The lapidarist problem

Test Data Scientist - EDT

Edison David Serrano Cárdenas

Maestría en Ciencias con Orientación en Matemáticas Aplicadas

Introduction and Exploratory Analysis

Context of the problem and dataset overview.

Estimating the Value of Stolen Diamonds

Context:

A robbery at Gringotts Wizarding Bank has resulted in the theft of valuable diamonds. The exact value of these diamonds needs to be determined for further investigation and recovery efforts.

Objective:

Accurately estimate the value of the stolen diamonds based on their characteristics using data-driven methods.

Diamond Dataset Overview

Categorical Data

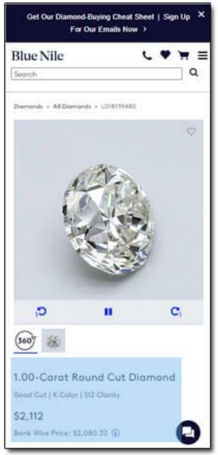
Cut, color, clarity

Numerical Data

carat, depth, table, price, x, y, z, longitude, latitude

Dataset Overview

The dataset contains information on 12 characteristics of diamonds. There is clean data on 50,934 diamonds in Krenk's dataset.

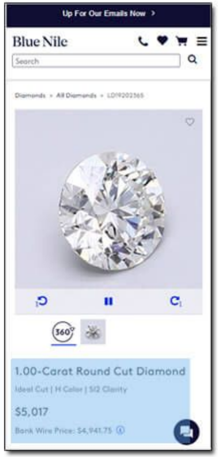

Get Our Diamond-Buying Cheat Sheet | Sign Up For Our Emails Now ›
Blue Nile
Search
Diamonds › All Diamonds › LD18739480
360°

1.00-Carat Round Cut Diamond

Good Cut | K Color | SI2 Clarity

$2,112

Bank Wire Price: $2,080.32

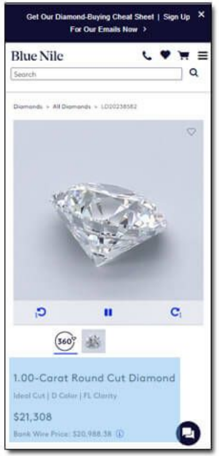


1.00-Carat Round Cut Diamond

Ideal Cut | H Color | SI2 Clarity

$5,017

Bank Wire Price: $4,941.75


Get Our Diamond-Buying Cheat Sheet | Sign Up For Our Emails Now ›
Blue Nile
Search
Diamonds › All Diamonds › LD20238582
360°

