

Predykcja cen nieruchomości za pomocą porównania trzech regresorów

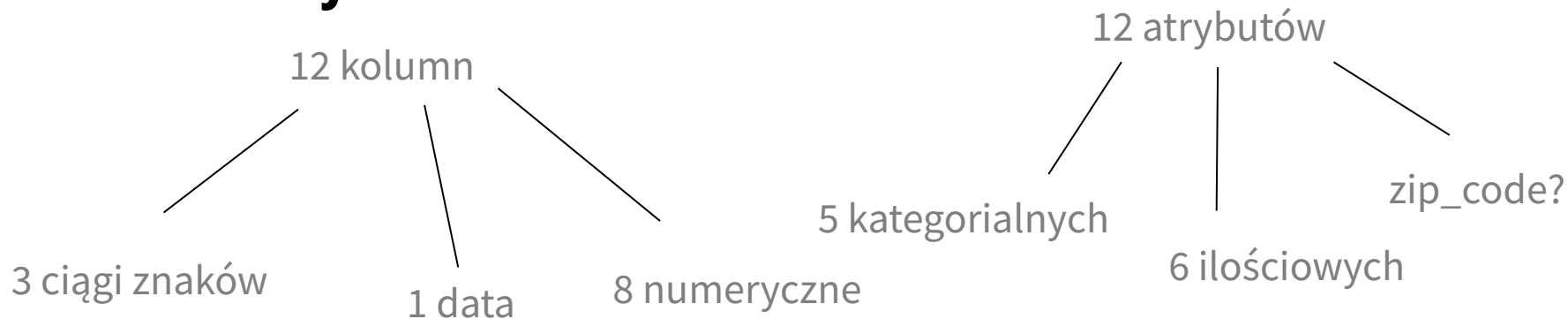
Badanie wpływu lokalizacji na cenę nieruchomości









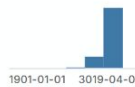
Plan prezentacji

1. Zbiór danych.
2. Analiza eksploracyjna i preprocessing danych.
3. Dane wykorzystywane do szkolenia.
4. Wybrane regresory.
5. Porównanie modeli.
6. Hipoteza – czy lokalizacja nieruchomości wpływa na cenę.


Zbiór danych



<https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset>

▲ brokered_by Broker / Agency encoded	▲ status Property sale status	# price House price	# bed Number of bedroom	# bath Number of bathroom	# acre_lot Total land size / lot size in acres	▲ street Street address encoded	📍 city City	📍 state State	📍 zip_code Zip code	# house_size House size / living space in square feet	📅 prev_sold_date Previously sold date
2226382 total values	for_sale 62% sold 36% Other (25067) 1%					2226382 total values	Houston 1% Chicago 1% Other (2184282) 98%		2226382 total values		
183378.0	for_sale	185000.0	3	2	0.12	1962661.0	Adjuntas	Puerto Rico	00601	920.0	
52707.0	for_sale	80000.0	4	2	0.88	1982874.0	Adjuntas	Puerto Rico	00601	1527.0	
183379.0	for_sale	67000.0	2	1	0.15	1484990.0	Juana Diaz	Puerto Rico	00795	748.0	

Zbiór danych

city	city_ascii	state_id	state_name	population
The name of the city/town.	city as an ASCII string.	The state or territory's USPS postal abbreviation.	The name of the state or territory that contains the city/town.	An estimate of the city's urban population. (2019).
19090 unique values	19085 unique values	TX 6% PA 6% Other (25010) 88%	Texas 6% Pennsylvania 6% Other (25010) 88%	
New York	New York	NY	New York	18713228
Los Angeles	Los Angeles	CA	California	12758887
Chicago	Chicago	IL	Illinois	8684283

<https://www.kaggle.com/datasets/sergejnuss/united-states-cities-database>

Połączono nazwę miasta i stanu z populacją.

