Diamond Price Prediction

WISDOM IZUCHUKWU ADIKE

Content

- Dataset Description
- Main objectives of the analysis.
- Applying various regression models.
- Machine learning analysis and findings.
- Models flaws and advanced steps.

Dataset Description

Introduction

This project involves predicting the price of approximately 54,000 diamonds based on various attributes. These attributes include the weight (carat), dimensions (length, width, and depth), depth percentage (%Depth), table width, and quality, color, and clarity ratings. The goal is to build a predictive model that can accurately estimate the price of diamonds based on these features, which are crucial determinants in the diamond pricing market. This prediction task holds significant value for both buyers and sellers in the diamond industry, aiding in informed decision-making and price assessment.

Dataset Description 01

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
5	6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
8	9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
9	10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

You'd notice 'Unnamed: 0' column similar to the indexing. We will drop the column.

Dataset Description 02

- **carat:** weight of the diamond.
- **x:** Length of the diamond.
- **y:** Width of the diamond.
- **z:** Depth of the diamond.
- **depth:** Depth percentage. Formula: z / mean(x, y) = 2 * z / (x + y)
- **table:** Width of top of the diamond relative to the widest point.
- **cut:** Quality of the diamond. Possible values (from best to worst)
- **color:** Color of the diamond. Possible values: from D (best) to J (worst).
- **clarity:** Measurement of how clear the diamond is. Possible values (from best to worst): IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1.
- **price:** price of the diamond in US dollars.

Dataset Description 03

	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

Mean:
Carat: 0.8
Depth: 62
Price: 3932

Min:
Carat: 0.2
Depth: 43
Price: 326

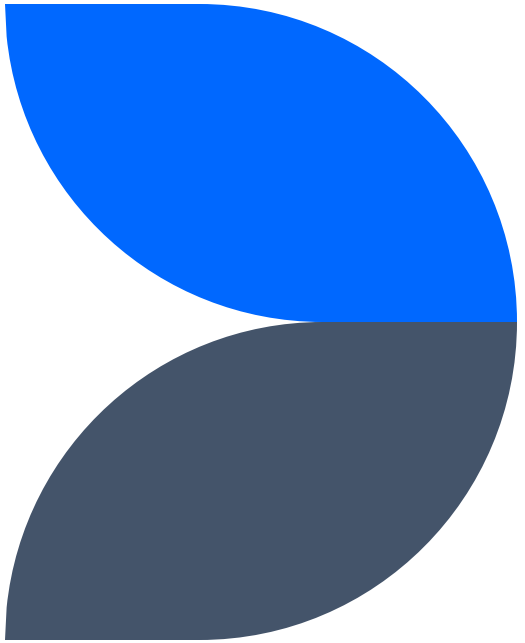
Max:
Carat: 5.0
Depth: 79
Price: 18823

Dataset Description 04

```
carat      0
cut        0
color      0
clarity    0
depth      0
table      0
price      0
x          0
y          0
z          0
dtype: int64
```

You can see we have no missing values in our dataset. Hence, no need for handling missing values.

Data Analysis



Main Objective of the Analysis

In this section, I'm visualizing feature correlations through pairplots to identify relevant features. Following that, I'll construct various regression models using advanced techniques, including GridSearch, ML pipelines, and hyperparameter tuning. The objective is to obtain the most accurate predictive model while highlighting the shortcomings of each model.

Data Analysis 01

	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
5	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
8	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
9	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	2	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	3	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	1	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	3	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	1	J	SI2	63.3	58.0	335	4.34	4.35	2.75
5	0.24	4	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	0.24	4	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	0.26	4	H	SI1	61.9	55.0	337	4.07	4.11	2.53
8	0.22	0	E	VS2	65.1	61.0	337	3.87	3.78	2.49
9	0.23	4	H	VS1	59.4	61.0	338	4.00	4.05	2.39

