Valuation of Stolen Diamonds from Gringotts Wizarding Bank Report: Predictive Analysis of the Lapidarist Problem

Alan García Zermeño

8/8/2024

Abstract

This report provides a detailed analysis of the *lapidarist problem*, focusing on valuing stolen diamonds from Gringotts Wizarding Bank. The study involved exploring and cleaning a dataset of 53,930 diamonds. Advanced predictive models, including Epsilon-Support Vector Regression (SVR), were developed, with the SVR model achieving the best performance (RMSE of 594.79). The estimated total value of the stolen diamonds was **$ 43,630.37 ± $ 1,880.88**, nearly double the average value of a random selection of diamonds. The report highlights the complexity of the theft and suggests that those involved had expert knowledge of diamond valuation.

1 Data Exploration and Cleaning

The dataset provided for analysis contains information on 53,930 diamonds. An initial examination revealed several key characteristics used to describe and value diamonds (A more in-depth analysis of these features can be accessed in the notebook attached to this report.):

1. **Carat**: A measure of the diamond's weight, with 1 carat equaling 0.2 grams.
2. **Cut**: A categorical measure of the diamond's cut quality, which affects its brilliance and light reflection.
3. **Color**: Classified using letters of the alphabet, with 'D' representing the clearest and most expensive diamonds, progressing through the alphabet as the diamond becomes more yellowish.
4. **Clarity**: Indicates the level of impurities in the diamond, directly impacting its value.
5. **Depth**: A proportion of the diamond's height relative to its size, with optimal values between 59% and 62%.
6. **Table**: The relative proportion of the diamond's top facet to its width, ideally between 55% and 60%.
7. **Geological coordinates**: Latitude and longitude on the planet for each diamond. This appears to indicate the origin or place of manufacture of the diamond.

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...

Figure 1: Diamonds database sample.

1.1 Data Cleaning Process

1. **Character Removal**: Random characters were found dispersed throughout the categories. A regular expression was applied to clean the data, retaining only letters and numbers relevant to the analysis.

2. **Handling Missing Values**: NAN values were identified in the depth, x, and y columns. To preserve the entire dataset for analysis, these empty spaces were filled with the median value of each respective characteristic, minimizing the impact on our estimates.

This cleaning process ensures a more accurate and comprehensive analysis of the diamond dataset, providing a solid foundation for further investigation and modeling.

2 General Statistics and Data Analysis

After cleaning the dataset, we conducted a comprehensive statistical analysis to understand the key features and their relationships with diamond prices.

Figure 2: Price distribution.

Price Distribution

The majority of diamonds in the dataset are priced between $300 and $5,000, with some exceptional pieces reaching up to $18,000. This wide range indicates the complexity of diamond valuation and the need for a multifaceted analysis.

Correlation Analysis

A correlation study of numerical variables revealed:

- Strong positive correlations between price and the dimensions (x, y, z) and carat weight of the diamonds.
- Surprisingly weak correlations between price and the depth and table variables, suggesting that the scales used for these measurements might be affecting their relationship with price.

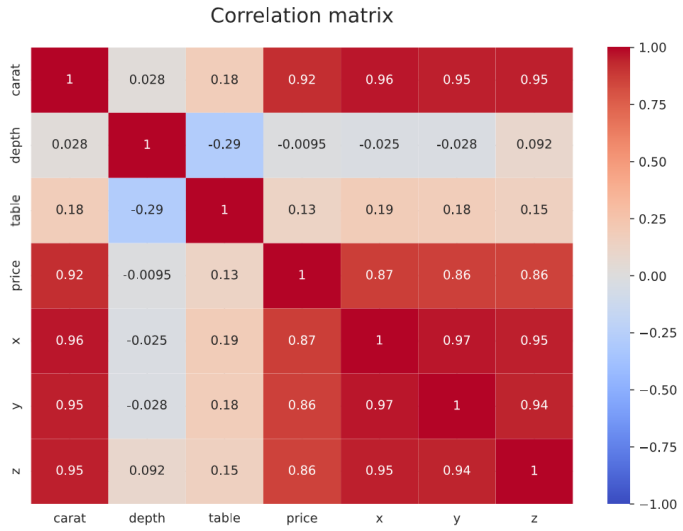

Figure 3: Correlation between numeric variables.

Categorical Variable Analysis

We examined the relationship between categorical variables and price using boxplots and ANOVA:

1. **Cut**: The premium cut showed a considerably higher price distribution compared to other cuts. Other cut categories displayed similar price distributions.
2. **Color**: Contrary to initial expectations, more yellowish diamonds (higher alphabet letters) tend to be more expensive in the Gringotts market. This trend differs from traditional valuation methods and warrants further investigation.