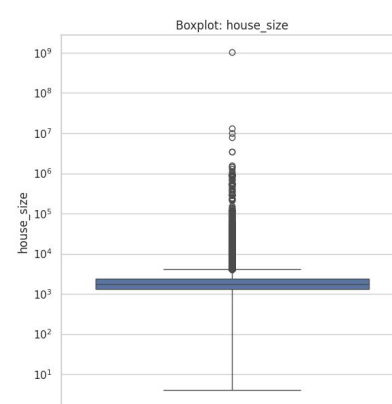
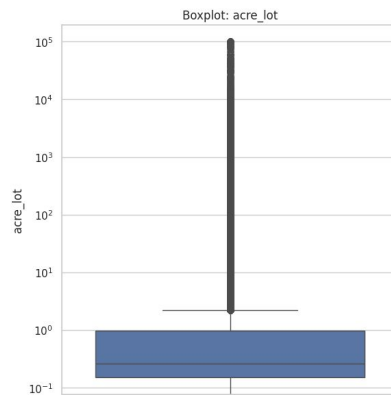
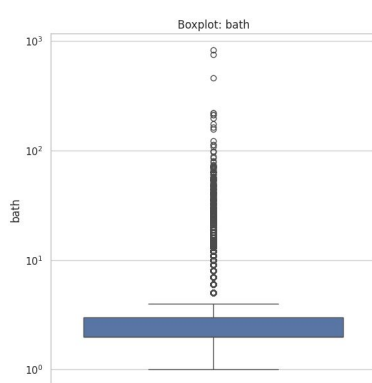
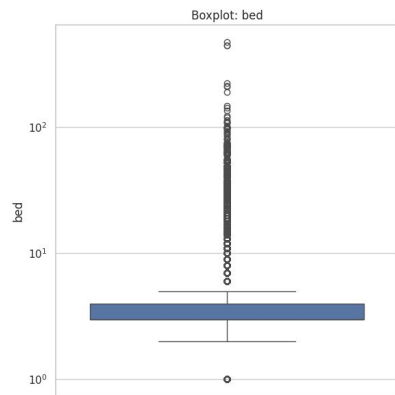
Lokalizacja to nie tylko współrzędne geograficzne, ale też wiele innych zmiennych – np. populacja, która może podnieść popyt.

Analiza eksploracyjna i preprocessing danych

Nieużyte atrybuty:

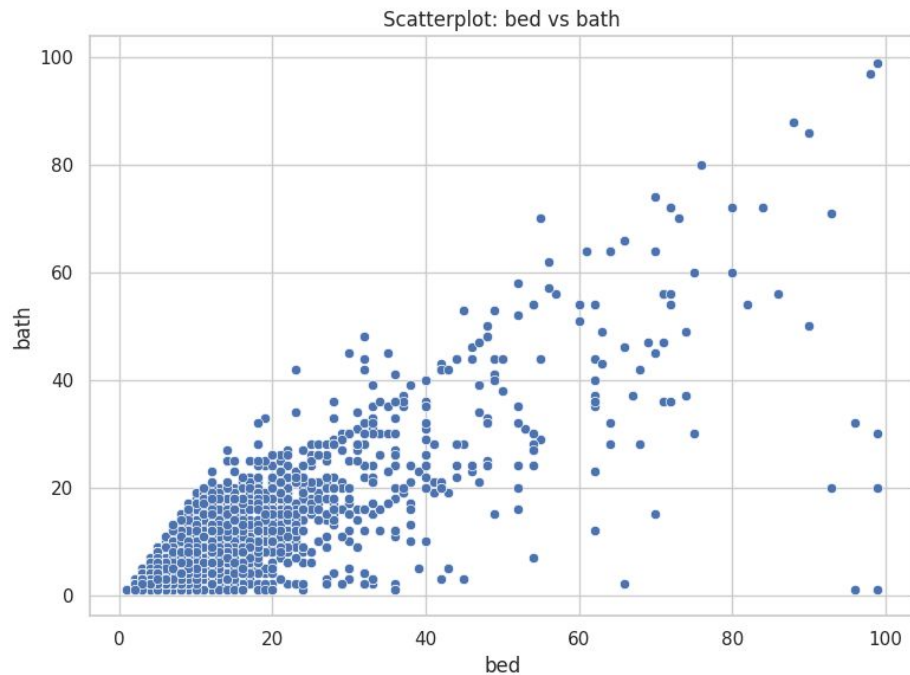
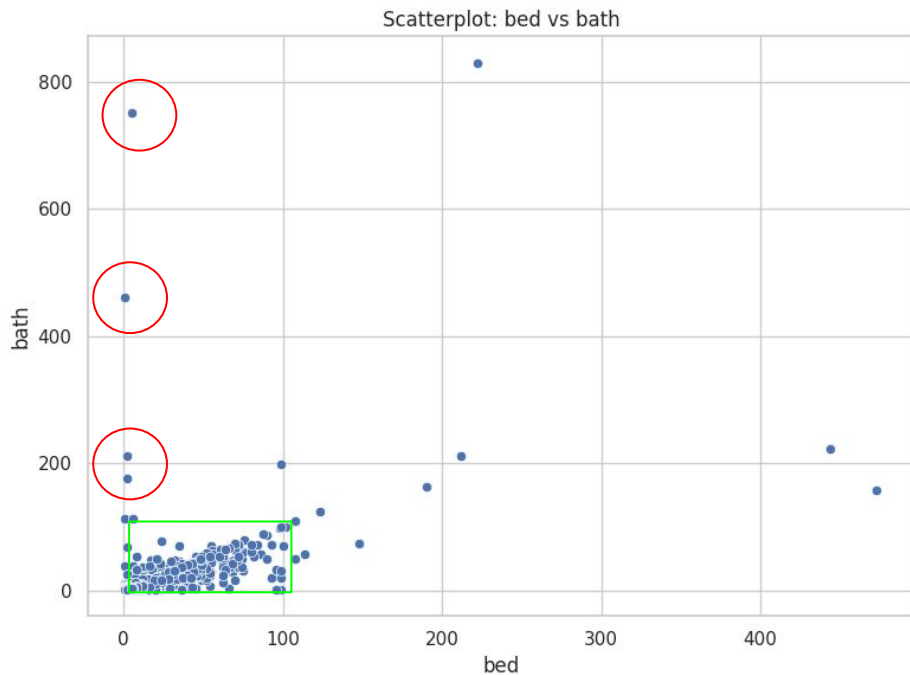
- brokered_by – nie ma to żadnego związku z lokalizacją, jest to tylko identyfikator pośrednika
- status – redundantna kolumna (razem z prev_sold_date)
- street – zbyt specyficzne
- prev_sold_date – przekształcone w sold_before (bool) i years_since_sold (integer)

