# Diamond Price Prediction using Machine Learning

Vincent Rivière

July 2019

Figure 1: Diamond

***Abstract -*** *Diamonds are found as rough stones and must be processed to create a sparkling gem that is ready for purchase. The value of diamonds as an investment is of significant interest to the general public, because they are expensive gemstones, often purchased in engagement rings, due in part to a successful 20th century marketing campaign by De Beers. Tavernier's law is used to determine the price of a diamond. The formula is for basic calculation and demonstrates how the price of a diamond increases along with its size. Larger gemstones are rarer and go up rapidly in price.*

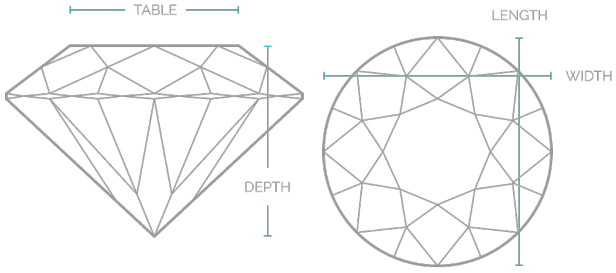
$$Price = W^2 * C$$

*where W is the weight in Carats and C is the basic price of a one-carat stone.*
*However, this method is very limited since there is a lot of diamond shape, color or clarity.*
*In this paper, we discuss about models which predict prices of diamonds given their properties. It is very important for diamond retailers or customers to appropriately estimate a diamond price by knowing just few features of the stone.*
*In this article, we will use a Kaggle dataset of 54.000 diamonds. For each diamond, the dataset include few features such as the weight in carat, the clarity, the color et cetera.*