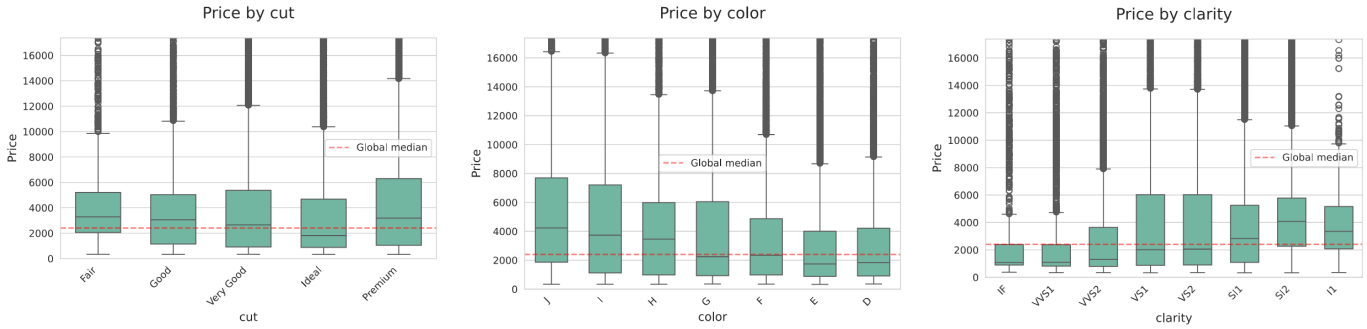

Figure 4: Diamond price boxplots by group

3. **Clarity**: Diamonds with intermediate impurity scales showed the most varied price distributions. Surprisingly, diamonds thought to be more valuable due to higher clarity displayed lower price distributions. This unexpected result may be related to the size of the diamonds and requires a deeper analysis of variable interactions.

ANOVA Analysis

An ANOVA test was performed for cut, color, and clarity about price. All three variables showed p-values very close to 0, confirming their significant role in explaining the variability of diamond prices.
These findings highlight the complex interplay of factors in diamond valuation and underscore the need for a sophisticated modeling approach to accurately predict diamond prices in the Gringotts market.

3 Predictive Model Development

After our comprehensive data analysis, we proceeded to develop predictive models to estimate diamond prices. We utilized all variables except geological coordinates, with some modifications to optimize model performance.

Data Preparation

1. Categorical variables (cut, clarity, and color) were transformed into numerical scales ranging from 0 to 1, with the best quality represented by values close to 1.

2. Depth and table variables were normalized, with their optimal values set to 1 and a decay factor applied for deviations from the optimum.

3. The dataset was randomly split into a training set (80%) and an evaluation set (20%).

4. Price values were log-transformed to work with more appropriate scales for model functionality.

3.1 Model 1: Simple Linear Regression

As a baseline, we implemented a simple linear regression model:

- **Root Mean Square Error (RMSE): 1827.21**. This provided a reasonable first approximation but left room for improvement.
- We also tested the model **including geological coordinates**, which resulted in a **worse RMSE of 2230.71**.

3.2 Model 2: Epsilon-Support Vector Regression (SVR)

To explore the potential of a more complex model, we implemented an SVR. We tested various parameter combinations (C and epsilon ε) to optimize performance.

	C	ε	RMSE
1	1000	0.1	620.22
2	500	0.2	638.95
3	**3000**	**0.01**	**594.79**

Table 1: RMSE error results of the SVR model with different values of C and epsilon ε parameters. A lower RMSE value indicates a better model.

The SVR model, particularly with $C = 3000$ and $\varepsilon = 0.01$, significantly outperformed the linear regression model. This substantial improvement in RMSE indicates that the SVR model is better suited to capture the complex relationships between diamond characteristics and price.

```
Real price: 855 ------> Our SVR prediction: 797.28
Real price: 736 ------> Our SVR prediction: 796.05
Real price: 2040 ------> Our SVR prediction: 2268.95
Real price: 1369 ------> Our SVR prediction: 1509.39
Real price: 716 ------> Our SVR prediction: 764.86
Real price: 1169 ------> Our SVR prediction: 1224.69
Real price: 10114 ------> Our SVR prediction: 10249.79
Real price: 5710 ------> Our SVR prediction: 5451.48
Real price: 1060 ------> Our SVR prediction: 1021.71
Real price: 8030 ------> Our SVR prediction: 7012.39
Real price: 15460 ------> Our SVR prediction: 15729.4
Real price: 4629 ------> Our SVR prediction: 4359.25
```

Figure 5: Radom examples of predictions.

The Epsilon-Support Vector Regression model with optimized parameters demonstrates the most promising results for estimating diamond prices. With an **RMSE of 594.79**, this model appears to be sufficiently powerful and consistent for the task of estimating the value of the stolen diamonds from Gringotts Wizarding Bank.

