Diamond Price Prediction by Physical Features based on Seven Machine Learning Algorithms

MAE University of California, Los Angeles

Author: Yinuo Song

Faculty Advisor: Randall R. Rojas

Date Completed: June 1, 2020

Table of Contents

1. Abstract

In this research, I'd like to do research about the diamond price prediction based on its nine physical features. I apply seven modelling algorithms to the dataset including Linear Regression, Lasso Regression, AdaBosst Regression, Ridge Regression, GradientBoosting Regression, Random Forest Regression and KNeighbours Regression. And I find that Random Forest Regression gives the highest R2-Score of [98.2%] while Ridge Regression gives the lowest R2-score of [75.4%]. The accuracy scores of Linear Regression, Lasso Regression and AdaBoost Regression are very close to each other; the accuracy scores of GradientBoosting Regression, KNeighbours Regression and RandomForest Regression are higher than others.

2. Introduction

"Diamonds last forever," a slogan of the century has changed the diamond industry. Diamond's characteristic chemical composition and crystal structure make it a unique member of the mineral kingdom.

Precious stone is the main diamond made of a solitary component: It is commonly about 99.95 percent carbon. Some follow components can impact its shading or gem shape. The manner in which a mineral structures decides its character. Precious stone structures under high temperature and weight conditions that exist just inside a particular profundity go underneath the world's surface. Precious stone's gem structure is isometric, which implies the carbon molecules are reinforced in basically a similar route every which way. With no one of these variables, precious stone may be simply one more mineral. Luckily, however, this uncommon blend of compound creation, precious stone structure, and development process gives jewels the characteristics that make them exceptional.

The unique identity and extraordinary features have made diamonds valuable in the market. According to the diamond report from Bain & Company, the rough diamond needs to be cut and polished to become diamond jewelry and the diamond jewelry needs to be designed and manufactured before it goes to retail sales. The price and revenue of diamond will improve throughout the process and this trend is expected to go up in the next few years. "Three patterns have the most elevated potential to influence the precious stone industry in the close to term: headways in advanced technology, the improvement of lab-developed diamonds and generational moves in consumer inclinations" (Bain & Company, 2018).

In this research, I'd like to do research about the diamond price prediction based on its nine physical features. I will do exploring dataset and examine what features affect the price of diamonds in the second part; and I will draw correlation between features in the third part; then I will make visualization of each feature and its relationship with price in the fourth part; Next I will apply seven modelling algorithms to the dataset including Linear Regression, Lasso Regression, AdaBosst Regression, Ridge Regression, GradientBoosting Regression, Random Forest Regression and KNeighbours Regression in the fifth part; I will make comparison and visualization of R2-Score of the seven algorithms I have applied in the sixth part; Lastly I will make conclusions in the final part.

3. Dataset Exploration

The classic dataset I use contains the prices and other attributes of almost 54,000 diamonds. Features includes carat, cut, color, clarity, depth, table, price, x, y, z. Among these features, cut, color and clarity are qualitative features while price is the target variable.

Table 1

Variable	Description
Carat	Carat weight of the diamond.
Cut	Describe cut quality of the diamond. Quality in increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Color of the diamond. With D being the best and J the worst.
Clarity	Diamond clarity refers to the absence of the inclusions and blemishes. (In order from best to worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3.
Depth	The height of a diamond, measured from the culet to the table, divided by its average girdle diameter.
Table	The width of the diamond's table expressed as a percentage of its average diameter.
X	Length of the diamond in mm.
Y	Width of the diamond in mm.
Z	Height of the diamond in mm.
Price	The price of the diamond.

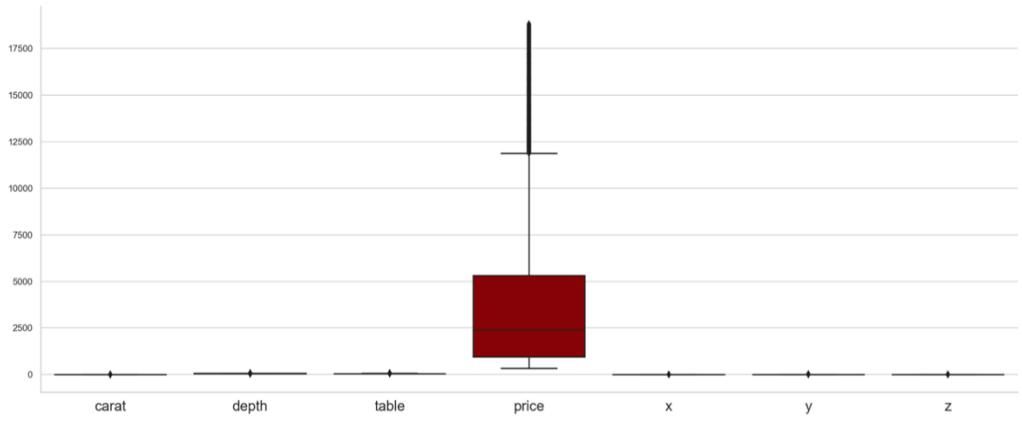
I have pre-processed the data by dropping all the missing value and null value. The description of the dataset with numerical variables are as follows in Table 2. We have 53940 samples in total, and the mean value of carat is 0.797940; the mean value of depth is 61.749405; the mean value of table is 57.457184; the mean value of price is 3932.79972; the mean value of x is 5.731157; the mean value of y is 5.734526; the mean value of z is 3.538734.

