



The Lapidarist* Problem

BRAULIO PIÑA AMAROS

The lapidarist* problem

- ▶ The Prime Minister needs help validating the claim from Mr. Krenk about the diamonds that seem to have been stolen.
- ▶ Create a model to value the missing diamonds.

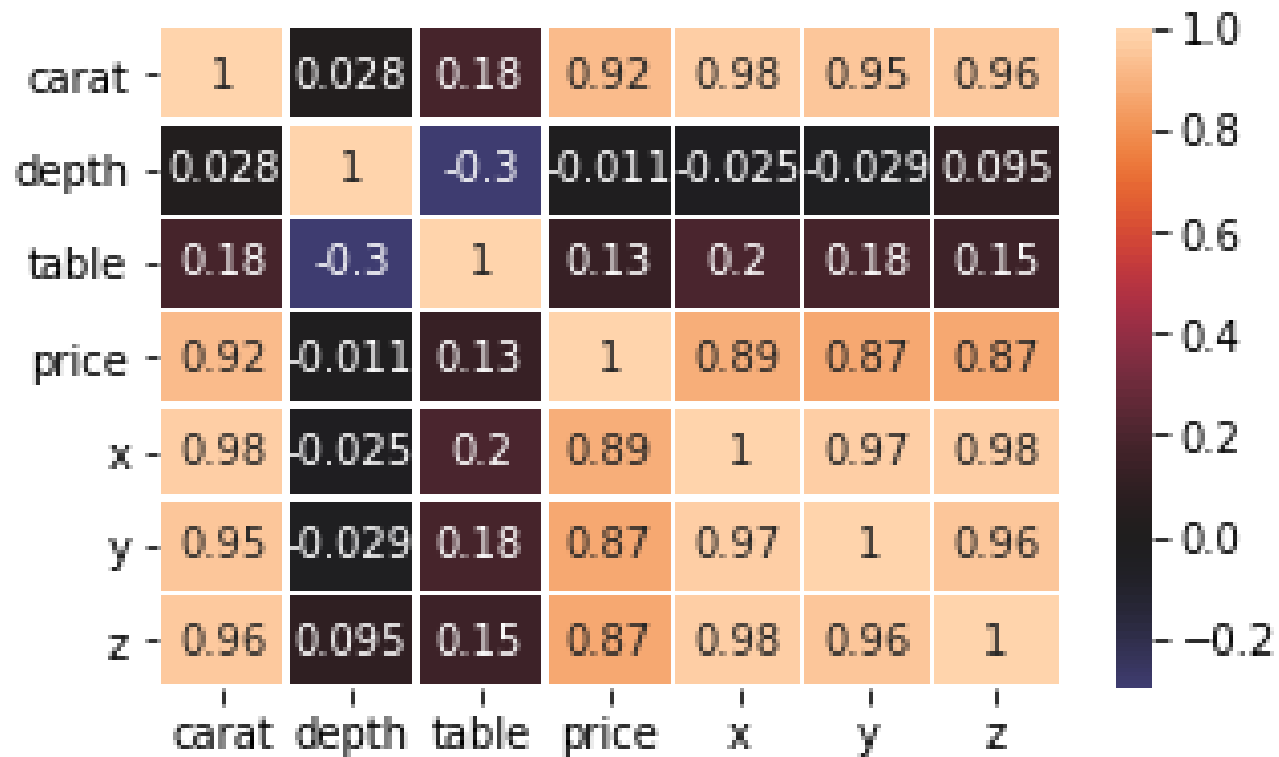
#	COLUMN	NON-NULL COUNT	DTYPE
0	Carat	53930 non-null	Float64
1	Cut	53930 non-null	Object
2	Color	53930 non-null	Object
3	Clarity	53930 non-null	Object
4	Depth	53930 non-null	Float64
5	Table	53930 non-null	Float64
6	Price	53930 non-null	Int64
7	X	53930 non-null	Float64
8	Y	53930 non-null	Float64
9	z	53930 non-null	float64

Diamonds data Preview

Measures of Central Tendency

	carat	depth	table	price	x	y	z
count	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000
mean	0.797976	61.749325	57.457328	3933.054942	5.731236	5.734601	3.538776
std	0.474035	1.432711	2.234578	3989.628569	1.121807	1.142184	0.705729
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	5325.000000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

