The lapidarist problem:
Price prediction.

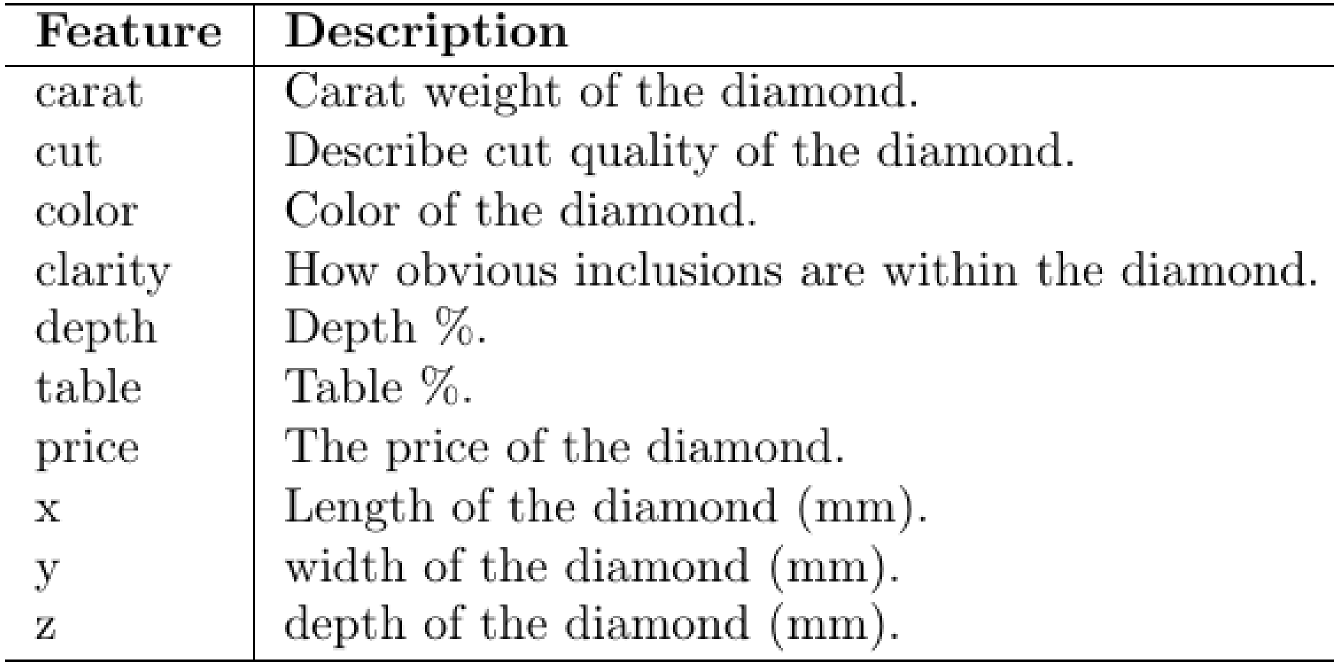
Roberto Carlos Vazquez Nava

Diamonds

Challenge

Create a model to value the stolen diamonds.

