House Prices: Advanced Regression Techniques

Paweł Lonca

Zadanie

Predykcja ceny nieruchomości w miejscowości Ames w stanie Iowa

Regresja

Root mean squared logarithmic error:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(x_i) - \log(y_i))^2}$$

Nie penalizujemy dużych różnic dla dużych wartości

Zestaw danych

train.csv

test.csv

1460 obserwacji, 79 zmiennych

data_description.txt

- objaśnienia dotyczące wartości
  - domyślne wartości
  - oznaczenia braków

Obróbka danych – braki(1)

	Total	Percentage
PoolQC	1453	0.995205
MiscFeature	1406	0.963014
Alley	1369	0.937671
Fence	1179	0.807534
FireplaceQu	690	0.472603
LotFrontage	259	0.177397
GarageType	81	0.055479
GarageCond	81	0.055479
GarageFinish	81	0.055479
GarageQual	81	0.055479
GarageYrBlt	81	0.055479
BsmtFinType2	38	0.026027
BsmtExposure	38	0.026027
BsmtQual	37	0.025342
BsmtCond	37	0.025342
BsmtFinType1	37	0.025342
MasVnrArea	8	0.005479
MasVnrType	8	0.005479
Electrical	1	0.000685

Większość braków uzupełniona
Za pomocą 0 lub „None”.

Obróbka danych – braki(2)

Zmienna ***LotFrontage*** informująca o długości odcinka ulicy przylegającego do posesji może być imputowana przy pomocy średniej dla poszczególnych poziomów zmiennej ***Neighborhood.***

Zmienna ***MSZoning*****,** czyli plan zagospodarowania przestrzennego wokół posesji może być określony na podstawie zmiennej ***MSSubclass*****,** czyli typu nieruchomości przeznaczonej na sprzedaż (np. dom jednorodzinny, mieszkanie w bloku).

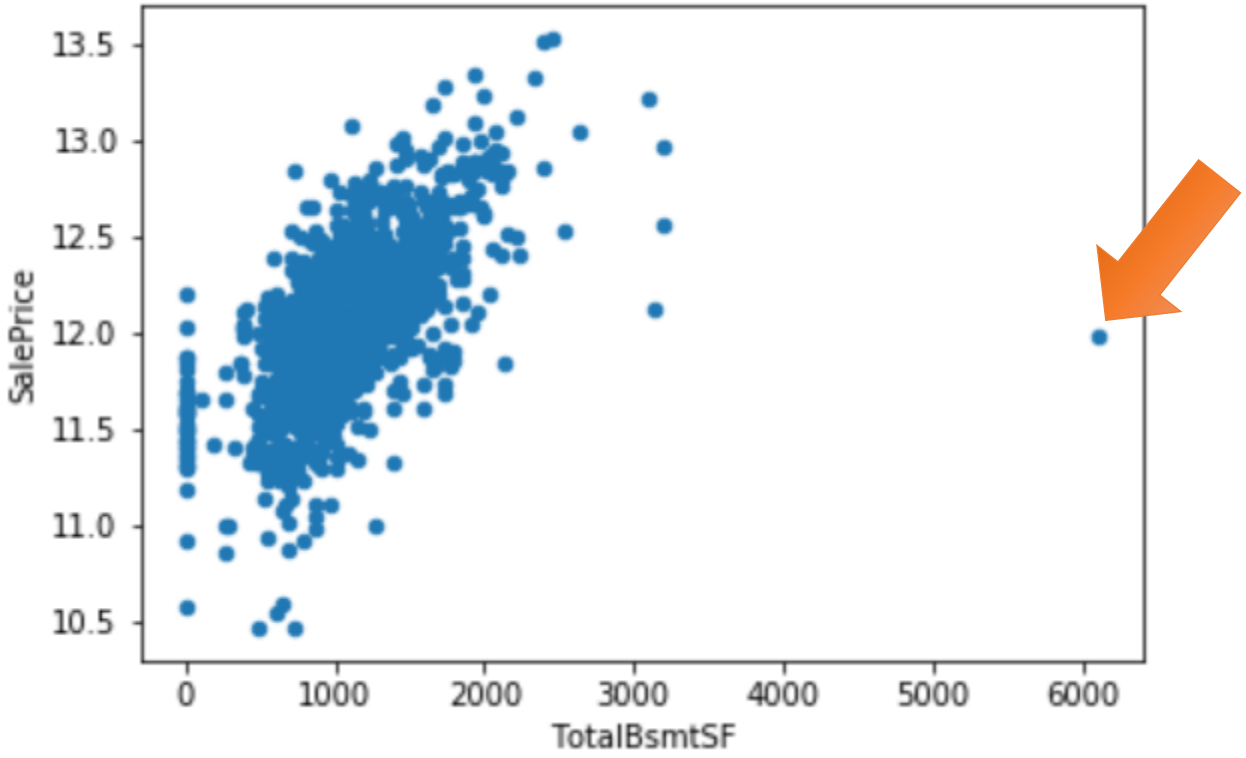
Braki dla zmiennych, dla których w legendzie nie przewidziano braków wartości wypełniam dominantą.

