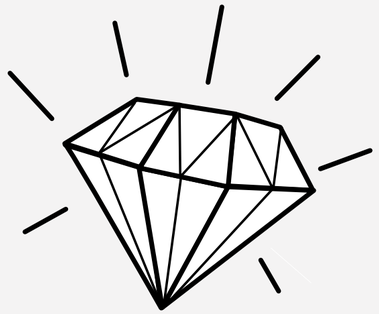


DIAMOND PRICE PREDICTION MODEL

A DATA-DRIVEN APPROACH TO PRICING ACCURACY

Data Science Programming Course Project

MEET THE TEAM

Luke Hartfield

Chris Breton

Lena Weisssman

Arjun Rajesh

Varsha Ramesh

AGENDA OVERVIEW

01

PROBLEM

02

EXPLORATORY DATA ANALYSIS

03

METHODOLOGY AND SOLUTION

04

RESULTS & INSIGHTS

PROBLEM STATEMENT

Diamond prices are difficult to predict accurately because they depend on multiple complex factors, causing challenges for consumers and sellers in determining fair value.

01

SCOPE OF THE STUDY

Dataset of 53,940 round-cut diamonds with 10 attributes (carat, cut, color, clarity, depth, table, dimensions)

02

RELEVANCE OF THE STUDY

Supports smarter decisions in the jewelry industry;
Help buyers avoid overpaying and sellers price competitively

03

RESEARCH QUESTION

How does quality and physical factors influence diamond price, and can we accurately predict its price using these features?

EXPLORING THE DATA

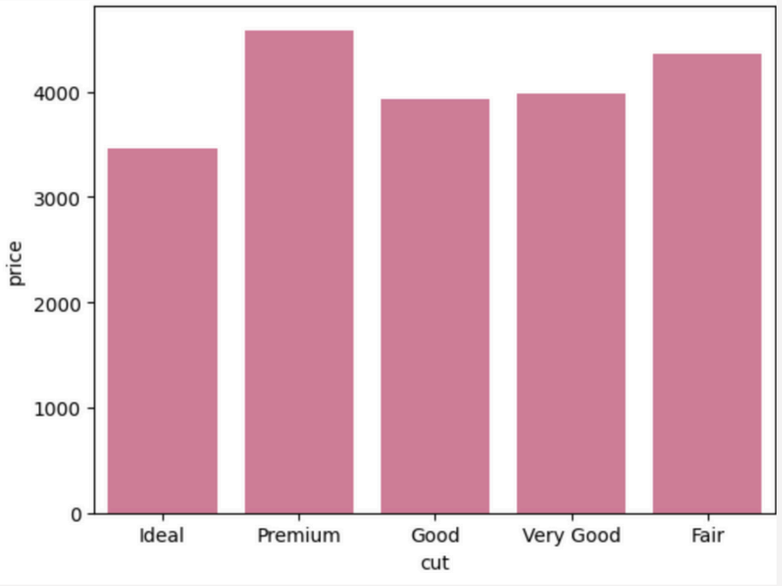
- Diamonds Dataset is recognized for having no missing values, however some rows have zero values in columns x, y, and z. (length (mm), width (mm), and depth (mm)
- Disclaimer: Cut (Premium more value) Color (D-F more value) Clarity (Fl, IF more value)