Table 2

	carat	depth	table	price	x	y	z
count	53940.00	53940.000	53940.000	53940.0000	53940.00	53940.00	53940.00
mean	0.797940	61.749405	57.457184	3932.79972	5.731157	5.734526	3.538734
std	0.474011	1.432621	2.234491	3989.43973	1.121761	1.142135	0.705699
min	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.00000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5324.25000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.0000	10.74000	58.90000	31.80000

Since the data range is very high, so I make the scaling of all the features. As seen from Figure 1, which shows the results after the scaling, all the qualities are conveyed over a small scope in the wake of scaling, which is much simple for investigation in the later part.

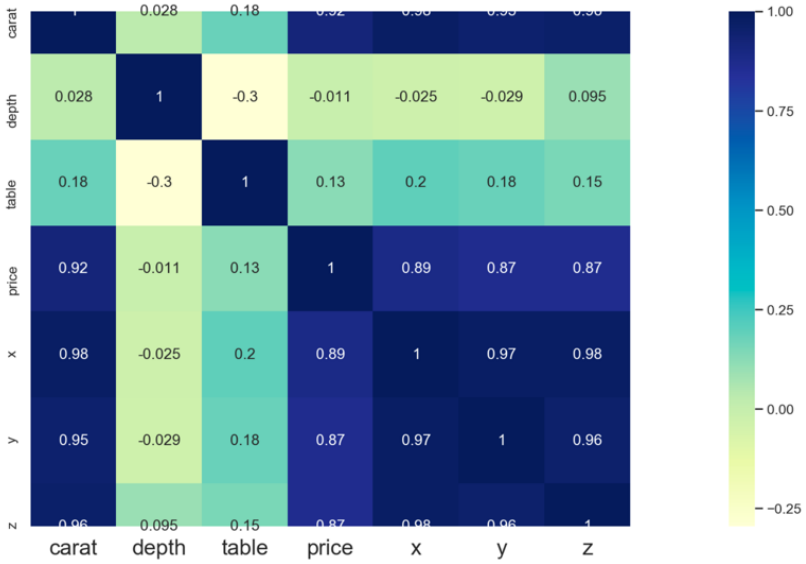
Figure 1: scaling of features



4. Correlation Between Features

Figure 2 shows the correlation of all features and price of diamond. According to the correlation plot from Figure 2, it can be seen that diamond depth is negatively related with diamond price. This is in such a case that a diamond's depth rate is too huge or little the diamond will become dark in appearance since it will not, at this point return an alluring measure of light; the price of the diamond is profoundly associated to carat and its measurements; the weight (carat) of a diamond has the most huge effect on its price. Since the bigger a stone is, the rarer it is, one 2 carat diamonds will be progressively 'costly' than the absolute expense of two 1 carat diamonds of a similar quality; the length(x) , width(y) and height(z) is by all accounts profoundly identified with price and even one another.

Figure 2: correlation map



5. Visualization of all features

5.1 Carat

Carat refers to the weight of the stone, not the size. The heaviness of a precious stone has the most essential effect on its cost. "KDE plot described as kernel density estimate is used for visualizing the probability density of a continuous variable". Figure 3 shows the kde plot of carat of diamond of this dataset, visualizing the probability density of the carat.

