The Lapidarist Problem

Matias Leoni[*]

(Dated: May 4, 2023)

We consider the Lapidarist Problem: a data scientist who, without much real knowledge of the value of precious stones, can compute the price of a given set of diamonds by only knowing their main physical characteristics. To do so we use a particular variation of the Diamond dataset.

I. INTRODUCTION: THE DATASET

The dataset given is a combination of two datasets with 53930 entries plus a set of 10 entries

- The first dataset is a variation of the traditional "Diamond" dataset which contains columns[1]

  `{carat, cut, color, clarity, depth, table, price, x, y, z}`

  Cut, color and clarity are cathegorical variables with possible values:

  ```
  cut: {Fair, Good, Very Good, Premium, Ideal}
  color: {D, E,...,J}
  clarity: {I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF}
  ```

  Carat is the weight, $\{x, y, z\}$ are the lengths in mm, depth and table are particular length ratios. Then there is Price, which is the only column that speaks of no physical characteristic but of its economical value. We shall call it the Train-test dataset.

- The second dataset on the other hand contains only two colums: Latitude and Longitude expressed in degrees. This is the most radical departure from the traditional diamond dataset in the sense that a location for each diamond does not belong to the original formulation of the problem. Since one cannot find any documentation regarding these locations we will assume they are the places where the diamonds were originally extracted or found. We shall call it the g-Train-test dataset.

- Last there is a table of 10 entries which combines the columns of the previous two datasets except for the price which we aim to predict. We shall call it the target dataset.

The first two datasets which we will use to train and test models to make predictions are pretty dirty and have to be thoroughly cleaned.

[*]Electronic address: matiasleoni@gmail.com

[1] see [1] for more on "the 4 C's of a diamond"

We start by noticing one typo in one of the 53930 entries of the g-Train-test dataset which prevents us from having a clean import of float numbers because a letter was wrongly inserted. It is curious that the corrected location points to a historical marker in Google Maps corresponding to the "Mescalero Apache Reservation" in the state of New Mexico.

The Train-test dataset on the other hand is much dirtier, it having four types of problems:

- The values of the three cathegorical variables are corrupted. We identified the different type of errors and we maped the corrupt values to correct ones in our code.
- Some of the $\{x, y, z\}$ lengths are negative, a fact easily corrected by taking the module.
- The depth column has 0s and NaNs. Since this column can be written as a function of $\{x, y, z\}$ we can reconstruct its correct value[2].
- There are $\{x, y, z\}$ values which are set to 0 or nonexistent (NaNs). In order not to throw these records we make the hypothesis that, on average, the ratio of the x radius to z and the y radius to z is normally distributed. In this way we can reconstruct the missing xs and ys without losing registries and only adding gaussian noise. We make the same assumption about the ratio between z and the *carat* in order to reconstruct null zs. These hypothesis seem plausible because one can imagine that given the different shapes chosen by lapidarists to cut the raw diamonds, they tend to preserve similar relative sizes in such a way that one will not find a cut gem which is two orders of magitude larger when measured in one dimension relative to other dimension. The same reasoning applies to the ratio z/carat: since diamonds have approximately the same density, their weight is completely determined by their size and the same ratio guess is considered.

II. THE LOCATION PROBLEM: ROUTE 66

With our datasets clean we start our analysis. We first consider the problem of location. It could well be that the price is conditioned by the location where the diamond was found. While one would think price has only have to do with physical characteristics of the diamond, pricing has a subjective irrational component which is difficult to grasp if one is not interested in that particular item.

Since spatial dependence of a variable is a particular type of dependence which has to be cared with different tools we begin by studying it separately. For this we will perform a Moran I test on our data[2] by taking price as the variable of interest and studying its dependence with location. This is, we perform a statistical test which has a null hypothesis given by spatial randomness of the price. The results we obtain for the z-score and the p-value of the statistical test are

$$z - score = 9.7 \times 10^{-4} \tag{1}$$

$$p - value = 0.9992 \tag{2}$$

this is, a small z-score and a p-value very close to 1. This gives evidence that there is no spatial-autocorrelation.