	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

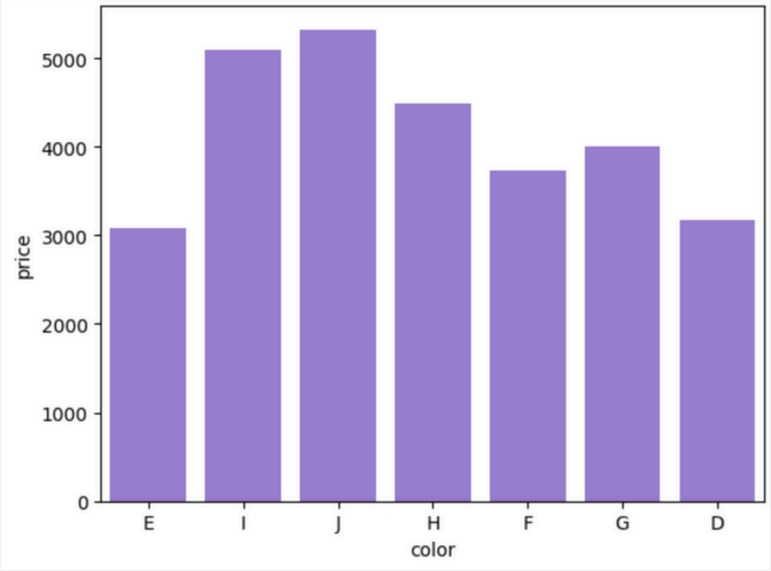
EXPLORING THE DATA

- Diamonds Dataset Premium cut diamonds are the most expensive, followed by fair and very good. J colored diamonds seem to be the most expensive, followed by I and H. Finally, SI2 clarity diamonds have the highest price followed by SI1.

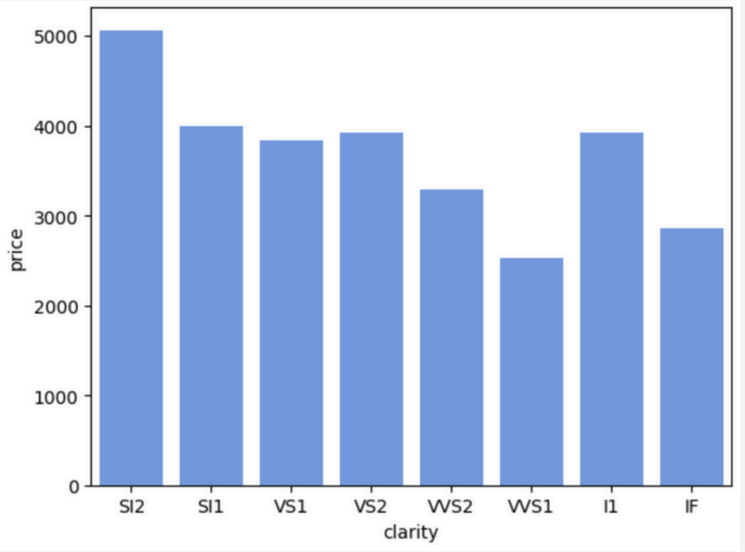
Price Vs. Cut



Price Vs. Color



Price Vs. Clarity



SUMMARY STATISTICS

	carat	depth	table	price	x	y	z
count	53943.000000	53943.000000	53943.000000	53943.000000	53943.000000	53943.000000	53943.000000
mean	0.797935	61.749322	57.457251	3932.734294	5.731158	5.734526	3.538730
std	0.473999	1.432626	2.234549	3989.338447	1.121730	1.142103	0.705679
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	5324.000000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