Figure 3: carat kde plot

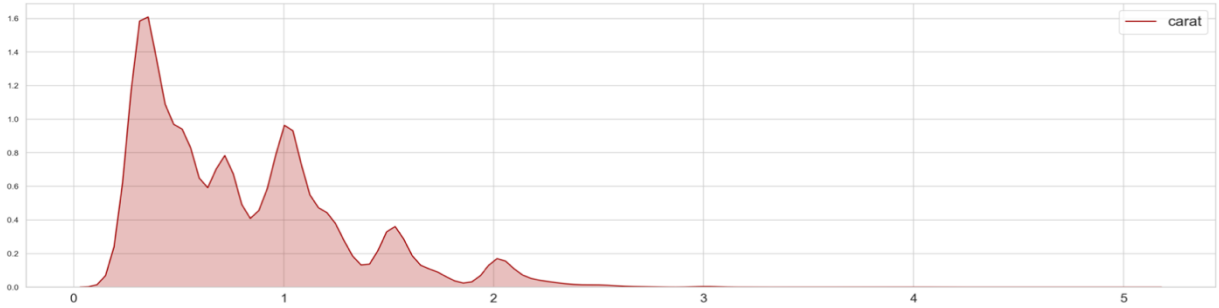
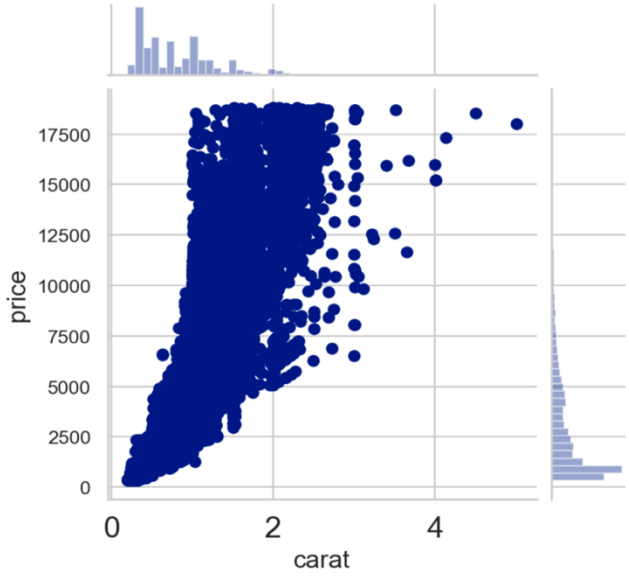


Figure 4 shows the relationship between carat and price of diamond. It seems that carat varies with price exponentially. The more the carat is, the higher the price gets which makes sense.

Figure 4: carat & price



5.2 Cut

Figure 5 shows the distribution of diamonds with different cuts. The cut can now definitely increment or decline its worth. With a higher cut quality, the diamond's expense per carat expands (Figure 5).

Figure 5: cut plot

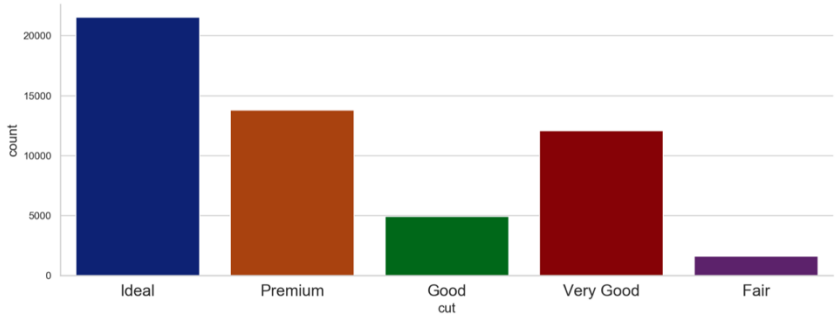
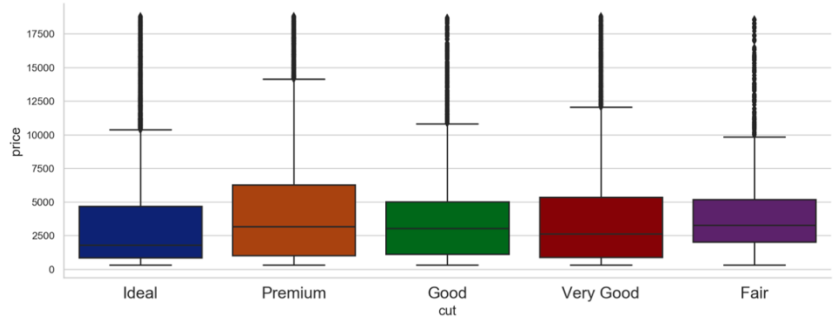


Figure 6 shows the relationship between diamond price and its cut. The bottom line shows the minimum price value; the upper line shows the maximum price value; the middle line in the box is the median value. Premium cut on diamonds have the highest price which can be seen from the plot.

Figure 6: cut & price



5.3 Color

Figure 7 shows the distribution of diamonds with different colors. The color of diamond can go from lackluster to a yellow or a swoon caramel hued tint. Colorless diamonds are rarer and much more valuable on the grounds that they seem whiter and more splendid (Figure 7).

