Diamond Price Prediction Model Using Data Science

Ms. Arpita Wagulde
Computer Engineering Department
Vidyalankar Institute of Technology
Mumbai, India
arpita.wagulde@vit.edu.in

Ms. Shruti Dhande
Computer Engineering Department
Vidyalankar Institute of Technology
Mumbai, India
shruti.dhande@vit.edu.in

Dr. Umesh Kulkarni
Computer Engineering Department
Vidyalankar Institute of Technology
Mumbai, India
umesh.kulkarni@vit.edu.in

Ms. Sajani Ghosh
Computer Engineering Department
Vidyalankar Institute of Technology
Mumbai, India
sajani.ghosh@vit.edu.in

Abstract—In this popular and growing gem industry, diamond holds a very special place. But along with this popularity comes severe competition for retailers and conmanship towards customers. We discuss and evaluate models to predict the prices of diamonds given their properties. This is important for diamond retailers to appropriately set prices and for customers to estimate prices for diamonds by knowing just a few features about each stone. Using a Kaggle dataset with around diamonds recorded properties for each example

We propose to extract features using variety of approaches in python for purpose of Data mining and visualization and then apply machine learning algorithm. We show that we can build an extremely successful model to predict a diamond's price given the properties and suggest a few diamonds based on the user's choices. We describe our method of choosing this model, optimizing results, and discussing implications of our state-of-the-art method compared to the previous best method.

Keywords—Diamonds, Regression, Prediction, Data Science, Machine Learning, Recommendation

I. Introduction

Diamonds are valued for reasons beyond aesthetics including their hardness (they are the hardest naturally sourced mineral), abrasive nature and ability to disperse light, about one-fifth of diamonds are used in industry, for lasers, drill bits and surgical equipment, largely in the auto and aerospace sectors. They also serve as insulation and have high heat conductivity, suggesting that diamonds may have applications in the semiconductor industry.

Nowadays precious metals around the globe are valued based on their weight, whereas diamonds involve many other factors that significantly affect their price; by the time a diamond reaches a retail store, it has undergone through a series of steps in the supply chain, adding more cost to the precious stone in every single iteration, this leads towards comparatively elevated prices for most available diamonds in the marketplace.

The rough gemstones must first be mined and cut. Once cut, diamonds are appraised to determine their value. Some of them undergo treatments to augment their appearance. Diamonds progress through these processes to be transformed into beautiful valuable pieces of personal adornment.

Some industry leaders like Rapaport, a major trade source for diamond prices, are working toward a standardization that would bring transparency to the pricing of diamonds. The diamond jewelry trade is fairly unified in claiming that this cannot be accomplished. Taking an adamant stance that diamonds cannot be considered a commodity, while by most definitions they can, the trade insists that each diamond is too unique and that standardization cannot account for the diamonds symbolic value of "enduring love and commitment" (which they believe and spend millions in advertising to convince consumers of this notion). The implied thought is that diamonds are nearly priceless and that their value would not hold should they be traded as a commodity.

Although the process of mining, cutting and polishing may set the baseline price for a diamond, the following features are the major defining factors to consider in order to obtain an accurate figure: carat weight, cut, color, clarity, length, width, depth, depth percentage and table width. The 4Cs describe the individual qualities of a diamond, and the value of an individual diamond is based on these qualities. The terms that people use to discuss the 4Cs have become part of an international language that jewelry professionals can use to describe and evaluate individual diamonds. Today, the descriptions of each of the 4Cs are more precise than those applied to almost any other consumer product. And they have a long history. Three of them—color, clarity, and carat weight—were the basis for the first diamond grading system established in India over 2,000 years ago.

The characteristics of the data can be observed to establish the factors associated with an outcome. Observation studies such as data mining, can reveal the association of the features to the target outcome. Data driven statistical research is becoming a common component to a plethora of areas such as stock market, product and business development. The discovery of diamond value prediction is possible by extracting the insights in the data that are directly related to the mineral.