- X, Y and Z are Cartesian measures that cannot include values equal to zero.
- We take them out and now we will work with 53910 observations

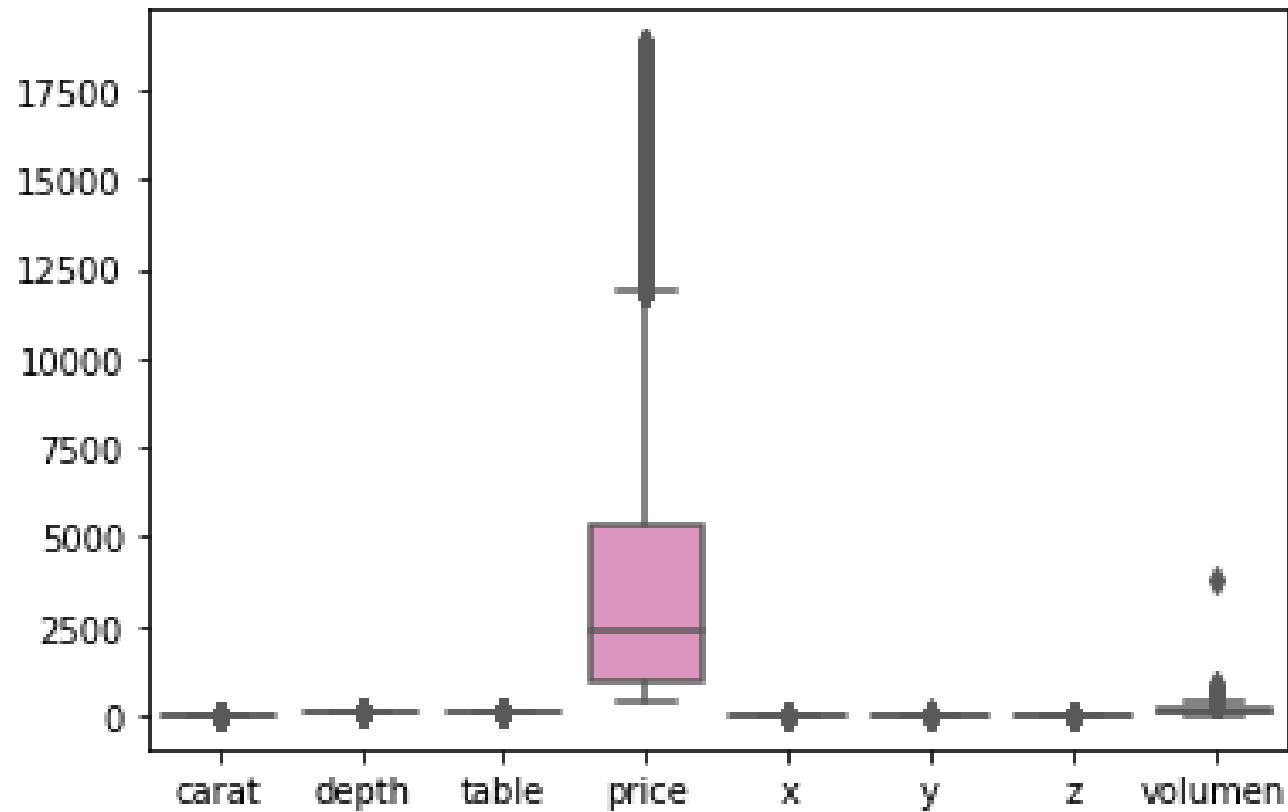


Correlation

We can see that x,y and z have a big correlation between each other

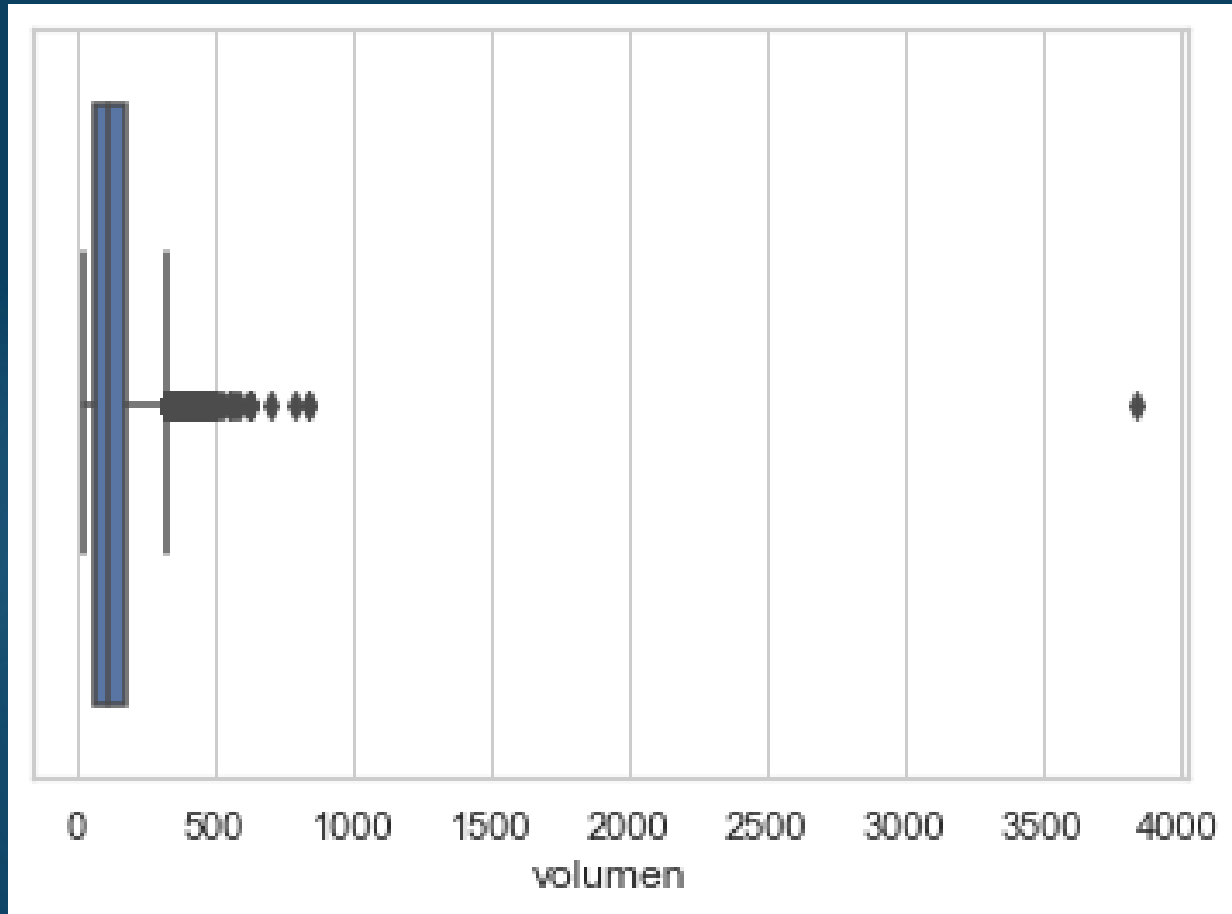
Carat is the most correlated variable to the Price

We have a problem of multicollinearity between x , y and z . So from $x*y*x$, we will create a new variable, the “volume” of the diamonds.



Qualitative
distribution of
the numeric
variables and
the scale in
which they are.

Distribution of the variable “volume”

