

Predykcja cen nieruchomości za pomocą porównania trzech regresorów

Badanie wpływu lokalizacji na cenę nieruchomości

Autorzy: Zuzanna Ławniczak 151835 & Jakub Brambor 151871

Streszczenie. Dla zbiorów danych [US real estate](#) połączonego z [informacjami o miastach](#) zaprojektowano 3 różne regresory (xgboost, głęboka sieć neuronowa i drzewo decyzyjne) by rozwiązać problem predykcji cen nieruchomości. Wstępnie przeanalizowano dane, usunięto wartości odstające i obsłużono wartości puste, dokonano przeskalowania danych. Następnie zaproponowano jak i porównano trzy różne sposoby podejścia do atrybutu zip_code (kod pocztowy). Ponadto, wykorzystując mechanizmy fine-tuningowania udostępniane przez bibliotekę scikit-learn, dobrano najlepsze parametry oraz wyszkolono modele. Dodatkowo, na podstawie charakterystyki danych i analizy eksploracyjnej postawiona została hipoteza badawcza – czy nieuwzględnienie danych lokalizacyjnych nie zmniejsza jakości predykcji ceny nieruchomości. Po przetestowaniu hipotezy na zaproponowanych modelach oraz odpowiednio sprofilowanych wariantach otrzymanych danych, wyciągnięto wniosek, że dane lokalizacyjne są składową (wraz z rodzajem wybranego do predykcji regresora), która wpływa na jakość predykcji.

Wprowadzenie.

Ceny nieruchomości od wielu lat są tematem debat i dyskusji zarówno w życiu codziennym, jak i w polityce. Dla niektórych jest to kwestia tego, czy będą mieli dach nad głową, a dla niektórych tego, czy firma zrobi wystarczający zysk na transformowaniu nieruchomości. Wszystko, zarówno od najbardziej oczywistych czynników, takich jak powierzchnia, liczba sypialni, czy też na pierwszy rzut oka mniej oczywistych cech, chociażby osoba odpowiedzialna za sprzedać, może mieć wpływ na jej cenę. Ze względu na tak ogromną liczbę potencjalnych czynników, postanowiono zbudować modele uczenia maszynowego próbujące przewidzieć ostateczną cenę nieruchomości, w ramach procesu edukacyjnego przedmiotu Eksploracja Danych. Na podstawie dostępne publicznych danych dokonano analizy eksploracyjnej i przygotowano dane do przetworzenia przez modele. Zbadana została również hipoteza dotycząca cech geograficznych lokalizacji.

Zbiory danych.

Na platformie kaggle znaleziono wiele zbiorów danych, które spełniałyby założenia zadania, ale wybrano ten, którego jakość była najwyższa. Zbiór ten znajduje się pod wskazanym [adresem](#). Zawiera on następujące kolumny:

Następnie stwierdzono, że sama nazwa miasta nie jest wystarczająco informatywna i dołączono do zbioru danych również populację miasta, w którym znajduje się nieruchomość. Informacja o mieszkańcach danych miast znajduje się pod wskazanym [adresem](#).

Zbiór danych został wybrany w taki sposób, by były w nim zarówno cechy samej nieruchomości (wielkość posesji, wielkość powierzchni mieszkalnej, liczba sypialni i łazienek), ale również lokalizacja (stan oraz miasto). Dobrane również zostały dodatkowe atrybuty, takie jak kod pocztowy i populacja danego miasta. Takie zestawienie atrybutów pozwala na zbadanie prawdziwości hipotezy, czy na cenę nieruchomości wpływa jej lokalizacja. Hipoteza zostanie dokładniej opisana w późniejszej [sekcji](#).

Pierwszy zbiór danych.

Podstawowy zbiór danych składa się z 12 atrybutów, w tym 3 to ciągi znaków, 1 to data, 8 to wartości numeryczne. Dzieląc atrybuty inaczej, otrzymamy 6 kategoryalnych i 6 ilościowych atrybutów (jeśli zaliczymy atrybut zip_code jako kategoryalny).

# brokered_by	# status	# price	# bed	# bath	# acre_lot	# street	# city	# state	# zip_code	# house_size	# prev_sold_date
Broker / Agency encoded	Property sale status	House price	Number of bedroom	Number of bathroom	Total land size / lot size in acres	Street address encoded	City	State	Zip code	House size / living space in square feet	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		
183373.8	for_sale	185889.8	3	2	0.12	1962661.8	Adjuntas	Puerto Rico	88681	928.8	
52787.8	for_sale	88888.8	4	2	0.88	1982874.8	Adjuntas	Puerto Rico	88681	1527.8	
183379.8	for_sale	67888.8	2	1	0.15	1484998.8	Juana Diaz	Puerto Rico	88755	748.8	

Usunięte atrybuty.

brokered_by – gdyż jest wartością arbitralną, identyfikatorem pośrednika.

status – nie ma znaczenia, czy nieruchomość została już zakupiona, czy jeszcze nie

street – uznano atrybut za zbyt specyficzny – odchylenie standardowe wynosi połowę średniej

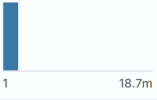
prev_sold_date – atrybut został przerobiony na dwa nowe: sold_before (bool) i years_since_sold (integer).

Describe

	brokered_by	price	bed	bath	acre_lot \
count	2.221849e+06	2.224841e+06	1.745065e+06	1.714611e+06	1.900793e+06
mean	5.293989e+04	5.241955e+05	3.275841e+00	2.496440e+00	1.522303e+01
std	3.064275e+04	2.138893e+06	1.567274e+00	1.652573e+00	7.628238e+02
min	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
25%	2.386100e+04	1.650000e+05	3.000000e+00	2.000000e+00	1.500000e-01
50%	5.288400e+04	3.250000e+05	3.000000e+00	2.000000e+00	2.600000e-01
75%	7.918300e+04	5.500000e+05	4.000000e+00	3.000000e+00	9.800000e-01
max	1.101420e+05	2.147484e+09	4.730000e+02	8.300000e+02	1.000000e+05

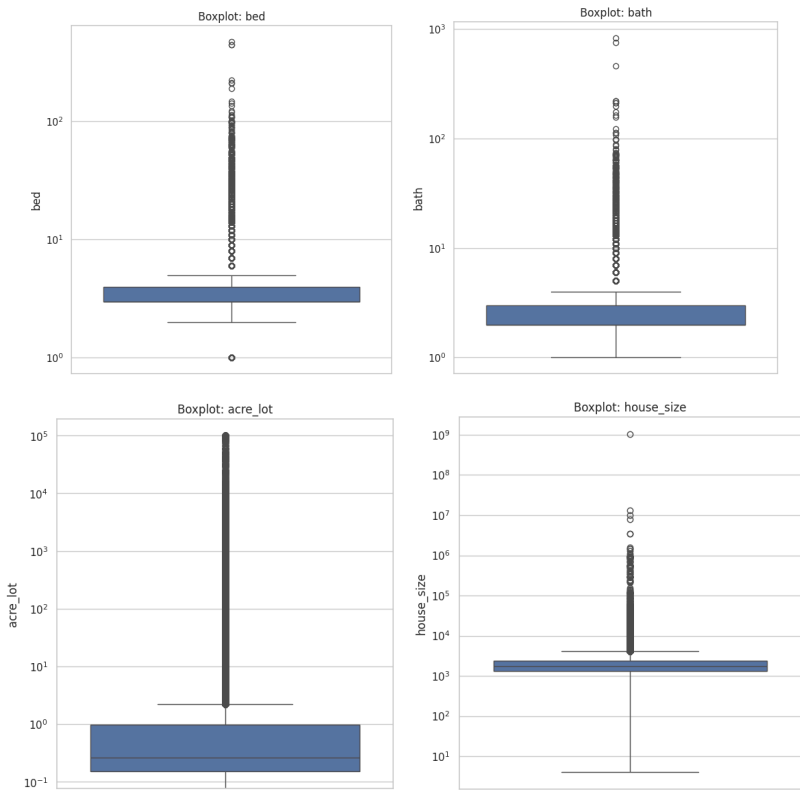
	street	zip_code	house_size
count	2.215516e+06	2.226083e+06	1.657898e+06
mean	1.012325e+06	5.218668e+04	2.714471e+03
std	5.837635e+05	2.895408e+04	8.081635e+05
min	0.000000e+00	0.000000e+00	4.000000e+00
25%	5.063128e+05	2.961700e+04	1.300000e+03
50%	1.012766e+06	4.838200e+04	1.760000e+03
75%	1.521173e+06	7.807000e+04	2.413000e+03
max	2.001357e+06	9.999900e+04	1.040400e+09

Drugi zbiór danych (populacja miast).

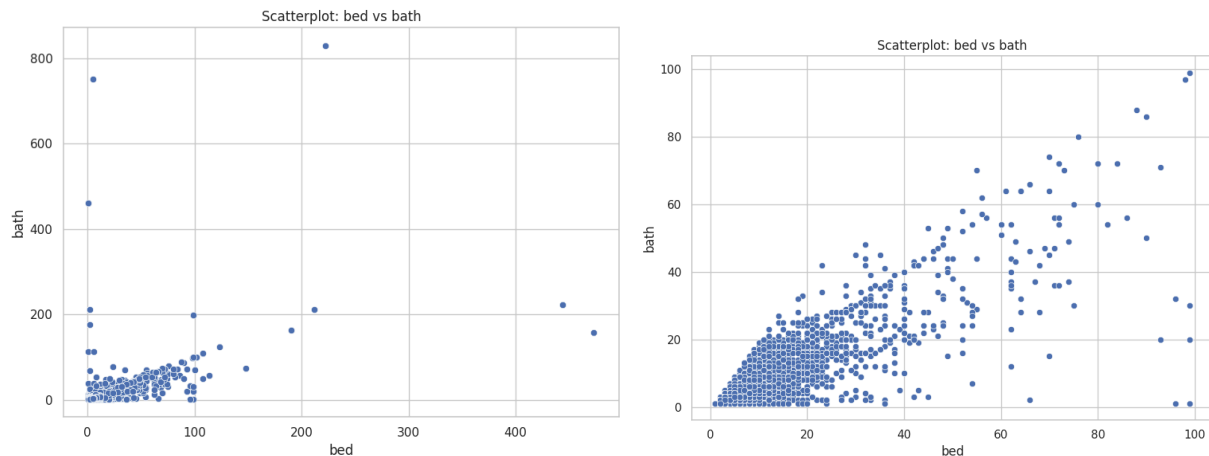
city	city_ascii	state_id	state_name	# population
The name of the city/town.	city as an ASCII string.	The state or territory's USPS postal abbreviation.	The name of the state or territory that contains the city/town.	An estimate of the city's urban population. (2019).
19090 unique values	19085 unique 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

Połączono nazwę miasta i stanu z populacją.

Outliery.



Większość nieruchomości posiada mniej niż tuzin sypialni i łazienek.



Usunięto rekordy, które stanowiły top 0,3% wartości `house_size`. Usunięto też rekordy, w których jest więcej niż 99 sypialni oraz te, w których przynajmniej jest dwa razy więcej łazienek, niż sypialni. Dodatkowo usunięto rekordy z niepoprawną i starą datą sprzedaży – sprzed 1900 roku.

Obsługa pustych wartości.

Rekordy z pustymi wartościami atrybutów `price`, `city`, `state` i `zip_code` zostały usunięte z dalszej analizy. W rekordach z pustymi wartościami atrybutów `acre_lot` i `house_size` wstawiono za wartość medianę dla danego atrybutu.

W rekordach z pustymi wartościami atrybutów `bath` i `bed` wstawiono za wartość modę (dominantę) dla danego atrybutu, gdyż chciano zachować całkowitoliczbowość.

Do rekordów z pustą wartością atrybutu `population` wstawiono średnią wartość.

Skalowanie.

Przeskalowano (min-max scaling) atrybuty `house_size`, `acre_lot`. Dla `house_size` następnie również zastosowano boxcox normalization.

Preprocessing.

Znaleziono 200 największych miast, a resztę nadpisano jako ‘`__other__`’ w rekordach. Oznaczono atrybuty `state` i `city` jako zmienne jakościowe i przetworzono je na zmienne typu One Hot Encoding. Atrybut `zip_code` został potraktowany na trzy różne sposoby, co można przeczytać w załączonej [sekcji](#).

Najważniejsze atrybuty.

Sprawdzono, jakie są najważniejsze atrybuty w zbiorze danych. Sprawdzono to dwoma metodami: `SelectKBest` oraz za pomocą Feature Importance modelu Random Forest.

`SelectKBest`:

<code>bath</code>	167435.796166
<code>house_size</code>	105755.866016
<code>bed</code>	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

Jak widać, wg jednego algorytmu najważniejszymi zmiennymi okazały się być liczba łazienek, rozmiar domu, liczba sypialni oraz to czy nieruchomość znajduje się w Kalifornii lub w Nowym Jorku. Wg drugiego liczba sypialni też jest bardzo ważna, ale też zip_code (kod pocztowy). Najprawdopodobniej dlatego, że kod pocztowy określa zarówno stan, ale też region w stanie, co jest dokładniejsze niż tylko sama zmienna state. Wg poprzednich obserwacji, istnieje korelacja między atrybutem bed a bath, więc najprawdopodobniej dlatego bath jest trochę niżej w ważności.

Propozycja rozwiązania.

Przeanalizowano dane, wybrano informatywne kolumny i spreparowano parę regresorów– głęboką sieć neuronową, Extreme Gradient Boosting (xgboost) oraz drzewo decyzyjne (Decision Tree). Dla regresora Extreme Gradient Boosting zastosowano RandomCV Search – w ten sposób znaleziono najlepsze hiperparametry modelu.

Regresory były szkolone na tych samych danych treningowych i testowych. Stosunek train-test split wynosił 0.2.

Przyjęte parametry dla XGBoosta:

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

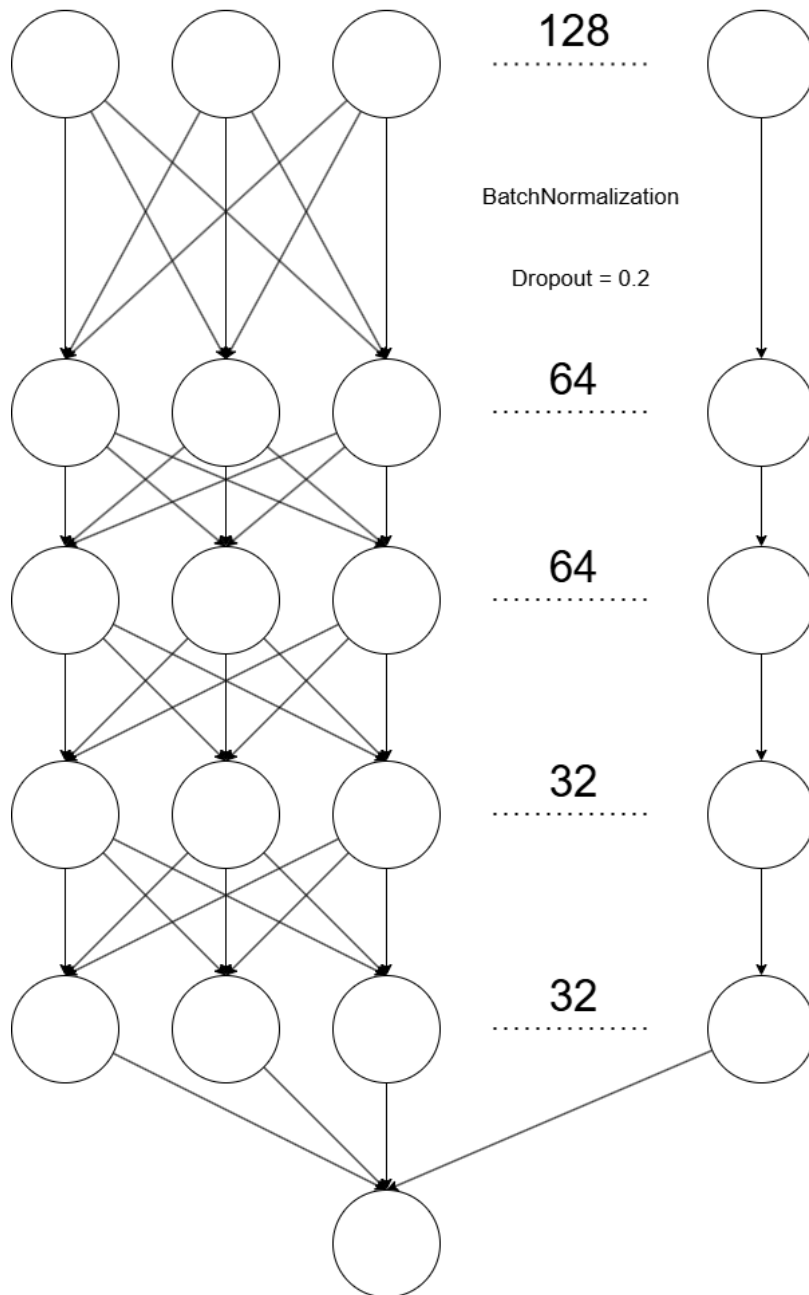
Przyjęte parametry dla Decision Tree:

max_depth	None
min_samples_split	2

Przyjęte parametry oraz architektura sieci neuronowej:

	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



Wyniki i konkluzje.

Atrybut zip_code.

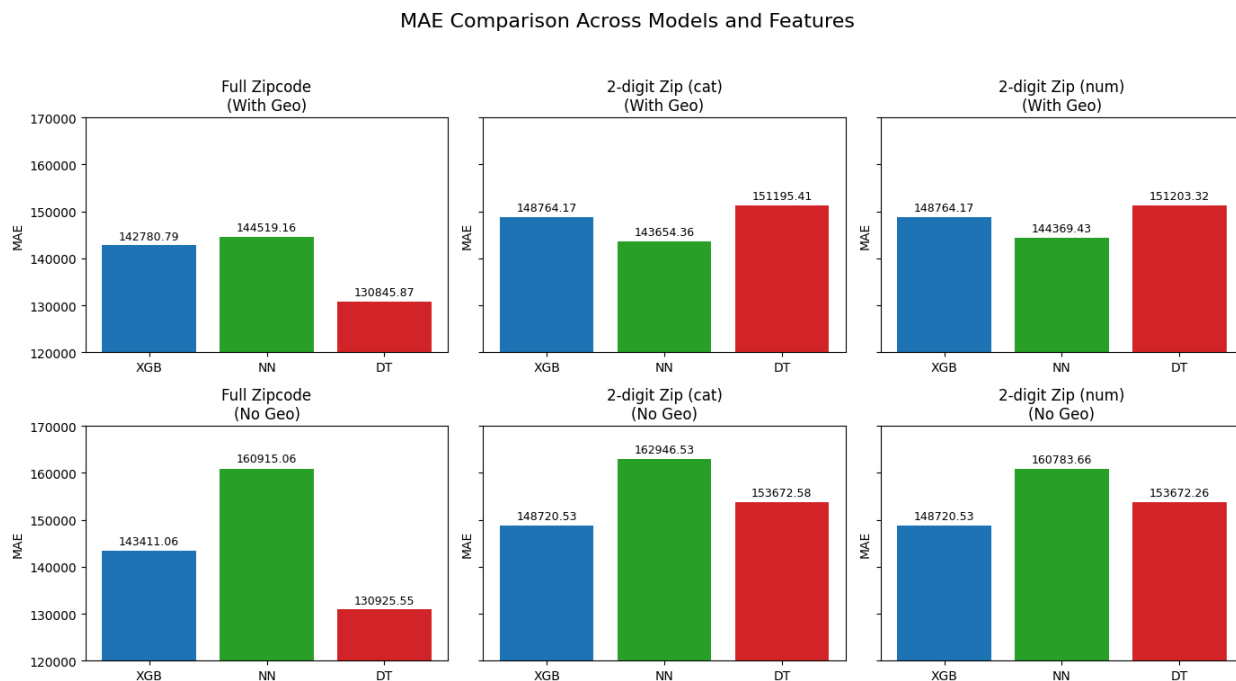
W pewnym momencie wystąpił problem z obsługą atrybutu zip_code. Po przeanalizowaniu mapy wynikło, że dwie-trzy pierwsze cyfry oznaczają region w USA. Część z tych regionów występowała obok siebie, zgodnie z numeracją. Postanowiono przetestować trzy różne podejścia do tego atrybutu.