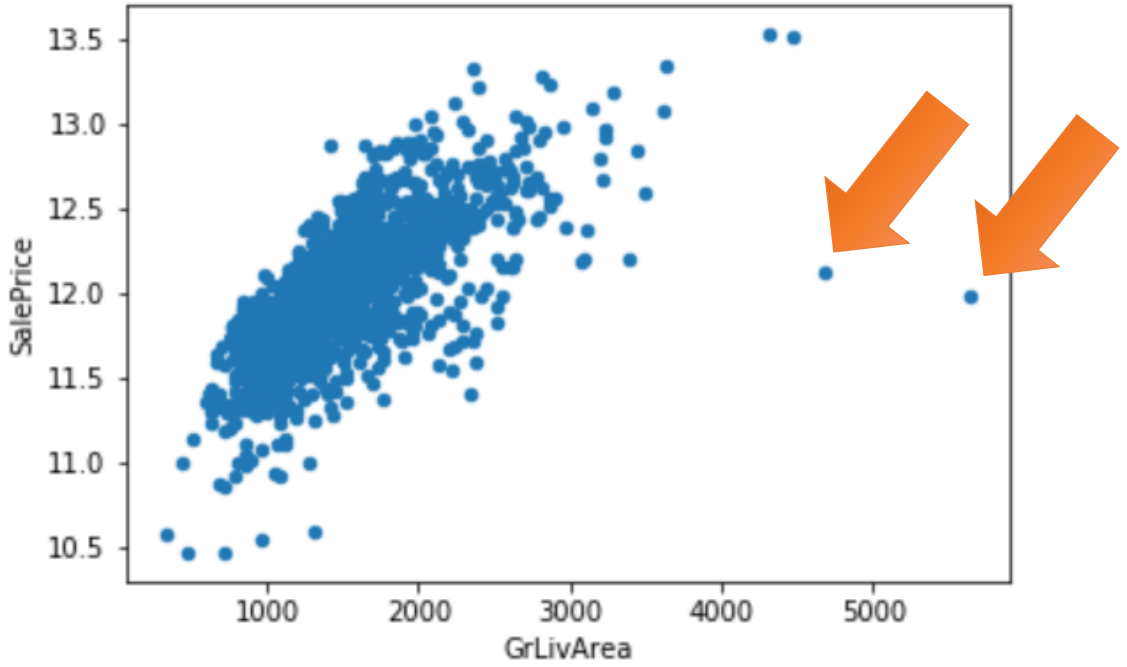
Obróbka danych – pozostałe kroki

- Niektóre zmienne mimo, że zapisane jako numeryczne tak naprawdę nimi nie są, np. miesiące

