IM-Analytics Qualification Test

The Lapardist problem

objective:
Validate the stolen diamonds worth

Index

Executive Summary

The final prediction for the prices of the stolen diamonds is given in figure 20, this prediction has an accuracy of that 96% of the variance is accounted for and (using a simple percentage of the difference between predicted price and real price) that the real price of the diamond should be around the predicted price ∓ 10%.

This values were reached using a decision tree regression algorithm from the well-known data science library SkLearn. The data was pre-processed and analysed manually before feeding it to the model to give the best results possible.

Introduction

The objective of this case is to validate the stolen diamonds worth (price), of which we have certain characteristics.

	carat	cut	color	clarity	depth	table	x	y	z
1	0.71	Good	I	VVS2	63.1	58	5.64	5.71	3.58
2	0.83	Ideal	G	VS1	62.1	55	6.02	6.05	3.75
3	0.5	Ideal	E	VS2	61.5	55	5.11	5.16	3.16
4	0.39	Premium	J	VS1	61.6	59	4.67	4.71	2.89
5	0.32	Premium	G	VS1	62.1	56	4.43	4.4	2.74
6	0.9	Good	F	SI2	63.3	57	6.08	6.14	3.87
7	0.51	Ideal	D	VS1	60.9	57	5.2	5.17	3.16
8	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13
9	0.4	Ideal	G	VVS2	62.4	56	4.72	4.74	2.95
10	0.36	Premium	I	VS2	62.7	59	4.54	4.58	2.86

fig 1. stolen diamonds characteristics

To do this a dataset has been given with the characteristics and price of over 50’000 diamonds, it is expected that a relationship between the characteristics of the diamond and its price can be found and with this a prediction of the prices can be made.

Profiling of data completeness and quality

First look at the data

To profile the data firstly it was opened in a spreadsheet software to review its format and its consistency and to check if there are any clear patterns of incompleteness.

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67

fig 2. basic format of the data.

In this simple view of the data, no incongruencies or incompleteness can be seen.

Data completeness test

A simple analysis of the data integrity can be made to confirm the data completeness, to do this, the data was described with an statistical package, giving the following results:

	carat	depth	table	price	x	y	z
count	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000
mean	0.797976	61.749325	57.457328	3933.054942	5.731236	5.734601	3.538776
std	0.474035	1.432711	2.234578	3989.628569	1.121807	1.142184	0.705729
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	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5325.000000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

fig 4. description of the data.

Most of this values make sense, but there are a few incongruencies and outliers:

- There is a diamond with a 5.01 carat, when the 75 percentile is of 1.04 and with a standard deviation of .47, it is clear that this value is too high and should be removed, because of this an outlier analysis of the carat should be made.
- There are values of zero in the x,y,z column, which is physically impossible since this represent the physical dimensions of the (length, width, depth) of the diamond, therefore this data points should be removed.

It should also be noted that since the compiler gave no errors or warnings, the dataset has no missing or “NaN” values. (this is also confirmed with the isnull() function):

```
In [17]: diamonds_data.isnull().sum()
Out[17]:
carat      0
cut        0
color      0
clarity    0
depth      0
table      0
price      0
x          0
y          0
z          0
dtype: int64
```

fig 5. isnull() function results

A box plot of the carat of the diamonds was built to determine if there are any outliers that have to be removed:

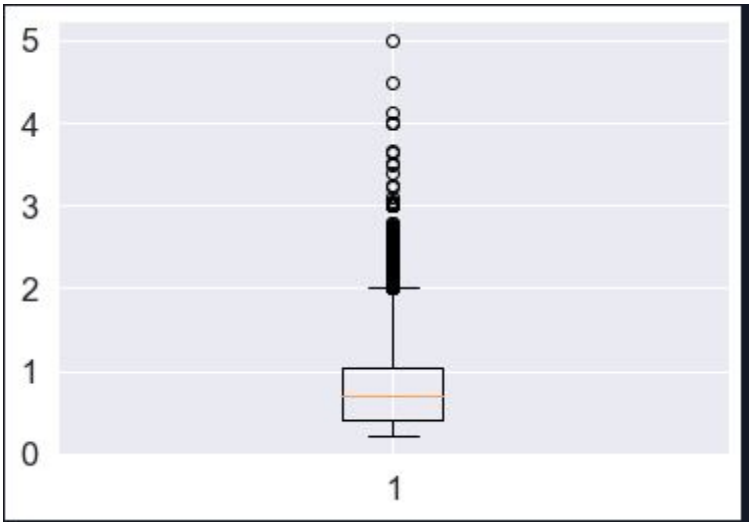

fig 6. Box plot of carat values

There are plenty of outliers in the data, but the values start dwindling down significantly after carat = 3, which means that this diamonds are really rare, and since every diamond in the stolen table has a carat lower than 1.2, it is safe to remove this data.