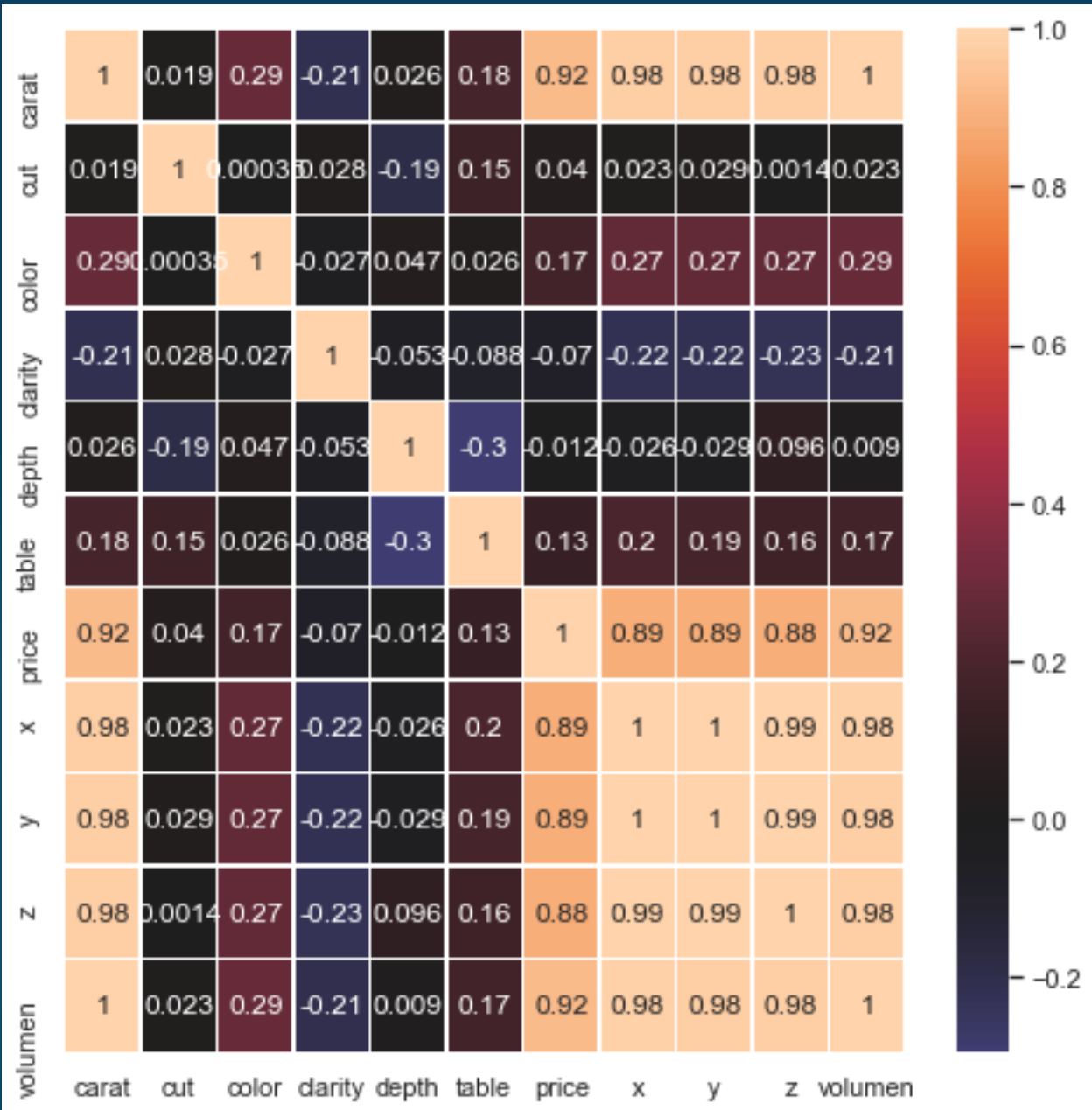


We remove:

17 observations that have volume greater than 500

1 observation that have volume greater than 1000

We have now 53893 observations



Correlation of the all variables

Most correlated variables with the Price are Carat and Volume

The bigger the diamond, higher the price

Preparing data for the model

- ▶ Divide variables in explanatory and objective.
- ▶ Divide variables: 80% training and 20% test
- ▶ Standardize variables

Metrics for evaluation:

- ▶ R-Squared: explanation percentage that the explanatory variables give to the output variable
- ▶ Mean Absolute Error (MAE): the mean absolute distance between X and Y
- ▶ Mean Square Deviation (RMSE): the squared root of the squared average errors. The effect of each error is proportional to the size of the quadratic error.