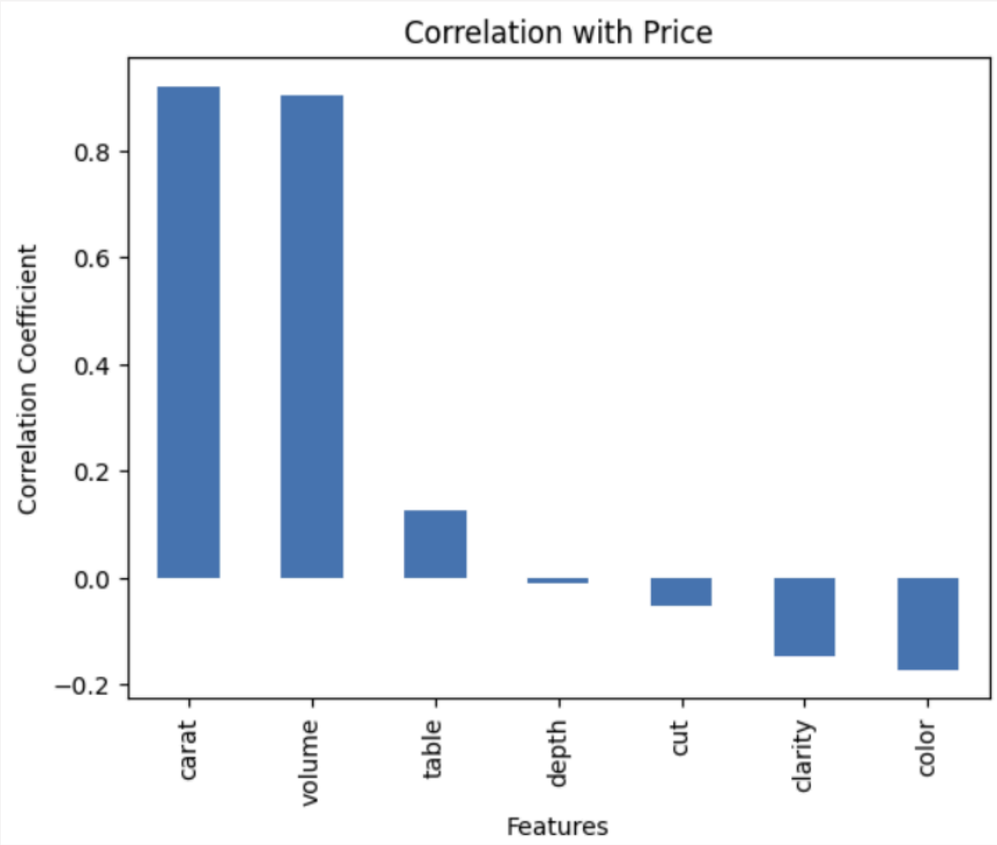
Carat: 0.20–5.01 (avg ~0.80)
Depth %: 43–79 (avg ~61.8)
Table %: 43–95 (avg ~57.5)
Price: $326–$18,000+
Outliers: Some extreme or zero values in dimensions

DATA PREPROCESSING

	carat	cut	color	clarity	depth	table	price	length(mm)	width(mm)	depth(mm)	volume
0	0.23	5	6	2	61.5	55.0	326	3.95	3.98	2.43	38.202030
1	0.21	4	6	3	59.8	61.0	326	3.89	3.84	2.31	34.505856
2	0.23	2	6	5	56.9	65.0	327	4.05	4.07	2.31	38.076885
3	0.29	4	2	4	62.4	58.0	334	4.20	4.23	2.63	46.724580
4	0.31	2	1	2	63.3	58.0	335	4.34	4.35	2.75	51.917250

We converted categorical quality grades into numeric rankings so the model could learn patterns.
We also created a 'volume' feature to capture diamond size more fully than carat alone.