Descriptive analysis and summary statistics

Some analysis and summary statistics were already made in the previous section (refer to fig 4), to continue with this analysis an histogram of each variable will be plotted:

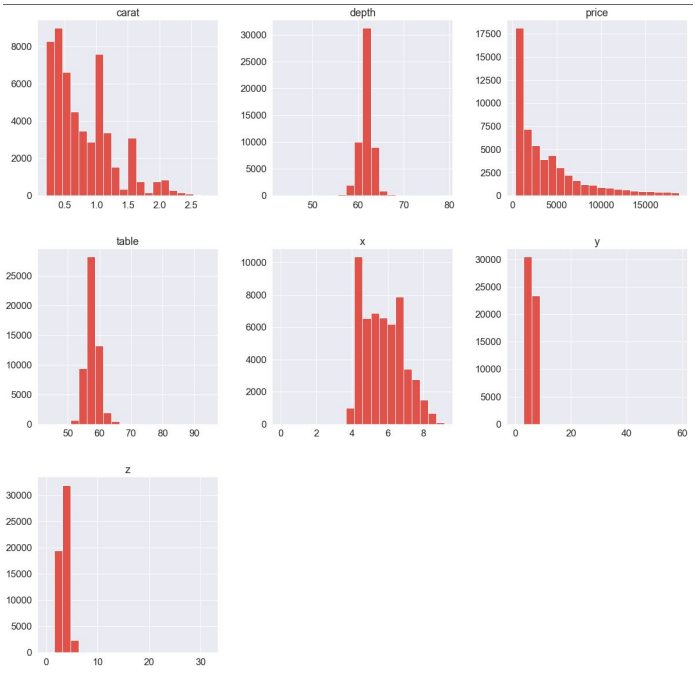

fig 7. Histogram of each category

Now the correlation of between variables will be analyzed with a heatmap

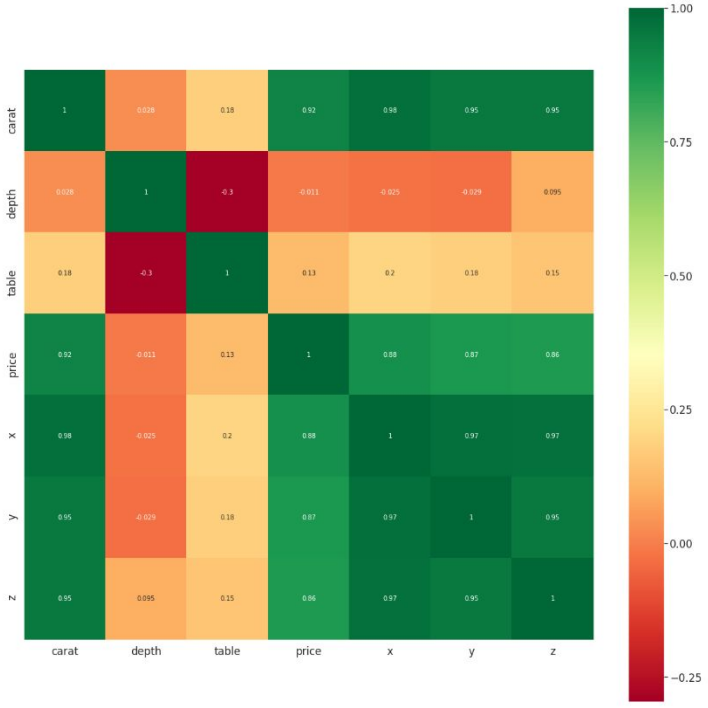

fig 8. Correlation heatmap

It can clearly be seen that there is a strong correlation between price and the dimensions of the diamond, as well as the carat.