The main purpose of this project is to predict the diamond prices by employing the Kaggle[1] diamond dataset and regression approaches to determine an accurate outcome. And then recommend a few diamonds based on the user's choices.

II. Dataset

The dataset for the model is obtained from Kaggle[1] having 53,940 instances and 10 attributes divided in three categories:

1. *Qualitative Features (Categorical attributes):*
   A categorical attribute is an attribute that cannot be numbered or given in digits or values. It is given in description.
   The categorical attributes in this dataset are cut, color and clarity.
2. *Quantitative Features (Numerical attributes):*
   Numerical attributes are the attributes that can be given in numbers or values.
   The numerical attributes in this dataset are carat, depth, table, X, Y, Z.
3. *Target variable:*
   Target variable is the variable which we want to predict.
   The target variable in our project is Price measured in US dollars ($).

Carat: Carat is the mass of the diamond. 1 carat (ct) is equal to 200mg. This is the only quantitative measure of the 4 Cs. Distribution of attribute Carat in dataset is shown in Fig 1.

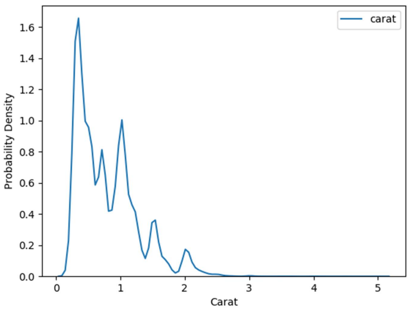

Fig. 1 Distribution of Carat in data

Cut: Cut refers to both the shape of the stone and the quality of its scintillation. The cut perfection is classified from "Fair" to "Ideal". Distribution of attribute Cut in dataset is shown in Fig 2.

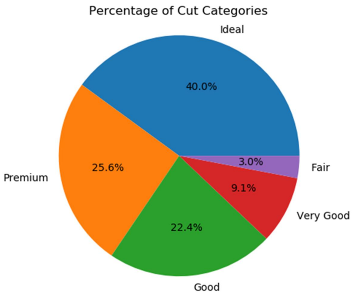

Fig. 2 Distribution of Cut in data

Color: Diamond colors vary from colorless to a light yellow. The more colorless a diamond is, the more expensive it is likely to be. The standard is a classification developed by the Gemological Institute of America and is the most used out of all the color grading schemes; it uses an alphabetical score, "D" being the most colorless and "Z" being a prominent yellow. Distribution of attribute Color in dataset is shown in Fig 3.

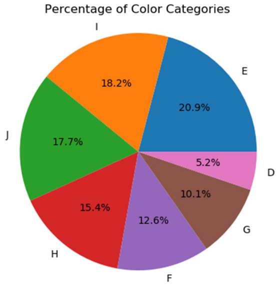

Fig. 3 Distribution of Color in data

Clarity: Diamonds may have internal blemishes and fractures which decrease their transparency, which in turn decreases their value. Clarity is graded on a scale from FL (Flawless) to I3 (Obvious Inclusions) based on the size, nature, position, and quantity of internal blemishes. Distribution of attribute Clarity in dataset is shown in Fig 4.

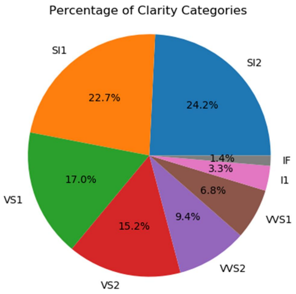

Fig. 4 Distribution of Clarity in data

There are several other properties that might affect the price of a diamond. While these might not be as popular as the commonly used 4 Cs, they still may measure some factors of the diamond that could affect the quality and consequently the price of the diamonds.

Depth: Depth is measured as the ratio between z and the average width of the top of the diamond.

Table: Table is measured as the width of the top of the diamond at its widest point.

III. Predictive task

Our goal is to predict the price of diamonds using features such as carat, clarity, color, cut, depth, table length, x, y, and z axis lengths in millimeters. The depth, table length, x, y, and z axis lengths were given as numerical data.