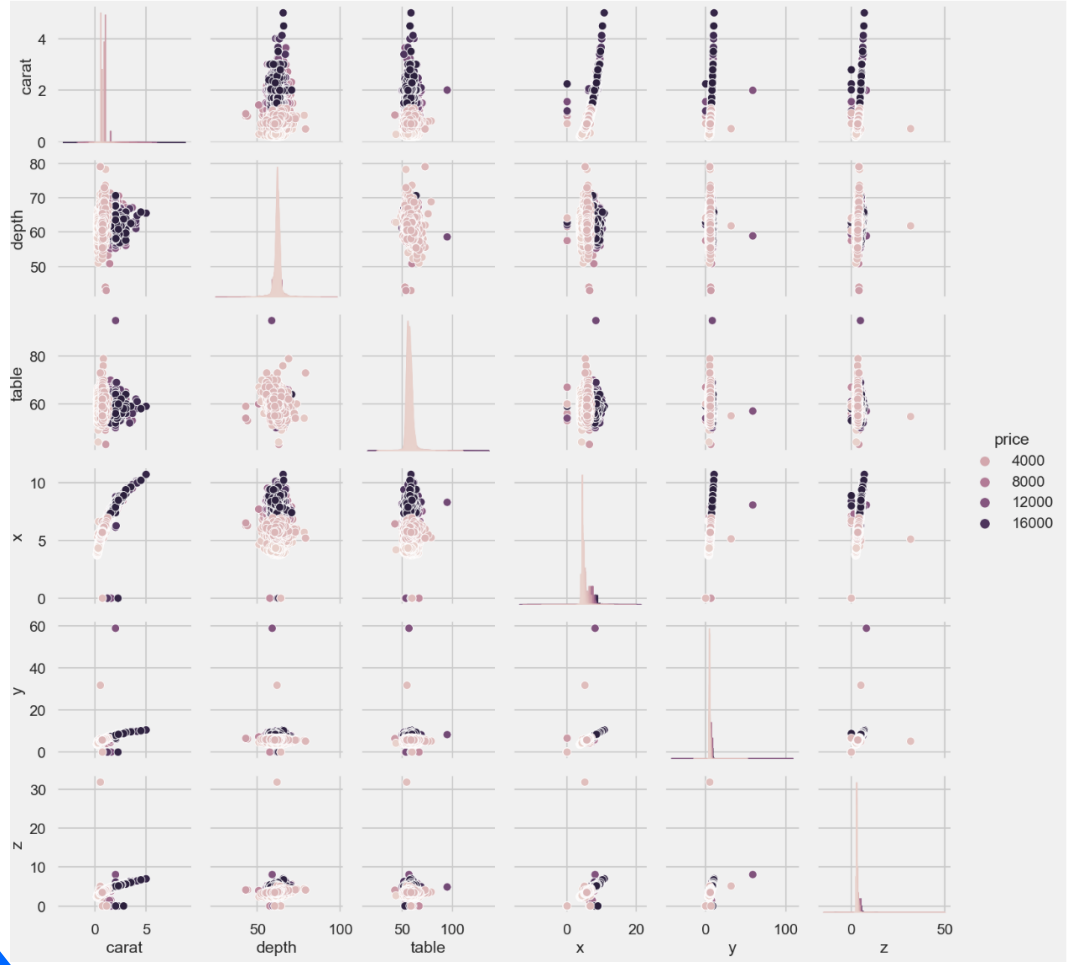
From the both tables, you will notice we converted the ordinal categorical variable 'cut' to a numerical variable.

Data Analysis 02

	carat	depth	table	price	x	y	z	cut_Fair	cut_Good	cut_Ideal	...	color_I	color_J	clarity_I1	clarity_
0	0.23	61.5	55.0	326	3.95	3.98	2.43	0.0	0.0	1.0	...	0.0	0.0	0.0	0
1	0.21	59.8	61.0	326	3.89	3.84	2.31	0.0	0.0	0.0	...	0.0	0.0	0.0	0
2	0.23	56.9	65.0	327	4.05	4.07	2.31	0.0	1.0	0.0	...	0.0	0.0	0.0	0
3	0.29	62.4	58.0	334	4.20	4.23	2.63	0.0	0.0	0.0	...	1.0	0.0	0.0	0
4	0.31	63.3	58.0	335	4.34	4.35	2.75	0.0	1.0	0.0	...	0.0	1.0	0.0	0
5	0.24	62.8	57.0	336	3.94	3.96	2.48	0.0	0.0	0.0	...	0.0	1.0	0.0	0
6	0.24	62.3	57.0	336	3.95	3.98	2.47	0.0	0.0	0.0	...	1.0	0.0	0.0	0
7	0.26	61.9	55.0	337	4.07	4.11	2.53	0.0	0.0	0.0	...	0.0	0.0	0.0	0
8	0.22	65.1	61.0	337	3.87	3.78	2.49	1.0	0.0	0.0	...	0.0	0.0	0.0	0
9	0.23	59.4	61.0	338	4.00	4.05	2.39	0.0	0.0	0.0	...	0.0	0.0	0.0	0