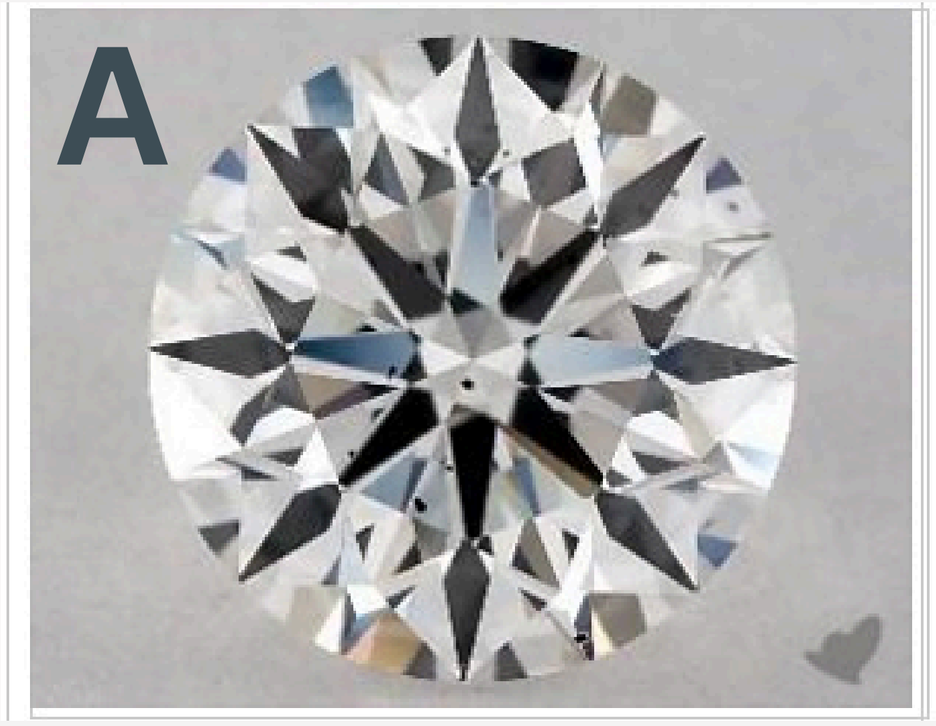
FEATURE IMPORTANCE

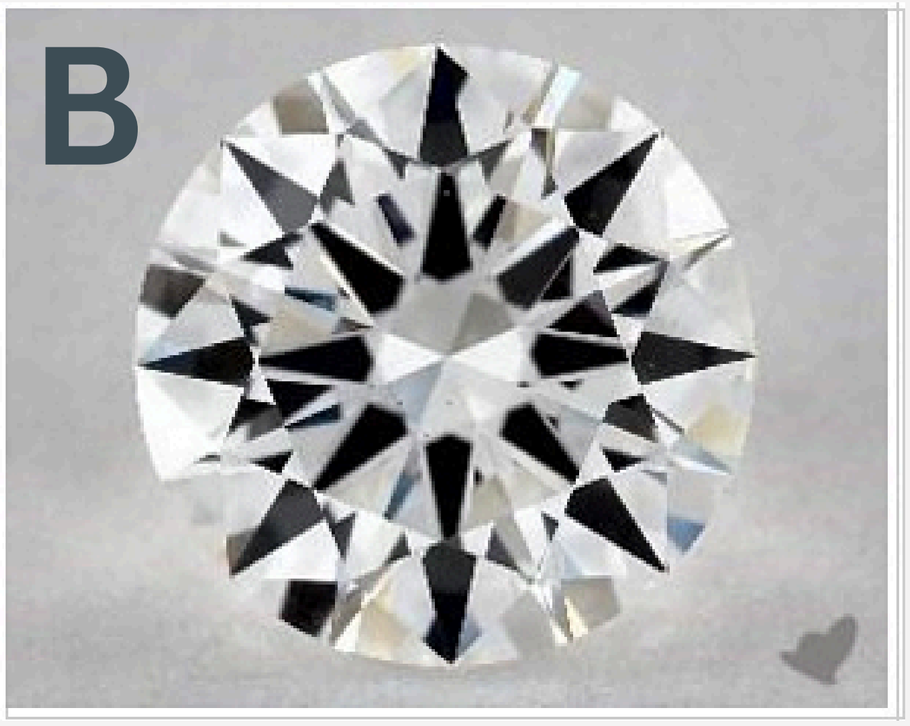


We identified the correlations between diamond features and price.

This analysis shows that volume and carat are the strongest predictors of diamond price, followed by clarity and color. Other features like depth, table, and cut have less impact.

WHICH DIAMOND COSTS MORE?





WHICH DIAMOND COSTS MORE?

ROUND 1 CARAT F SI1 EXCELLENT $5,920

ITEM#:	4891717
CARAT:	1
COLOR:	F
CLARITY:	SI1
CUT:	Excellent
DEPTH:	62.5
TABLE:	56
POL/SYM:	EX/ EX
FLUOR:	NN
GIRDLE:	Medium to Slightly Thick
CULET:	N
MEASUR':	6.39*6.38*3.99
LAB:	GIA

ROUND 1 CARAT F VS1 EXCELLENT $8,330

ITEM#:	5039682
CARAT:	1
COLOR:	F
CLARITY:	VS1
CUT:	Excellent
DEPTH:	59.1
TABLE:	60
POL/SYM:	EX/ EX
FLUOR:	NN
GIRDLE:	Medium to Slightly Thick
CULET:	VS
MEASUR':	6.55*6.51*3.86
LAB:	GIA

If you guessed B, you were correct!