Finally a pairplot was built to visualize the relationships between the data and confirm the heatmap data.

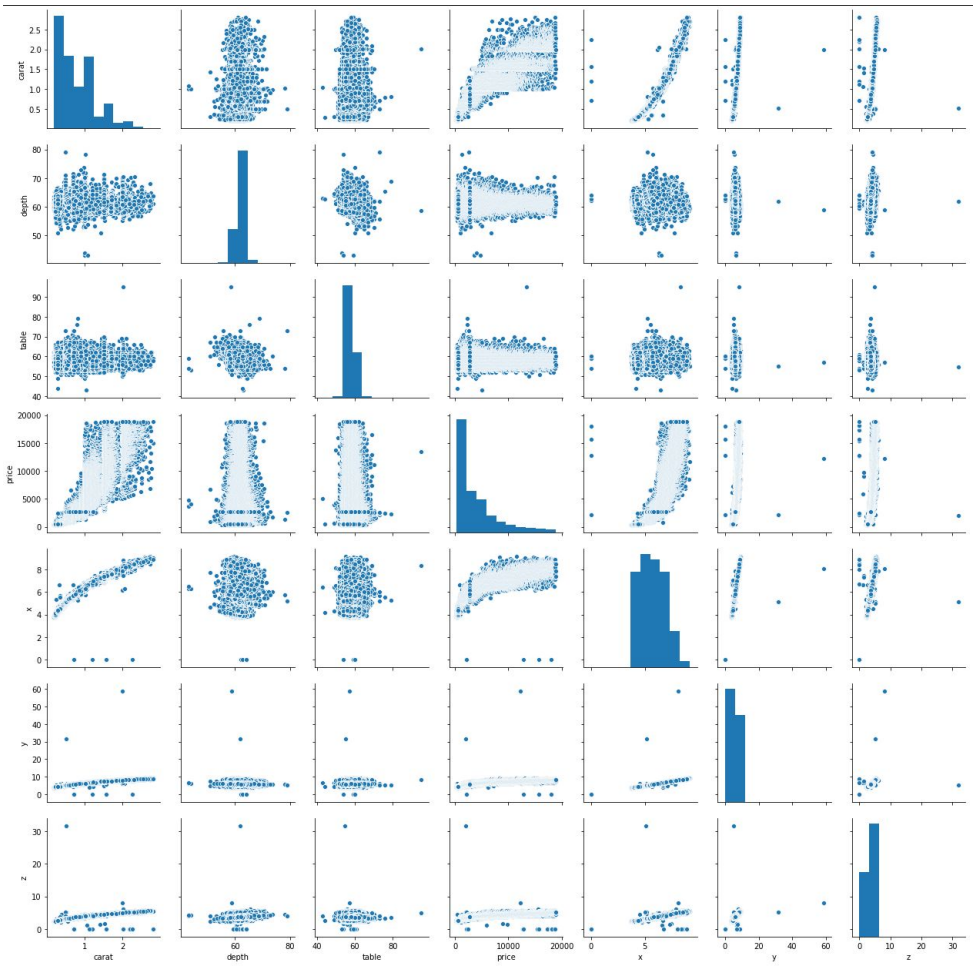

fig 9. characteristics pairplot

Hypotheses and Modeling

Hypothesis

The hypotheses in this case is that a linear regression model can be used to value the missing diamonds out of their characteristics, Ridge and Lasso regression will also be implemented as to compare results and select the best model, a prediction of the values of the stolen diamonds could also be made with each model to have a better range of their real value.

Modeling

Pre-treatment for categorical data

To treat the categorical data so as to be able to properly feed it to the model, one-hot-encoding was used, this consists on transforming categorical data into a series of ones and zeros, an example of this is shown in the next image.