Figure 7: color plot

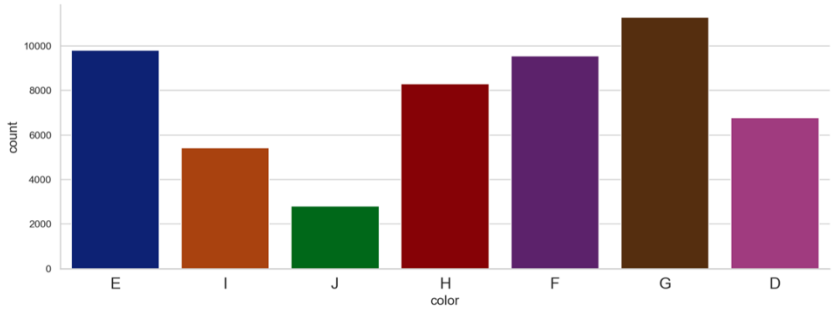
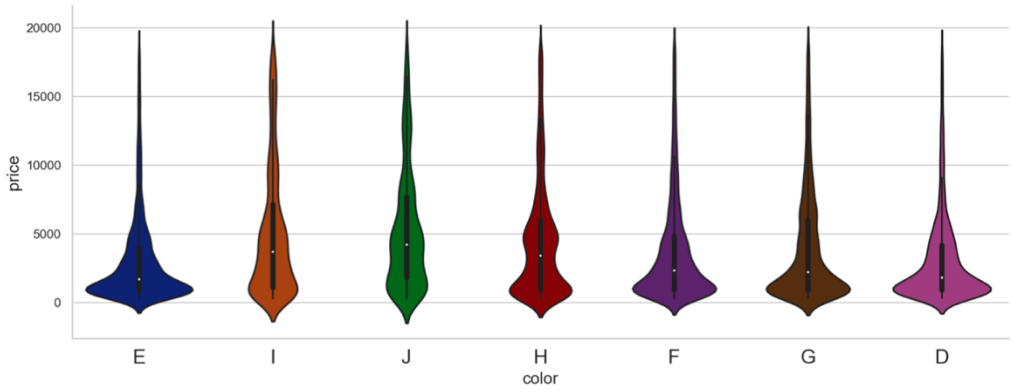


Figure 8 shows the relationship between diamond price and its color. The colorless the diamond is, the higher its price will be. And its price will range higher than more colorful diamonds.

Figure 8: color & price



5.4 Clarity

Figure 9 shows the distribution of diamonds with different percentages of clarities. Diamond clarity refers to the absence of the inclusions and blemishes. An inclusion is a flaw situated inside diamond. Blemishing is an aftereffect of cleaning process incorporating scratches and so on.

Figure 9: percentage of clarity categories

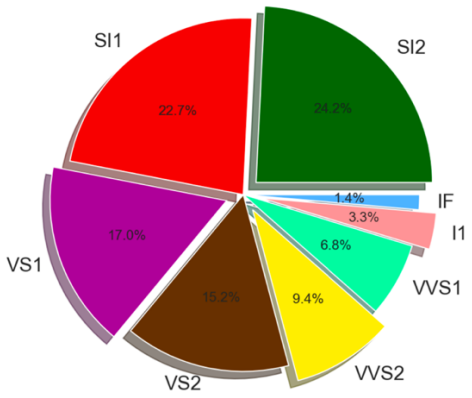
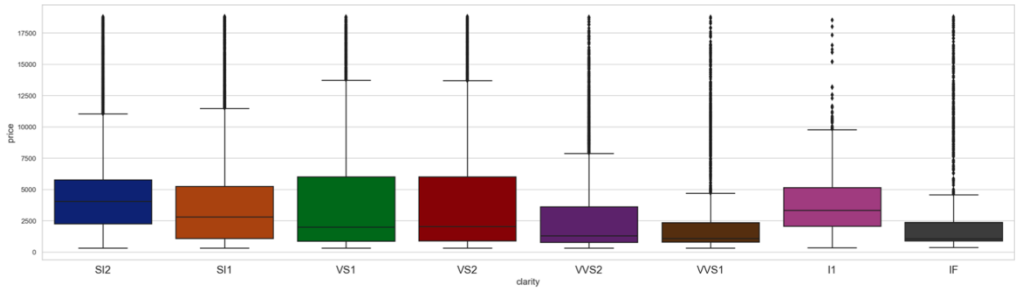


Figure 10 shows the relationship between the diamond price and its clarity. From the correlation plot below, it seems that VS1 and VS2 have the similar effect on the diamond's price, and VS1 and VS2 have quite high price margin than other features (Figure 10).