MODELING APPROACH

Multilinear Regression

Grow-Prune

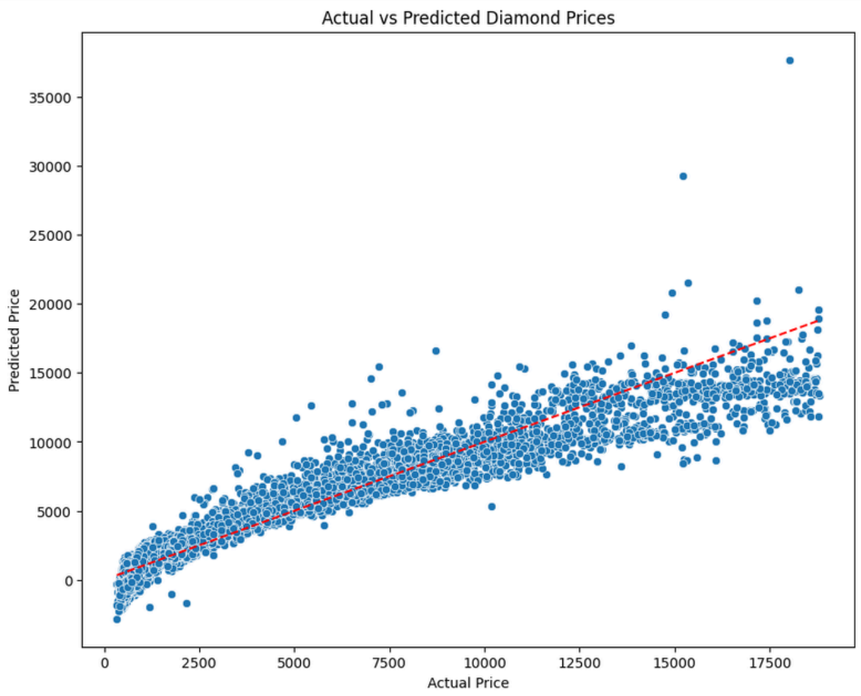
Bagging

Random Forest

Boosting

MODELING

MULTIPLE LINEAR REGRESSION

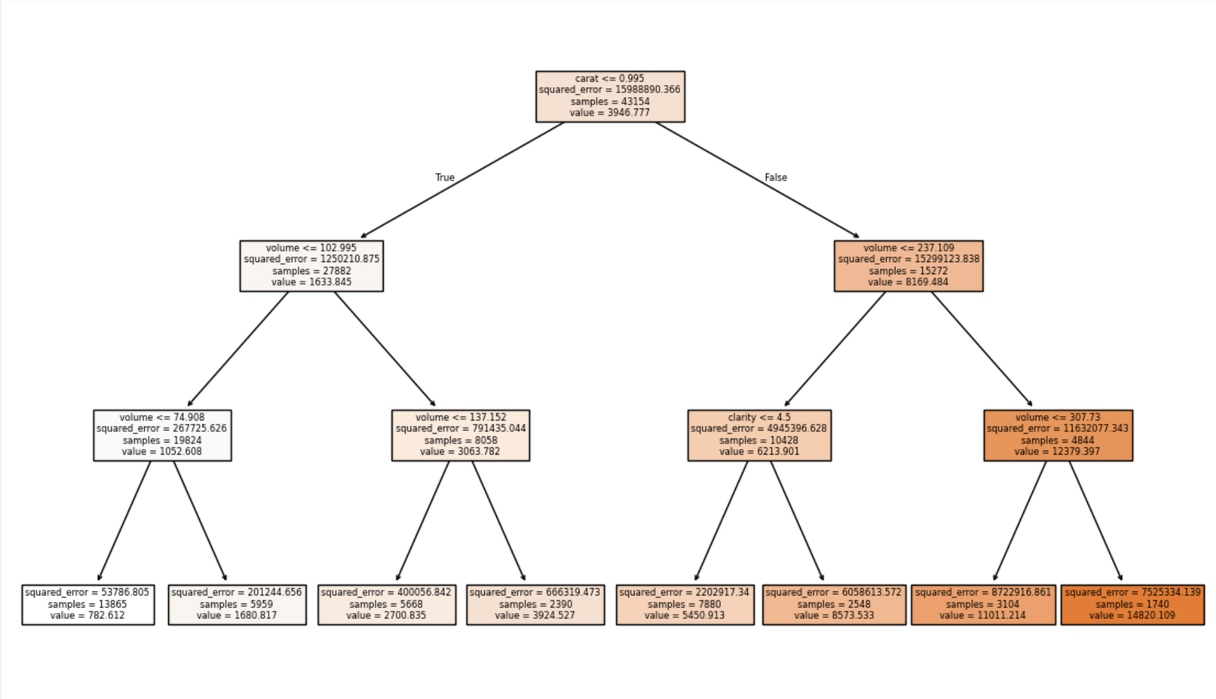