4 Geographical Analysis

Data Cleaning

Before the geographical analysis, it was necessary to clean the geographical database. Some latitude entries contained letters, which were removed using a regular expression. This ensured that only numerical values were used for the coordinates.

Analysis of Prediction Outliers

To investigate any potential geographical patterns in our model's performance, we focused on the test prices that our model failed to predict accurately. We used the *reverse_geocoder* library to extract the country of origin for each diamond based on its coordinates.
Findings:

1. The majority of diamonds with the worst price predictions were from the USA, Great Britain, and Canada.
2. This distribution closely mirrored the overall distribution of diamonds in the global database.
3. No clear relationship was observed between the geolocation of diamonds and their price or the accuracy of price predictions.

This geographical analysis **did not reveal any significant patterns** that could explain price variations or prediction inaccuracies. The distribution of worst-predicted diamonds across countries appears to be proportional to the overall distribution of diamonds in the dataset. This suggests that geographical factors may not play a crucial role in determining diamond prices or in the performance of our predictive model.

5 Conclusions and Final Verdict

Model Application

We applied our best-performing model (Epsilon-Support Vector Regression with C=3000 and ε=0.01) to estimate the value of the stolen diamonds. The data was transformed using the same methodology as in our training process.

Geographical Origin

Using geolocation data, we determined that all the stolen diamonds originated from the United States of America.

Economic Impact Assessment

Based on our model's predictions, we can conclude that:

1. **Estimated Total Value**: The stolen diamonds amount to an **approximate economic loss of $ 43,630.37 ± $ 1,880.88**.

2. **Comparison to Average**: The average value of a random diamond in our dataset is $2,401. Ten random diamonds would typically cost approximately $24,010.

3. **Significance of the Theft**: The stolen collection of 10 diamonds is valued at nearly double the cost of 10 average diamonds from our dataset.

	carat	cut	color	clarity	depth	table	x	y	z	latitude	longitude	price
0	0.71	Good	I	VVS2	63.1	58.0	5.64	5.71	3.58	35.02636	-114.38351	4879.61
1	0.83	Ideal	G	VS1	62.1	55.0	6.02	6.05	3.75	35.0035	-109.78961	8929.52
2	0.5	Ideal	E	VS2	61.5	55.0	5.11	5.16	3.16	35.10544	-106.669673	1943.69
3	0.39	Premium	J	VS1	61.6	59.0	4.67	4.71	2.89	34.94666	-104.6473	701.83
4	0.32	Premium	G	VS1	62.1	56.0	4.43	4.4	2.74	35.18864	-101.98602	771.18
5	0.9	Good	F	SI2	63.3	57.0	6.08	6.14	3.87	35.26611	-99.63874	2604.69
6	0.51	Ideal	D	VS1	60.9	57.0	5.2	5.17	3.16	35.51572	-97.6708	2676.45
7	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13	36.163605	-95.7595	19313.03
8	0.4	Ideal	G	VVS2	62.4	56.0	4.72	4.74	2.95	37.689186	-92.6473	1227.56
9	0.36	Premium	I	VS2	62.7	59.0	4.54	4.58	2.86	38.66303	-90.21808	582.82

Figure 6: Provided lists of stolen diamonds with their respective price prediction.

Expert Insight

The substantial difference between the value of the stolen diamonds and that of average diamonds suggests that **the thieves possessed significant knowledge about diamond valuation**. They managed to select 10 diamonds that collectively are worth almost twice as much as a random selection would be.

Final Verdict

Based on our comprehensive analysis and advanced predictive modeling, we can confidently state that the economic impact of the diamond theft from Gringotts Wizarding Bank is substantial. **The precision of the thieves' selection indicates a sophisticated operation**, potentially carried out by individuals with insider knowledge or expertise in the field of diamond valuation.

Additional Experimentation

It's worth noting that we also experimented with a 2-layer neural network model in our quest for the most accurate prediction method. While we observed a gradual reduction in Mean Squared Error (MSE) over 100 epochs of training, the results fell short of our expectations. The performance of this neural network model was significantly inferior to our SVR model, and thus, we decided not to include it in the main body of this report. This experiment underscores the importance of model selection and highlights that more complex models don't always yield better results, especially with datasets of this nature and size.

Conclusions and future work

This concludes our analysis of the lapidarist problem. We stand ready to provide any further clarification or assistance as needed in this matter. We believe that it is possible to delve deeper into this research by proposing new predictive models and by conducting further tests with different parameter variations or by adding the geolocations of the diamonds directly into the analysis in a more efficient way. We also believe that further evaluating the locations of the cities from which the stolen diamonds originate could provide more relevant clues.