1.00-Carat Round Cut Diamond

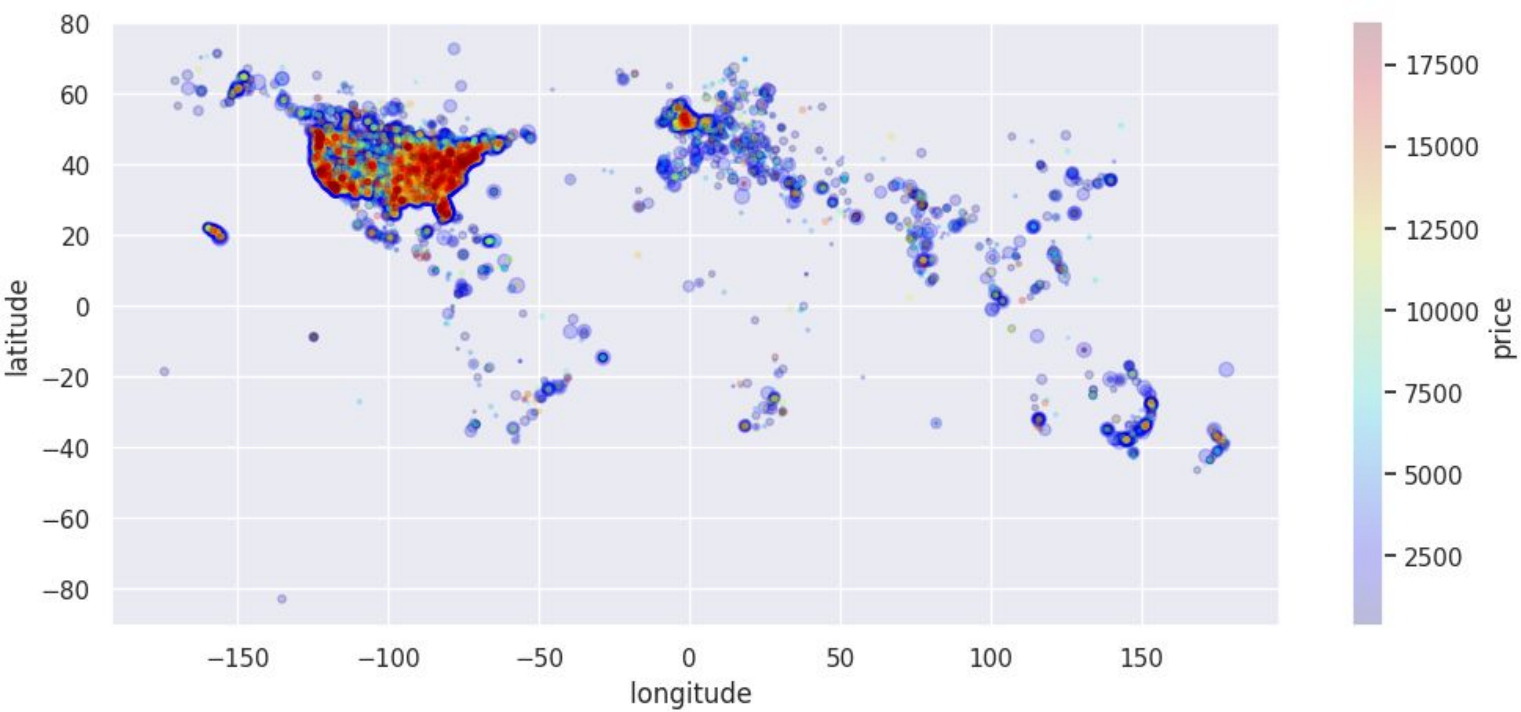
Ideal Cut | D Color | FL Clarity

$21,308

Bank Wire Price: $20,988.38