We take off our predictive task by studying correlation between attributes. We can explore the data by taking a heatmap of the correlations between each of the potential features (Fig 5).

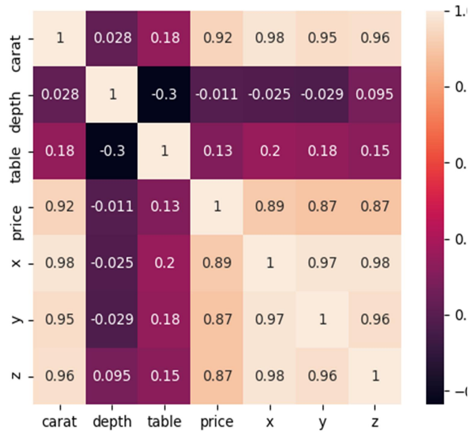

Fig. 5 Heatmap

Conclusions drawn from the heat map:

- Depth is inversely related to Price.
- The Price of the Diamond is highly correlated to Carat, and its Dimensions.
- The Weight (Carat) of the Diamond has the most significant impact on its Price.
- The length(x), width(y) and height(z) seems to be highly related to the Price and even each other.
- Self-Relation that is of a feature to itself is 1 as expected.

It's worthwhile to identify that price is strongly linearly correlated with the carat of a diamond, while not perfectly correlated. We know now that we can use carat as a strong predictor for the price of a diamond. We can observe the same in Fig 6.

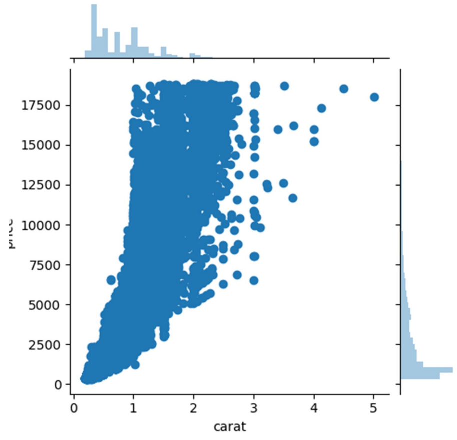

Fig. 6 Correlation between Carat and Price

The heatmap does not reveal everything about the data, however. Other than carat and the price of the diamond, there does not seem to be any linear correlation, but if we plot the data visually while labelling the categorical properties, we can see there is some sort of pattern.

By studying the correlation between cut and price, it can be observed that the cut can drastically Increase or Decrease value of price. With a Higher Cut Quality, the Diamond's Cost per Carat Increases. As seen from Fig. 7, premium cut diamonds have the highest price.

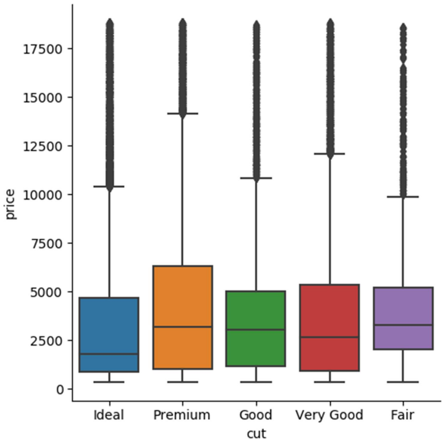

Fig. 7 Correlation between Cut and Price

While it's a little difficult to tell if there is some sort of relationship between the colour of a diamond and its mass and its price, there is a little bit of structure that suggests that higher priced diamonds with less mass tend to be of higher colour grade. The plots also suggest that higher priced diamonds with less mass tend to be of better-quality cut, and very strongly so for higher clarity. (Fig. 8)

