Challenge Regression

Nathanaël Mariaule

3 June 2021

The goal of the challenge was to build a machine learning model that predict price for housing in Belgium using a previously collected data set.

3 steps:

- Features engineering and Preprocessing of the datas
- model selection
- hyperparameters tuning

Extra-Features

I collected datas with median prices for house in each belgian commune and add it to the data set.
Source: trends.knack.be (probably Fednot)

Pre-processing

- Remove unnecessary column
- preliminary filling of nan value e.g. fill garden_area=0 for house without garden, add 'unkown' category for state_of_building

Pre-processing

- Remove unnecessary column
- preliminary filling of nan value e.g. fill garden_area=0 for house without garden, add 'unkown' category for state_of_building

Pre-processing

- Fill nan with median or mean (also tried KNN algorithm)
- replace prize by log(prize)
- OneHotEncoding, MinMaxScaler and StandarScaler

Pre-processing

- Fill nan with median or mean (also tried KNN algorithm)
- replace prize by log(prize)
- OneHotEncoding, MinMaxScaler and StandarScaler

Pre-processing

- Fill nan with median or mean (also tried KNN algorithm)
- replace prize by log(prize)
- OneHotEncoding, MinMaxScaler and StandarScaler

Pre-processing

- Fill nan with median or mean (also tried KNN algorithm)
- replace prize by log(prize)
- OneHotEncoding, MinMaxScaler and StandarScaler

Model Selection

Various possibilities tested: Ridge, LASSO, Lars, RandomForest,...

XGBoost was the clear winner.

Model Selection

Various possibilities tested: Ridge, LASSO, Lars, RandomForest,...
XGBoost was the clear winner.

Hyperparameters tuning

Using CV, tuning of the following hyperparameters: max_depth, min_child_weight, learning_rate, subsample, colsample_by_tree and n_estimators.

Results

RMSE: 239000

R2 score: 0.77

Results

RMSE: 239000

R2 score: 0.77