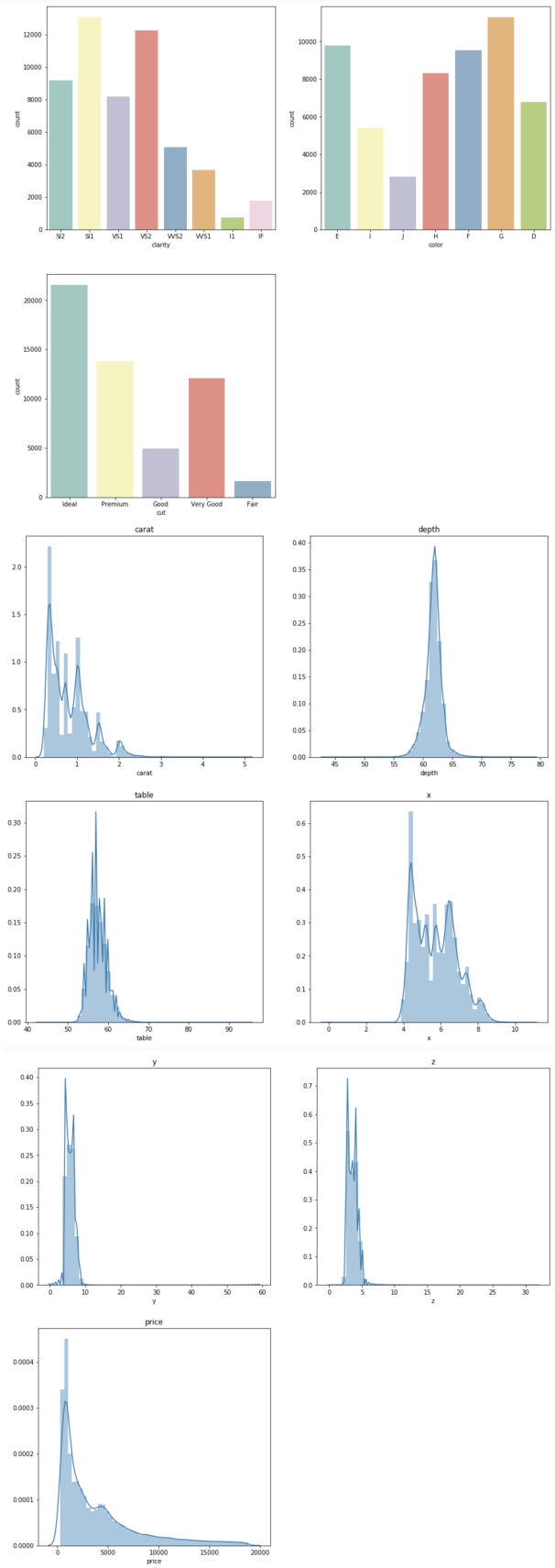
# 1 Introduction

For centuries, diamonds have been a sign of power, wealth and status. The stone was a rare find and therefore was worth more.
However, in the 1800s, a veritable diamond trove was unearthed in Kimberly, South Africa. This new found mine had the potential to flood the market with diamonds and bring down the cost for the precious stone. To prevent too many diamonds from hitting the market, De Beers quickly intervened, bought up the mine and maintained tight control over the global diamond supply.
For centuries, diamonds have been worn by royalty and noblemen as a status symbol. And diamond engagement rings have been traced all the way back to 1477 when the Austrian Archduke Maximilian proposed to Mary of Burgundy with a diamond ring.
Since scientists have highlighted diamond property that can distinguish its value. Also, the diamond cut is very important in the price prediction. The diamond cut is a style or design guide used when shaping a diamond for polishing such as the brilliant cut. The cut of a diamond greatly affects a diamond's brilliance, this means if it is cut poorly, it will be less luminous.
Our dataset contains the column carat, cut, color, clarity, depth, table, x, y, z and the price.
**Carat** is the weight of the diamond in carat. Currently, one carat is a unit of mass equal to 200 mg.
**Cut** is the quality of the cut. This column can be Fair, Good, Very Good, Premium or Ideal.
**Color** is the color of the diamond from J (worst) to D (best).
**Clarity** Clarity of the diamond from I1 (worst) to IF (best). This column can be I1, SI2, SI1, VS2, VS1, VVS2, VVS1 or IF.
**Depth** is the depth of the diamond. In terms of cut quality, depth is described in percentages. It can be calculated by dividing the diamond's physical depth measurement by its diameter. In our case, depth = z / mean(x, y) = 2 * z / (x + y).

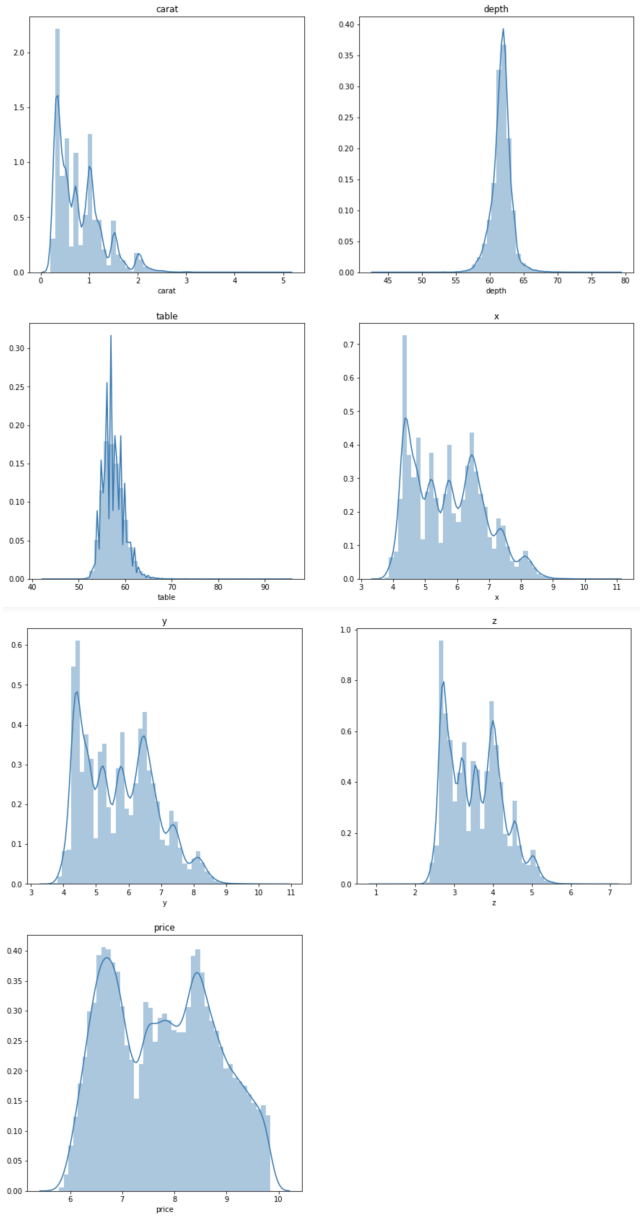

**Table** is calculated by dividing the diamond's physical table measurement by its diameter.
**X** is the length in mm.
**Y** is the width in mm.
**Z** is the physical depth measurement in mm.