Database description



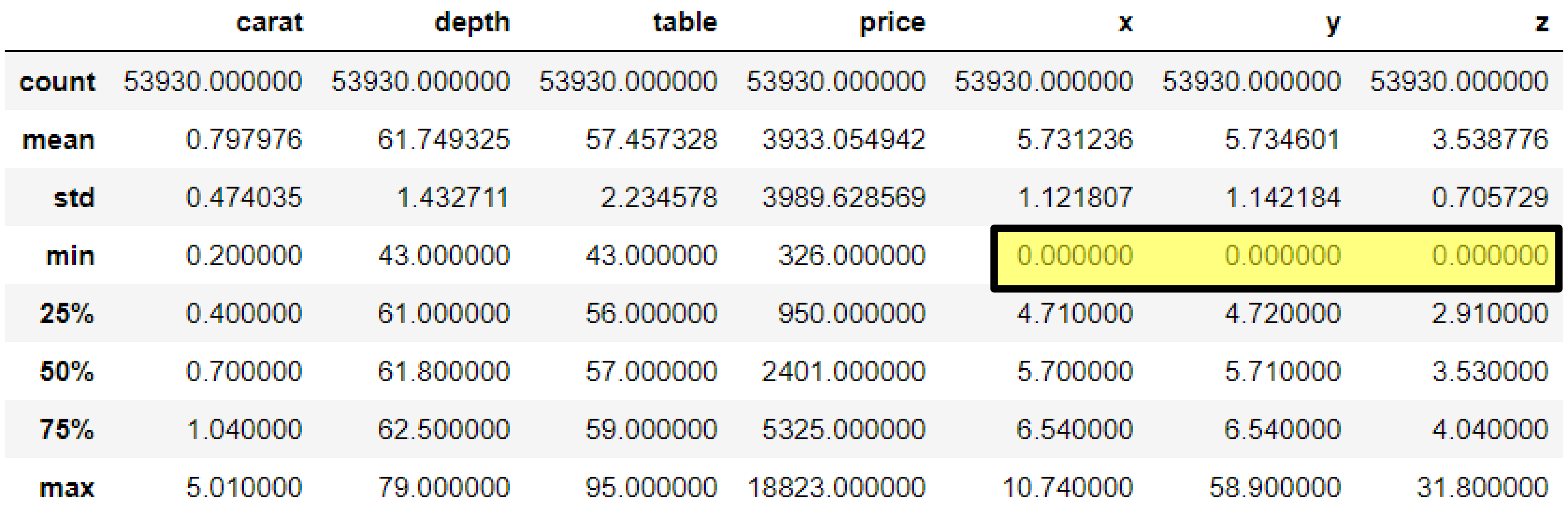
Feature	Description
carat	Carat weight of the diamond.
cut	Describe cut quality of the diamond.
color	Color of the diamond.
clarity	How obvious inclusions are within the diamond.
depth	Depth %.
table	Table %.
price	The price of the diamond.
x	Length of the diamond (mm).
y	width of the diamond (mm).
z	depth of the diamond (mm).

The database has 53,910 instances and 9 features. The target is the price.

Data Profiling

Database description

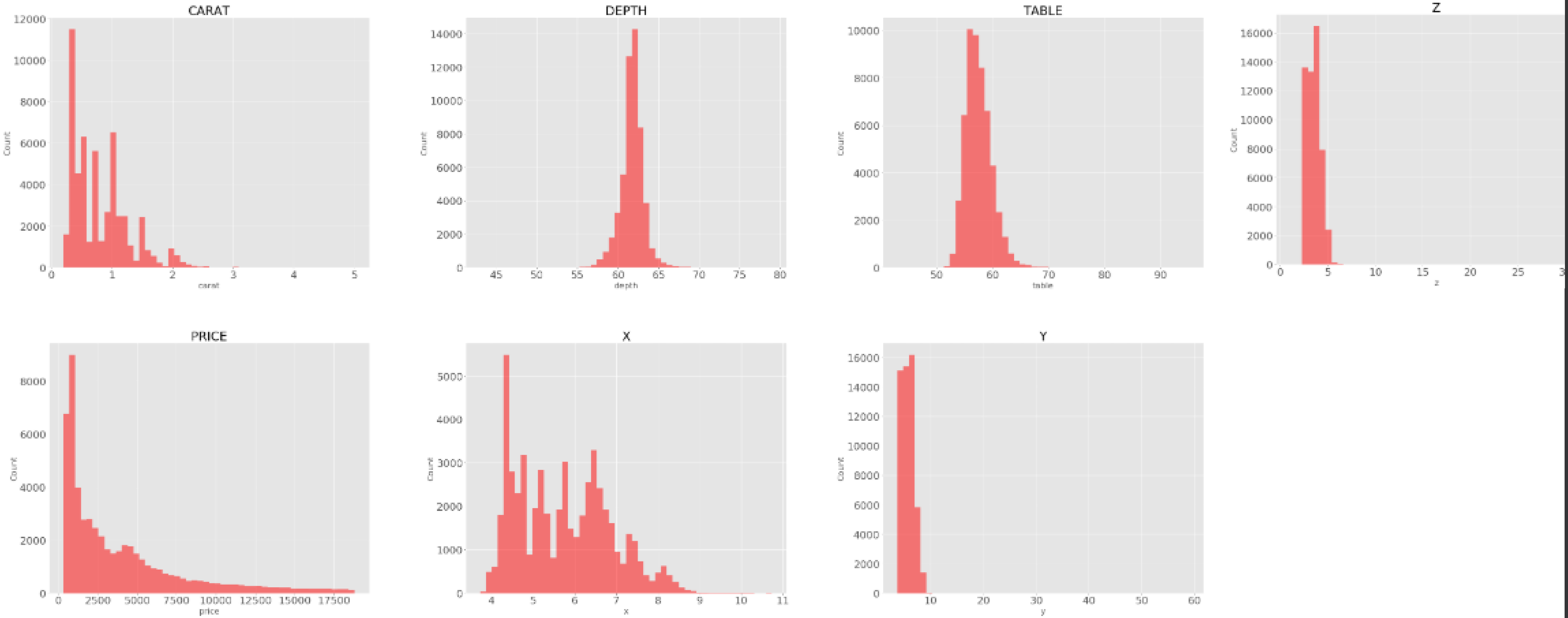
The database does not have missing values. However, it presents an incoherence with x, y and z features. Those instances with value equal 0 are removed.



