between carat and price

Then we made other factor plots between the features to understand the data better.





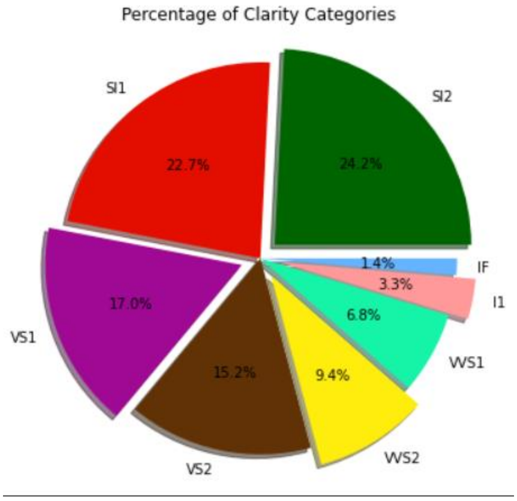


Fig 5. Percentage of clarity categories

III. CONVERT THE TEXT ATTIBUTES

Machine learning prefers to deal with number values, so before applying the algorithms we convert the text attributes to numbers. We used Ordinal Encoder to replace the cut feature values [Fair, Good, Very Good, Premium, Ideal], the color, and the clarity to numbers.

Then we split the data into 80% train set and 20% test set, then after that we used standard scalar to fit and transform the data.

IV. METHODS AND RESULTS

Machine learning algorithms offer to us many approaches to find the accuracy. We used linear logistic, and decision tree. We used these classification algorithms because they are often used for these problems, and they give a good result.

A. Linear Regression

Some regression algorithms can be used for classification problems and vice versa. Logistic regression model make probability based on binary dependent variable. Logistic regression predicts the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities. We use these approach in our dataset which perform a 1525.83 MSR in the test data, and we found that the accuracy score is 88%.

B. Decision Tree Regressor

Next, We tested a Decision tree classifier which is one of the predictive modeling approaches used in machine learning and data mining and is used to go from observations to conclusions about the item's target value. We found that the dataset performs a 736.94 MSR in the test data.

C. Random Forest

Random Forest is a machine learning algorithm that can be used for a variety of tasks either regression or classification. It consists many of small decision trees, called estimators. Each estimator tries to create a model that its prediction is more accurate than any individual tree. The Random Forest I fit my dataset performed comparably to the other models and we found that it performs a 535.57 MSR in the test data with accuracy equals to 99.5%.

V. CONCLUSION

TABLE I. COMPARISON BETWEEN THE MODELS

Data	**LR**	**Decision Tree**	**Random Forest**
Accuracy	88%	98%	99.5%
MSR	1525.83	736.94	535.57

Overall, the study was successful in achieving a slight improvement over existing prediction methods for Diamonds price using the same feature set. Future improvement will require further data collection with a bigger feature space.