MSE 1,451,708
R-squared: 0.91
RMSE: **$1204.86**

remove outliers →

MSE: 747,798
R-squared: 0.81
RMSE: **$864.75**

GROW-PRUNE



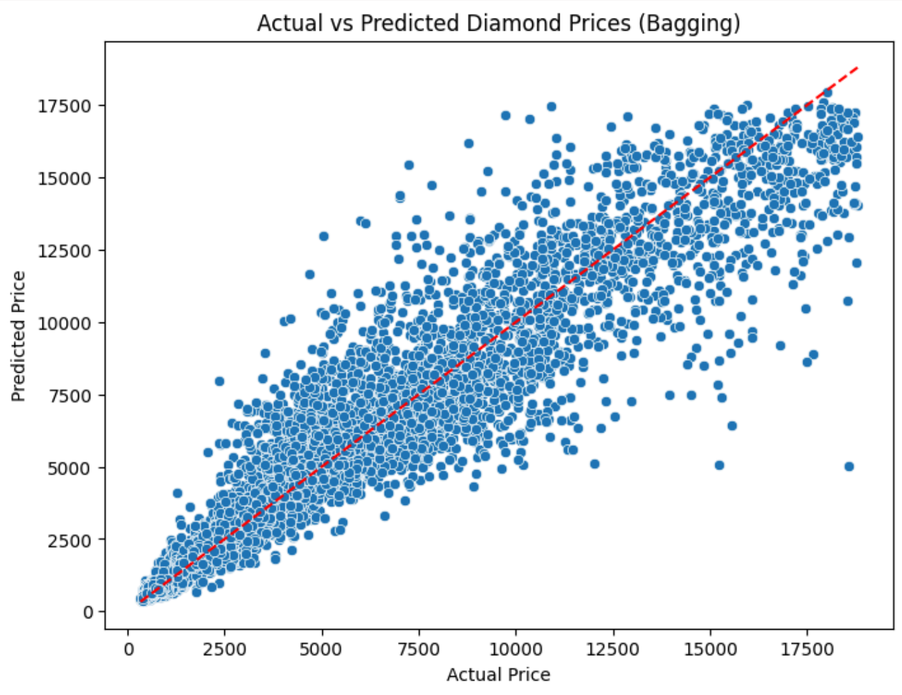
MSE: 511,684
R-squared: 0.97
RMSE: **$711.50**