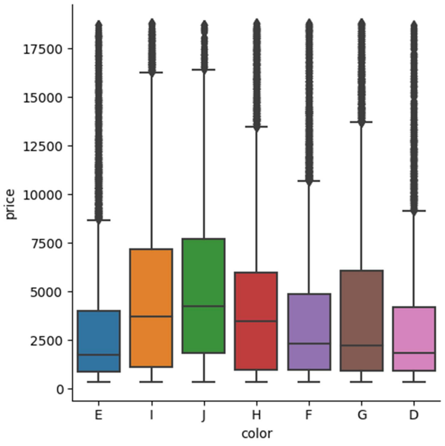

Fig. 8 Correlation between Color and Price

From the heatmap, we observed that the length(x), width(y) and height(z) seems to be highly related to each other. Hence we create a new feature which is Volume which will be product of all the three x, y, z attributes. By studying the graph (Fig. 9) between Volume and price it is observed that the features are linearly correlated with a steep graph.

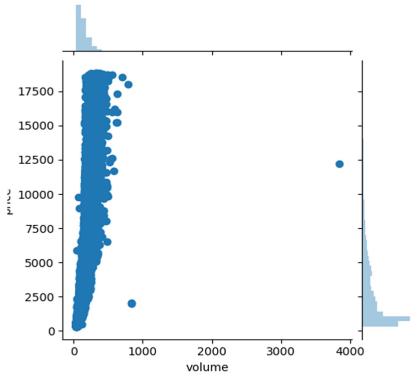

Fig. 9 Correlation between Volume and Price

As a takeaway however, in reference to our predictor, we can visually conclude that there is some positive relationship between cut, clarity, color and volume with price, and we can base a predictor based off these features.

IV. Design

We have made a machine learning predictor using Supervised learning technique in order to train the model on a dataset. Machine Learning model deployment usually involves two phases: training phase and testing phase.

A. *Training Phase*

It is the first step. In the training phase the algorithm used in the model learns to predict the correct output. Since we are using Supervised learning we trained our model on a dataset. Training phase is more important than the testing phase as the training phase directly determines the accuracy of our machine learning model.

B. *Testing Phase*

In this phase our trained model predicts the output of unlabeled input. It is called the testing phase because the accuracy of our trained model is calculated in this phase.

To provide a uniform experience across multiple platforms, a web GUI is designed, which will enable the users to put reviews that are to be classified and display the corresponding results in an organized and responsive way. The website is powered by PHP on the back end.

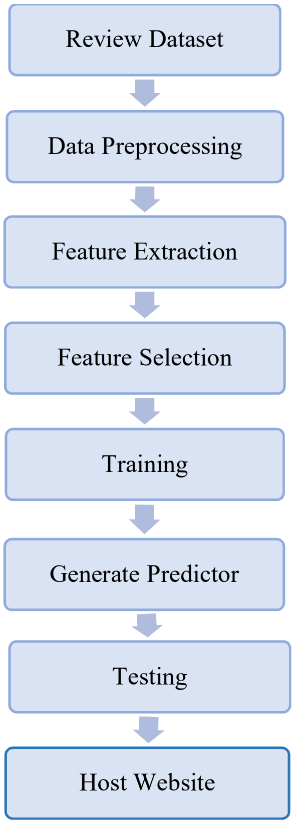

Fig. 10 Data Flow Diagram

V. Model

The price of the diamond was predicted by training 4 regression models as explained below.

The first baseline model we used was *Linear regression*. Linear regression is used for finding a linear relationship between the target and one or more predictors.

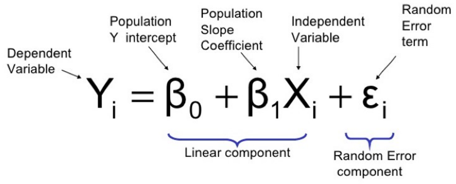

Fig. 11 Linear Regression

The second model we used was *Random Forest regression*. Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which aggregates many decision trees.