Before going further, here is an analysis of the distribution of the data. This allows us to better understand the data that we will be processing, but also to detect the noise that we will have to manage. Indeed, the dataset we are going to process contains noise like zero width / length.
Understanding the underlying data distribution before applying any machine learning or statistical modelling approach is one of the most important concept.



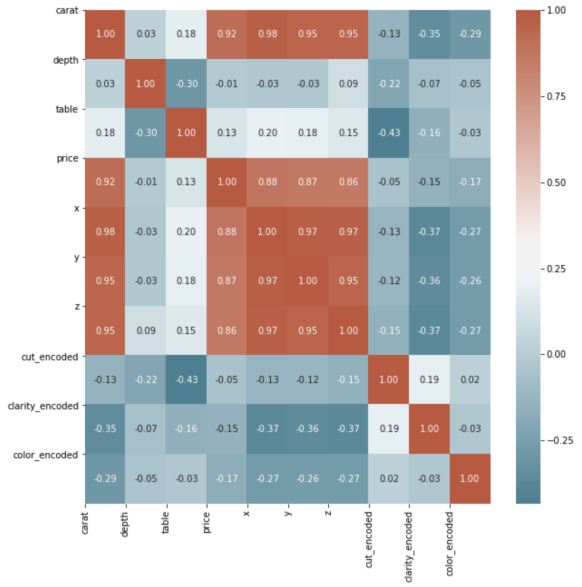

Through these distributions, we can see many problems. First, there are most likely false values (noise) in our data set. Notably zero or too big values. This can be seen by looking at the minimum and maximum values chosen by seaborn for Y and Z columns.
Then we can also see that the distribution of the prize is completely skewed to the right.
In order to get better results, we will remove false data and apply logarithm to the prices.
Once the noise is removed, we find that the distribution of Y, Z and price are much better now.



## 2 Data analyze

Before going further, we will first look at our data. More specifically the relationships that each feature has with each other.

Correlated features do not affect classification accuracy. For a fixed number of training examples, increasing the number of features typically increases classification accuracy to a point but as the number of features continue to increase, classification accuracy will eventually decrease because we have a limited dataset relative to the large number of features.

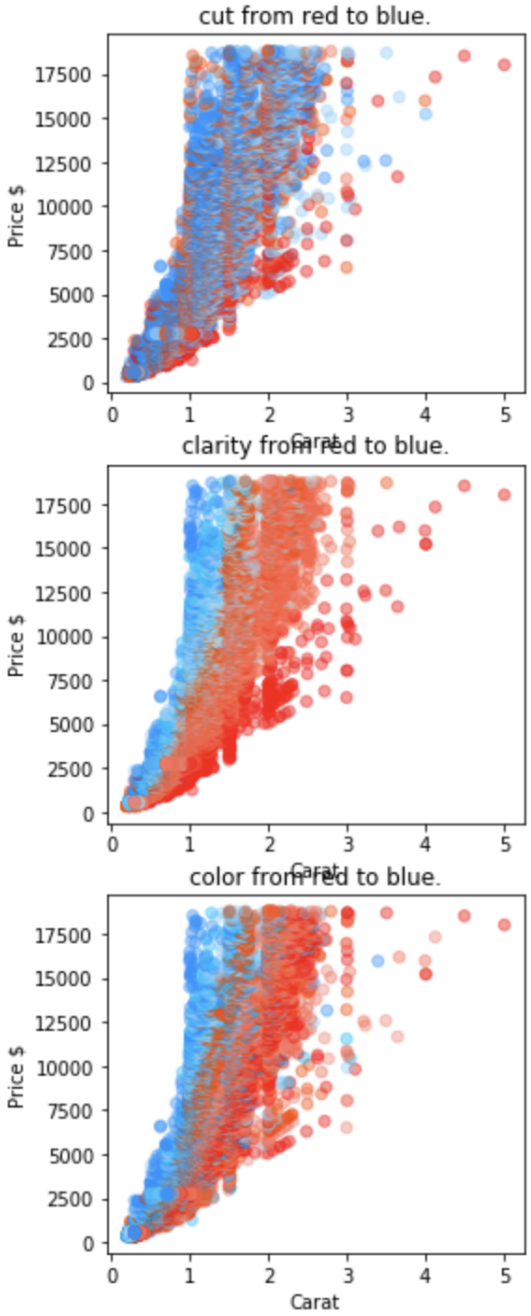



We first find that X, Y, Z and Carat are extremely corelated. This is completely normal since the diamonds are all cut in the same proportions and the weight depends directly on the dimensions. In fact, we will not use the colons X, Y and Z. This limits the number of features we will pass to the model and thus save time during training.

