Section2 Project

Diamond의 가격 예측

AI_14_송일수

목차

- 문제 정의
- EDA
- Base model
- Linear model
- Boosting model
- Randomforest model
- PDP / SHAP

문제 정의

- Diamond의 price에 영향을 주는 특성은 무엇이 있을까?
- Diamond의 각 특성이 price와 어떤 관계를 이루고 있을까?

EDA

	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
...	...	...	...	...	...	...	...	...	...	...	...
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64
53940	53941	0.71	Premium	E	SI1	60.5	55.0	2756	5.79	5.74	3.49
53941	53942	0.71	Premium	F	SI1	59.8	62.0	2756	5.74	5.73	3.43
53942	53943	0.70	Very Good	E	VS2	60.5	59.0	2757	5.71	5.76	3.47

53943 rows × 11 columns

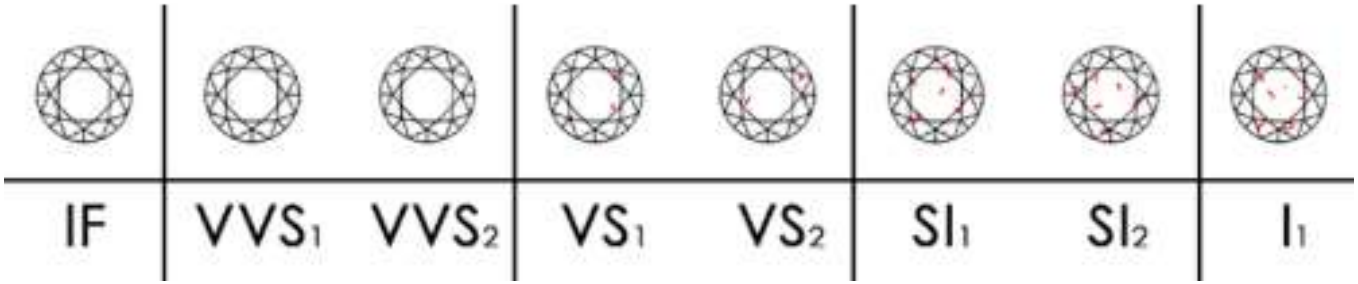
EDA

carat : 보석의 질량을 재는 단위, 1 carat = 200 mg

cut : cutting quality

color & clarity

GOOD←----------→BAD



EDA

x, y, z

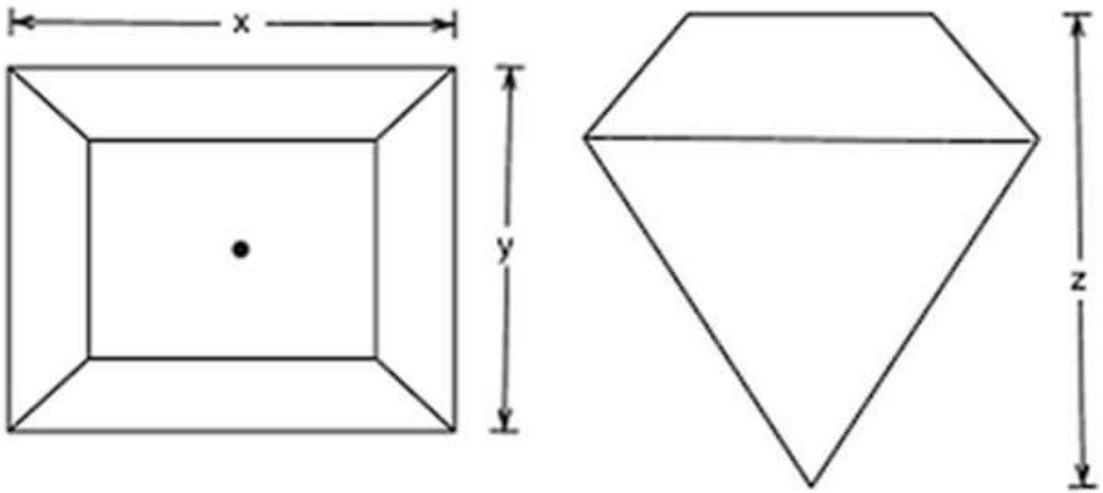


table : 다이아몬드에서 가장 넓은 폭에 대한 상부 면적 너비의 비

depth : z / {(x + y) / 2}

price : **target**

EDA

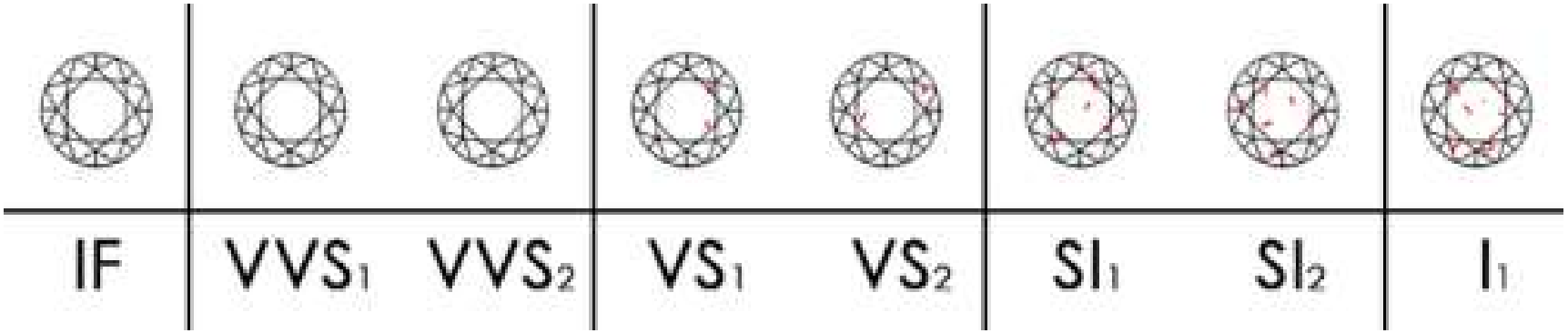
	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...	...
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64
53940	53941	0.71	Premium	E	SI1	60.5	55.0	2756	5.79	5.74	3.49
53941	53942	0.71	Premium	F	SI1	59.8	62.0	2756	5.74	5.73	3.43
53942	53943	0.70	Very Good	E	VS2	60.5	59.0	2757	5.71	5.76	3.47