Figure 10: clarity & price



5.5 Depth

Figure 11 shows the diamond's depth plot of this dataset. Diamond depth is its tallness (in millimeters) estimated from the culet to the table. In the event that a diamond's profundity rate is too enormous or little the diamond will get dim in appearance since it will not, at this point return an appealing measure of light (Figure 11).

Figure 11: depth plot

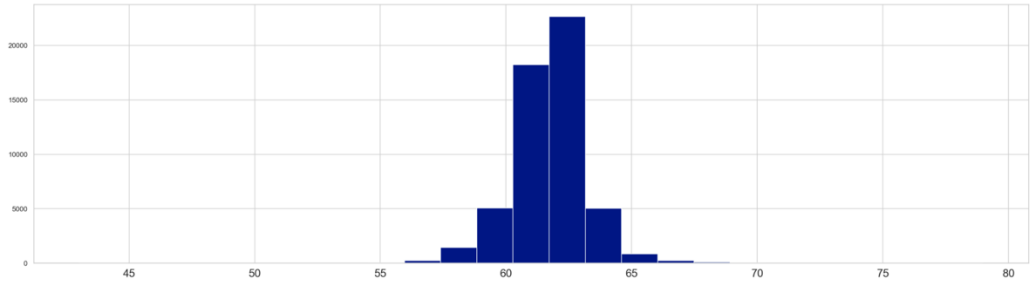
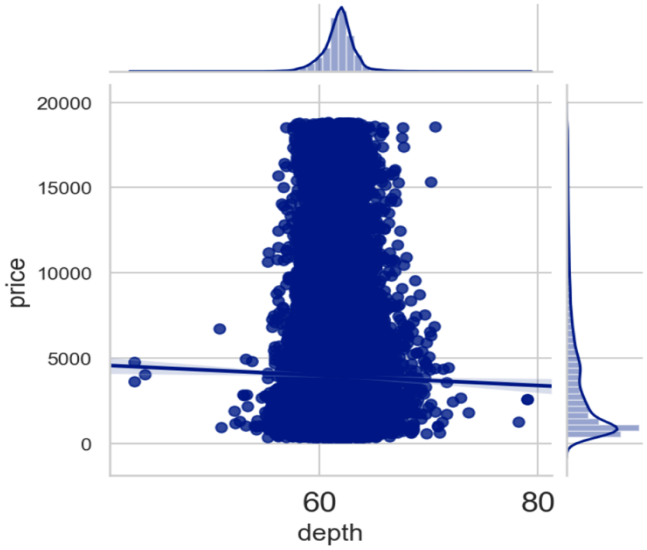


Figure 12 shows the relationship between the diamond price and its depth. According to Figure 12, it can be seen that the diamond price can vary much heavily at the same diamond depth. And the pearson's correlation shows that there's a slightly negative relationship between these two features.

Figure 12: depth & price



5.6 Table

Figure 13 shows the plot of diamond table. Table is the width of the diamond's table communicated as a level of its normal distance across. In the event that the table is too enormous at that point light won't play off of any of the crown's edges or features and won't make the shimmering rainbow hues. In the event that it is excessively little, at that point the light will get caught and that shaft of light will never come out however will "spill" from different places (Figure 13).

Figure 13: table plot

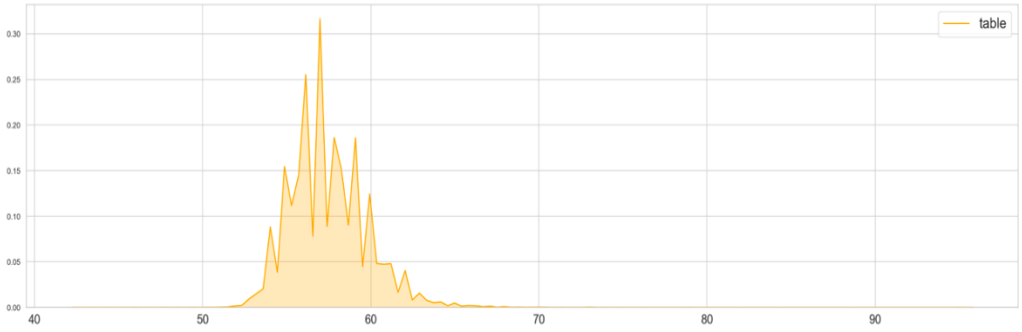
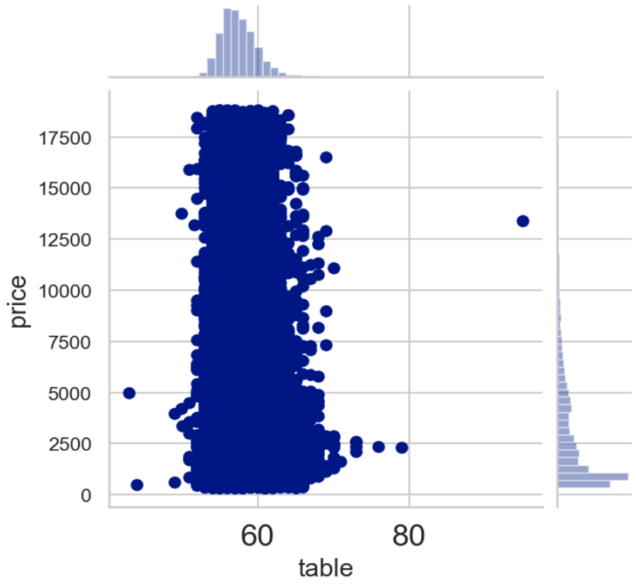


