Better Prediction of House Prices

An example of data improvement and model selection

By Luiz Augusto Carneiro
Sydney - Mar 2020

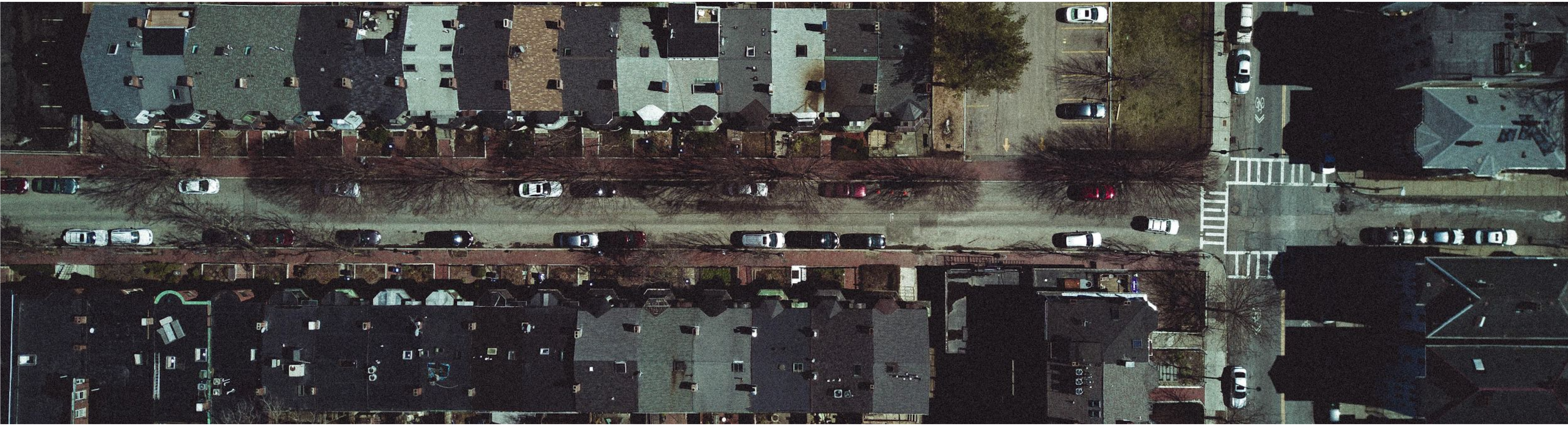


Photo by Ryan Mercier on Unsplash

Objective

To show how predictions can be enhanced by improving the original data set and choosing the most appropriate model to deal with the specific data characteristics.

Summary

ETL

- For sake of simplicity and comparison, we use a well-known data set: the Boston House Prices Data Set
- The data set has 506 observations, with 13 features and a target (median house price)
- There are no missing values

EDA

EDA shows:

- censored data
- outliers

Regression diagnostics show that the data is not best suited for OLS:

- non-linear patterns
- non-normal residuals
- non-constant variance
- leverage points/outliers

Insights

- Regression results improve a lot after removing censored data and outliers:

  Using **OLS**, MSE goes from 21.89 to 10.27!

- Lower MSEs, with better models:
  - **MLP regressor**: 8.86
  - **GLM**: 8.52 (the best!)

The Data Set

The target variable

MEDV: Median value of owner-occupied homes in $1000’s

The features

CRIM: per capita crime rate by town (kept in the final data set)

ZN: proportion of residential land zoned for lots over 25,000 sq.ft. **(not significant, p>0.1)**

INDUS: proportion of non-retail business acres per town **(not significant, p>0.1)**

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) **(not significant, p>0.1)**

NOX: nitric oxides concentration (parts per 10 million). **(not significant, p>0.1)**

RM: average number of rooms per dwelling (kept in the final data set)

AGE: proportion of owner-occupied units built prior to 1940 (kept in the final data set)

DIS: weighted distances to five Boston employment centres (kept in the final data set)