53943 rows × 11 columns

EDA


D
E
F
G
H
I
J
Colorless
Near Colorless

GOOD←----------→BAD


IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1

EDA

```
df.duplicated().value_counts()
```

```
False    53794
True       149
dtype: int64
```

중복값의 개수

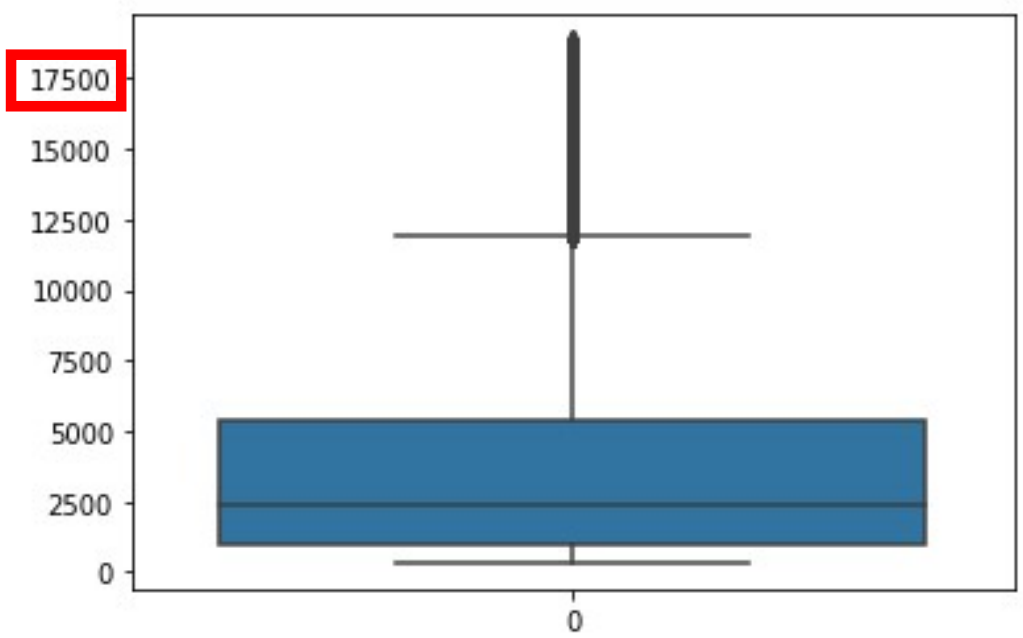
```
df=df.drop_duplicates(ignore_index=True)
```

```
df.duplicated().value_counts()
```