Figure 14 shows the relationship between the diamond price and its table. According to Figure 14, it can be inferred that the diamond price can vary much heavily at the same diamond table. And the pearson's correlation shows that there's a slightly negative relationship between these two features.

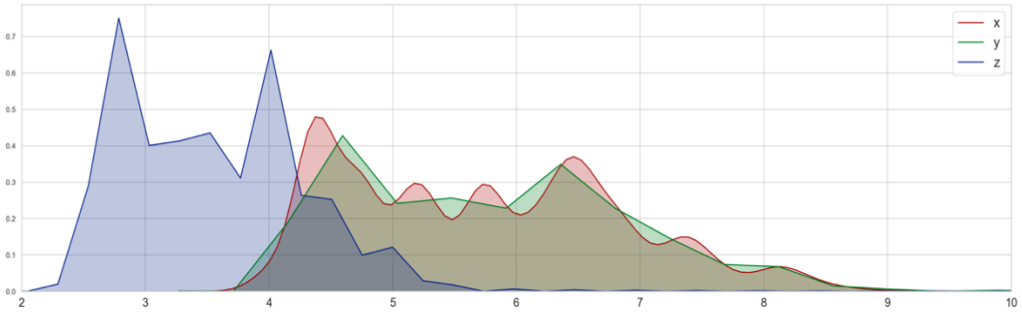
Figure 14: table & price



5.7 Dimensions

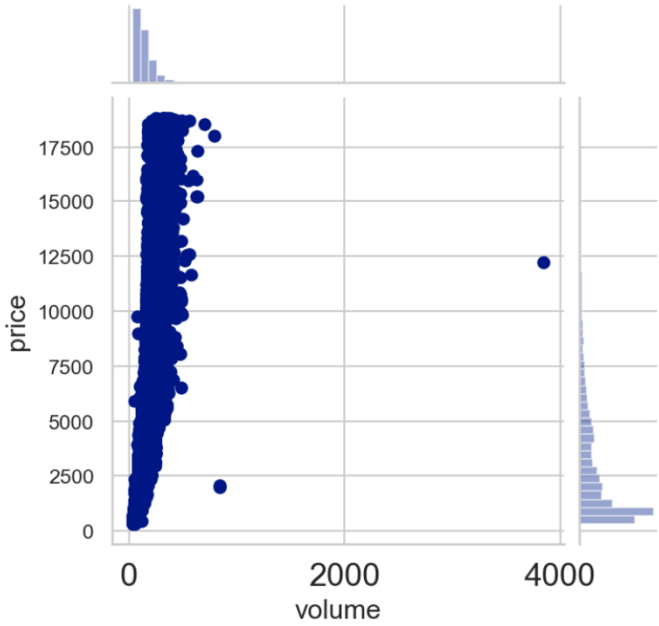
Figure 15 shows the kde plot of dimensions (x, y, z) of diamond of this dataset, visualizing the probability density of the carat.

Figure 15: dimension kde plot



In order to combine these three factors into one category, I create a new feature based on the dimensions called 'volume' and make visualization about how it will affect the diamond price. Figure 16 shows the relationship between the diamond price and its dimensions. As the dimensions increment, clearly the price of diamond ascends as an ever increasing number of common assets are used. As indicated by Figure 16, it appears that there is straight connection between diamond price and its volume (x * y * z).

Figure 16: dimension & price



6. Modelling Algorithms

I divide the dataset into train and test, so that I can fit the train for modelling algorithms and predict on test. Next I apply feature scaling to better make predictions more accurate. I apply seven modeling algorithms to the dataset to do the prediction, including Linear Regression, Lasso Regression, AdaBosst Regression, Ridge Regression, GradientBoosting Regression, Random Forest Regression and KNeighbours Regression. I try to collect the accuracy score of each modeling algorithms and make comparison.