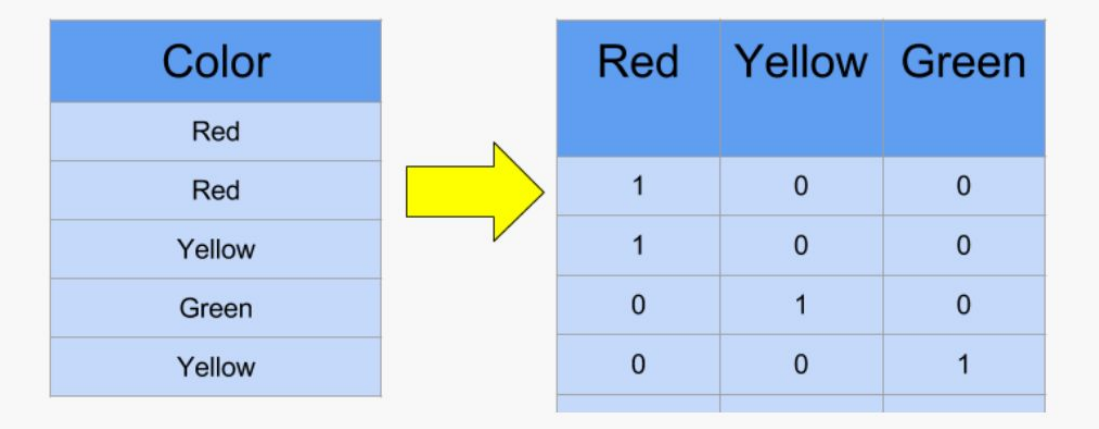

fig 10. one-hot-encoding example
(taken from https://blog.datascienceheroes.com/content/images/2019/07/one-hot-encoding.png)

In the case of our data, that would mean making a new column for each categorical value (cut_Fair, cut_Good, cut_Ideal, cut_Premium would be some of the new columns for example).

cut_Good	cut_Ideal	cut_Premium	cut_Very Good	color_D	color_E	color_F	color_G	color_H	color_I	color_J	clarity_I1
0	1	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	1	0

fig 11. preview of the new columns in the dataset.

Pre-treatment for numerical data

For the numerical data an standard scaler will be implemented, which simply consists of applying the following formula to every value in a column:

$$z = (x - u)/s$$

where :

$u = means\ of\ the\ samples.$

$s = standard\ deviation$

The scaled values of some columns are shown next:

	carat	depth	table	price	x	y	z
0	-1.206507	-0.173941	-1.099752	326.0	-1.592603	-1.539802	-1.578708
1	-1.249134	-1.362100	1.586675	326.0	-1.646326	-1.662844	-1.749744
2	-1.206507	-3.388959	3.377626	327.0	-1.503066	-1.460703	-1.749744
3	-1.078627	0.455084	0.243462	334.0	-1.368760	-1.320084	-1.293648
4	-1.036000	1.084109	0.243462	335.0	-1.243407	-1.214619	-1.122611

fig 12. Scaled data.

Model training and results

To train the models a training and testing dataset was made out of the diamond database, in all cases 30% of the data was taken randomly to test the accuracy of the model.

To measure the accuracy of the model an R^2 analysis was made, this analysis measures the proportion of the total deviation of Y from the mean value given by the model, and its calculated with the following formula:

$$R^2 = RSS / TSS$$

$$Where:$$

$$RSS = Sum(Y'i - Y)^2$$

$$Where: \quad Y' = estimated\ value\ of\ Y$$

$$Y = mean\ of\ dependent\ variables$$

$$i = number\ of\ variations$$

$$TSS = Sum(Yi - Y)^2$$

$$Where: \quad Yi = dependent\ variables$$

$$Y = mean\ of\ dependent\ variables$$

$$i = number\ of\ variations$$

The mean absolute error and mean squared error of the whole dataset will also be displayed, but its importance is minor compared to the R^2.

Linear regression

As for the linear regression the model was trained using the Sklearn statistical package *linear_model.LinearRegression(),* the results of this analysis were the following:

```
LINEAR REGRESSION
accuracy: 92.68223617444734%
Mean absolute error: 714.8090247259531
Mean squared error: 1144195.5712142165
R Squared: 0.9268223617444734
Adjusted R Squared: 0.9267089924096288
```

fig 13.Linear regression accuracy.