[2] It turns out that $depth = 100\frac{z}{mean(x,y)}$.

The question is therefore why the Prime Minister supplied the usual diamond dataset with the location information which seems irrelevant in order to understand the pricing of the ten lost diamonds. To dig into this question we proceed to map the 10 lost diamonds besides the location given in the g-Train-test dataset (see figure (1)).

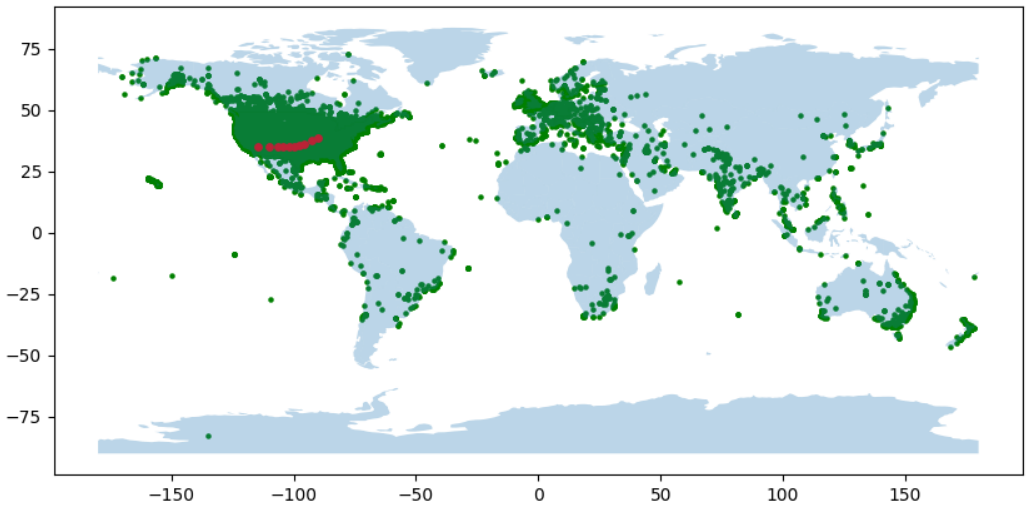

Figure 1: Green dots: g-Train-test dataset locations. Red dots: the 10 lost diamond locations.

We see from figure (1) that the lost diamonds form a bent line crossing the Midwest region of the United States. In fact, a closer look on the locations reveal a very curious fact: they are all 10 located along the old Route-66, "The mother road". While irrelevant as an information to price the diamonds, I would certainly inform the Prime Minister about this fact since it seems a relevant clue about the robbery of the diamonds.

III. APPROACH 1: BRUTE FORCE

Having discarded location as a relevant factor to price the diamonds we concentrate our efforts in the Train-test dataset which has 9 characteristics of thousands of diamonds and will help us to construct models to price the 10 lost diamonds.

We map the cathegorical variables to integer sets and we construct the covariance heatmap to get a feeling of the possible relation among variables. The obvious observation from the figure (2) is that price is strongly correlated to carat and $\{x, y, z\}$. But notice also that, as expected, carat is strongly correlated to $\{x, y, z\}$ because diamonds have an almost fixed density[3]. We keep these observations in mind for our second approach to the pricing problem.

Our first approach is to use the full power of the scikit-learn [3]. We code four pipelines classes in order to train and cross-validate four different class models: Random Forest, K-nearest neighbors, Decision trees and Linear Regressors. We train the four classes to obtain four different models and we use the criteria of lowest root mean squared error to choose among the four.

[3] Diamonds having an almost fixed density implies that their weight is proportional to the product of their dimensions, this is carat $\sim x\,y\,z$.