6.1 Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It plays out a regression task. Regression models an objective forecast esteem dependent on independent factors.

Table 3 shows the accuracy score and some residual score of Linear Regression method. It can be seen from Table 3, the accuracy of Linear Regression is 88%. And MSE of this algorithm is 1911398.80; the MAE of this algorithm is 926.72; the RMSE of this algorithm is 1382.53.

Table 3

Accuracy	Linear Regression
Score	0.8814
MSE	1911398.80
MAE	926.72
RMSE	1382.53
R2	0.88

6.2 Lasso Regression

Lasso is a regression investigation technique that performs both variable choice and regularization so as to improve the expectation precision and interpretability of the factual model it produces. Shrinkage is the place information esteems are contracted towards an essential issue, similar to the mean.

Table 4 shows the accuracy score and some residual score of Lasso Regression method. As can be seen from Table 4, the accuracy of Lasso Regression is 87%. And MSE of this algorithm is 2162331.94; the MAE of this algorithm is 909.60; the RMSE of this algorithm is 1470.49.

Table 4

Accuracy	Lasso Regression
Score	0.8659
MSE	2162331.94
MAE	909.60
RMSE	1470.49
R2	0.87

6.3 AdaBoost Regression

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm defined by Yoav Freund and Robert Schapire. It can be utilized together with numerous other learning algorithms to improve performance.

Table 5 shows the accuracy score and some residual score of AdaBosst Regression method. As can be seen from Table 5, the accuracy of AdaBosst Regression is 85%. And MSE of this algorithm is 2340156.64; the MAE of this algorithm is 1290.77; the RMSE of this algorithm is 1529.76.

Table 5

Accuracy	AdaBoost Regression
Score	0.8548
MSE	2340156.64
MAE	1290.77
RMSE	1529.76
R2	0.85

6.4 Ridge Regression

Ridge Regression is a strategy for relieving the issue of multicollinearity in multiple regression, which usually happens in models with enormous quantities of parameters.

Table 6 shows the accuracy score and some residual score of Ridge Regression method. As can be seen from Table 6, the accuracy of Ridge Regression is 75%. And MSE of this algorithm is 3970442.17; the MAE of this algorithm is 1346.18; the RMSE of this algorithm is 1992.60.

Table 6

Accuracy	Ridge Regression
Score	0.7537
MSE	3970442.17
MAE	1346.18
RMSE	1992.60
R2	0.75

6.5 GradientBoosting Regression

Gradient boosting is a machine learning strategy for regression and classification issues, which delivers a prediction model as an outfit of powerless forecast models, normally decision trees.

Table 7 shows the accuracy score and some residual score of GradientBoosting Regression method. As can be seen from Table 7, the accuracy of GradientBoosting Regression is 91%. And MSE of this algorithm is 1518030.06; the MAE of this algorithm is 720.72; the RMSE of this algorithm is 1232.08.

Table 7

Accuracy	GradientBoosting Regression
Score	0.9058
MSE	1518030.06
MAE	720.72
RMSE	1232.08
R2	0.91

6.6 RandomForest Regression

Random forests or random decision forests are an outfit learning strategy for classification, regression and tasks that work by developing a large number of decision trees at training time and yielding the class that is the method of the classes or mean prediction of the individual trees.

Table 8 shows the accuracy score and some residual score of RandomForest Regression method. It can be seen that the accuracy of RandomForest Regression is 98%. And MSE of this algorithm is 319186.39; the MAE of this algorithm is 286.40; the RMSE of this algorithm is 564.97.

Table 8

Accuracy	RandomForest Regression
Score	0.9802
MSE	319186.39
MAE	286.40
RMSE	564.97
R2	0.98

6.7 KNeighbours Regression

The k-nearest neighbors algorithm (k-NN) is a non-parametric strategy utilized for classification and regression. k-NN is a kind of instance based learning, where the function is just approximated locally and all calculation is conceded until work evaluation.

Table 9 shows the accuracy score and some residual score of KNeighbours Regression method. As can be seen from Table 9, the accuracy of KNeighbours Regression is 96%. And MSE of this algorithm is 660416.40; the MAE of this algorithm is 424.98; the RMSE of this algorithm is 812.66.