Performed One-Hot encoding on the continuous categorical variables.

Data Analysis 03



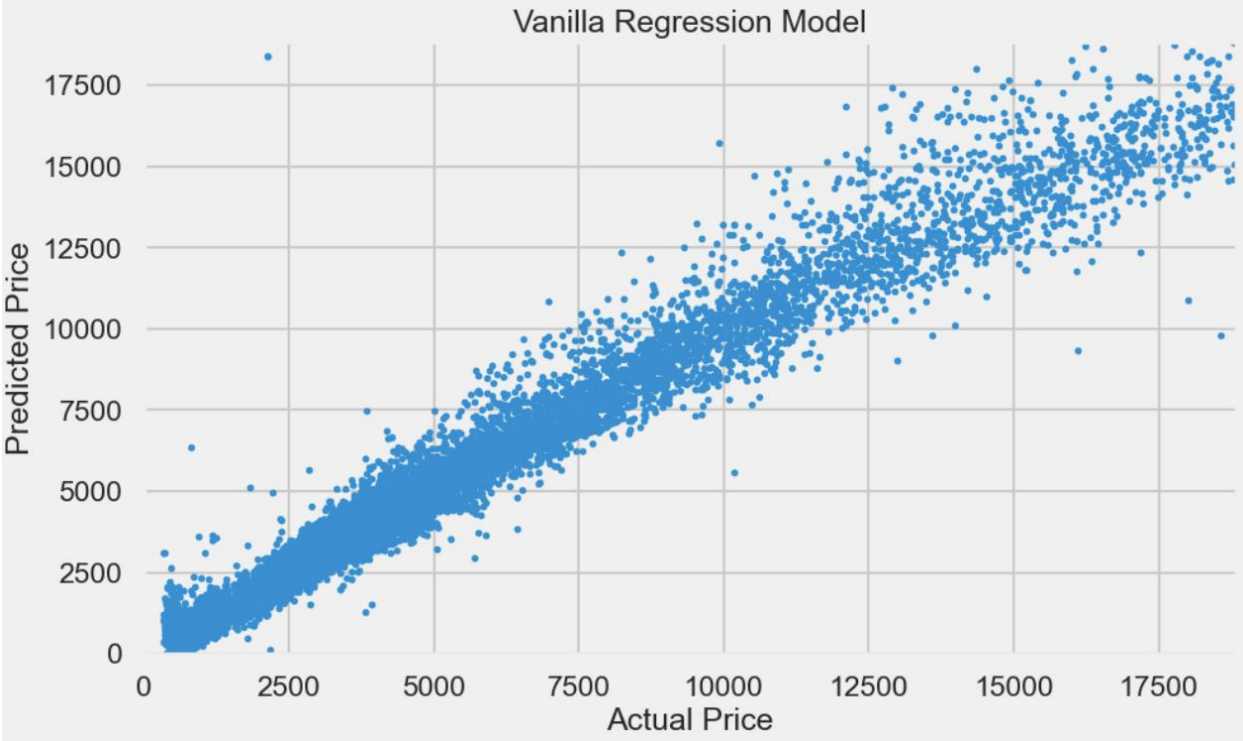
A pairplot showcasing the relationship and the trend between the diamond features.

Machine Learning Analysis and Findings

Machine Learning Analysis and Findings.