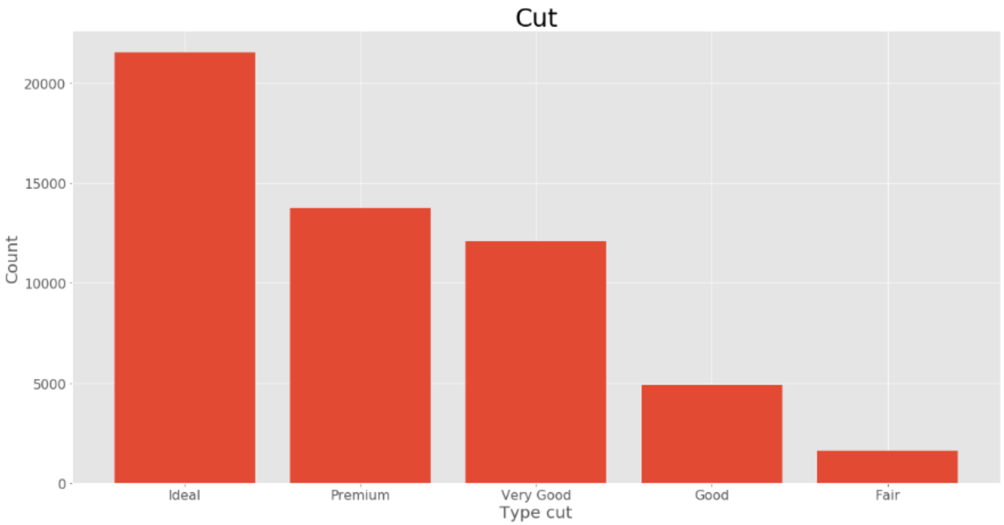
	carat	depth	table	price	x	y	z
count	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000	53930.000000
mean	0.797976	61.749325	57.457328	3933.054942	5.731236	5.734601	3.538776
std	0.474035	1.432711	2.234578	3989.628569	1.121807	1.142184	0.705729
min	0.200000	43.000000	43.000000	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