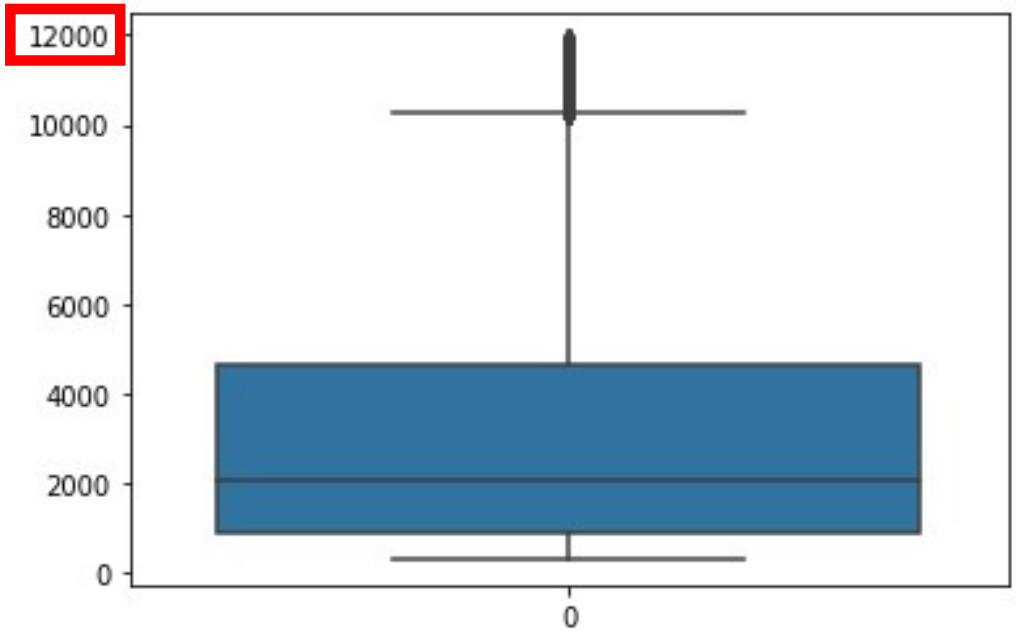
```
False    53794
dtype: int64
```

데이터 가공을 통해 사라진 모습

EDA

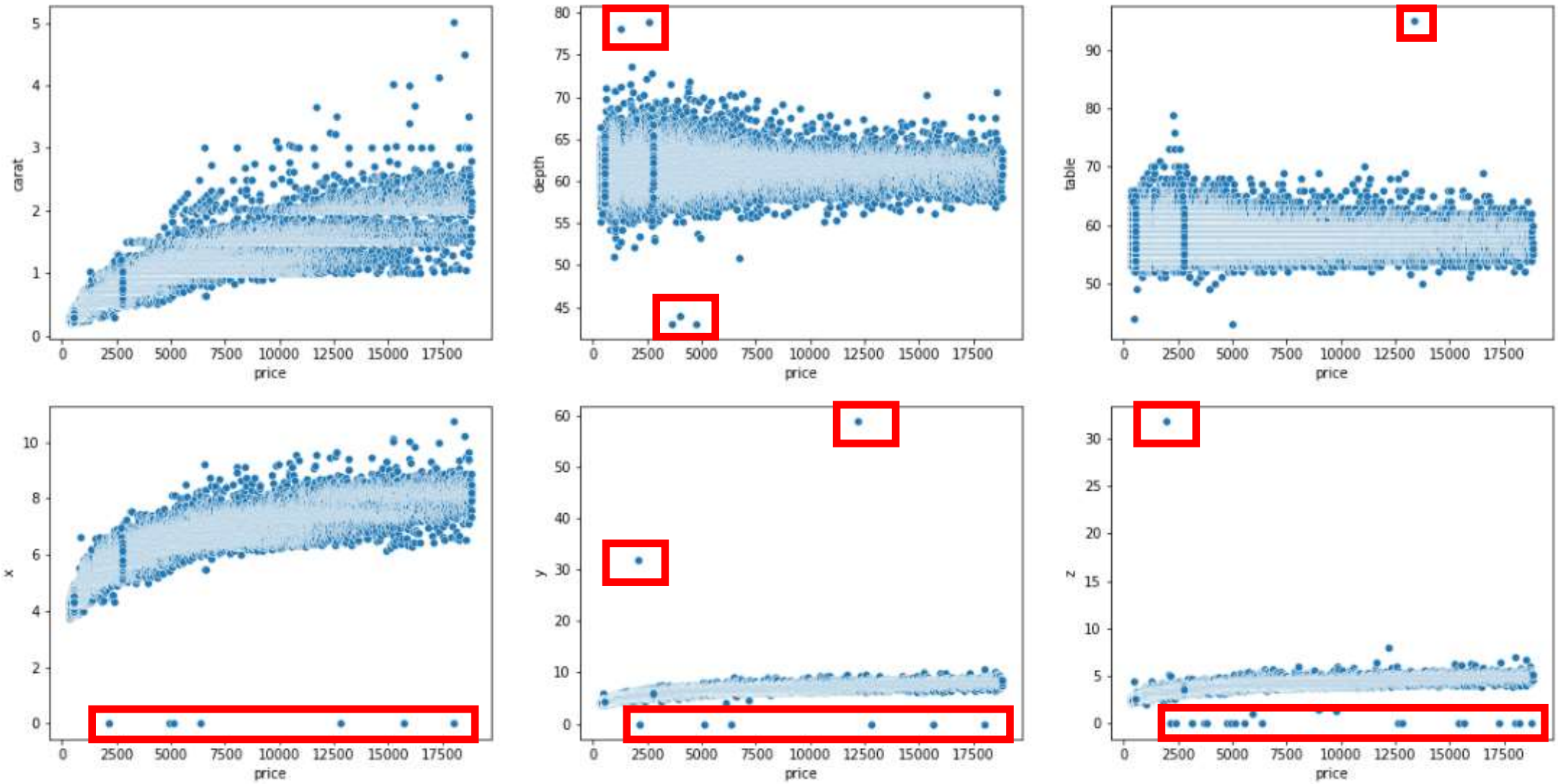


Outlier 제거 전

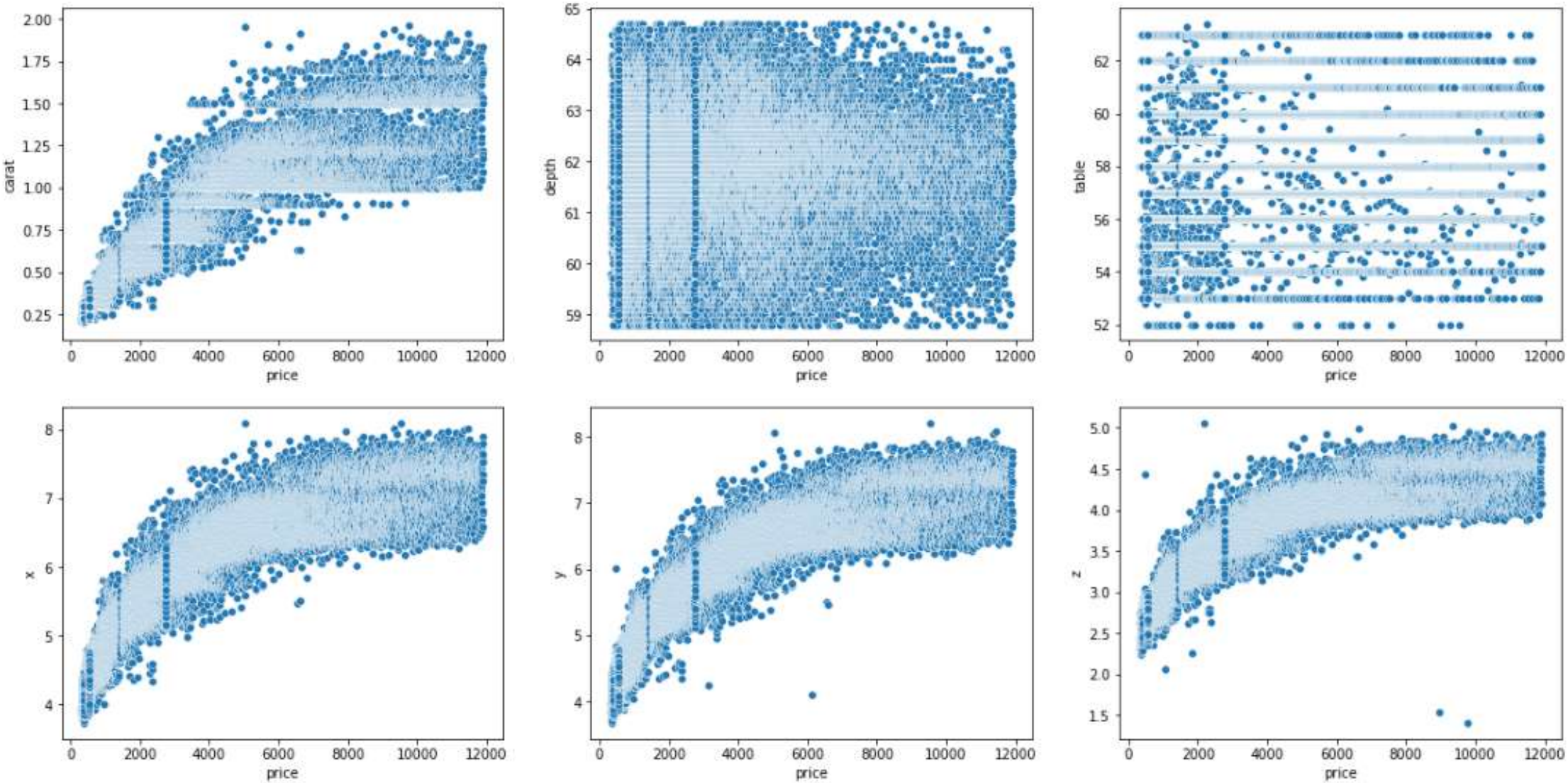


Outlier 제거 후

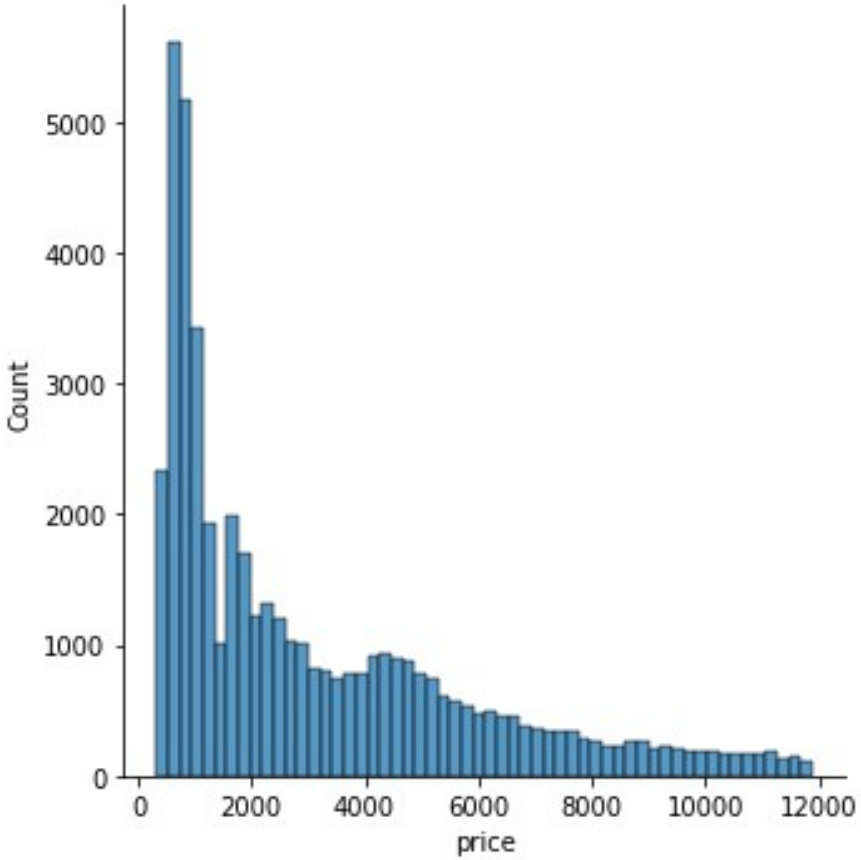
EDA



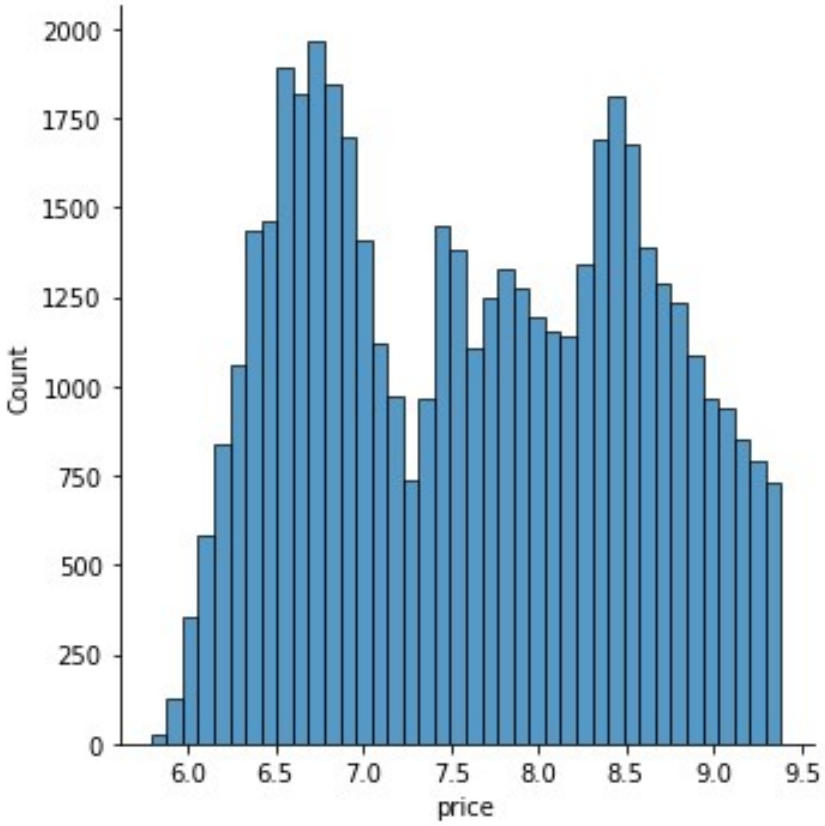
EDA



EDA



로그 변환 전



로그 변환 후