In the upcoming analysis, I will conduct a comparison among four distinct regression models: Vanilla, Lasso, Ridge, and ElasticNet. The primary focus will be on assessing their accuracy in predicting Diamond prices. To achieve robust modeling, I'll employ various techniques including standard scaling, polynomial effects, regularization regression, cross-validation, and grid search. I'll also evaluate model performance using metrics such as RMS and R2 Score.

Machine Learning Analysis 01



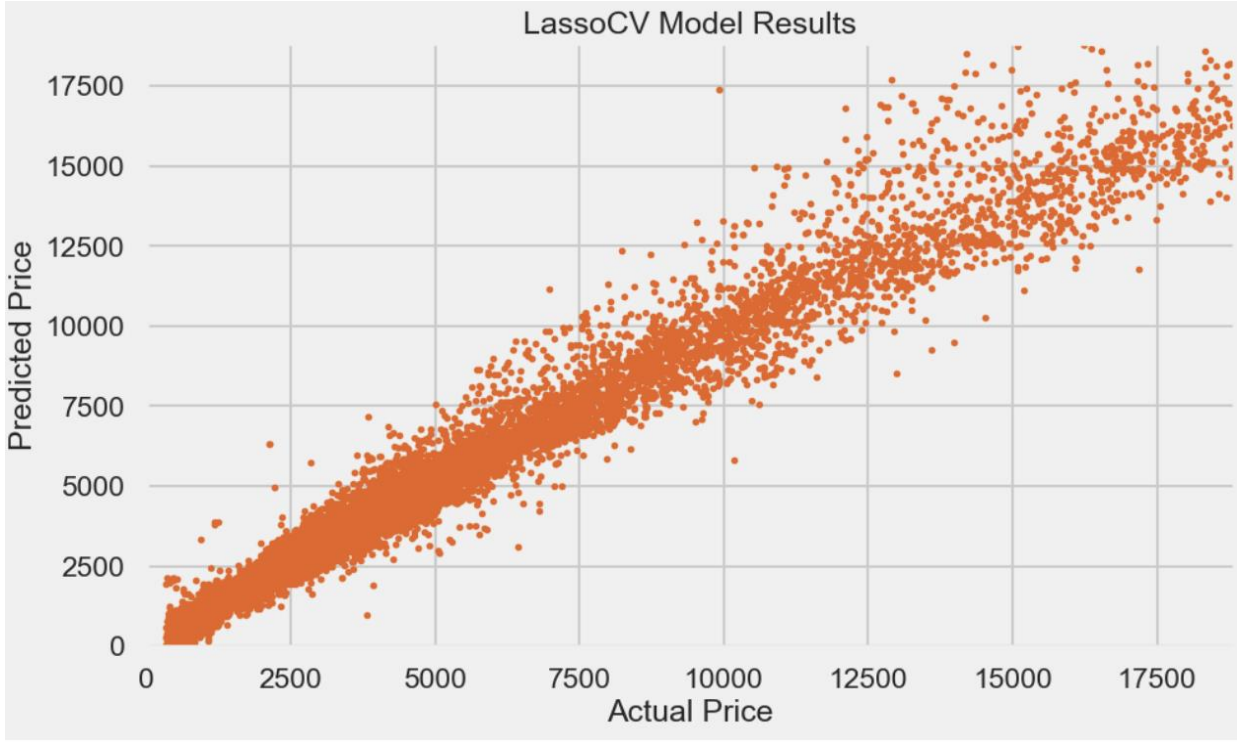
Vanilla Regression Model:

Model features and parameter:

- Model: LinearRegression()
- Polynomial features: Degree = 2
- StandardScaler

RMSE Score	R2 Score
1588.8188353340868	0.8369282829489186

Machine Learning Analysis 02



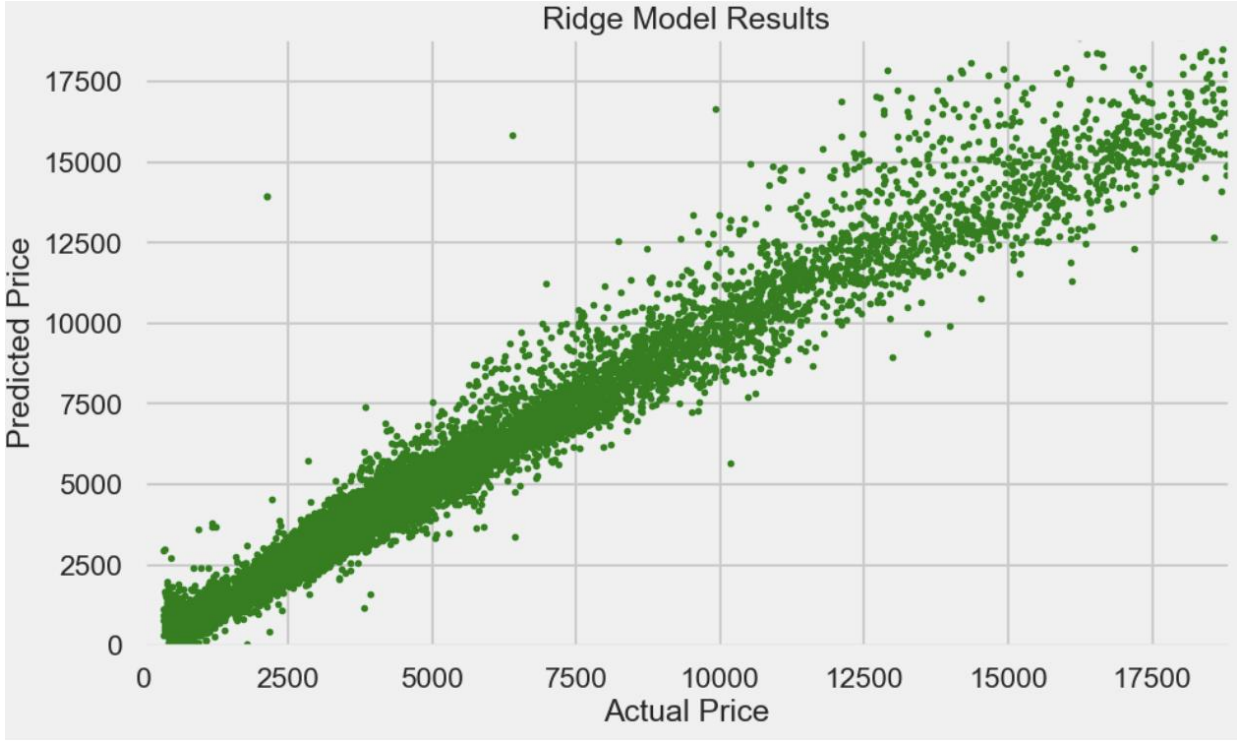
Lasso Regression Model:

Model features and parameter:

- Model: LassoCV()
- Polynomial features: Degree = 2
- StandardScaler
- Alpha = 5.3792
- max_iter = 10,000