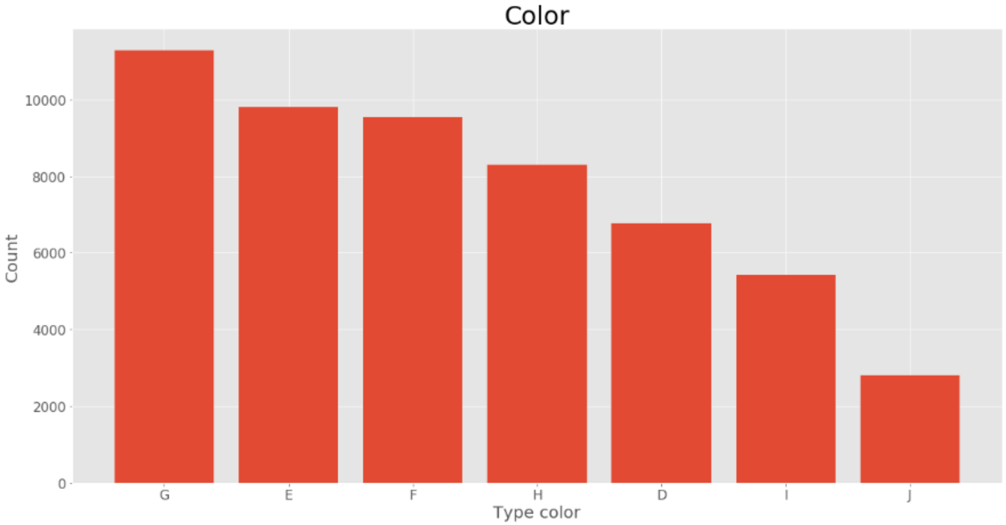
Data analysis

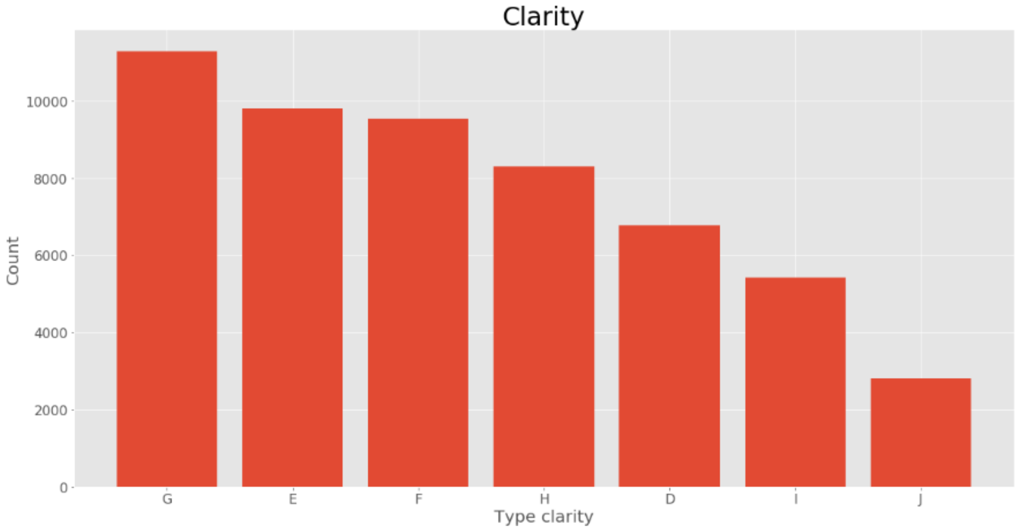
Histograms of numeric features



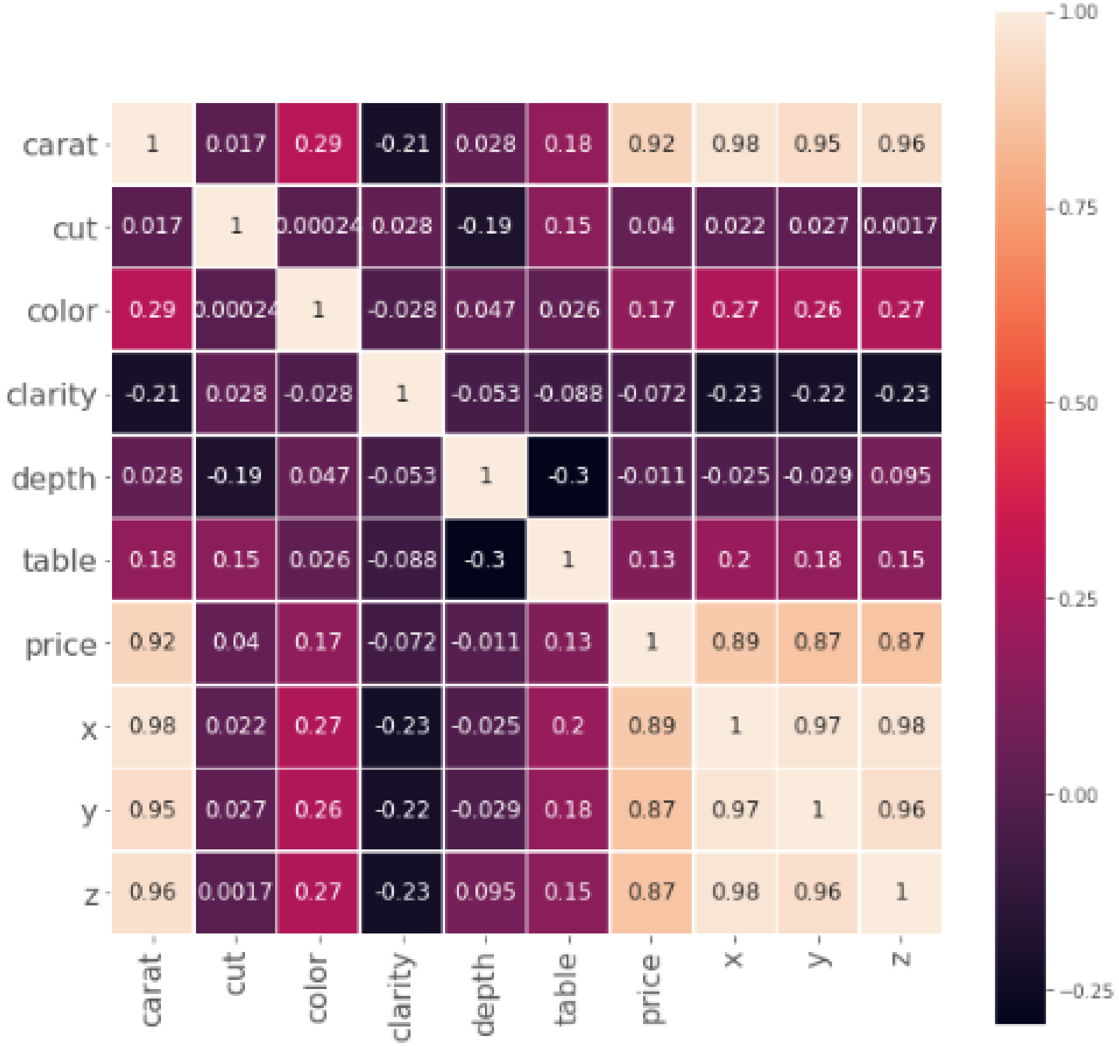
Categorical features







Correlation plot



	carat	cut	color	clarity	depth	table	price	x	y	z
carat	1	0.017	0.29	-0.21	0.028	0.18	0.92	0.98	0.95	0.96
cut	0.017	1	0.00024	0.028	-0.19	0.15	0.04	0.022	0.027	0.0017
color	0.29	0.00024	1	-0.028	0.047	0.026	0.17	0.27	0.26	0.27
clarity	-0.21	0.028	-0.028	1	-0.053	-0.088	-0.072	-0.23	-0.22	-0.23
depth	0.028	-0.19	0.047	-0.053	1	-0.3	-0.011	-0.025	-0.029	0.095
table	0.18	0.15	0.026	-0.088	-0.3	1	0.13	0.2	0.18	0.15
price	0.92	0.04	0.17	-0.072	-0.011	0.13	1	0.89	0.87	0.87
x	0.98	0.022	0.27	-0.23	-0.025	0.2	0.89	1	0.97	0.98
y	0.95	0.027	0.26	-0.22	-0.029	0.18	0.87	0.97	1	0.96
z	0.96	0.0017	0.27	-0.23	0.095	0.15	0.87	0.98	0.96	1

Hypotheses and modeling

Hypotheses

Based on the correlation plot, the dimensions of the diamonds are strongly correlated with each other and with the price. In the same way, it is carat too. Therefore, the cut, depth, color, clarity and table features with any of the three remaining ones are enough to make the price prediction. It must be considered that the mentioned features do not have a strong correlation with the target (price). Experimentation will decide it.

Models

- In my experience, the Linear Regression algorithm has given good results in the problems I have faced. Besides, it is an algorithm widely used in the literature when a regression problem is being solved. So, I will use it with this database.

- On the other hand, the SVM algorithm has a regression format (SVR). However, the main disadvantage is the time it takes to train it. Also, if the data is not scaled, its performance decreases. Although if the parameters are chosen correctly, good results are obtained.