Models

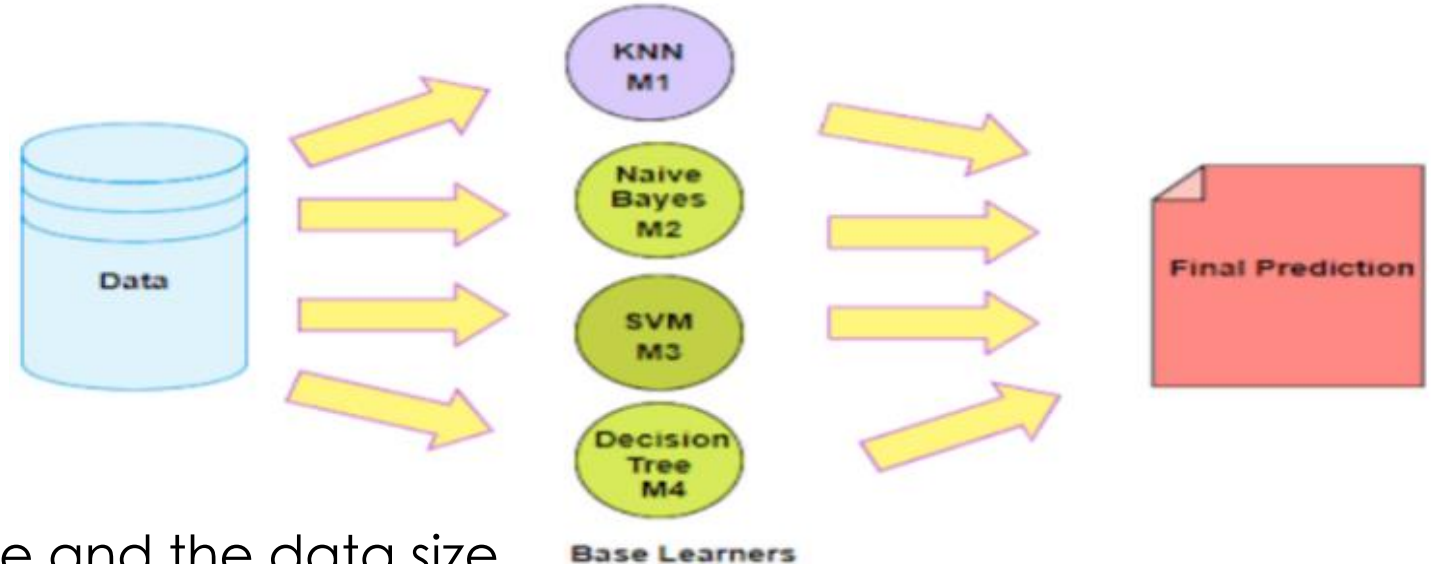
Algorithms of the ensembling type.
Two boosting and one bagging

Ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model

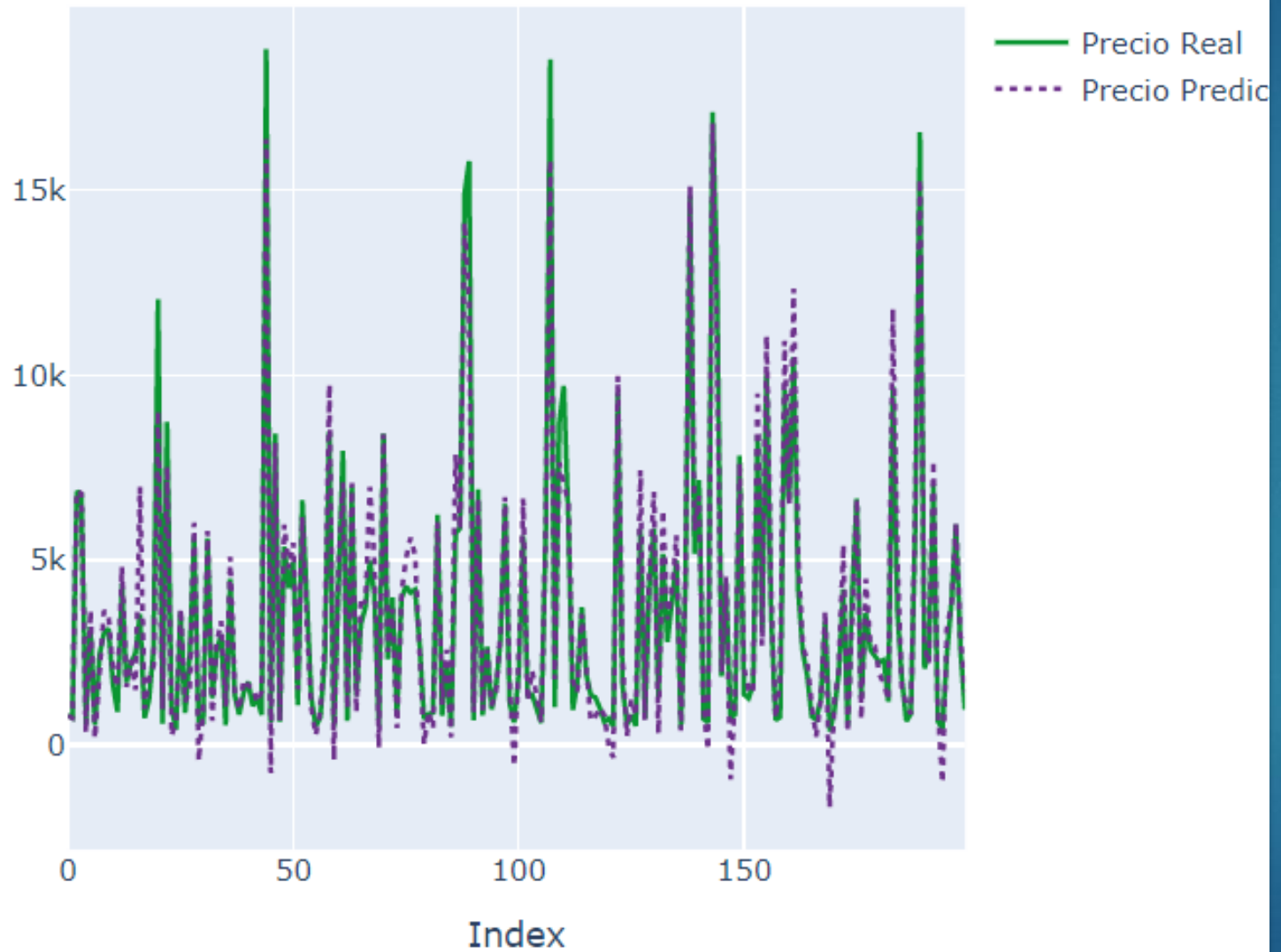
- ▶ GradientBoosting
- ▶ RandomForest
- ▶ XGBOOST