RAD: index of accessibility to radial highways **(high correlation - VIF)**

TAX: full-value property-tax rate per $10,000 **(high correlation - VIF)**

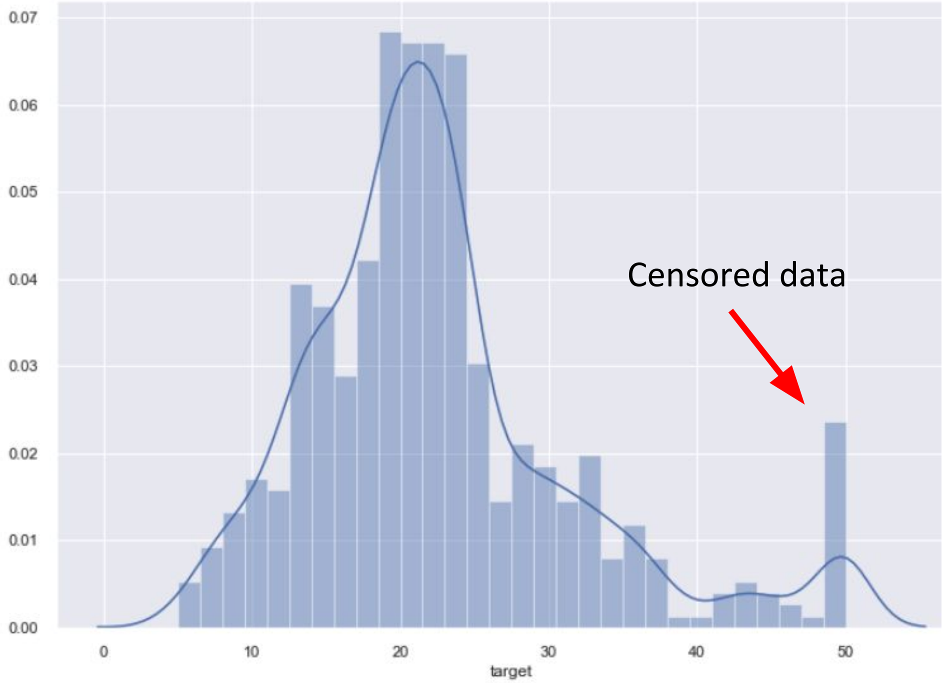
PTRATIO: pupil-teacher ratio by town (kept in the final data set)

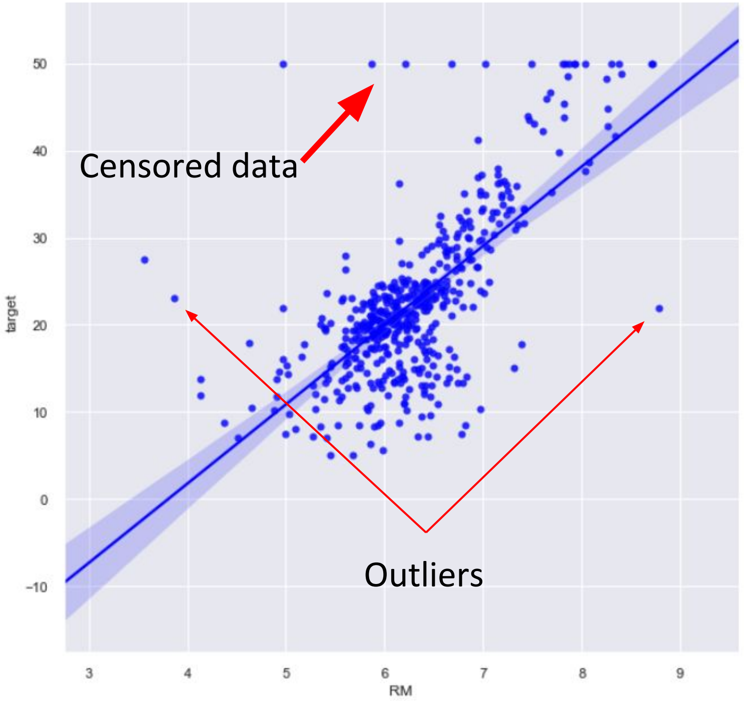
B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. **(not significant, p>0.1)**

LSTAT: % lower status of the population (kept in the final data set)

Insights from EDA

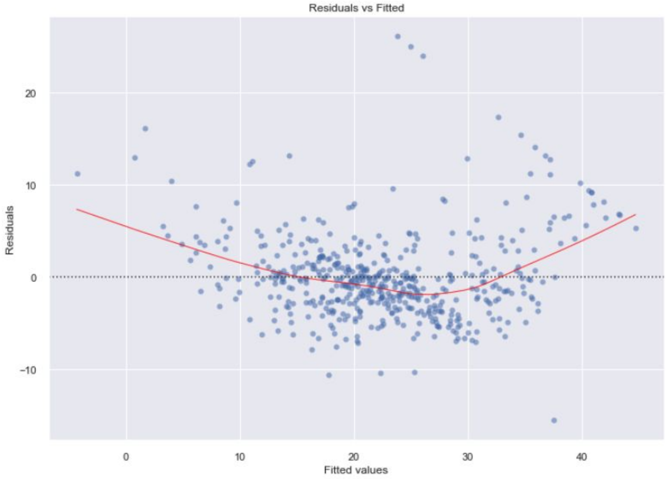
Data analysis show the presence of censored data and outliers