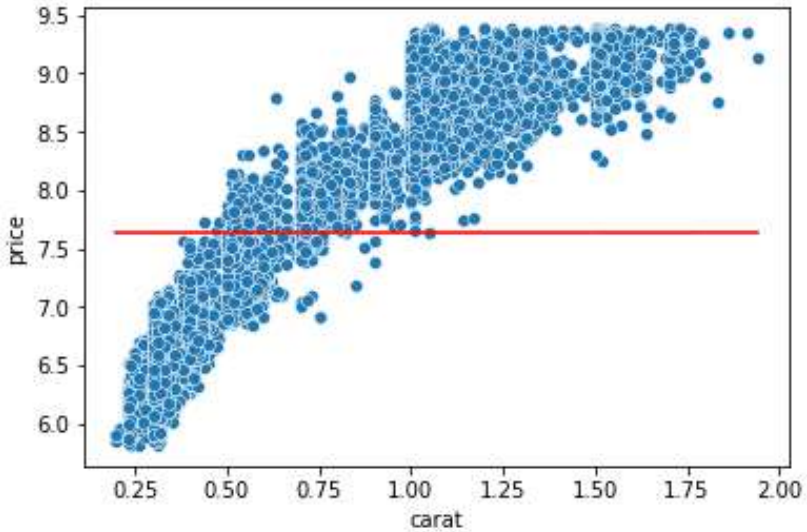
Base model

```
basepredict=y_test.mean()
y_base = [basepredict] * len(y_test)

print('mse : ',mean_squared_error(y_test,y_base))
print('mae : ',mean_absolute_error(y_test,y_base))
print('r2 : ',r2_score(y_test,y_base))
```

```
mse :  0.8646829258354961
mae :  0.8150777244989953
r2 :  0.0
```

```
sb.lineplot(x=X_test.carat, y=basepredict, color='red')
sb.scatterplot(x=X_test.carat, y=y_test);
```



*mse : mean square error,
평균 제곱 오차,
오차의 제곱에 대한 평균,
낮을수록 성능이 좋음을 나타냄

*mae : mean absolute error,
평균 절대 오차,
오차의 절댓값에 대한 평균,
낮을수록 성능이 좋음을 나타냄

*r2 : r-squared,
결정계수,
독립변수가 종속변수를 얼마만큼
설명하는지를 나타냄,
0 to 1의 값을 가지며, 1에 가까울수록
성능이 좋음을 나타냄