Advantages:

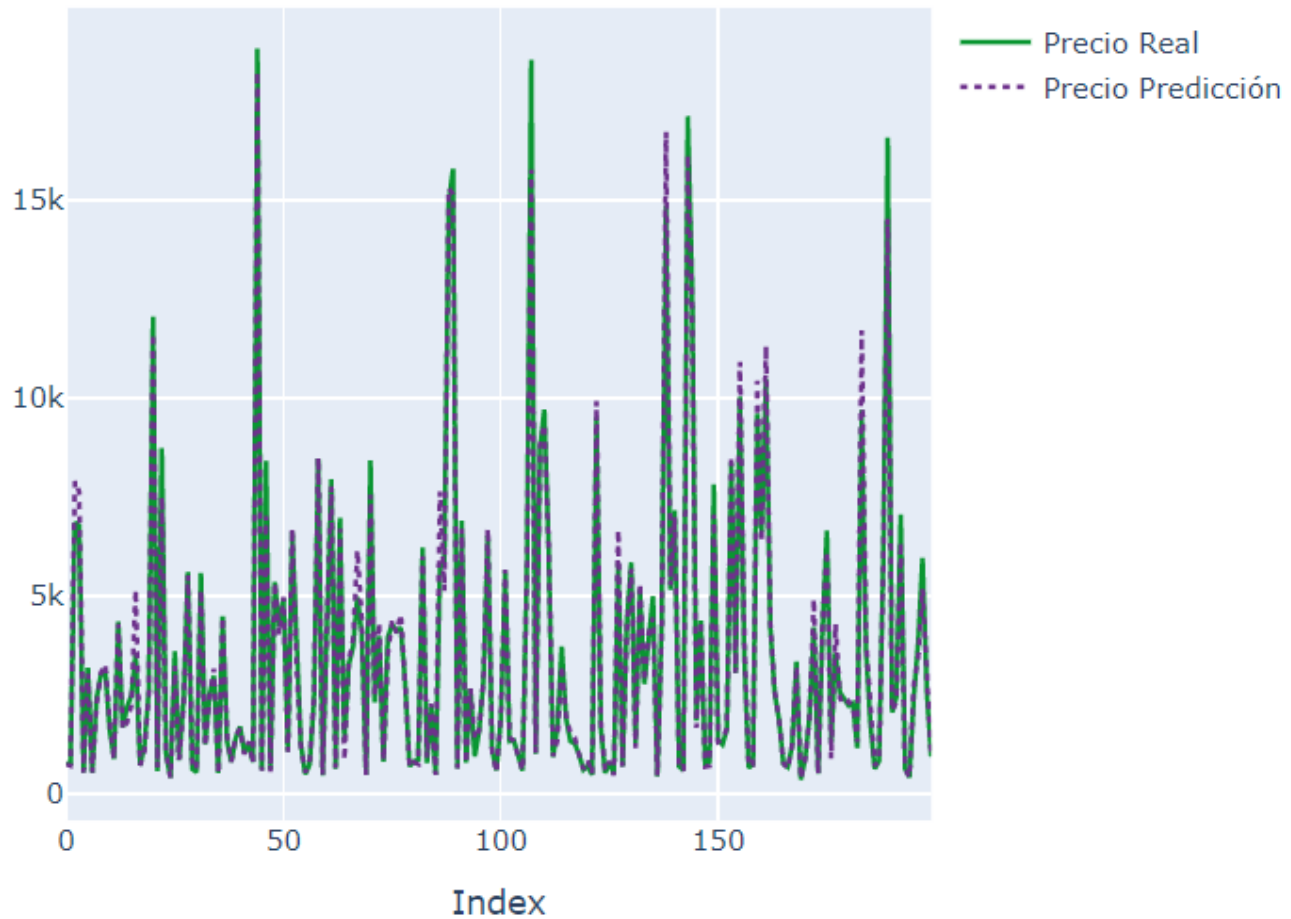
- Great predictive power
- Efficient because of the RAAM use and the data size
- Custom evaluation metric
- Previous experience