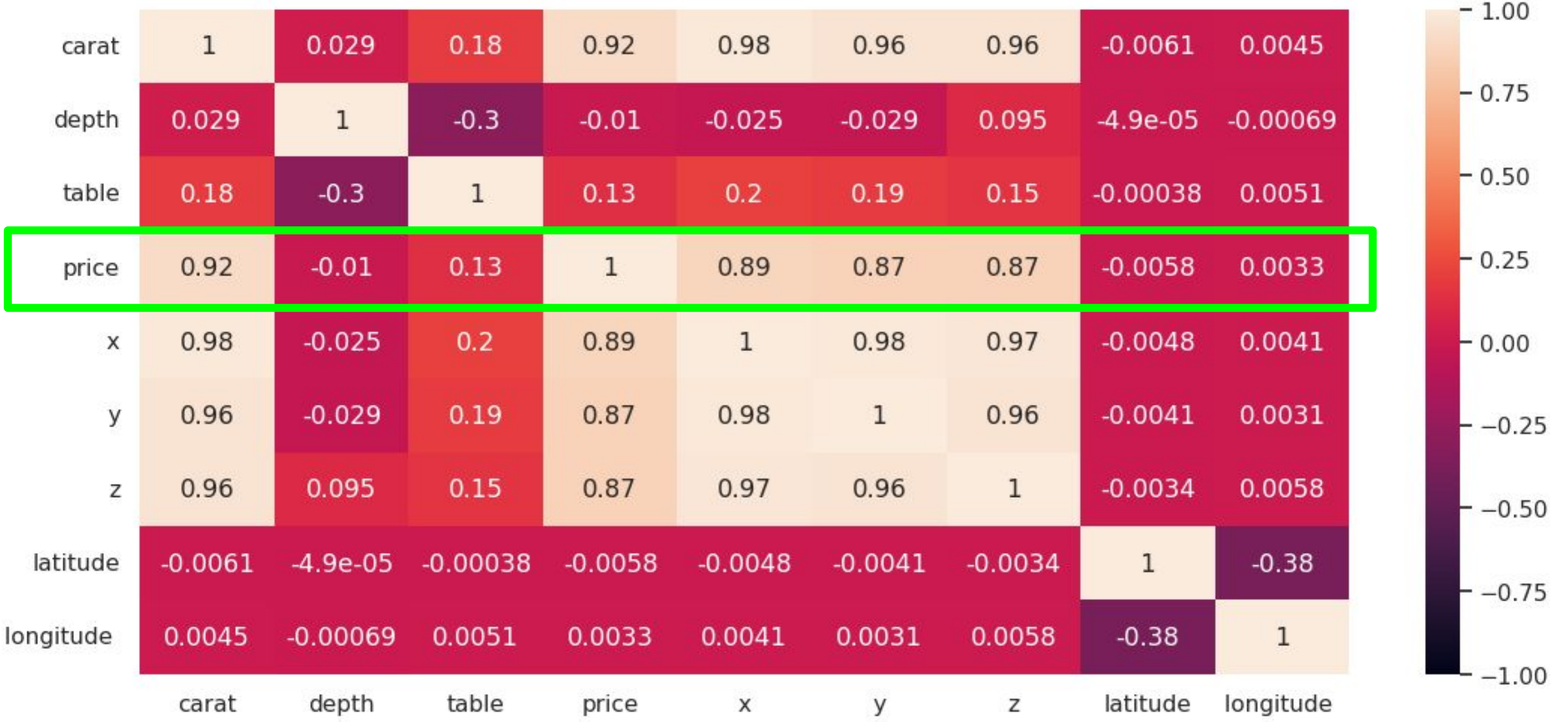
• • •

Geographic Distribution of Data



93.47% of the dataset is sourced from the American continent.