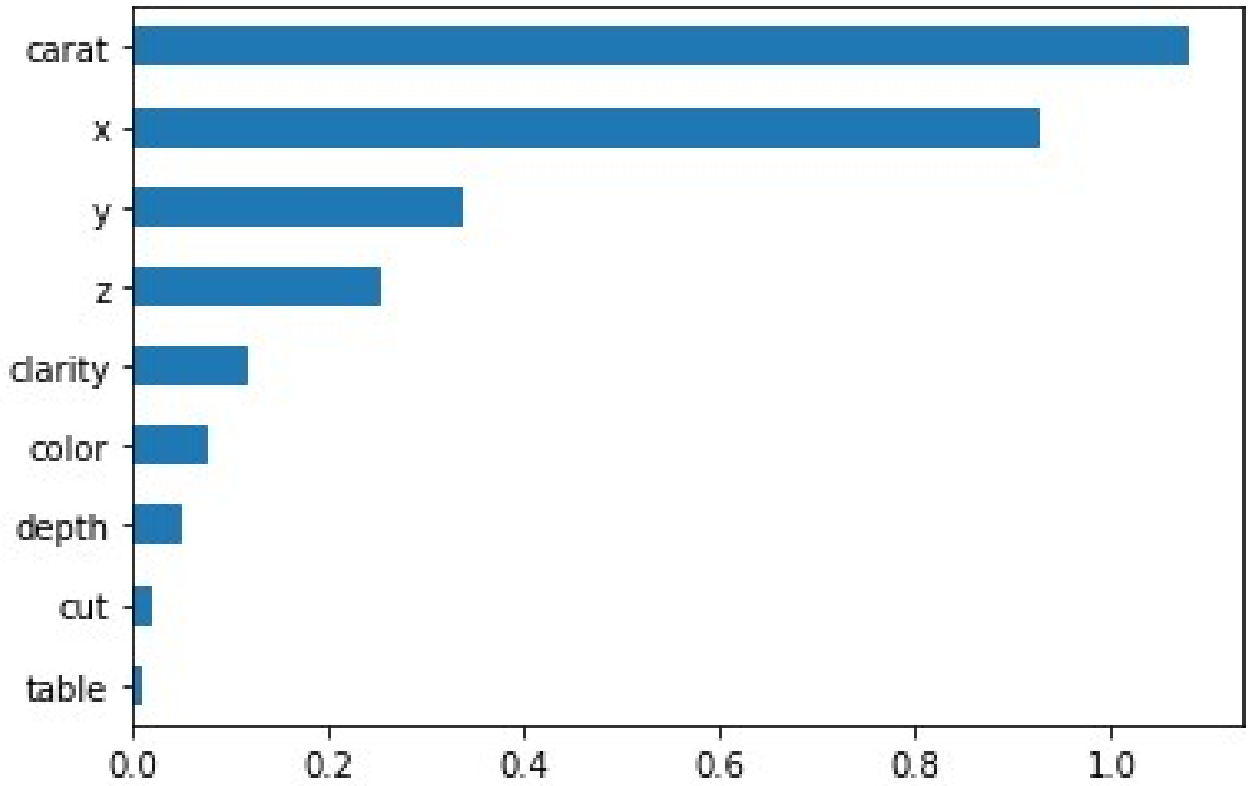
Linear model

검증 점수

```
mse :  0.0190767503693539 47
mae :  0.10773964477140249
r2 :  0.9775141721012658
```

평가 점수

```
mse :  0.019465418367734084
mae :  0.10985001173962149
r2 :  0.9772557200906326
```



* mse, mae는 낮을수록,
  R2는 1에 가까울 수록 좋은 모델

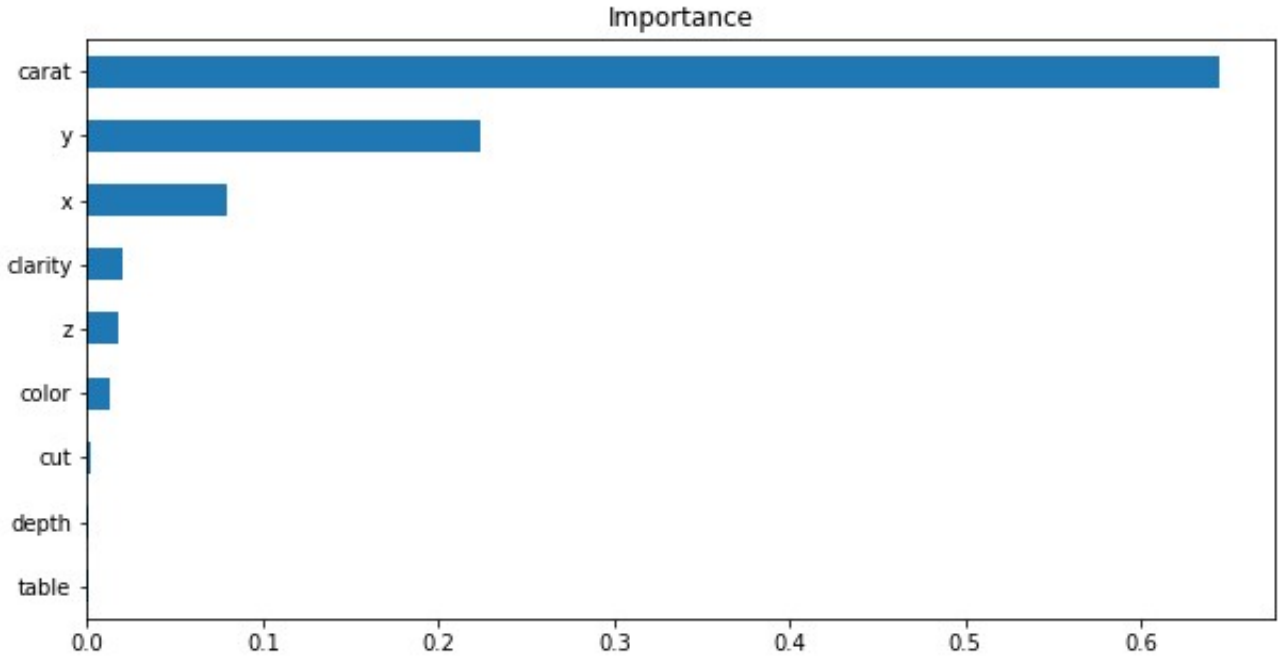
Boosting model

검증 점수

```
mse :  0.008574750765885968
mae :  0.070267116952764 04
r2 :  0.9898929132969106
```

평가 점수

```
mse :  0.008480081161474678
mae :  0.06983068327094831
r2 :  0.990091487583414
```