- Also, there are algorithms based on tree ensembles. Examples of them are Random Forest Regressor, Gradient Boosting Regressor, and Extra Trees Regressor. The advantages of these algorithms are the low computation time to be trained in comparison to the SVR. Besides, train a model with a small number of samples and have good results. One of the disadvantages is that they can reach overfitting.

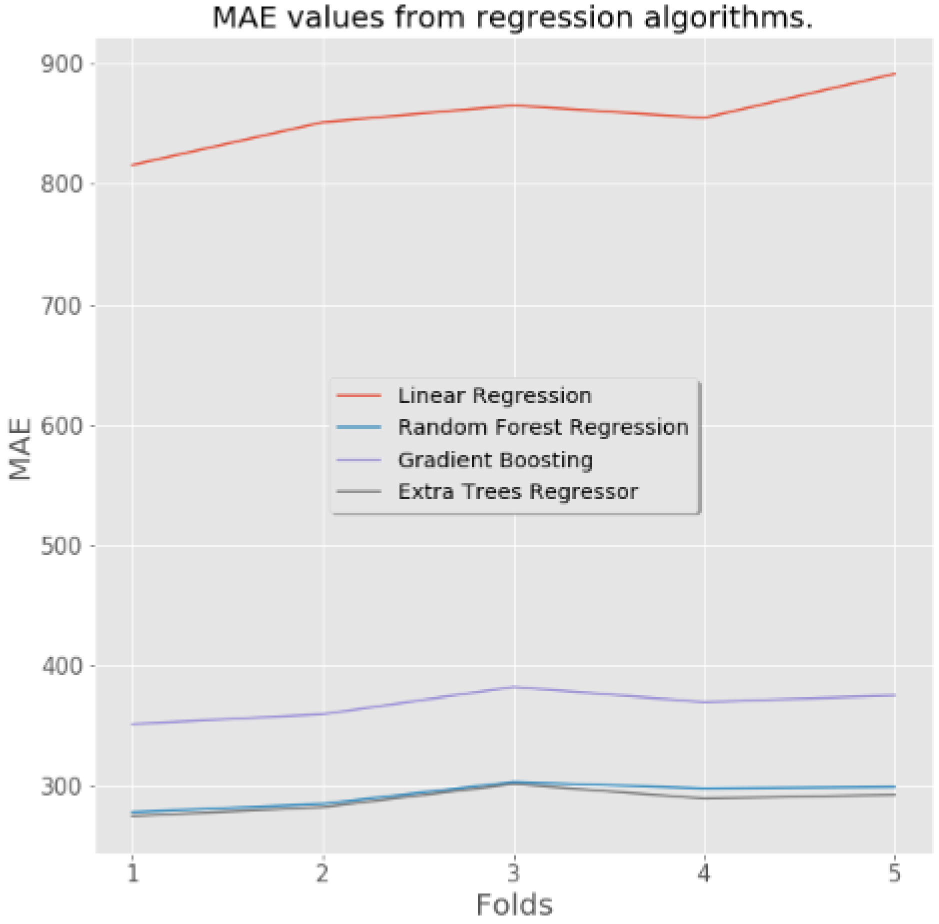
Experimentation setup

- To select the best proposed algorithm(s) I used 5-cross validation.
- The metric used to compare the performance of all models is MAE.
- The default parameters models given by the library were used.
- All the features of the dataset were used.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Robust to outliers

Result of 5-CV



Random Forest and Extra Tree algorithms obtained the best results.

Results of test dataset

The original dataset was two parts, the first one for the training process with 70% of instances and the second one for the testing process with 30% of instances. The three best models were used.