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.

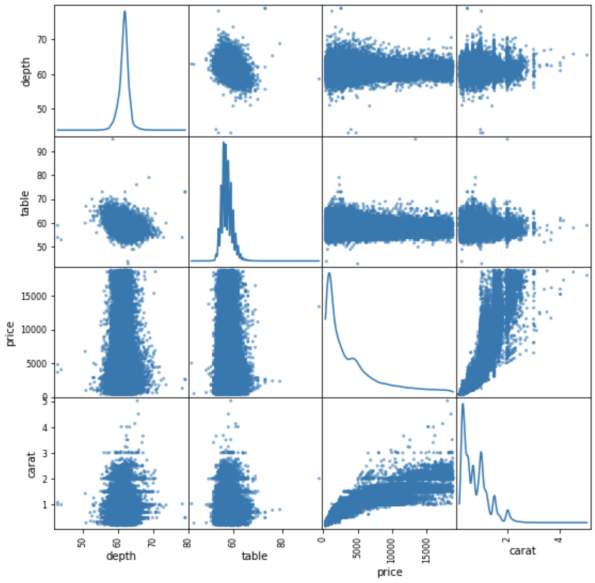
The heatmap does not reveal everything about the data, but except carat, the size, and the price of the diamond, there does not seem to be any linear correlation. Therefore we will plot the data visually while labeling with color the categorical properties.

This plots show that higher priced diamonds with less mass tend to be of higher color grade, better quality cut, and very strongly higher clarity.




Then we find that the corelation between depth, table and price are close to 0 (-0.01 and 0.13). When the correlation value of two features is close to zero, generally between -0.1 and +0.1, the variables are said to have no linear relationship or a very weak linear relationship.

To know if the depth is important we will use different approach of plotting.




This graph suggests that the higher the price, the more the table and depth converge to 60

## 3 Features selection

Now that we have done the analysis of our dataset, here are the features we will keep for our model.

$$f(carat, cut, color, clarity, depth, table) = price$$

The features cut, clarity and color were given as categorical. We decided to transform them into numeric value using an encoder and keeping their order.

Then depth, table and carat axis were given as numerical data. Thus, we normalized the data by subtracting the mean from each data point and dividing by the standard deviation.

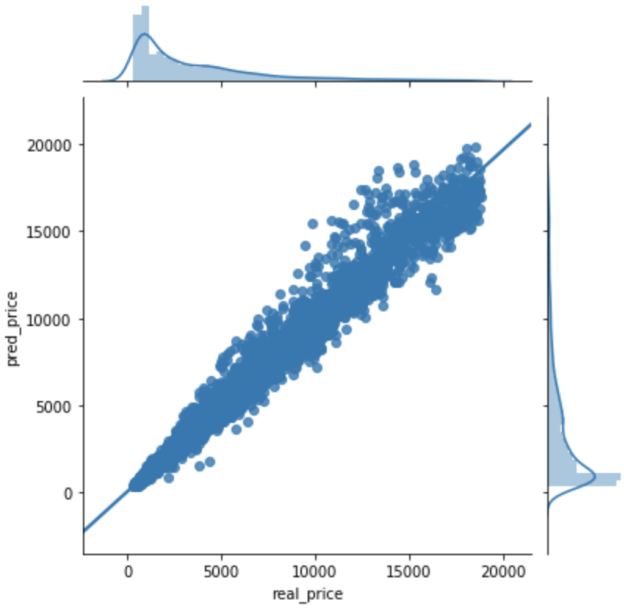
This will result in having a data with a mean of zero and a unit variance. This method is widely used for normalization in many machine learning algorithms.

## 4 Chose the model

Since we want to predict the price of the diamonds, we will use a regression model over classification. To evaluate our prediction, we will split the 54.000 row we needed to split the data into training, validation, and test sets.

We used 20% of our dataset as test set. Also, we used 10% of the remaining training dataset as validation set.

We used a simple multi layer scikit learn regressor model and obtained encouraging results.