* mse, mae는 낮을수록,
  R2는 1에 가까울 수록 좋은 모델

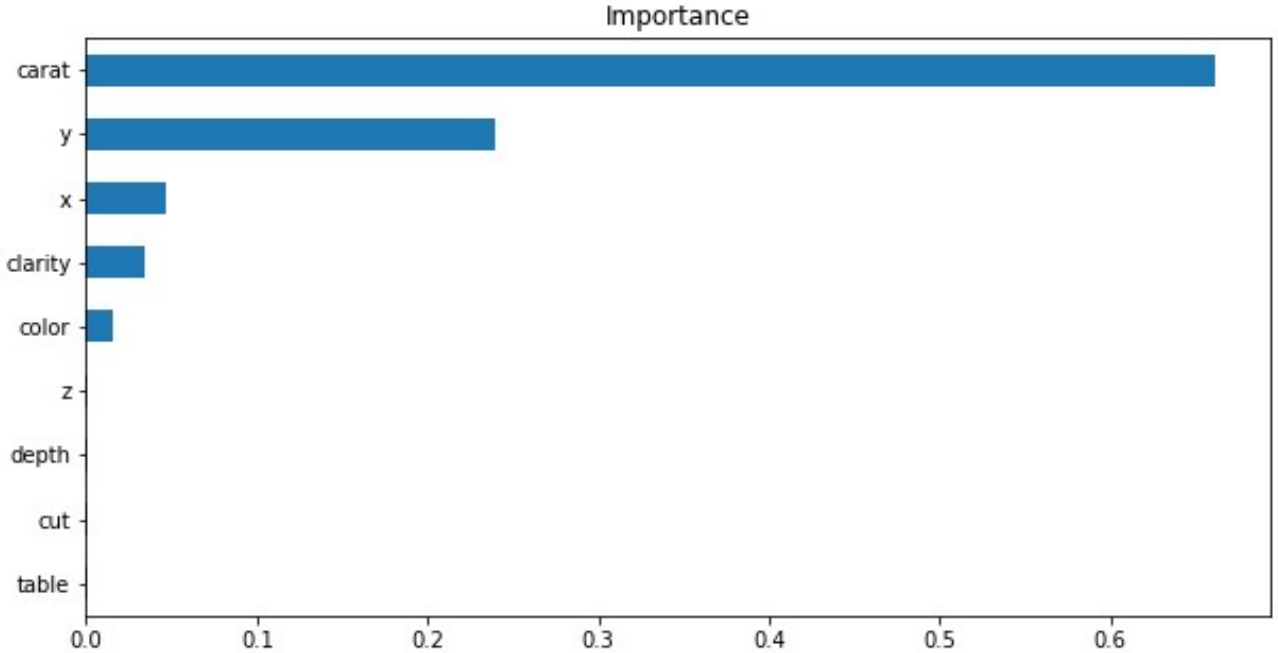
Randomforest model

검증 점수

mse : 0.007308447211523517
mae : 0.06120830860222889
r2 : 0.9913305366310332

평가 점수

mse : 0.007142484171319 76
mae : 0.06092930517398254
r2 : 0.9916411331359548



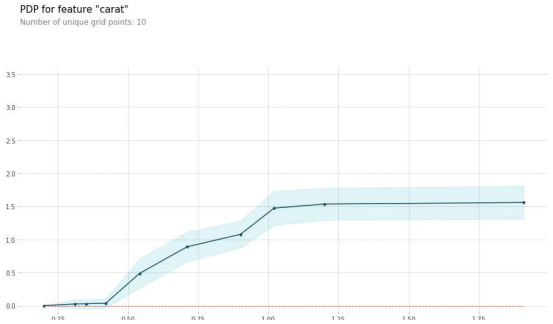
* mse, mae는 낮을수록,
R2는 1에 가까울 수록 좋은 모델

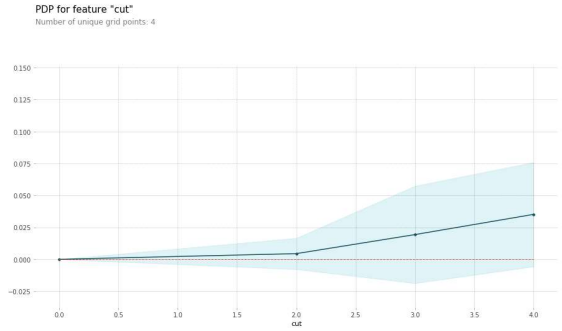
Summary

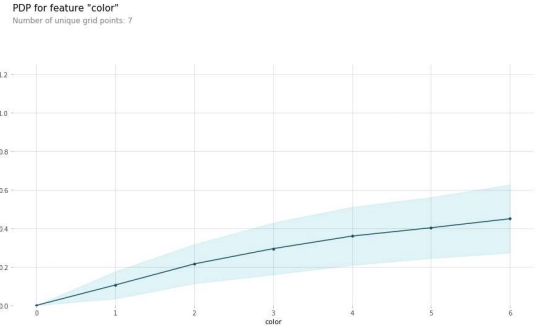
	mse ↓	mae ↓	r2 ↑
base	0.854480	0.808493	0.000000
linear	0.019602	0.109096	0.977060
boosting	0.008556	0.070091	0.989987
randomforest	0.007142	0.060929	0.991641