Apartamentowce



Większość mieszkań ma mniej niż tuzin łóżek, łazienek itd.

Łazienki i sypialnie...



Outliery

- Usunięto rekordy z top 0,3% house_size.
- Usunięto rekordy, w których jest 100 lub więcej sypialni
- Usunięto rekordy, w których jest przynajmniej dwa razy więcej łazienek, niż sypialni
- Usunięto rekordy, w których data poprzedniej sprzedaży jest sprzed 1900 roku.

Puste wartości

Usunięto rekordy z pustymi wartościami atrybutów:

- price, city, state, zip_code

Wstawiono medianę do rekordów z pustymi wartościami atrybutów:

- acre_lot, house_size

Wstawiono dominantę (modę) do rekordów z pustymi wartościami atrybutów:

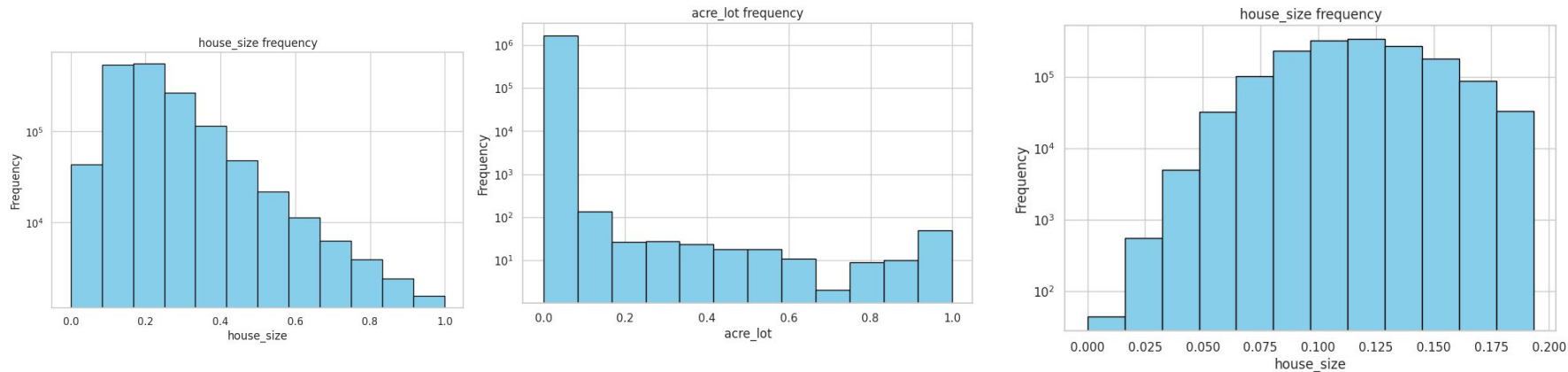
- bed, bath

Wstawiono średnią wartość do rekordów z pustymi wartościami atrybutów:

- population

Min-max scaling (1,2), boxcox normalization (3)

Po przeskalowaniu:



Dodatkowo zastosowano min-max scaling dla danych o populacji miasta.

Dane wykorzystywane do szkolenia

Oznaczono atrybuty city i state jako kategoryjne.

Wybrano 200 najpopularniejszych miast, resztę oznaczono jako ‘__other__’.

Dokonano one-hot encoding tych danych.

Do atrybutu zip_code zastosowano trzy podejścia: domyślne, pierwsze dwie cyfry jako liczba, pierwsze dwie cyfry jako kategoria. W przypadku traktowania ich jako kategorii, zastosowano Ordinal Encoding.

Oprócz tego, do danych lokalizacyjnych (miasto i stan) zastosowano dwa podejścia: uwzględnienie oraz nieuwzględnienie. Na podstawie tych dwóch podejść zaproponowano hipotezy

Najważniejsze atrybuty

SelectKBest:

bath	167435.796166
house_size	105755.866016
bed	46826.495396
state_California	28604.914328
city_New York City	23311.510279
city_New York	19622.305296
city_population	11552.355466
state_New York	7103.574394
city_Los Angeles	5405.444726
city_Miami Beach	4275.899600
zip_code	4185.718504
city_San Francisco	3564.830256
state_Ohio	3177.636633
state_Hawaii	2158.581209
...	

Feature importance (Random Forest):

bed	0.236969
zip_code	0.187631
house_size	0.167859
bath	0.137261
acre_lot	0.079442
city_population	0.077886
city	0.037794
state	0.036868
years_since_sold	0.034561
sold_before	0.003728

Wybrane regresory

Problemy: ograniczenia sprzętowe, czasowe i zasobowe. Po zakodowaniu cech kategoryalnych - około 300 kolumn.

- XGBoost (eXtreme Gradient Boosting) - wybrany ze względu na skalowalność i duże dane
- Sieć neuronowa - wybrana ze względu na uniwersalność i możliwość dostosowywania architektury
- Drzewo decyzyjne - wybrane ze względu na prostotę i szybkość działania

Parametry regresorów

XGBoost

enable_categorical	TRUE
subsample	0,8
n_estimators	1000
min_child_weight	3
max_depth	7
learning_rate	0,01
gamma	0
cosample_bytree	0,5