- 1) Numeryczna wartość całego kodu pocztowego.
- 2) Kategoryczna wartość dwóch pierwszych cyfr kodu pocztowego.
- 3) Numeryczna wartość dwóch pierwszych cyfr kodu pocztowego.

Każdy z trzech klasyfikatorów wyszkolono z tak sprecyzowaną kolumną zip_code i porównano wyniki. Wykresy pokazujące wpływ tych trzech podejść znajdują się w następnej sekcji.

Hipoteza: cechy lokalizacyjne wpływają na cenę nieruchomości.

Przy analizie eksploracyjnej postawiono hipotezę zerową: **Nieuwzględnienie cech lokalizacyjnych nieruchomości nie zmniejsza jakości predykcji ceny nieruchomości.** Wybrane dane geograficzne to kateryczne cechy usytuowania - miasto oraz stan. Przyjmujemy $\alpha = 0.05$. Przetestowano łącznie 18 modeli - 3 regresory na 6 zbiorach danych (dane geograficzne i brak danych, oraz trzy różne podejścia do kwestii kodów pocztowych).



Na każdym wykresie słupki odpowiadają mean average error (MAE) każdego z podejść - XGB - XGBoost, NN - głęboka sieć neuronowa, DT - drzewo decyzyjne. W macierzy sześciu wykresów, pierwszy wiersz odpowiada wykorzystaniu wszystkich danych, a drugi odpowiada odrzuceniu danych lokalizacyjnych. Pierwsza kolumna odpowiada domyślnemu wykorzystaniu cechy kodu pocztowego, druga, traktowanie jego pierwszych dwóch cyfr jako kategorii, a trzecia, traktowanie jego pierwszych dwóch cyfr jako atrybutu numerycznego

Przetestowano hipotezy dla każdego rodzaju regresora w każdej wersji.

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

Przyjmujemy hipotezę zerową dla par modeli: XGBoost z dwucyfrowym kategorialnym kodem pocztowym, XGBoost z dwucyfrowym numerycznym kodem pocztowym oraz Decision Tree z nieprzetworzonym kodem pocztowym.

Warunek $p < \alpha$ (odrzućcie hipotezy zerowej i przyjęcie alternatywnej) spełniony jest dla następujących modeli:

- XGBoost z pełnym kodem pocztowym
- Każda sieć neuronowa (niezależnie od doboru cech)
- Drzewo decyzyjne (wersje z dwucyfrowym kodem pocztowym)

Łącznie przyjęto hipotezę zerową dla trzech regresorów, a odrzucono ją i przyjęto alternatywną dla sześciu.

Otrzymane wyniki z hipotez nie są jednoznaczne. Można posunąć się do wniosku, że na jakość predykcji wpływają dwie rzeczy - dane (lub ich brak) oraz sam typ regresora. Dla braku danych lokalizacyjnych regresory takie jak sieci neuronowe mogą mieć problem z celną predykcją, gdy te dane rzeczywiście są statystycznie istotne (co było zweryfikowane za pomocą modułu SelectKBest z biblioteki scikit-learn). Dla bardziej "schludnych" danych (2-cyfrowy kod pocztowy) XGBoost nie zanotował statystycznie ważnego spadku jakości. Sytuacja ma się zupełnie na odwrót w przypadku drzew decyzyjnych. Natomiast każda wersja sieci neuronowych odnotowała istotny spadek w jakości predykcji.

Bibliografia.

1. Ho, W. K. O., Tang, B. S., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48–70.
<https://doi.org/10.1080/09599916.2020.1832558>

2. Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, Housing Price Prediction via Improved Machine Learning Techniques, Procedia Computer Science, Volume 174, 2020, Pages 433-442, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.06.111>
<https://www.sciencedirect.com/science/article/pii/S1877050920316318>
3. Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur
https://thesai.org/Downloads/Volume12No12/Paper_91-Advanced_Machine_Learning_Algorithms.pdf
4. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324