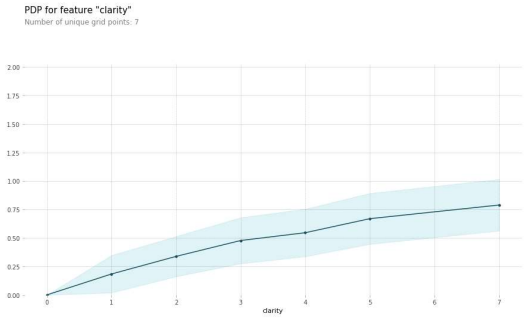
* mse, mae는 낮을수록,
  R2는 1에 가까울 수록 좋은 모델

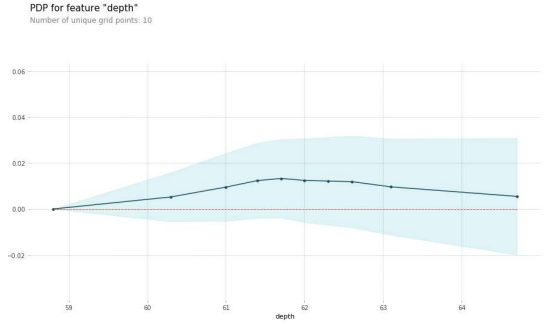
PDP

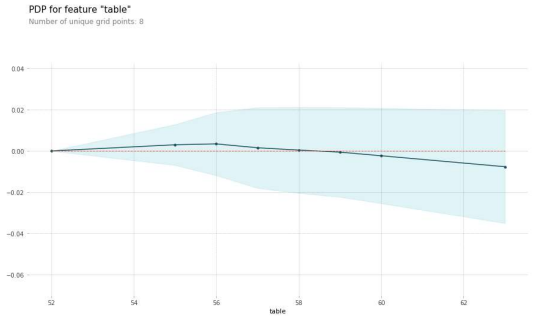

PDP for feature "carat"
Number of unique grid points: 10

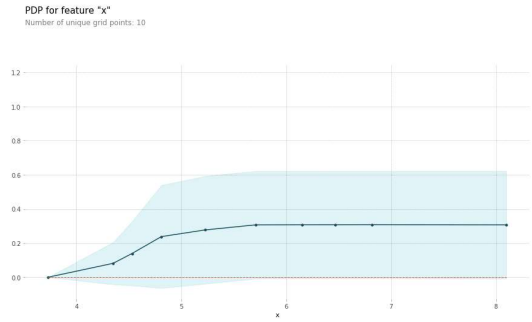

PDP for feature "cut"
Number of unique grid points: 4
cut

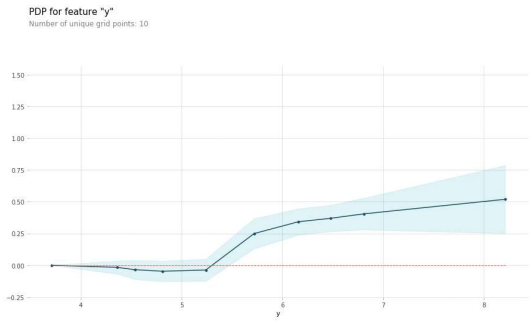

PDP for feature "color"
Number of unique grid points: 7
color

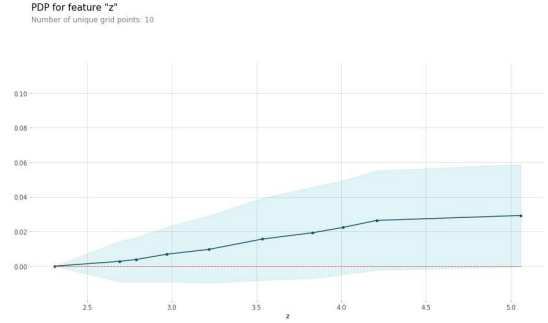

PDP for feature "clarity"
Number of unique grid points: 7
clarity


PDP for feature "depth"
Number of unique grid points: 10
depth

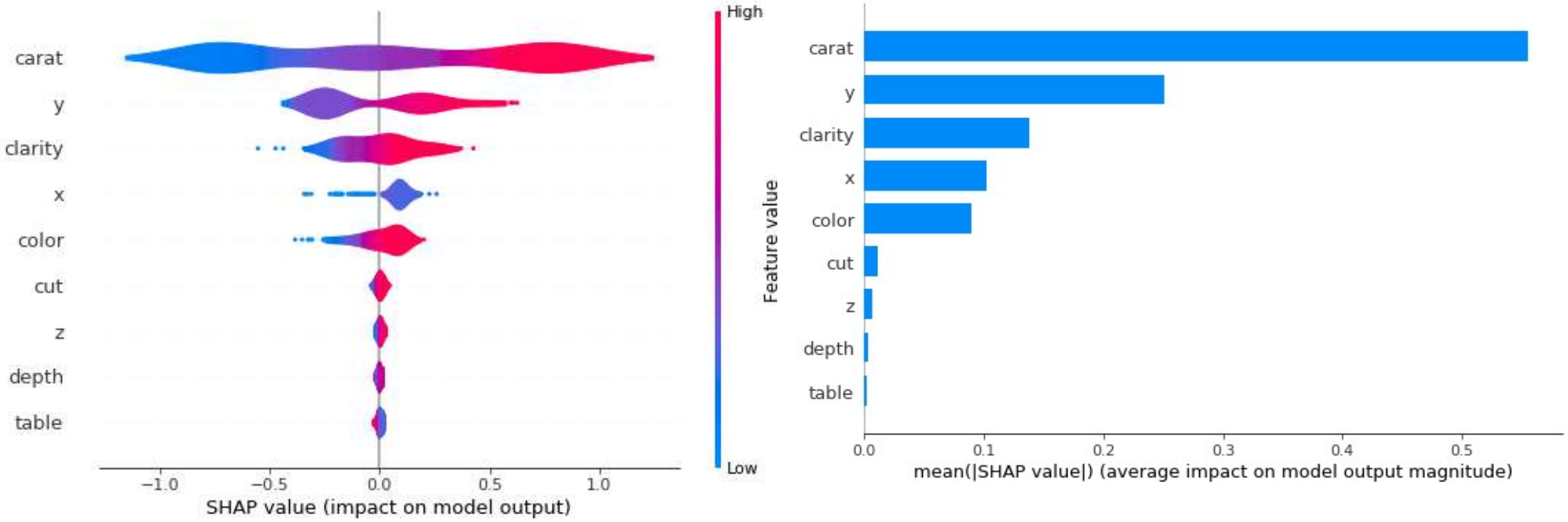

PDP for feature "table"
Number of unique grid points: 8
table


PDP for feature "x"
Number of unique grid points: 10
x


PDP for feature "y"
Number of unique grid points: 10
y


PDP for feature "z"
Number of unique grid points: 10
z

SHAP


carat
y
clarity
x
color
cut
z
depth
table
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low
carat
y
clarity
x
color
cut
z
depth
table
0.0
0.1
0.2
0.3
0.4
0.5
mean(|SHAP value|) (average impact on model output magnitude)