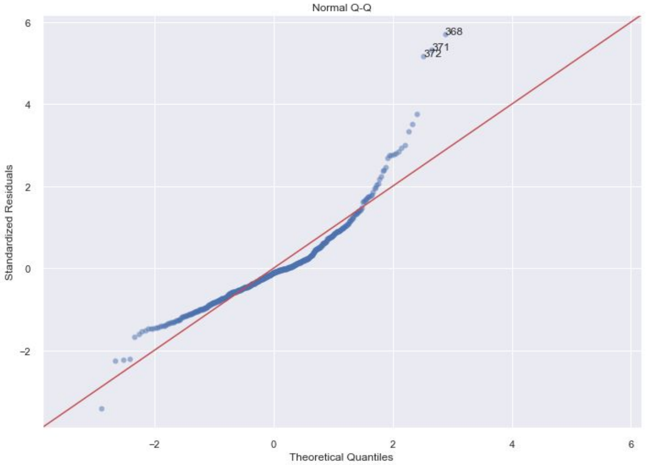


The data is not well suited for OLS Models

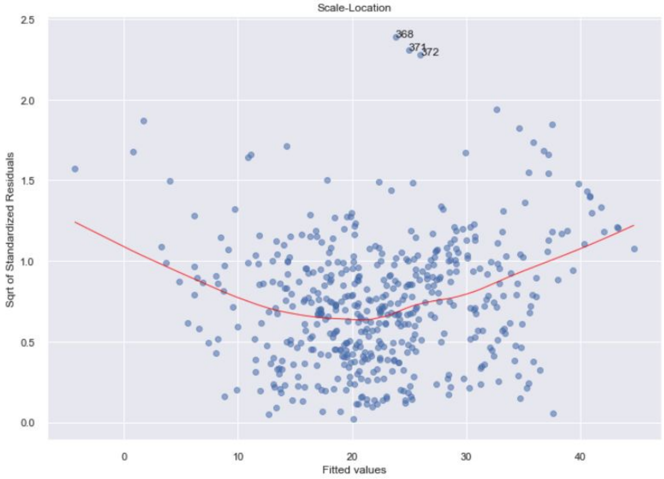
Non-linear relationships



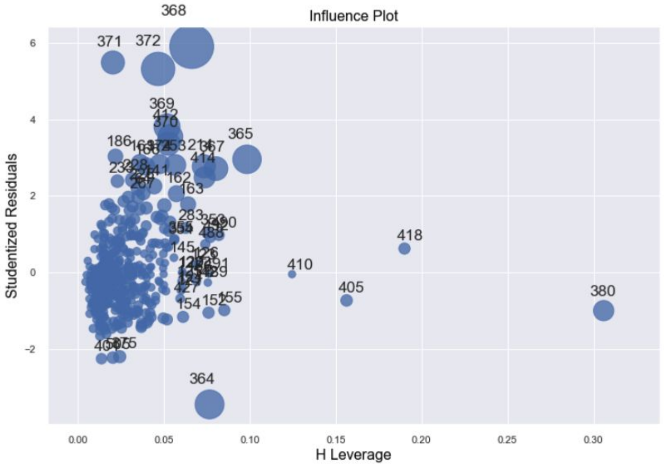
Residuals not normal



Residuals with no constant variance



Leverage points to tilt regression line



Final data set and results

The final set of features is:

CRIM
RM
AGE
DIS
PTRATIO
LSTAT

The final data set has 391 observations.

80/20 train/test split: 312 (train) and 79 (test)

MSE from OLS, GLM and MLP Regressor

GLM has the lowest MSE

Model \ Data	No shuffling	Shuffle 1	Shuffle 2	Shuffle 3	Shuffle 4	Shuffle 5	Average
OLS	10.27	10.84	12.38	14.5	10.25	10.5	**11.46**
GLM	7.71	7.29	7.15	11.75	8.66	8.52	**8.51**
MLP Regressor	9.99	8.16	10.03	13.42	9.35	8.86	**9.97**

This **data set was shuffled five times** before splitting into train and test again, to confirm the consistency of the results.

The End

Any questions?!!

Thank you!

Luiz Augusto Carneiro
carneiro_aus@yahoo.com.au