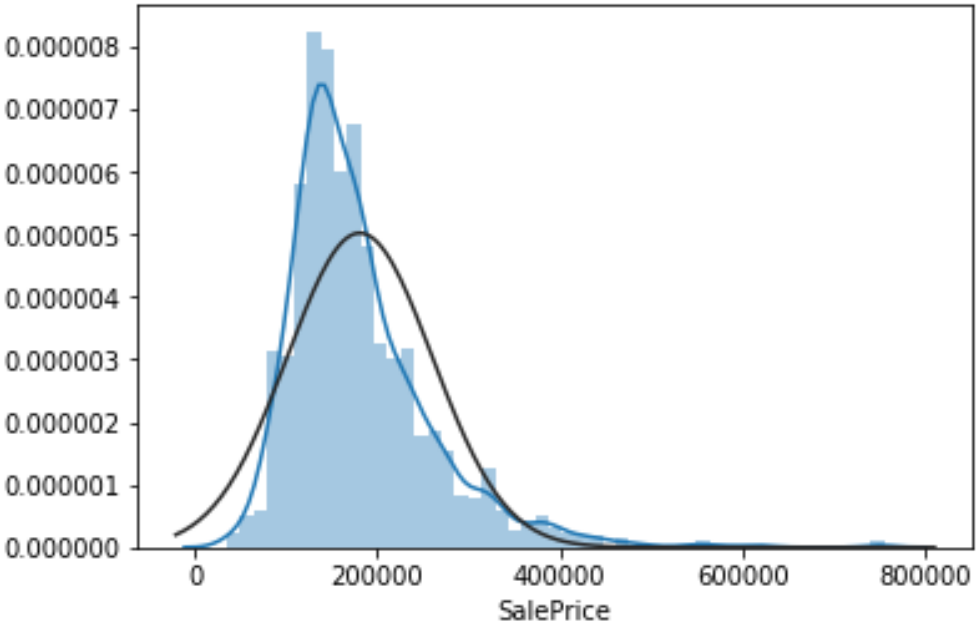
- Tworzę dodatkowe zmienne, np. ***hasPool, hasGarage, hasBasement, hasFirePlace, totalSF, TotalBathrooms***

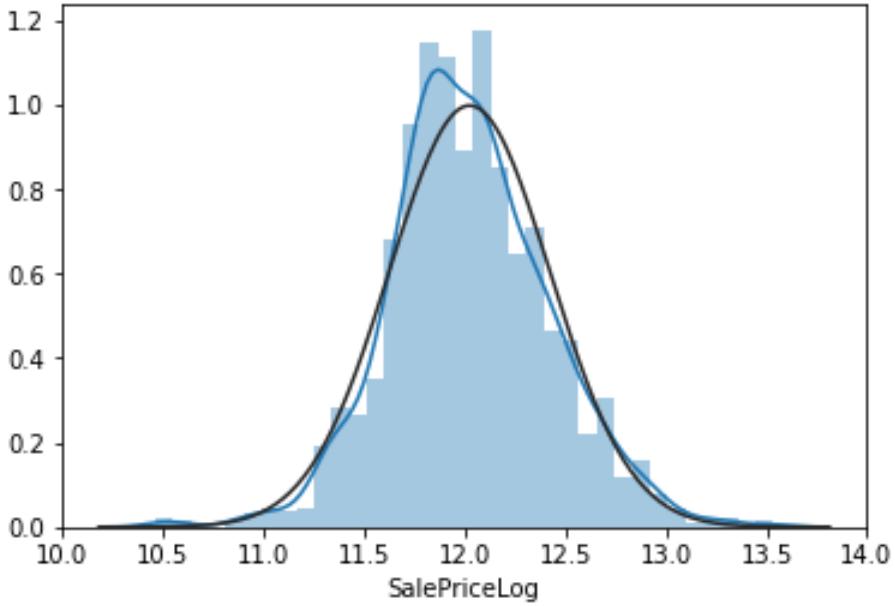
Obserwacje odstające





Zmienna objaśniana

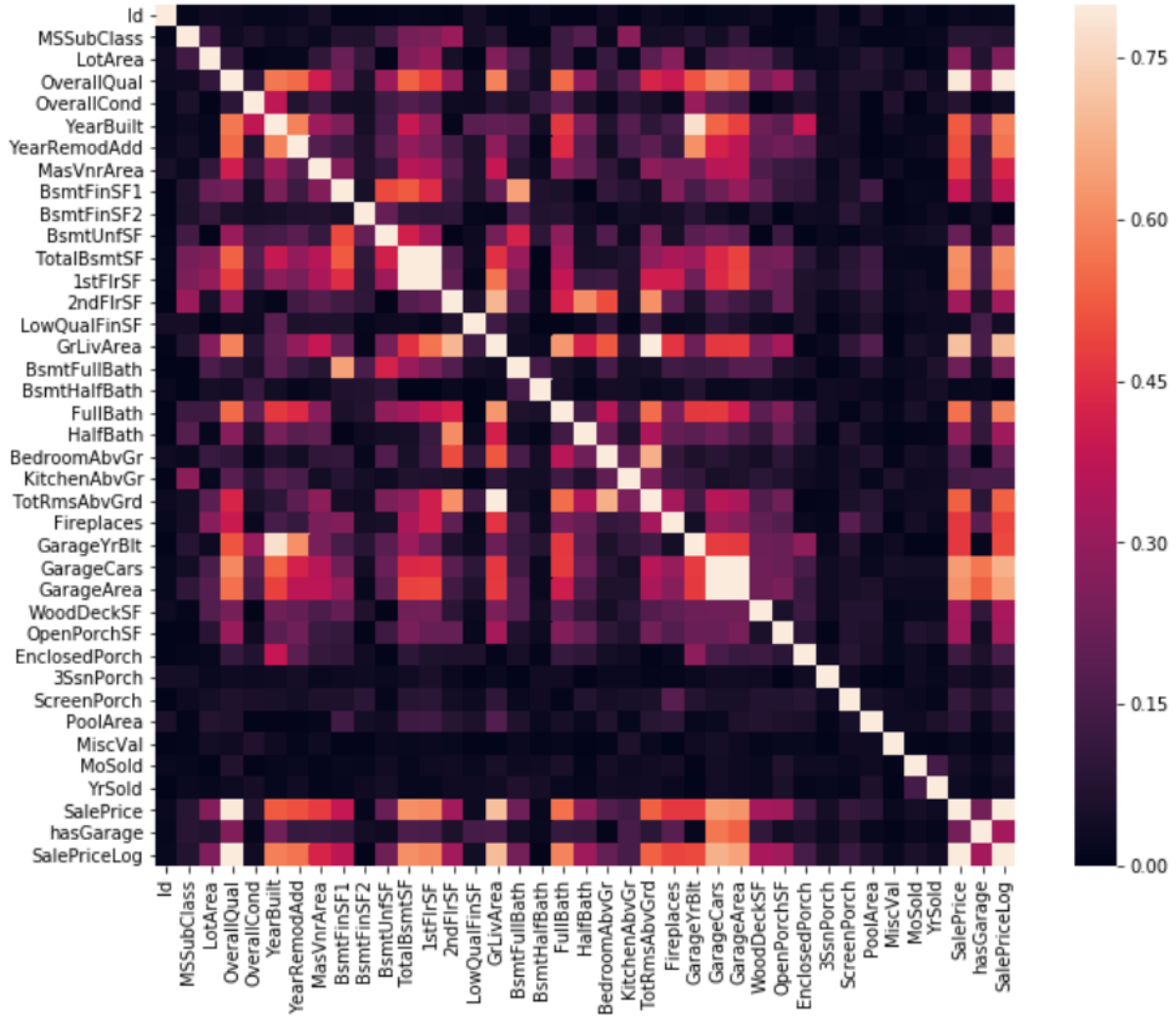




Brak rozkładu normalnego zmiennej objaśnianej

Transformacja logarytmem

Podejście nr 1



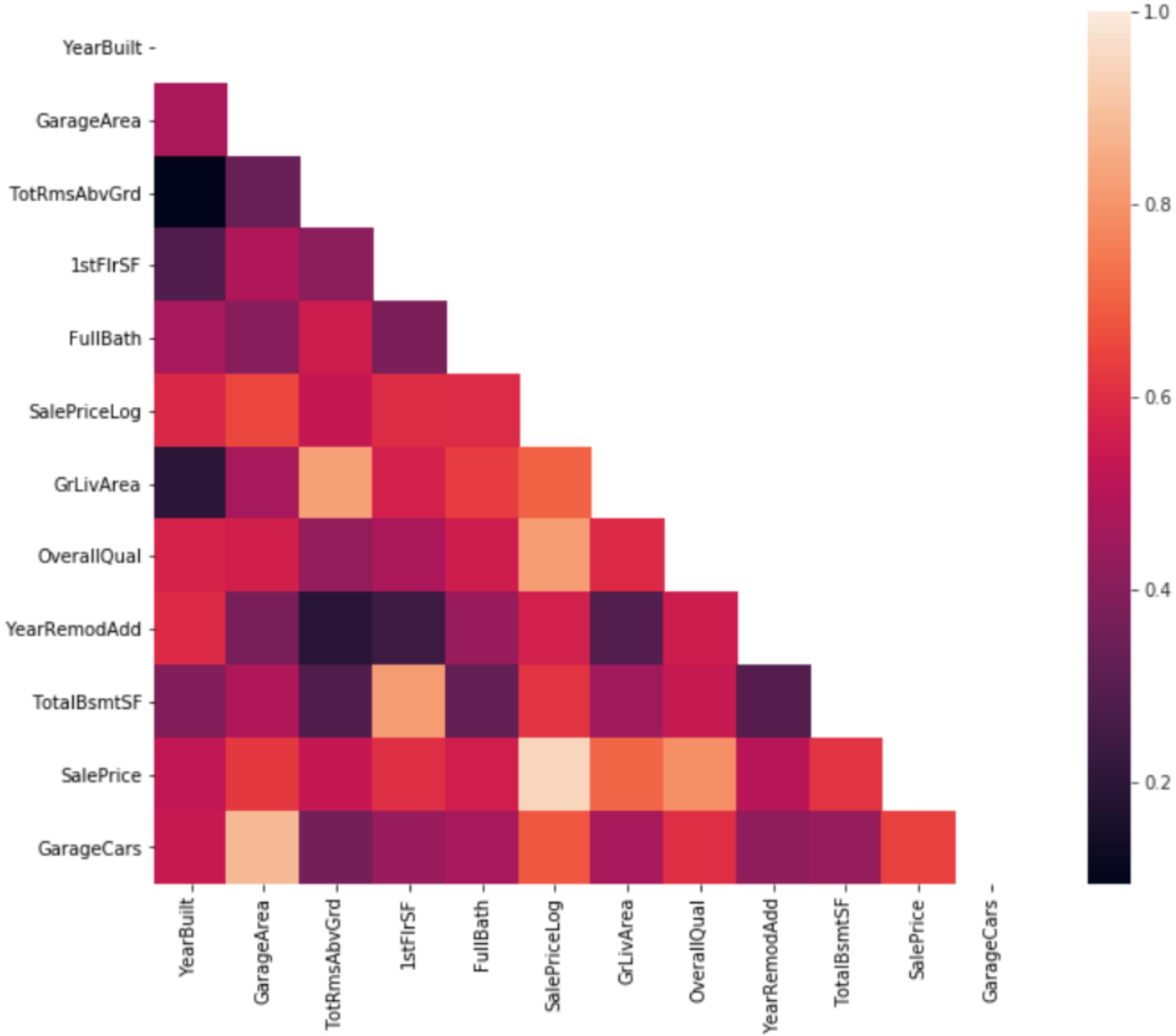
- Regresja liniowa kilku zmiennych
- Konieczny wybór zmiennych, które silnie korelują ze zmienną objaśnianą, ale nie ze sobą nawzajem

Poniższe zmienne wyjaśniające przekazują podobną informację:

1. `GarageCars` oraz `GarageArea` - im większa powierzchnia garażu tym więcej samochodów możemy trzymać w garażu.
2. `GrLivArea` oraz `TotRmsAbvGrd` - im więcej pokojów jakimś standardzie tym większa ich łączna powierzchnia.
3. `TotalBsmtSF` oraz `1stFlrSF` - im większa powierzchnia pierwszego piętra tym większa powierzchnia całości nieruchomości.
4. `YearBuilt` oraz `GarageYrBlt` - dom i garaż zapewne wybudowane w tym samym roku.