Table 9

Accuracy	KNeighbours Regression
Score	0.9590
MSE	660416.40
MAE	424.98
RMSE	812.66
R2	0.96

7. Comparison and Visualization of Seven Algorithms' R2-Scores

7.1 Comparison

Random Forest Regression gives the highest R2-Score of [98.2%]; Follows by KNeighbours Regression with [95.9%]; GradientBoosting Regression gives R2-score of [90.6%]; Linear Regression gives R2-score of [88.1%]; Lasso Regression gives R2-score of [86.6%]; AdaBoost Regression gives R2-score of [85.5%]; Ridge Regression gives the lowest R2-score of [75.4%].

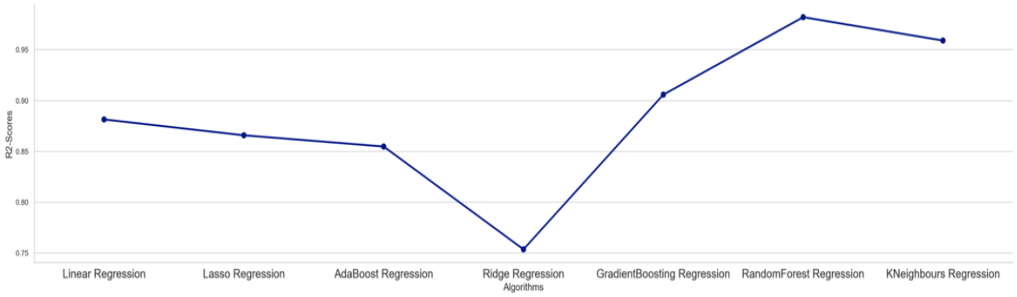
Table 10

	Algorithm	R2-Score
5	RandomForest Regression	0.982025
6	KNeighbours Regression	0.959033
4	GradientBoosting Regression	0.905833
0	Linear Regression	0.881432
1	Lasso Regression	0.865866
2	AdaBoost Regression	0.854835
3	Ridge Regression	0.753705

7.2 Visualization

From the plots of Figure 17, we can find that the accuracy scores of Linear Regression, Lasso Regression and AdaBoost Regression are very close to each other; the accuracy score of Ridge Regression has the lowest accuracy; the accuracy scores of GradientBoosting Regression, KNeighbours Regression and RandomForest Regression are higher than others with RandomForest Regression having the highest accuracy.

Figure 17: R2 score factor plot



8. Conclusion

Random Forest Regressor gets the highest accuracy score than other regression methods while the Ridge Regression gets the lowest accuracy score. The accuracy scores of Linear Regression, Lasso Regression and AdaBoost Regression are very close to each other; the accuracy scores of GradientBoosting Regression, KNeighbours Regression and RandomForest Regression are higher than others.

Actually, the price of diamond is determined by various factors other than its physical features. Diamond shows the love, commitment and adornment, which appeals people continuously strong throughout the past decade. A recession of any duration, as in the early 1980 would also affect demand for diamonds. Technology is another factor that undoubtedly will have an impact on the industry, but this will be minimized if the technology of gem identification can keep pace with the technology of treatments and synthesis (William, 1988).

The diamond industry has endured the extreme variances of the previous decade well. Each sign is that it will have the option to confront future difficulties much more productively and successfully (William, 1988). As ways of life globally keep on ascending, as De Beers keeps on elevating to old and new markets the same, as provisions keep on being solid, the future for the diamond business does without a doubt show up brilliant.

9. Reference:

Bain & Company. (2018). The Global Diamond Industry

Geeksforgeeks. kde. https://www.geeksforgeeks.org/kde-plot-visualization-with-pandas-and-seaborn/

Geeksforgeeks. Linear regression. https://www.geeksforgeeks.org/ml-linear-regression/

Geeksforgeeks. Lasso. https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/

Paperspace. https://blog.paperspace.com/implementing-gradient-boosting-regression-python/

Statisticshowto. Lasso. https://www.statisticshowto.com/lasso-regression/

Statisticshowto. Ridge. https://www.statisticshowto.com/ridge-regression/

Towards. https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f

William E. Boyajian. (Fall 1988). An economic review of the decade in diamonds. Gems & Gemology.

Wikipedia. (n.d.). Lasso. https://en.wikipedia.org/wiki/Lasso_(statistics)

Wikipedia. (n.d.). AdaBoost. https://en.wikipedia.org/wiki/AdaBoost

Wikipedia. (n.d.). Ridge. https://en.wikipedia.org/wiki/Tikhonov_regularization

Wikipedia. (n.d.). Gradient_boosting. https://en.wikipedia.org/wiki/Gradient_boosting

Wikipedia. (n.d.). Random_forest. https://en.wikipedia.org/wiki/Random_forest

Wikipedia. (n.d.). K-NN. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm