BOSTON HOUSE PREDICTOR

Iván jiménez

THE DATA FRAME

MAIN FEATURES:

Information about the variables involved in Boston house price

Small dataset (506 rows, 14 columns)

Only discrete and continuous values

GOAL:

Create a ML model to predict Median value of owner-occupied homes based on determined variables

Cleaning process

Problems with null values:

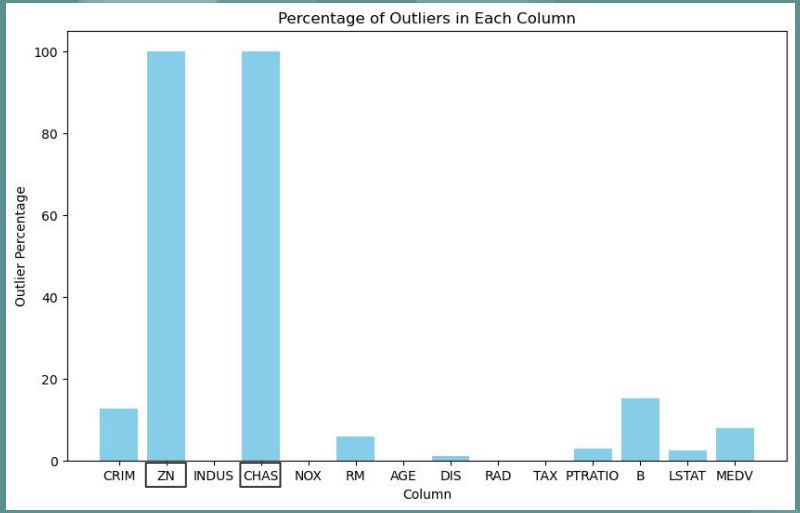
- ⅕ of the total rows were null values
- Due to their different nature, it is necessary to act in isolation on each of them.
- Methodologies followed: Mean and proportions in binary columns