Correlation Heatmap

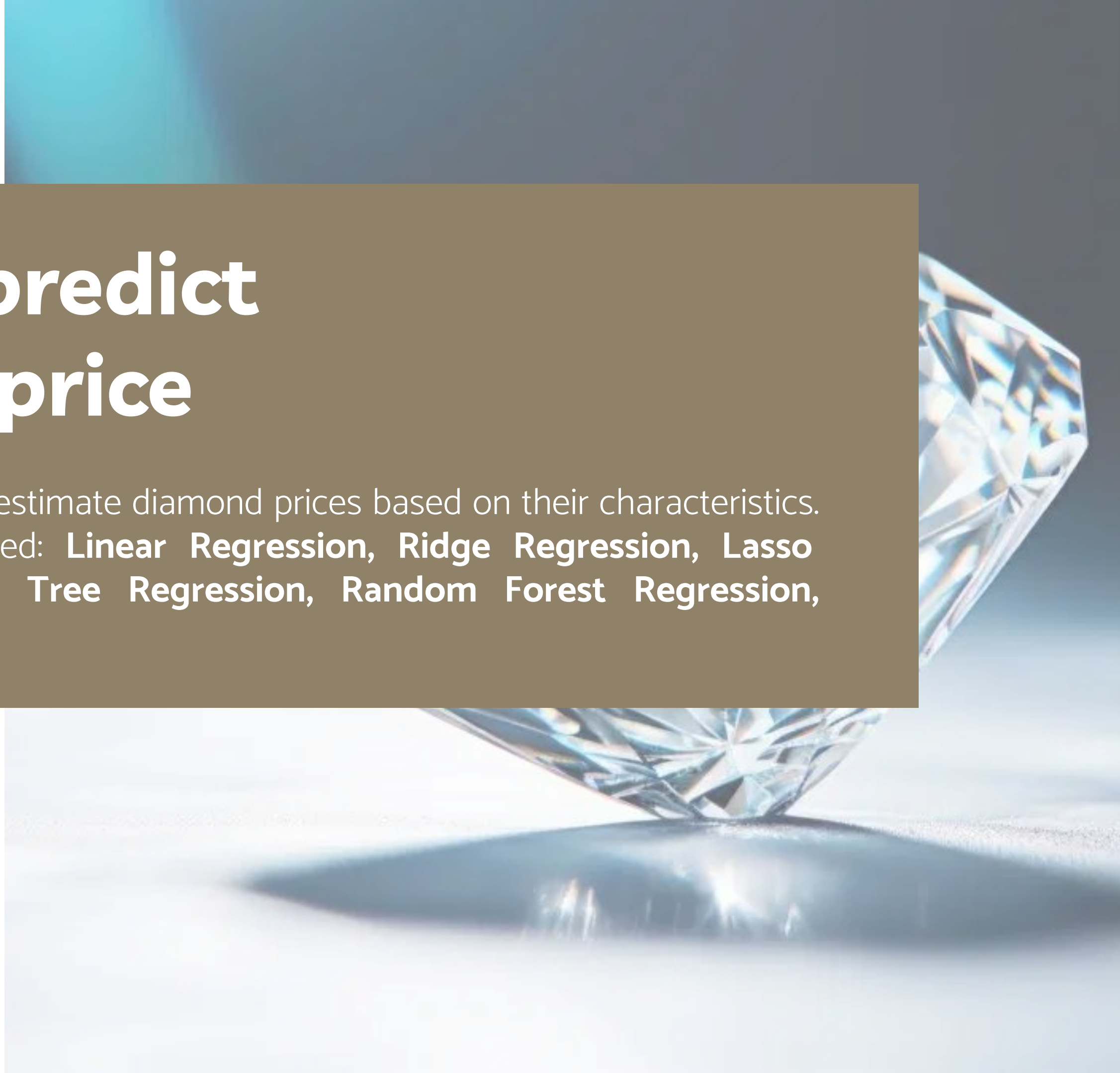


	carat	depth	table	price	x	y	z	latitude	longitude
carat	1	0.029	0.18	0.92	0.98	0.96	0.96	-0.0061	0.0045
depth	0.029	1	-0.3	-0.01	-0.025	-0.029	0.095	-4.9e-05	-0.00069
table	0.18	-0.3	1	0.13	0.2	0.19	0.15	-0.00038	0.0051
price	0.92	-0.01	0.13	1	0.89	0.87	0.87	-0.0058	0.0033
x	0.98	-0.025	0.2	0.89	1	0.98	0.97	-0.0048	0.0041
y	0.96	-0.029	0.19	0.87	0.98	1	0.96	-0.0041	0.0031
z	0.96	0.095	0.15	0.87	0.97	0.96	1	-0.0034	0.0058
latitude	-0.0061	-4.9e-05	-0.00038	-0.0058	-0.0048	-0.0041	-0.0034	1	-0.38
longitude	0.0045	-0.00069	0.0051	0.0033	0.0041	0.0031	0.0058	-0.38	1

The most correlated variable is carat with a negative coefficient of 0.9214. Geographical data may not be relevant for predicting diamond prices in this context.

Models to predict diamond’s price

This is a regression model to estimate diamond prices based on their characteristics. Eight models were considered: **Linear Regression, Ridge Regression, Lasso Regression, SVR, Decision Tree Regression, Random Forest Regression, XGBoost, and LightGBM.**



Model Performance Metrics

Mean Squared Error (MSE):

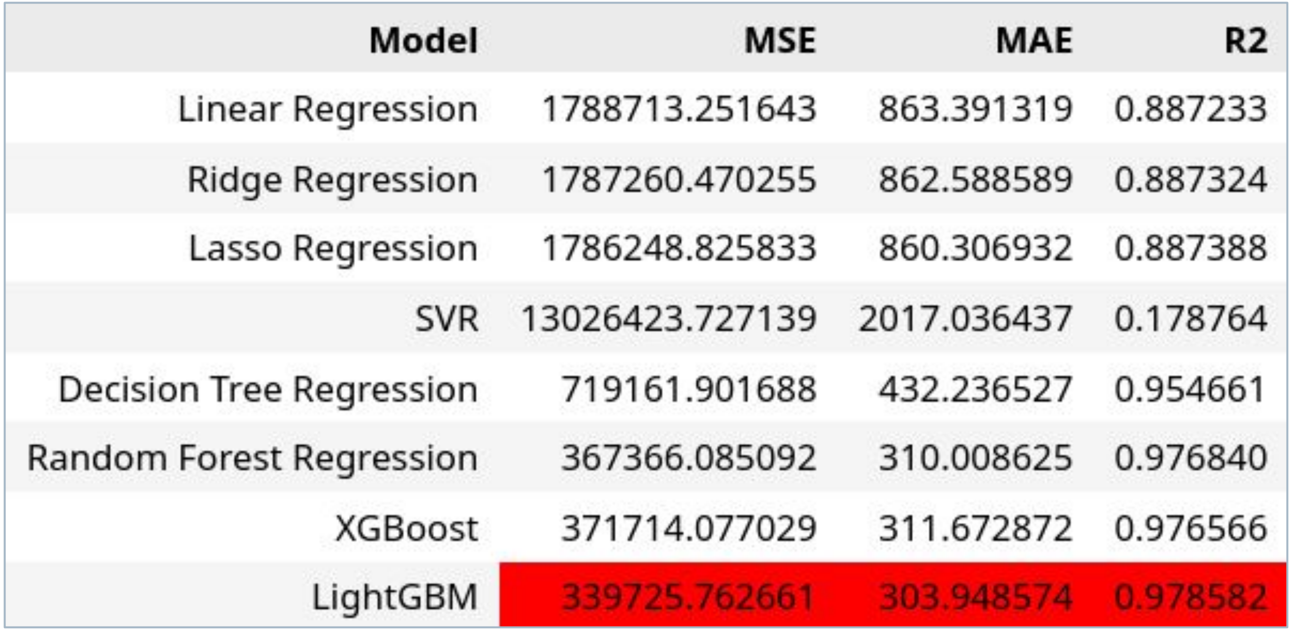
Mean Absolute Error (MAE):

These metrics provide complementary insights into model accuracy, with MSE being more sensitive to outliers and MAE offering a more straightforward interpretation.

R-Squared (R^2):

This metric gives an overall indication of how well the model fits the data, helping to assess the model's explanatory power. **An R^2 close to 1 indicates that the model explains a large portion of the variance in the data**