By using all features

Random Forest
MAE: 286.26837954497944

Gradient Boosting
MAE: 371.1343471051312

Extra Trees
MAE: 284.6502473257899

Ensemble: 282.31020423798265

By using all features proposed in the in the hypothesis

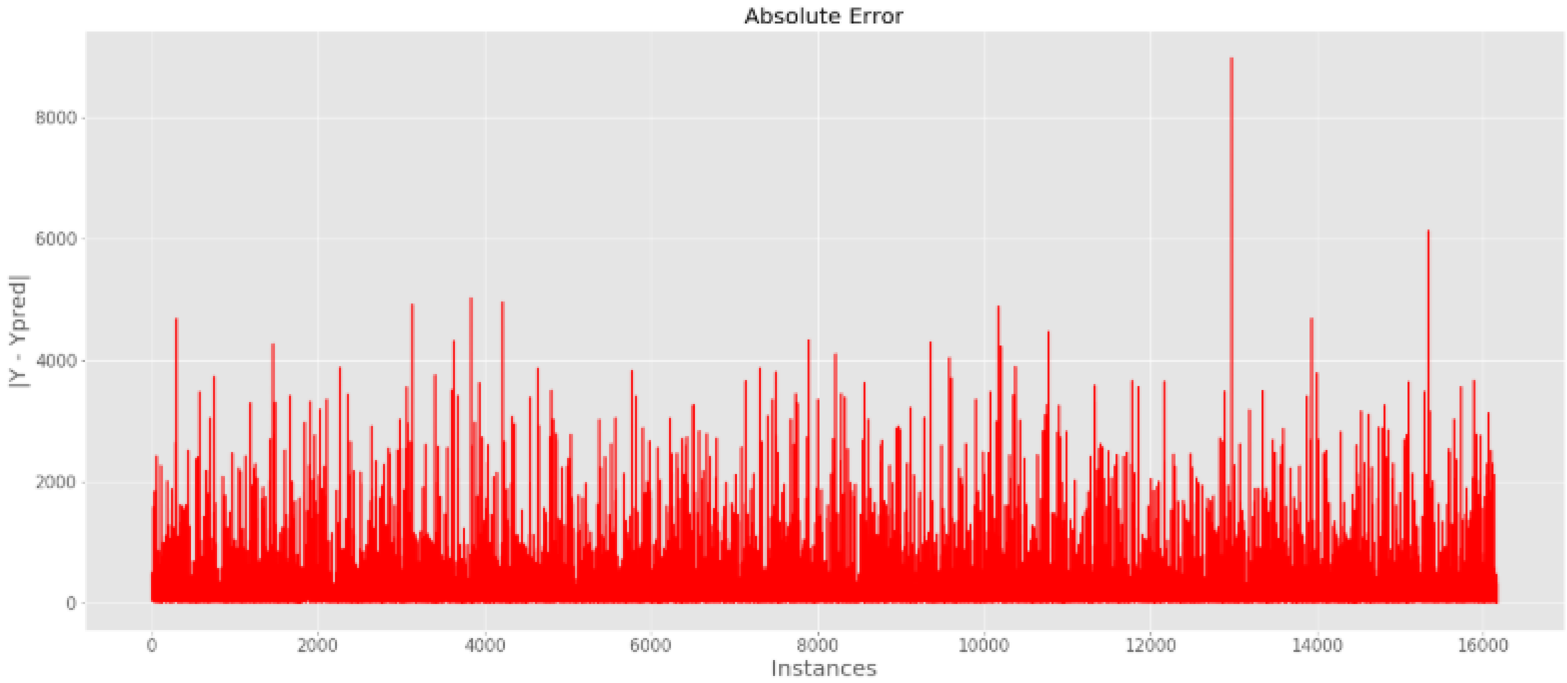
Random Forest
MAE: 659.679268754803

Gradient Boosting
MAE: 653.823862866317

Extra Trees
MAE: 673.4322533440508

Ensemble: 628.6054671661772

Plot of the error (|Y - Yp|)



It is observed two outliers from the error plot. Besides, the model can be improved. The plot is obtained by using the ensemble model with a MAE of **282.31.**

Conclusions and future work

Conclusions

- The hypothesis proposed was rejected. The best result was obtained by using all the features given by the database.
- The ensemble created by the three regression algorithms slightly improved the result obtained by Extra Trees algorithm.

Future work

- Improve feature extraction process.
- Remove outliers from dataset.
- Tuning of the proposed algorithms to improve performance.
- Try cubist algorithm that gets the best performance in (Fernández-Delgado, 2018).

Fernández-Delgado, M., Sirsat, M. S., Cernadas, E., Alawadi, S., Barro, S., & Febrero-Bande, M. (2018). An extensive experimental survey of regression methods. *Neural Networks*.