prune →

MSE: 1,777,274
R-squared: 0.89
RMSE: **$1,333.20**

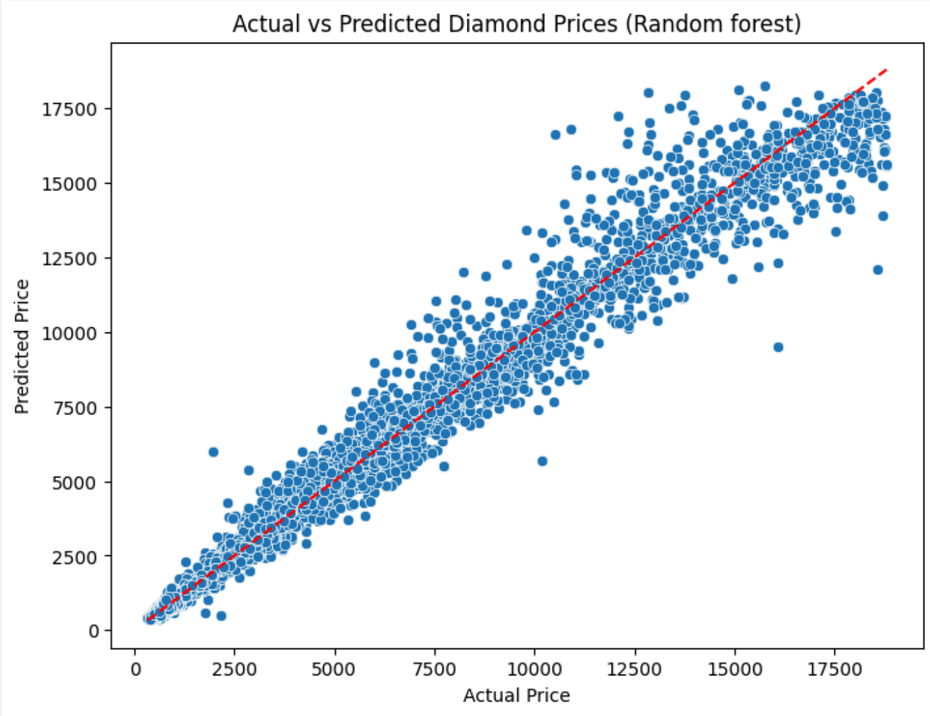
MODELING

BAGGING



Mean Squared Error: 1,332,001
R-squared: 0.91
RMSE: **$1154.10**

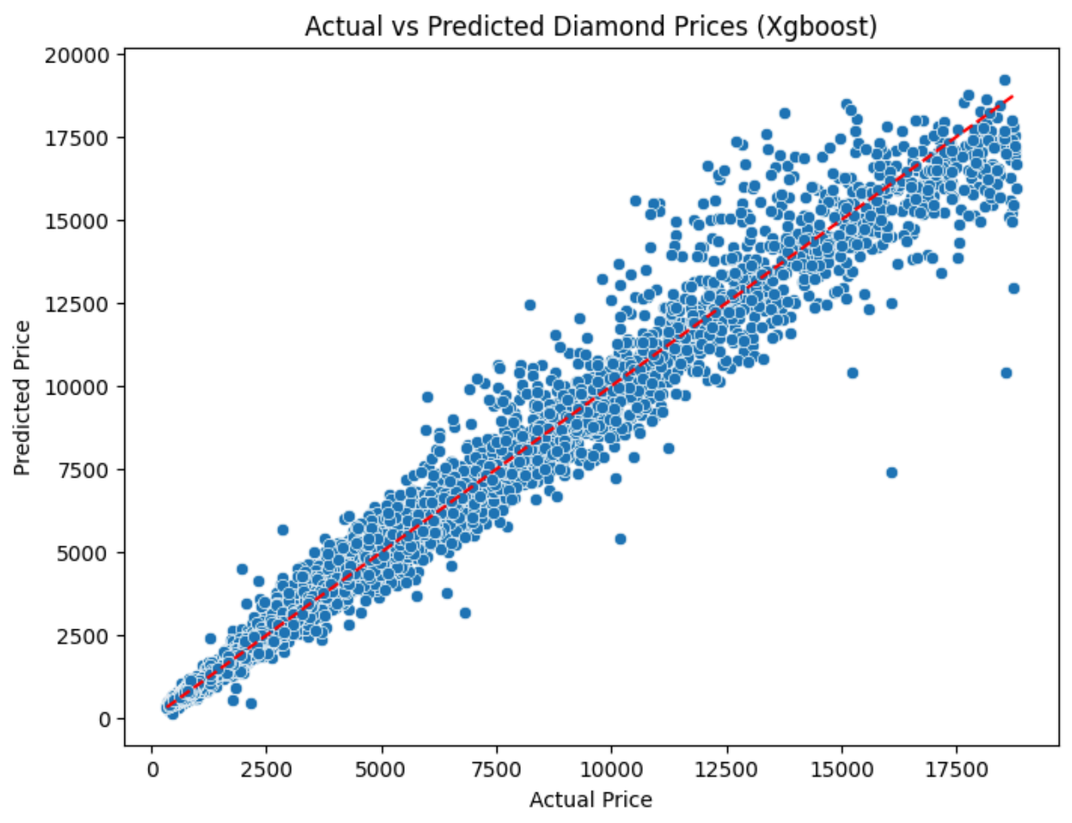
RANDOM FOREST



Mean Squared Error: 286,563
R-squared: 0.98
RMSE: **$535.10**

MODELING

BOOSTING

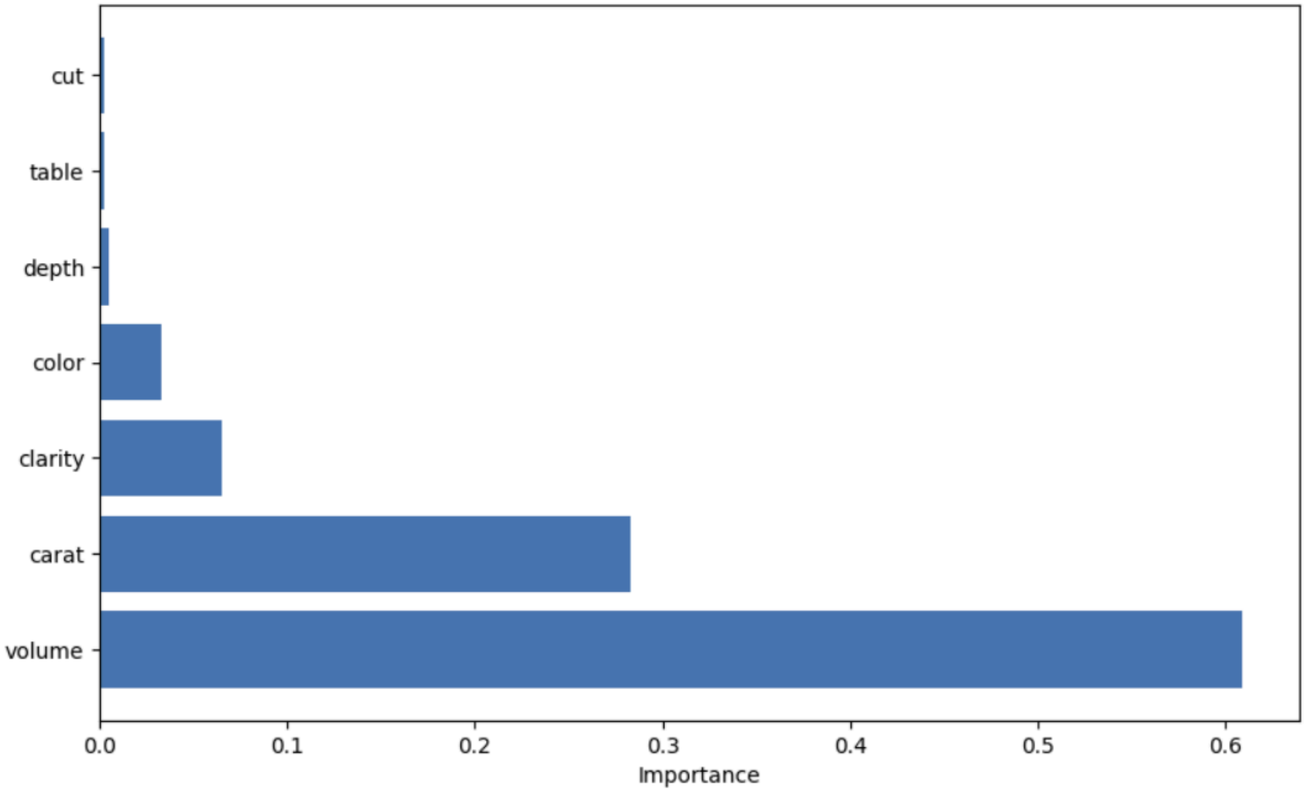


Similar Results to Random Forests

MSE: 294,682
R-squared: .88
RMSE:**$542.8**

- Capture Pattern
- Reduction of Noise in the Model
- Averaging Out Via Iteration of Residuals

RESULTS



Random Forest gave the lowest error, meaning it's the most reliable model for prediction in this dataset.

RMSE of **$535.14** and R^2 of **98%**

The size and the carat (weight) of the diamond are the most important features

CONCLUSION

Consumers – Spot overpriced diamonds; know top price drivers (volume, carat, clarity, color)

Jewelers – Competitive pricing; quick quotes for custom orders

Online Marketplaces – Real-time suggested pricing for listings

Data Science Programming

THANK YOU

Questions?