RMSE Score	R2 Score
702.6860954874735	0.968102796835838

Machine Learning Analysis 03



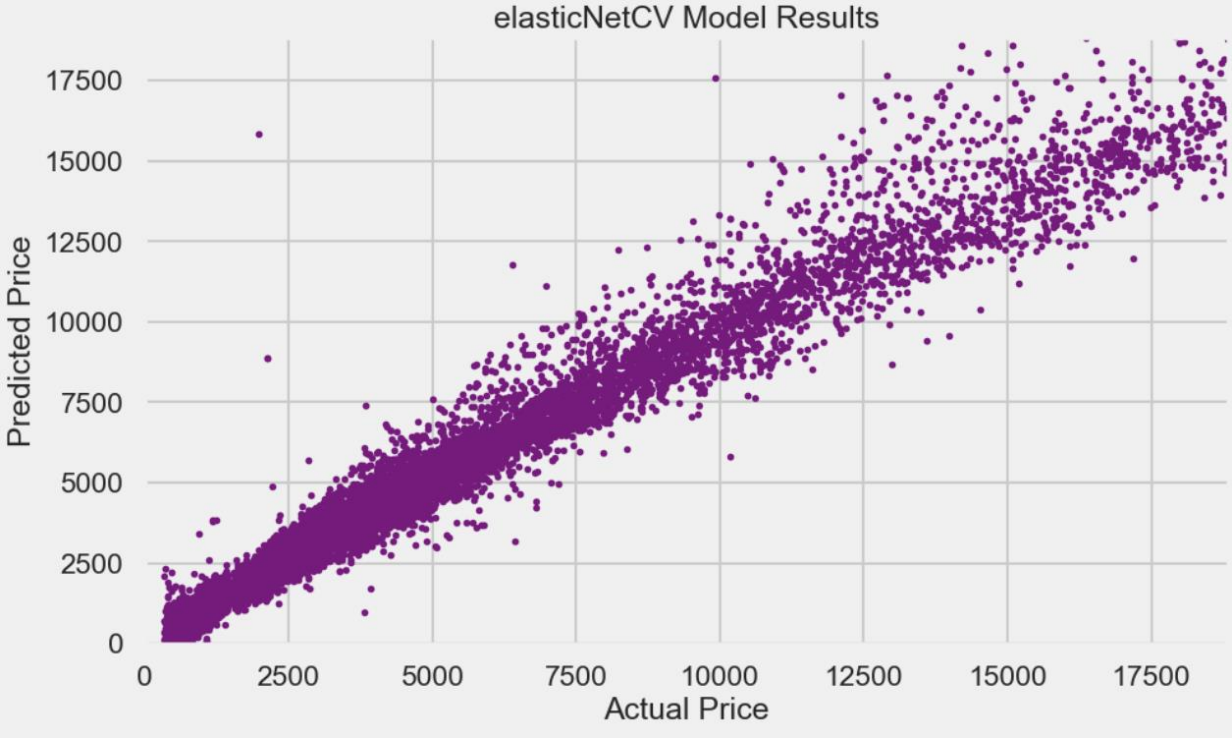
Ridge Regression Model:

Model features and parameter:

- Model: RidgeCV()
- Polynomial features: Degree = 2
- StandardScaler
- Alpha = 20.0
- max_iter = 10,000

RMSE Score	R2 Score
791.7394738480549	0.959505645454924

Machine Learning Analysis 04



ElasticNet Regression Model:

Model features and parameter:

- Model: ElasticNetCV()
- Polynomial features: Degree = 2
- StandardScaler
- Alpha = 0.12328
- L1 Ratio = 0.9
- max_iter = 10,000

RMSE Score	R2 Score
717.0028927778961	0.959505645454924

Machine Learning Analysis 05

	RMSE	R2
Linear	1588.818835	0.836928
LassoCV	702.686095	0.968103
RidgeCV	791.739474	0.959506
ElasticNetCV	717.002893	0.966790

As depicted in the data frame, all models yield excellent prediction results, with very minimal differences between them aside Linear which is a bit lower but with a high RMSE. However, the final model selection hinges on identifying the one with the highest R2 score outcome.

I have arranged the models in ascending order as follows:

1. Linear
2. RidgeCV
3. ElasticNetCV
4. LassoCV

Model Flaws and Strength and advanced steps

Model Flaws and Strength and Advanced steps:

In terms of simplicity, it's worth noting that vanilla linear regression delivered respectable predictive results, although not the absolute best. It stands out as the simplest and fastest model in terms of parameters. On the other hand, models like Lasso, Ridge, and ElasticNet yielded higher results in accuracy, but they introduced complexity and slowed down the training process, particularly when employing grid search to fine-tune parameters. This presents a tradeoff scenario: for larger datasets, these more complex models may offer superior performance, albeit with longer training times, whereas opting for the vanilla model might involve a slight accuracy sacrifice but significantly faster training.

In terms of the best model, LassoCV comes with the highest R2 score and we can't ignore that despite its training process is slow.

Advanced Steps:

To boost the model's effectiveness, we utilize regularization techniques such as L1 (Lasso) and L2 (Ridge) to manage overfitting and stabilize coefficients. Furthermore, the inclusion of k-fold cross-validation assists in evaluating the model's resilience and fine-tuning hyperparameters, ensuring its ability to generalize effectively across various data subsets. These methods together result in heightened model accuracy and dependability in regression applications.

Thank you

WISDOM IZUCHUKWU ADIKE