And the predicted values of the stolen diamonds is shown in the following table

	Predicted_price	carat	cut	color	clarity	depth	table	x	y	z
2	2789.4402758176	0.71	Good	I	VVS2	63.1	58	5.64	5.71	3.58
3	4780.5004358857	0.83	Ideal	G	VS1	62.1	55	6.02	6.05	3.75
4	2047.335325587	0.5	Ideal	E	VS2	61.5	55	5.11	5.16	3.16
5	-743.9066429876	0.39	Premium	J	VS1	61.6	59	4.67	4.71	2.89
6	687.1413433208	0.32	Premium	G	VS1	62.1	56	4.43	4.4	2.74
7	3445.7168236038	0.9	Good	F	SI2	63.3	57	6.08	6.14	3.87
8	2561.6152793701	0.51	Ideal	D	VS1	60.9	57	5.2	5.17	3.16
9	7800.7526183562	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13
10	1645.4631201091	0.4	Ideal	G	VVS2	62.4	56	4.72	4.74	2.95
11	-425.5860997752	0.36	Premium	I	VS2	62.7	59	4.54	4.58	2.86

fig 14 .Linear regression predictions.

Lasso Regression

For the lasso regression the model was trained using the Sklearn statistical package *linear_model.Lasso()*, the results of this analysis were the following:

```
LASSO REGRESSION
accuracy: 92.67553192052233%
Mean absolute error: 712.8377826204952
Mean squared error: 1145243.8392141368
R Squared: 0.9267553192052232
Adjusted R Squared: 0.926641846005752
```

fig 15 .Lasso regression accuracy.

And the predicted values of the stolen diamonds is shown in the following table

	Predicted_price	carat	cut	color	clarity	depth	table	x	y	z
2	2808.0989812496	0.71	Good	I	VVS2	63.1	58	5.64	5.71	3.58
3	4780.3717608205	0.83	Ideal	G	VS1	62.1	55	6.02	6.05	3.75
4	2050.9904040406	0.5	Ideal	E	VS2	61.5	55	5.11	5.16	3.16
5	-731.8426199166	0.39	Premium	J	VS1	61.6	59	4.67	4.71	2.89
6	676.2804261292	0.32	Premium	G	VS1	62.1	56	4.43	4.4	2.74
7	3460.3225943415	0.9	Good	F	SI2	63.3	57	6.08	6.14	3.87
8	2549.3849006743	0.51	Ideal	D	VS1	60.9	57	5.2	5.17	3.16
9	7791.7880825938	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13
10	1638.0328758007	0.4	Ideal	G	VVS2	62.4	56	4.72	4.74	2.95
11	-413.0479950792	0.36	Premium	I	VS2	62.7	59	4.54	4.58	2.86

fig 16 .Lasso regression predictions.

Ridge Regression

For the ridge regression the model was trained using the Sklearn statistical package *linear_model.Ridge()*, the results of this analysis were the following:

```
RIDGE REGRESSION
accuracy: 92.68168167679349%
Mean absolute error: 714.8071489303456
Mean squared error: 1144282.2717111043
R Squared: 0.9268168167679349
Adjusted R Squared: 0.9267034388426203
```

fig 17 .Ridge regression accuracy.

	Predicted_price	carat	cut	color	clarity	depth	table	x	y	z
2	2790.2230718621	0.71	Good	I	VVS2	63.1	58	5.64	5.71	3.58
3	4781.0940003558	0.83	Ideal	G	VS1	62.1	55	6.02	6.05	3.75
4	2047.5683925966	0.5	Ideal	E	VS2	61.5	55	5.11	5.16	3.16
5	-743.0779569272	0.39	Premium	J	VS1	61.6	59	4.67	4.71	2.89
6	686.5644841009	0.32	Premium	G	VS1	62.1	56	4.43	4.4	2.74
7	3446.2586065349	0.9	Good	F	SI2	63.3	57	6.08	6.14	3.87
8	2561.77919581	0.51	Ideal	D	VS1	60.9	57	5.2	5.17	3.16
9	7800.8564230345	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13
10	1645.2933906974	0.4	Ideal	G	VVS2	62.4	56	4.72	4.74	2.95
11	-425.446158127	0.36	Premium	I	VS2	62.7	59	4.54	4.58	2.86

fig 18 .Ridge regression predictions.

Decision Tree Regression

For the ridge regression the model was trained using the Sklearn statistical package *DecisionTreeRegressor,* the results of this analysis were the following:

```
DecisionTreeRegressor
accuracy: 96.42074723353376%
Mean absolute error: 360.9868836230898
Mean squared error: 559647.0808946359
R Squared: 0.9642074723353375
Adjusted R Squared: 0.9641520213102637
```

fig 19 .Decision Tree regression accuracy.