Model Comparison Performance

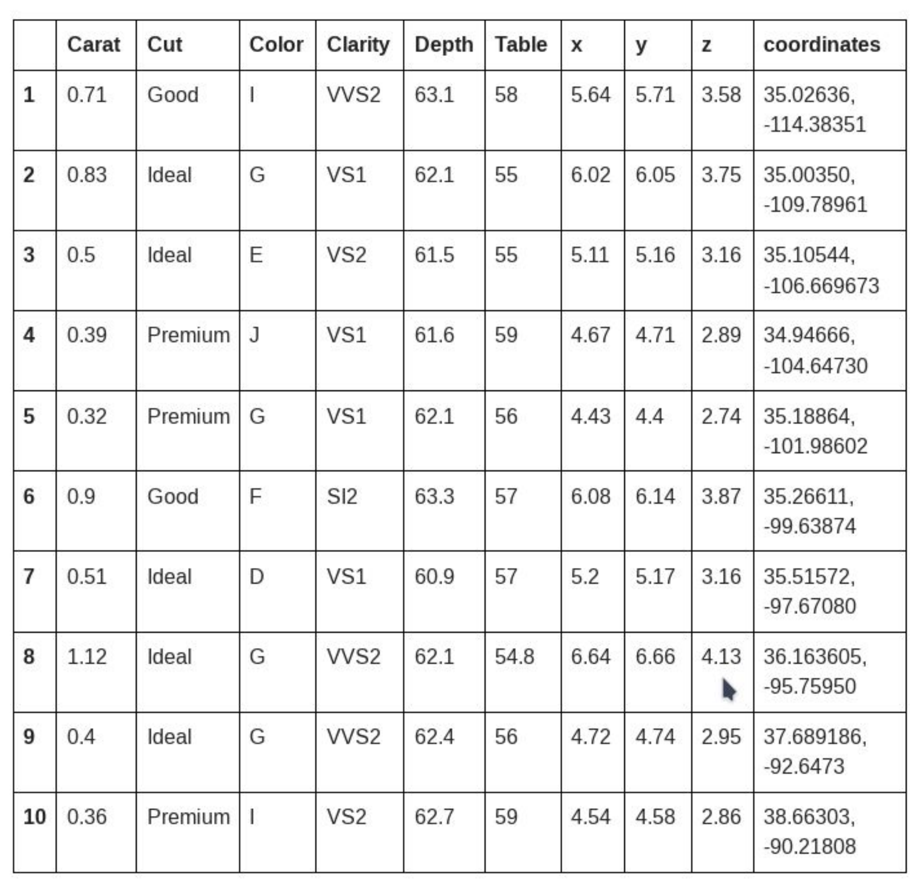


Model	MSE	MAE	R2
Linear Regression	1788713.251643	863.391319	0.887233
Ridge Regression	1787260.470255	862.588589	0.887324
Lasso Regression	1786248.825833	860.306932	0.887388
SVR	13026423.727139	2017.036437	0.178764
Decision Tree Regression	719161.901688	432.236527	0.954661
Random Forest Regression	367366.085092	310.008625	0.976840
XGBoost	371714.077029	311.672872	0.976566
LightGBM	339725.762661	303.948574	0.978582

The best model among the eight is the LightGBM model, which has the best performance across all the metrics considered.

CONCLUSIONS

It shows the best model performance and predicts the price of the stolen diamonds.

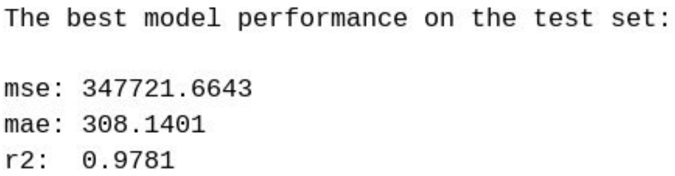


	Carat	Cut	Color	Clarity	Depth	Table	x	y	z	coordinates
1	0.71	Good	I	VVS2	63.1	58	5.64	5.71	3.58	35.02636, -114.38351
2	0.83	Ideal	G	VS1	62.1	55	6.02	6.05	3.75	35.00350, -109.78961
3	0.5	Ideal	E	VS2	61.5	55	5.11	5.16	3.16	35.10544, -106.669673
4	0.39	Premium	J	VS1	61.6	59	4.67	4.71	2.89	34.94666, -104.64730
5	0.32	Premium	G	VS1	62.1	56	4.43	4.4	2.74	35.18864, -101.98602
6	0.9	Good	F	SI2	63.3	57	6.08	6.14	3.87	35.26611, -99.63874
7	0.51	Ideal	D	VS1	60.9	57	5.2	5.17	3.16	35.51572, -97.67080
8	1.12	Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13	36.163605, -95.75950
9	0.4	Ideal	G	VVS2	62.4	56	4.72	4.74	2.95	37.689186, -92.6473
10	0.36	Premium	I	VS2	62.7	59	4.54	4.58	2.86	38.66303, -90.21808

How much was stolen?

$24777.30

The most valuable stolen diamond was the eighth one, which is estimated to have a price of **$8,681.98**



```
The best model performance on the test set:

mse: 347721.6643
mae: 308.1401
r2:  0.9781
```