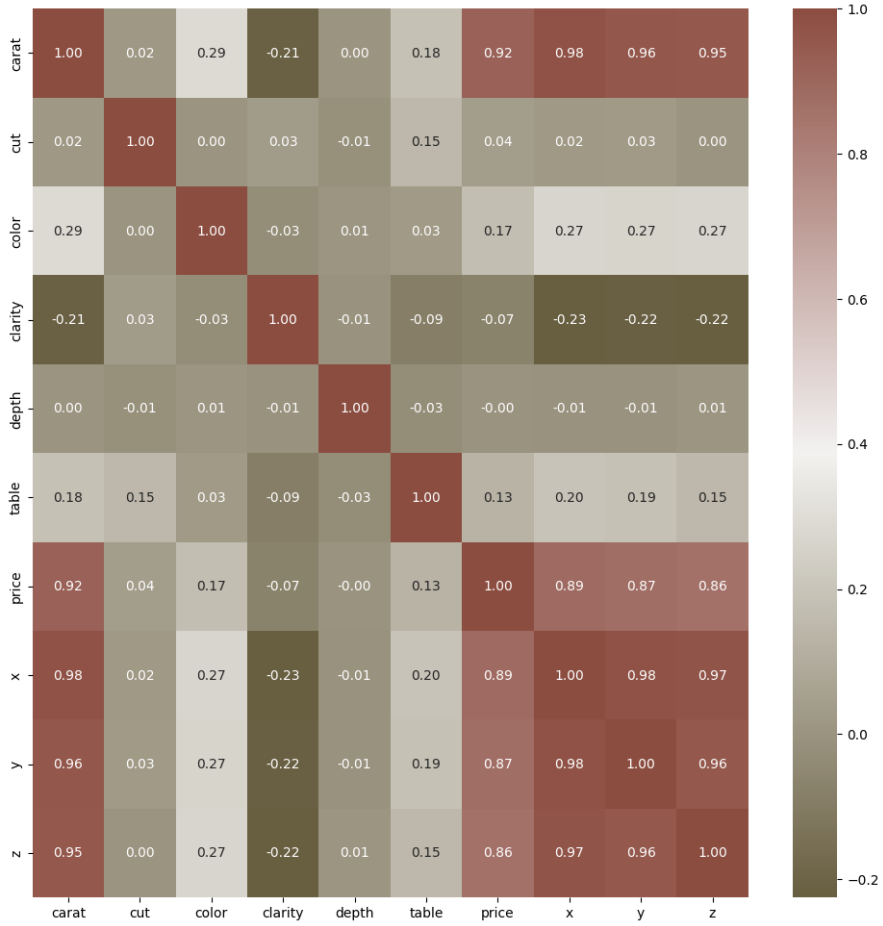

Figure 2: Covariance heatmap of the train-test dataset.

The best model we obtained we got it by training the Random Forest class. After computing the coefficient of determination on the test sample we obtain $R^2 = 0.97$. We can interpret therefore that 97% of the variability of our data can be explained by the Random Forest model we trained. We therefore use this trained model to predict the price of the 10 lost diamonds and we obtain:

Diamond	**Price**
1	$2,370.78
2	$3,872.71
3	$1,661.12
4	$661.47
5	$805.73
6	$3,711.87
7	$1,973.51
8	$9,306.73
9	$978.35
10	$549.70

IV. APPROACH 2: QUICK AND DIRTY

While our fancy and sophisticated 9 variable Random Forest trained model to predict the price of the diamonds could well be our final answer to the problem, it has an evident disadvantage: it is difficult to interpret and to use that interpretation in order to explain the model to the layman.

In this sense, it would be nice to have a simple model which, while less precise than our previous Random Forest, it is easier to interpret and to use to explain to Prime Ministers and Magicians. For this we continue with our previously explained hypothesis and obvervations. Firstly, that price is strongly correlated to the lengths $\{x, y, z\}$ and Carat while all the other characteristics have mild if not negligible correlation with price. Secondly, that Carat information (weight) is already contained in the lengths $\{x, y, z\}$.