CRIM	20
ZN	20
INDUS	20
CHAS	20
NOX	0
RM	0
AGE	20
DIS	0
RAD	0
TAX	0
PTRATIO	0
B	0
LSTAT	20
MEDV	0

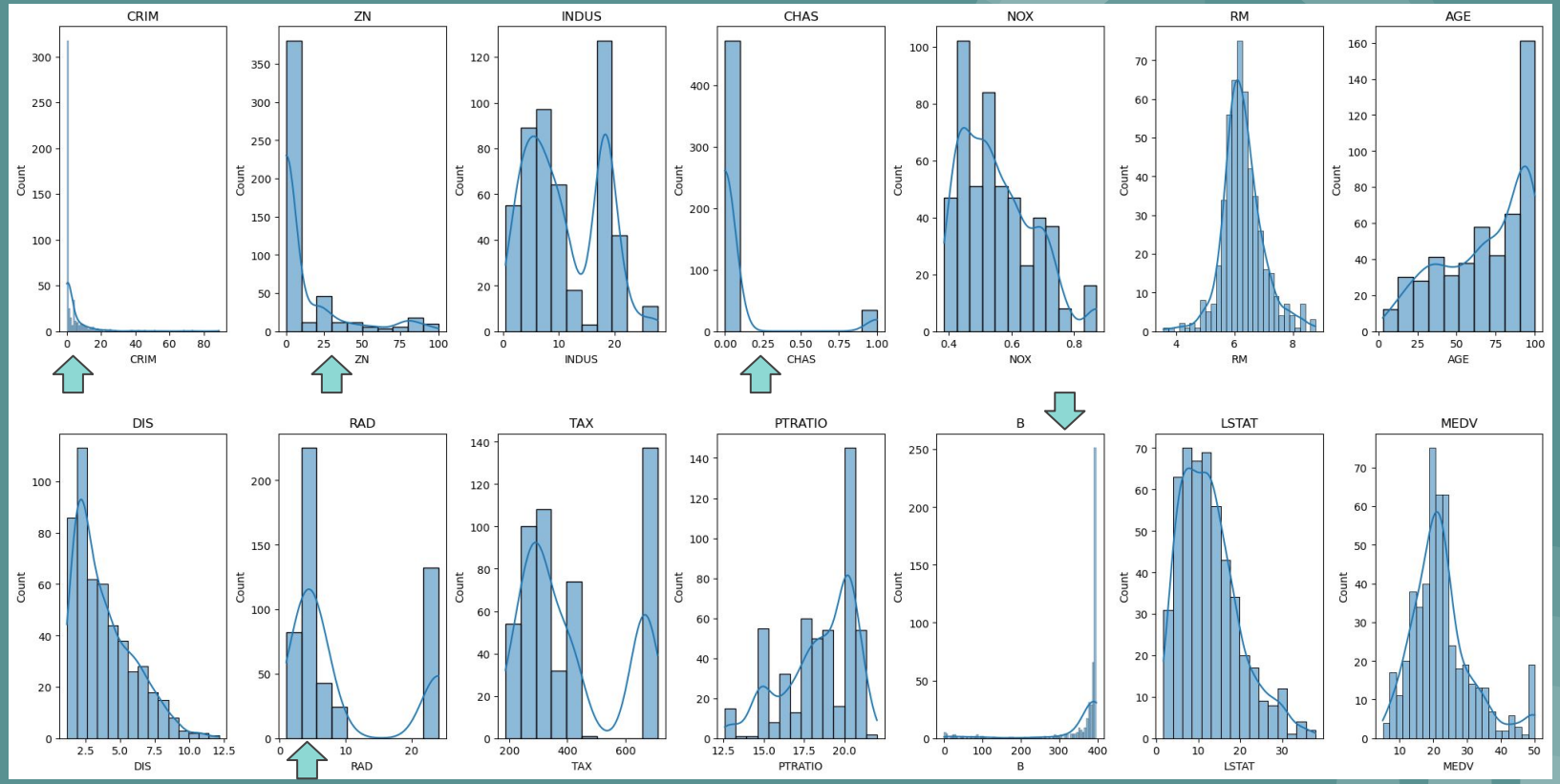
Cleaning process

Columns with tendencies of outliers:

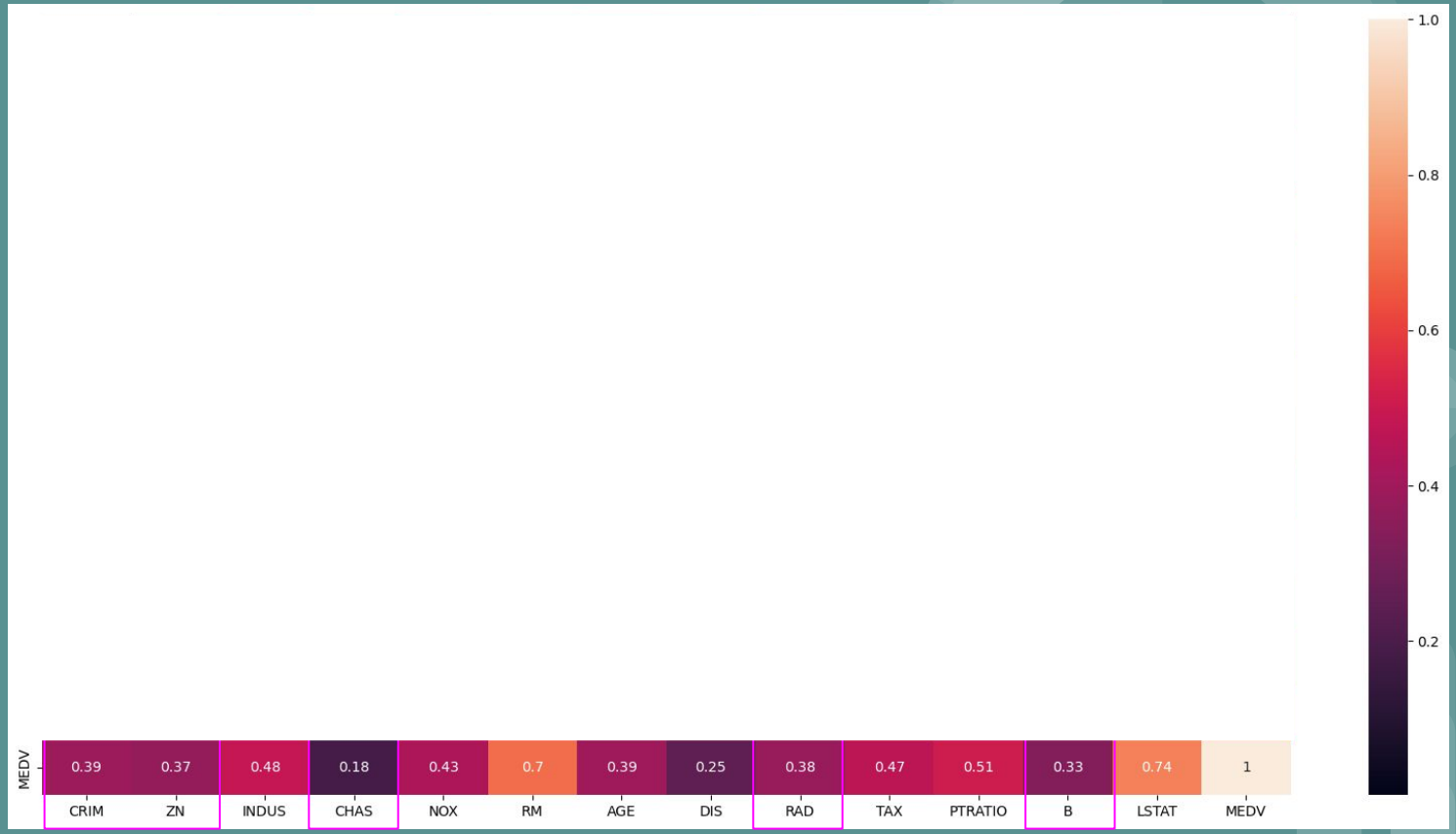
- Extract the percentages of outliers
- Create hist plots to understand better the skews and tendencies
- Variables skewed may produce biased or inaccurate estimation, leading to poor predictions



