Part 2: SciKit-Stack and Deep Learning

All notebooks available at https://github.com/jandroi/3_2_BD

Management Report A – SKLearn

Data understanding

The data consist in 53.940 observations and 10 features. Each feature characterizes the valuation of a diamond according to the data description provided.

It is important to note that x, y and z represent physical measurement, while there can be a 0 in any dimension (due to the data description) a value of 0 in the 3 features will be considered as missing value in further steps and hence, eliminated.

Variables have a high correlation specially with carat in 0.92 to price. The table and depth variables do not correlate and will be eliminated in the data preparation stage.

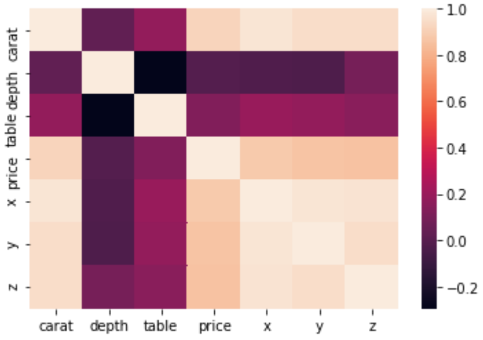


Figure 1 Correlation Between variables

Data Preparation

For better modeling. The data was categorized using modeling methodologies that would ensure us to input variables such as color and clarity, which are not numerical, into the model.

A total of seven observations were removed since their X Y and Z values were 0

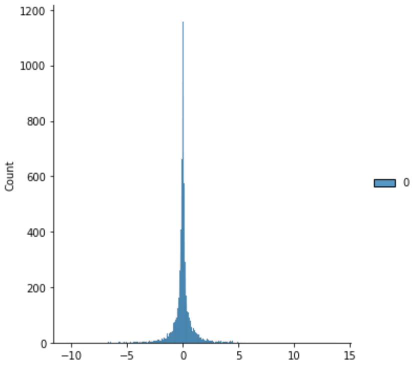
Other techniques such as scaling were used to ensure a better model fit.

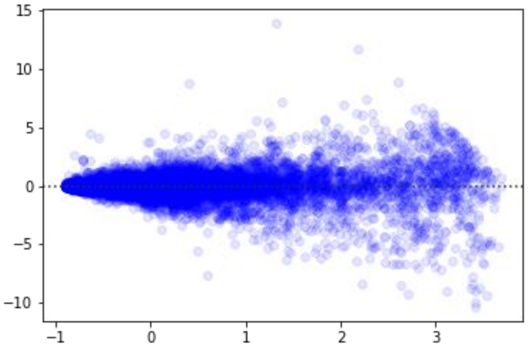
Modeling and Evaluation

Two approaches were made to model the prediction of the price of a diamond according to its qualities. The first approach is linear regression with a R^2: 0.918 which means that the price is explained in 91% by the model.

Another model Random Forest was trained to pursue a better approximation of price prediction. This random forest achieved an R^2 of 0.979. At first hand this seems a very good approach but further tests should be done to search for overfitting.

A visual inspection of the residuals lets us see that the bell shape is suspiciously sharp in the middle which can indicate a lack of normality. Also plotting the standardized residuals against the prediction might indicate Heteroscedasticity and further analysis should be commended.





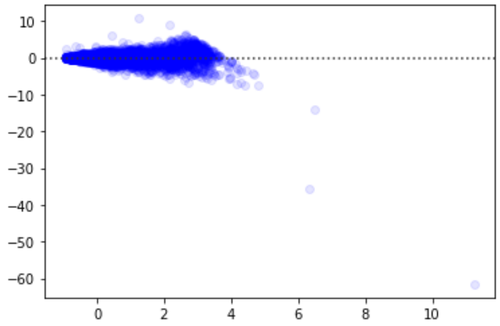
Management Report B – Deep Learning

Modeling

In order to optimize and search for better outcomes, two deep learning models were trained under the following assumptions:

1. Simple: 1 layer = mae: 770.0211 - mape: 33.8937
2. Multiple: 4 layers = mae: 339.1227 - mape: 9.4457

Heteroscedasticity and residual outlier demands more in depth analysis of the model .


10
0
−10
−20
−30
−40
−50
−60
0
2
4
6
8
10