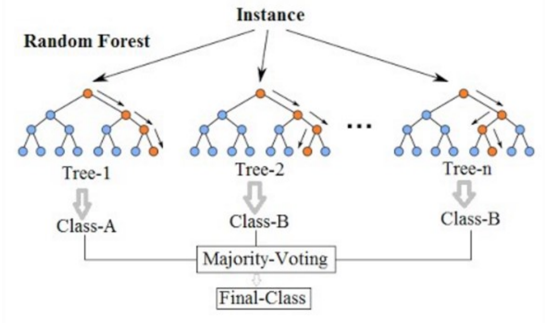

Fig. 12 Random Forest Regression

The third model was *Gradient Boost regression*. Decision trees are used as the weak learners in gradient boosting. Decision Tree solves the problem of machine learning by transforming the data into tree representation. Each internal node of the tree representation denotes an attribute and each leaf node denotes a class label. The loss function is generally the squared error (particularly for regression problems). The loss function needs to be differentiable.

The fourth model used is *K Neighbors regression*. K Neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure.

Table 1 r2 scores

Model Name	**r2 score**
Linear Regression	0.88
Random Forest Regression	0.98
Gradient Boost Regression	0.97
KNeighbors Regression	0.96

The best r2 score we obtained was in Random Forest Regression model and henceforth our model will be using Random Forest Regression.

VI. Recommendation Engine

To give an edge to any normal prediction model, we recommend user diamonds based on the choices they provided. If a user has asked for the price of diamond of *x carats* then he is more likely to buy diamond with their carats near x. This approach can be used for Quantitative attributes like carat, table, volume, price. If a user has asked for the price of diamond with a certain quality of cut then he is more likely to buy diamond with their cuts of almost same quality. This approach can be used for Qualitative attributes like cut, color, clarity.

This recommendation engine shows user the diamonds he will like from the dataset with its prices also nearer to the diamond's price the user asked for.

VII. Future Scope

The dataset we have used does not contain an attribute *Rarity,* which can be used to make our model better because many a times the price varies depending upon how rare the diamond was. By including Rarity attribute, model can be made better with increased accuracy.

Use of Visualization techniques to help user see for what diamond he/she has requested the price for. The model will showcase a picture of the diamond that has been requested by the user along with the price generated by the model so that the user can have a proper visual of what they have asked for.

Every user has a certain price range. A diamond that the person can afford and also matches his expectations of cut, color and clarity. This model helps such users by giving them suggestions which will aid the user in order to buy the diamond which they desire or closest to their preference and present the user with appropriate choices that will make their experience, a good and easy one.

VIII. Conclusion

In our system we predict the price of the diamond based on the values of the attributes provided by the user using Supervised Machine Learning Technique. The system also recommends user the diamonds based on his choices.

The accuracy of the model would depend on the number of features on which the model is trained and how well the data is pre-processed. We were able to achieve 98.07% accuracy by using Random Forest regression.

Our model will take diamond industry to another level by helping retailers widen their scope by deciding appropriate prices for their diamonds. Our model will help customers as well by helping them know an approximate price of the diamond they are looking to buy. This model will help create an authentic market in real world.

IX. References

[1] Diamonds. Diamonds — Kaggle, Unknown, 25 May 2017, www.kaggle.com/shivam2503/diamonds.

[2] Cardoso, Margarida G. M. S., and Luis Chambel. A Valuation Model for Cut Diamonds. *International transactions in Operational Research, Blackwell Publishing*, 7 July 2005, onlinelibrary.wiley.com/doi/10.1111/j.1475-3995.2005.00516.x/pdf.

[4] Nicolas Stier-Moses, Assaf Zeevi, Deconstructing the Price of Diamonds, *Columbia Business School*, February 24, 2008.

[5] Singfat Chu, National University of Singapore. Pricing the C's of diamond stones, *Journal of Statistics Education Volume 9, Number 2 (2001)*.

[6] R. Saravanan, Pothula Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification", June 2018

[7] Rabiul Islam Jony, Nabeel Mohammed, Ahsan Habib, Sifat Momen, Rakibul Islam Rony, "An Evaluation of Data Processing Solutions Considering Preprocessing and Special Features"

[8] Agata Nawrocka, Andrzej Kot, Marcin Nawrocki, "Application of machine learning in recommendation systems", May 2018