	Predicted_price	carat	cut	color	clarity	depth	table	x	y	z
2	2456	0.71	Good	I	VVS2	63.1	58	5.64	5.71	3.58
3	4181	0.83	Ideal	G	VS1	62.1	55	6.02	6.05	3.75
4	1621	0.5	Ideal	E	VS2	61.5	55	5.11	5.16	3.16
5	790	0.39	Premium	J	VS1	61.6	59	4.67	4.71	2.89
6	828	0.32	Premium	G	VS1	62.1	56	4.43	4.4	2.74
7	3401	0.9	Good	F	SI2	63.3	57	6.08	6.14	3.87
8	1835	0.51	Ideal	D	VS1	60.9	57	5.2	5.17	3.16
9	8295	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13
10	1050	0.4	Ideal	G	VVS2	62.4	56	4.72	4.74	2.95
11	648	0.36	Premium	I	VS2	62.7	59	4.54	4.58	2.86

fig 20 .Decision Tree regression predictions.

Model discussion and advantages/disadvantages

From the four models used, three models are part of the linear regression analysis branch of statistics (Linear, Lasso, Ridge), which is one of the main reasons they give similar accuracy values. on the other hand the decision tree regressor has a better accuracy and it avoids some of the disadvantages of linear regression problems.

advantages of the linear models

- Simple to build, this model can be built with pre-made tools taken from already well-developed statistical packages which makes it useful when time is of essence (I have to give results to the minister by tomorrow)
- Train speed, this model train speed is really fast compared to more complex statistical or machine learning models, this is because of the relative simplicity of the operations that have to be made in order to train the model.
- There is strong quantitative justification for the results of this model, since it's based on relatively simple formulas and has plenty of statistical background and proof

disadvantages of the linear models

- The biggest disadvantage of this model is that linear regression does not understand that some values cannot have certain values (The prediction gives negative values in some cases, which is impossible)
- Some relationships between the data could have been overlooked because of how linear the model is.

- The model could be hard to explain to someone without an statistics/math/data science background

advantages of the decision tree model

- Simple to build, this model can be built with pre-made tools taken from already well-developed statistical packages which makes it useful when time is of essence (I have to give results to the minister by tomorrow)
- Train speed, this model train speed is really fast compared to more complex statistical or machine learning models, this is because of the relative simplicity of the operations that have to be made in order to train the model.
- There is strong quantitative justification for the results of this model, since it's based on relatively simple formulas and has plenty of statistical background and proof.
- this model can understand some underlying relationships in the data (it does not go below zero for example)

disadvantages of the decision tree model

- Some relationships between the data could have been overlooked because of the simplicity of the model.
- The model could be hard to explain to someone without an statistics/math/data science background.

Conclusion and next steps

Conclusion

To conclude I would give the ministers the values obtained using the decision tree model and explain that it's accuracy means that 96% of the variance is accounted for and (using a simple percentage of the difference between predicted price and real price)that the real price of the diamond should be around the predicted price ∓ 10%.

Next steps

Regarding the next steps to analyze this data there are a couple recommendations to be made:

- The descriptive analysis can be extended to each variable
- Clean more outliers from the data.
- More robust statistical models could be applied and an exploration .
- Relatively new and experimental algorithms like Reinforcement learning models could be used to predict the price with higher accuracy (eg. deep-q-learning).
- There could be other factors outside the database that explain the valuation of the diamonds.

SOURCES

[1] A Data Analyst. (2019). *What Make A Really Good Diamond? - A Data Analyst*. [online] Available at: https://adataanalyst.com/data-analysis-resources/what-make-a-really-good-diamond/

[2] Shrutimechlearn. (2019, March 9). Types of Regression and Stats in depth. Retrieved from https://www.kaggle.com/shrutimechlearn/types-of-regression-and-stats-in-depth/notebook.

[3] Applying Predictive Analysis on Diamond Prices. (n.d.). Retrieved from https://thepythonguru.com/applying-predictive-analysis-on-diamond-prices/.

[4] Selvaraj, H. (2018, September 29). Regression in Supervised Machine Learning. Retrieved from https://medium.com/harinathselvaraj/regression-in-supervised-machine-learning-aab502e47957.