Concerning the metrics, we used Mean Absolute Error, Median Absolute Error and R2 Score.

| Metrics | Result |
|---|---|
| Mean Absolute Error | 288€ |
| Median Absolute Error | 119€ |
| R2 Score | 0.98034 |

Current state of the art metrics are

| Metrics | Result |
|---|---|
| R2 Score | 0.99241 |

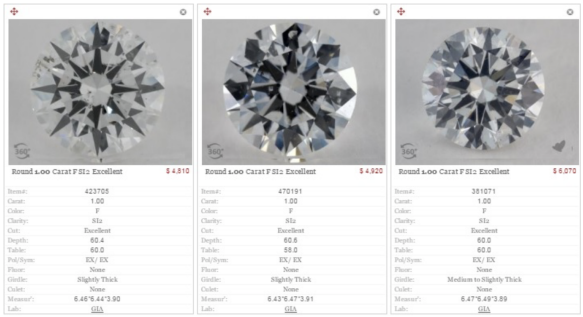
In a few minutes of programming, we obtained convincing results compared to the current state of the art.

## 5 Why machine learning ?

Now that we have a model capable of estimating an approximate price for a diamond based on the four 'C's (Carat, Color, Clarity and Cut), we will be interested in the visual differences. Indeed, the characteristics of a diamond are not enough to calculate its true value.

In the world of jewelery, the technical characteristics are not enough. Indeed the price of a jewel

is exclusively calculated according to its rarity, and the jewels have visual attributes that influence their rarity.



The more expensive diamond is eye-clean while the 2 cheaper options aren't. The reason is simple. Even if these three diamonds have exactly the same characteristics, the inclusions of the diamond on the right are much less disturbing to the naked eye than the two others.

Another attribute of diamonds that can change their price is fluorescence. In general, blue fluorescence lowers the value of colorless (D-F) diamonds. In near-colorless (G-J) diamonds, medium to strong blue fluorescence can actually add a slight premium to the stones. Other colors of fluorescence like yellow or green will cause the diamond to trade at a discounted price.

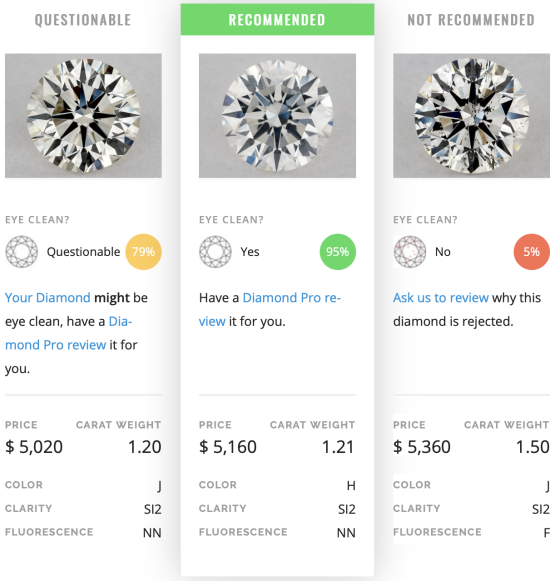
As you can see, the price of a diamond can never be calculated by a simple mathematical formula. Only real jewelery experts can estimate the true price of a diamond. But today it is possible to predict the true price of a diamond without having to move to a professional thanks to artificial intelligence.

Indeed, by combining the result of our previous model with a new model using deep learning through images, it is quite possible to accurately predict the price of a diamond in today's market.

Furethemore, artificial intelligence eliminates the human error factor. AI creates a non-biased standardized accuracy that can be leveraged by gem labs to improve the consistency of diamond grading, and adhere more strictly to industry grading standards. This increases the reliability and consistency of results, which is a benefit for all parties manufacturers, gem labs, retailers and consumers.

# 6 State of the art

During the realization of this work, I discovered a tool using the artificial intelligence to detect the imperfections of the diamonds.




Ringo is an AI tool that evaluates diamond inclusions and flaws. Ringo's patent-pending AI technology can spot inclusions – a particle or flaw visible to the naked eye. When inclusions or flaws are not visible to the naked eye, a diamond is considered eye-clean. Ringo's AI model is trained to use human-like visual perception to determine if each diamond is eye-clean, once the diamond is placed in a ring setting of the buyer's choice. This was accomplished by running several thousand eye-clean or not-eye-clean decisions by The Diamond Pros through Ringo's AI engine.

Also, in May 2018, LLC Announces an artificial intelligence software to grade diamonds. GemAppraiserAI is a machine-learning algorithm for evaluating the quality and authenticity of gemstones. In addition to instantly determining whether a stone is natural or synthetic, the algorithm outputs the four 'C's (Cut, Clarity, Carat and Color).

Using this technology, the consumer can instantly determine diamond grade while shopping at the jewelry store or at home with a smartphone or tablet.