Cleaning process



Cleaning process



ML training

- Scaler: MinMaxScaler
- Look for the optimal hyperparameters
- Train different models and compare R2:

MODEL	R2 RESULT
KNN	0.69
LinearRegression	0.67
BaggingAndPasting	0.85
RandomForest	0.89
AdaBoosting	0.86
GradientBoosting	0.84

Final variables

Weight of the different variables in the model

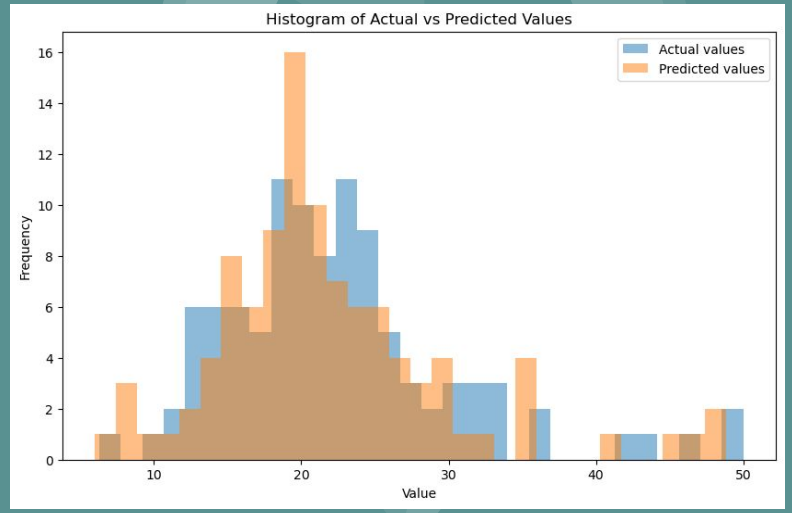
INDUS	proportion of non-retail business acres per town	0.9%
NOX	nitric oxides concentration (Close to industrial clusters or power station)	3.9%
RM	average number of rooms per dwelling	50%
DIS	weighted distances to five Boston employment centres	10%
TAX	full-value property-tax rate	2.2%
PTRATIO	pupil-teacher ratio	2.7%
LSTAT	% lower status of the population	31%

ML understanding

All models has similar overestimation and underestimation

More tendency to overestimate in mid values

Tendency to underestimate in values separated from the mean



TIME FOR A QUICK DEMO

Hope you all like the presentation

Thanks you all