Price

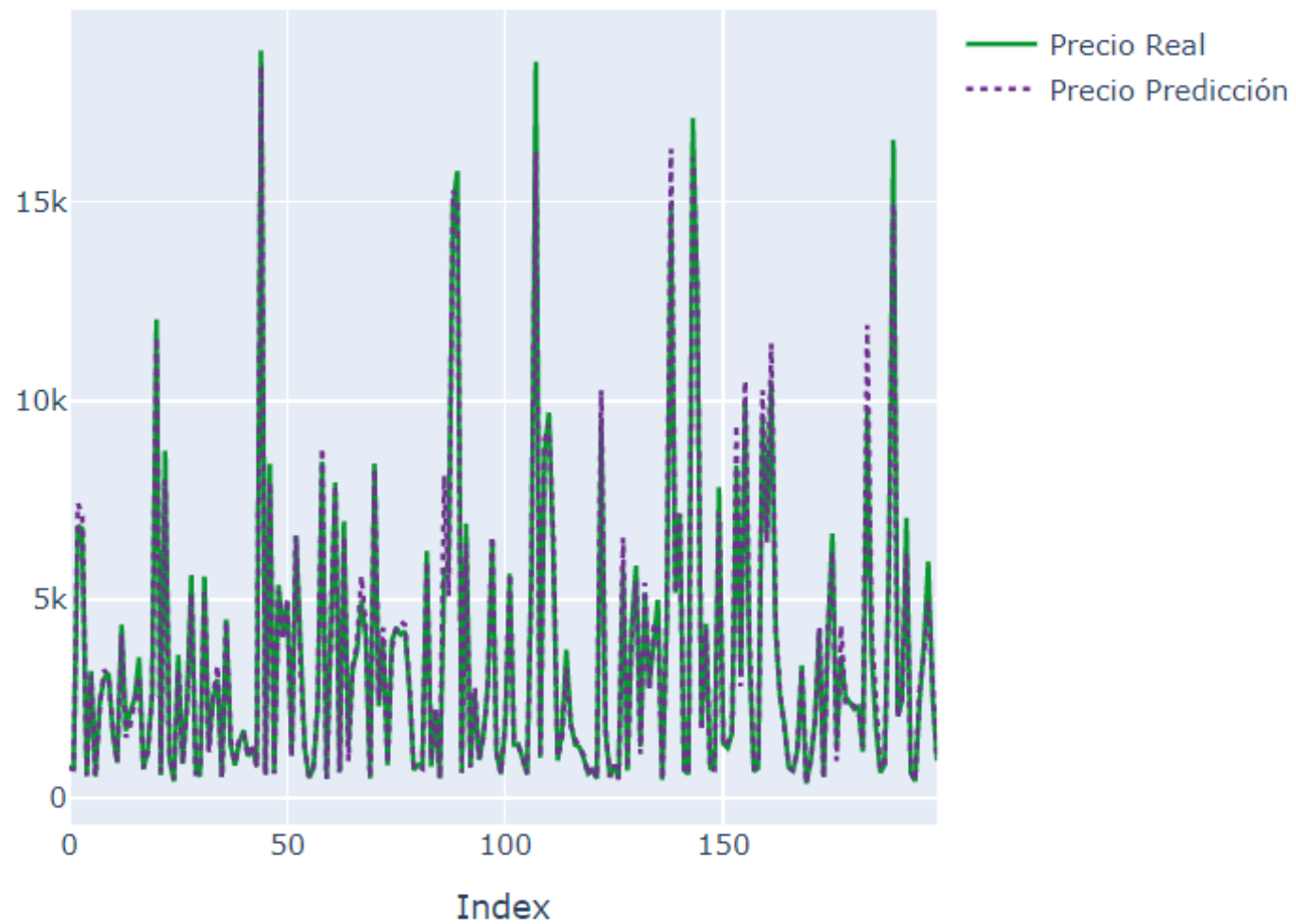


Forecasts versus
Real
GradientBoosting

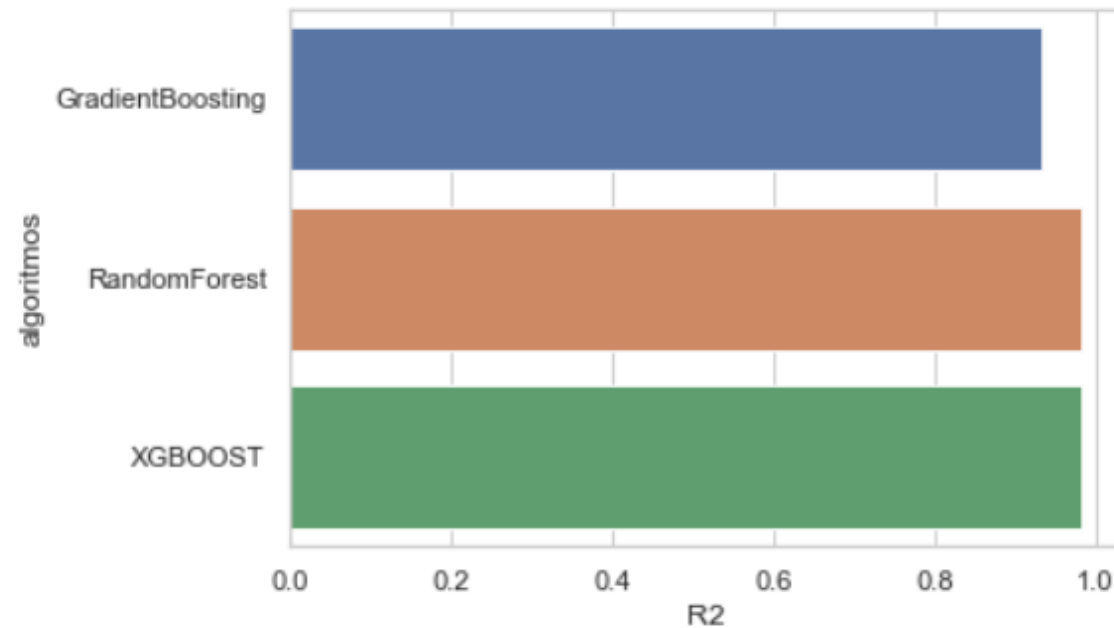


Forecast versus
Real
RandomForest

Price




Forecast versus
Real XGBOOST



	algoritmos	R2	MAE	RMSE
2	XGBOOST	0.981830	276.152443	541.437208
1	RandomForest	0.981536	270.576969	545.806459
0	GradientBoosting	0.933261	668.687769	1037.673156

Metrics Comparison

- 
- ▶ XGBOOST and RandomForest contain more similar values between the forecast and the real data
 - ▶ We choose XGBOOST model because its RMSE was the least and its R2 was the higher.
 - ▶ XGBOOST is also computationally more efficient than RandomForest
 - ▶ XGBOOST does not contains bias towards the categories with the greatest levels

Conclusion

- ▶ The model was based in creating quality features
- ▶ The problem was attack from the presence of outliers
- ▶ A new variable was created to solve the multicollinearity between x,y and z. That new variable "Volume" was useful for a better capture of the variability of the data set.
- ▶ Carat and volume were the characteristics which influence the most the Price of the diamonds.
- ▶ The price of the diamond was also determined in great measure from its size.
- ▶ Analysis should continue with a tuning of the hyperparameters so we can get a better development of the algorithms