We consider moreover that the main contribution to the price of the diamond comes from the volume of the diamond and volume can be estimated as the volume of a cube of sides x, y and z such as $V \sim xyz$. We would also like to fit variables without dimensions. For this we define the characteristic length of the diamond as its main diagonal:

$$l_0 = \sqrt{x^2 + y^2 + z^2} \tag{3}$$

and we define normalized x, y and z as

$$\hat{x} = \frac{x}{l_0}, \qquad \hat{y} = \frac{y}{l_0}, \qquad \hat{z} = \frac{z}{l_0} \tag{4}$$

Thus we propose a model where the price of a diamond is proportional to some power of its approximate volume:

$$P = \lambda \, (\hat{x}\, \hat{y}\, \hat{z})^{\alpha} \tag{5}$$

where λ is a constant of proportionality and α is an undetermined power. Since α is not knwon *a priori*, we determine it by making a linear fit of the log of the price vs the log of the product $\hat{x}\, \hat{y}\, \hat{z}$ and this fit is our trained model. We can see the scatter plot of the log variables and the linear fit in figure (3):

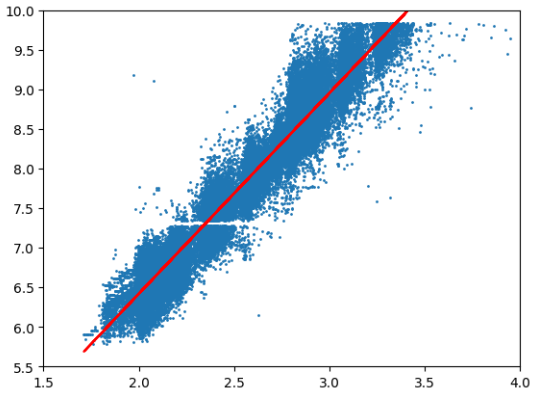

Figure 3: Scatter plot and linear fit of $\log(\hat{x}\, \hat{y}\, \hat{z})$ vs $\log$ price.

The coefficient of determination of this fit is $R^2 = 0.93$ which, taking in consideration the simplicity of our model compared to the random forest of the last section, is impressive. The prices prediction with this simple model is:

Diamond	Price	% change w/ RF
1	\$2,631.56	9.7%
2	\$3,481.89	-10.5%
3	\$1,503.59	-10.4%
4	\$952.18	41.8%
5	\$711.59	-11.2%
6	\$3,860.21	2.7%
7	\$1,547.07	-21.7%
8	\$5,689.56	-38.0%
9	\$1,019.96	1.8%
10	\$857.27	52.4%

where in the last column we state the percentual difference with the prediction given by the Random Forest model of the last section. We see differences that are $<\sim 10\%$ except for the fourth, seventh, eighth and tenth diamonds. Regarding the relative price hierarchy established by both methods we see they are similar up to a couple of permutations, this is: $\{8, 2, 6, 1, 7, 3, 9, 5, 4, 10\}$ vs $\{8, 6, 2, 1, 7, 3, 9, 4, 10, 5\}$.

V. CONCLUSIONS

We were challenged with a variation of the Lapidarist problem which, besides having a corrupt dataset it included location of the diamonds. After cleaning the dataset and discarding the relevance of location with some statistical tests, we faced the problem of pricing 10 missing diamonds given their characteristics.

We used four of the most used machine learning algorithms and we chose the best of those four which turned out to be of the Random Forest class. This model gave very good scores in its testing and we used it to price the gems.

While we could have ended our job at that point, we chose to construct one more trained model. Our motive was to find a much simpler, less precise model than the Random Forest we had found but which could allow us to better interpret the results.

---

[1] American Gem Society "4Cs of Diamonds," https://www.americangemsociety.org/4cs-of-diamonds/

[2] Moran, P. A. P. Notes on Continuous Stochastic Phenomena. Biometrika 37, no. 1/2 (1950): 1723.

[3] Pedregosa F. *et.al.*, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12, no. 85 (2011): 2825–2830.