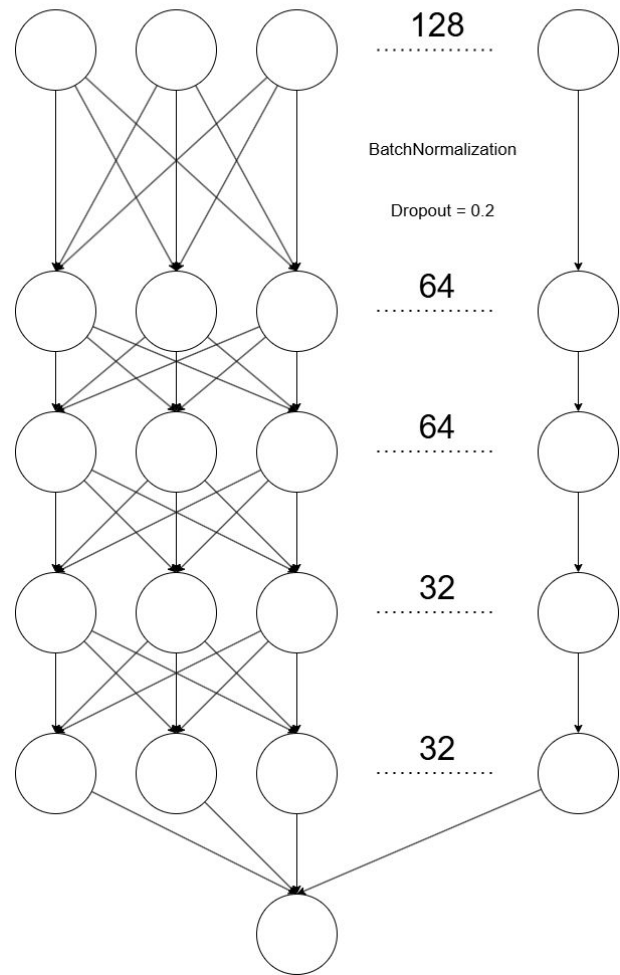
Decision Tree

max_depth	None
min_samples_split	2

Sieć neuronowa

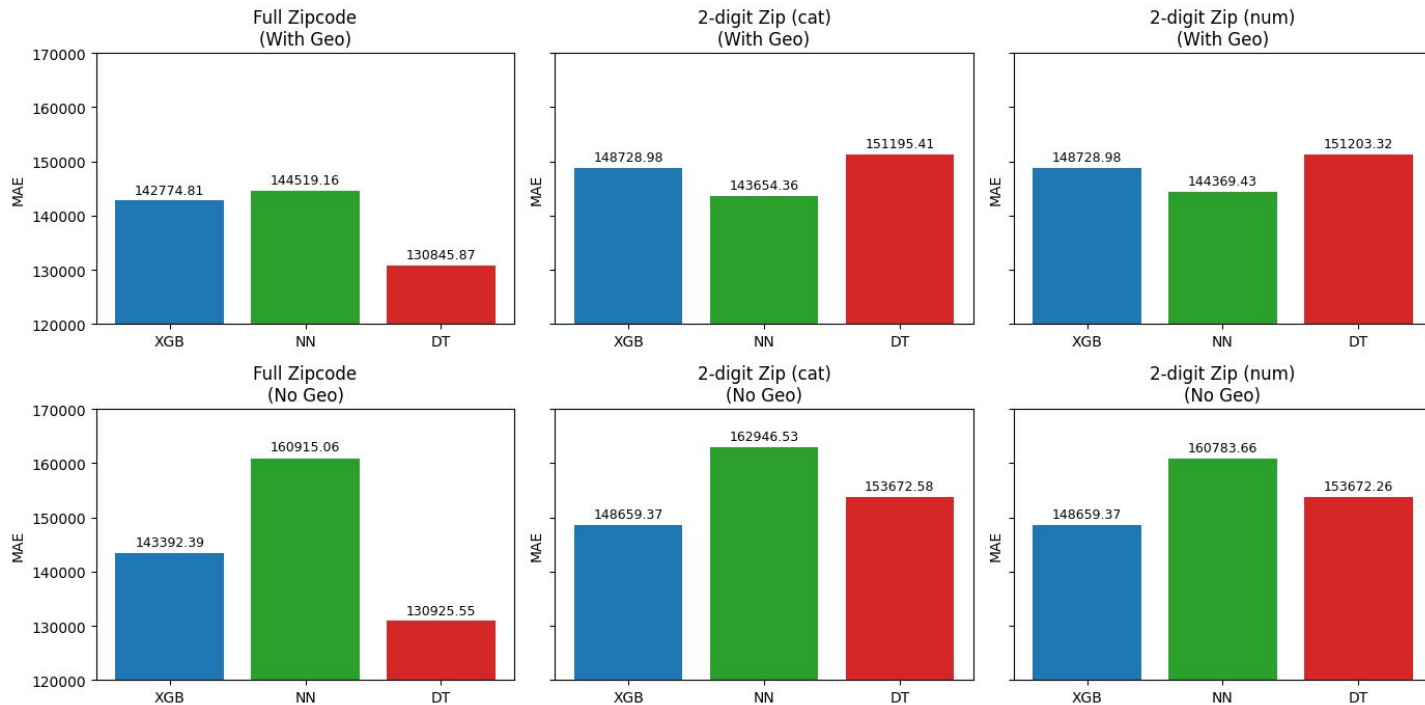
	neurons
activation	relu
	model compilation
optimizer	adam
loss	mae
metrics	mae
	early stopping
monitor	val_mae
patience	10
	model learning
epochs	50
batch_size	32
validation_split	0,2

Dane Wejściowe



Porównanie modeli

MAE Comparison Across Models and Features



Hipoteza

Hipoteza zerowa: Nieuwzględnienie cech lokalizacyjnych nieruchomości nie zmniejsza jakości predykcji jej ceny.

Hipoteza alternatywna: Nieuwzględnienie cech lokalizacyjnych nieruchomości zmniejsza jakość predykcji jej ceny.

Przyjęto $\alpha=0.05$

Hipoteza

- Przyjęto H_0 dla 3 z 9 modeli
- Odrzucono H_0 i przyjęto H_1 dla 6 z 9 modeli

Dla sieci neuronowych spadek jakości predykcji jest najbardziej widoczny

Porównanie par modeli	Wartość statystyki T	p-value
XGB - pełny kod p.	-6.1803	0.0000
XGB - 2-cyfrowy kod p. - kat.	0.7432	0.4573
XGB - 2-cyfrowy kod p. - num.	0.7432	0.4573
NN - pełny kod p.	-67.3706	0.0000
NN - 2-cyfrowy kod p. - kat.	-82.1985	0.0000
NN - 2-cyfrowy kod p. - num.	-75.2806	0.0000
DT - pełny kod p.	-0.2296	0.8184
DT - 2-cyfrowy kod p. - kat.	-6.2826	0.0000
DT- 2-cyfrowy kod p. - num.	-6.2611	0.0000

Zuzanna Ławniczak 151835 & Jakub Brambor 151871

**Dziękujemy za
uwagę**