5. `GrLivArea` oraz `2ndFlrSF` - im większa powierzchnia drugiego piętra tym większa powierzchnia przestrzeni mieszkalnej.
6. `TotRmsAbvGrd` oraz `BedroomAbvGr` - im więcej sypialni tym więcej pokoi.
7. `2ndFlrSF` oraz `HalfBath` - im większe drugie piętro pod względem powierzchnie tym więcej (pół)łazienek można tam wcisnąć.

1. `OverallQual`
2. `GrLivArea`
3. `GarageCars`
4. `GarageArea`
5. `TotalBsmtSF`
6. `1stFlrSF`
7. `FullBath`
8. `YearBuilt`
9. `YearRemodAdd`
10. `TotRmsAbvGrd`

Zmienne, które najsilniej korelują ze zmienną zależną



Spośród zmiennych, które najsilniej korelują ze zmienną zależną usuwam te, które zbyt mocno korelują ze sobą: ***GarageCars,*** ***1stFlrSF,*** ***TotRmsAbvGrd***

test.csv 0.15925

5 days ago by Pawel Lonca

First submission based on linear regression.

Podejście nr 2

XGBoost

LightGBM

lgb_sub.csv
5 hours ago by Pawel Lonca
add submission details

0.12063

Dziękuję za uwagę