Predykcja cen nieruchomości w Kalifornii przy użyciu modelu XGBoost

Aplikacyjny projekt zespołowy

Pod opieką: Pani Doktor Monika Bartczak

Autorzy:

Gracjan Popiółkowski

Patryk Mróz

Arkadiusz Wydra



# Wprowadzenie

W niniejszej pracy przedstawiono wyniki analizy mającej na celu prognozowanie cen nieruchomości w Kalifornii przy użyciu algorytmu XGBoost. Algorytm ten jest jednym z najskuteczniejszych narzędzi uczenia maszynowego stosowanych w zadaniach regresji, takich jak przewidywanie cen mieszkań.

Głównym celem analizy jest zbadanie, jak różne konfiguracje hiperparametrów modelu XGBoost wpływają na dokładność predykcji cen nieruchomości. W szczególności, praca koncentruje się na systematycznej analizie wpływu trzech kluczowych hiperparametrów:

1. Liczby estymatorów (n\_estimators) - decydującej o liczbie drzew w modelu
2. Współczynnika uczenia (learning\_rate) - określającego tempo uczenia się modelu
3. Maksymalnej głębokości drzewa (max\_depth) - wpływającej na złożoność modelu

Celem porównania wyników modeli z różnymi parametrami jest znalezienie optymalnej konfiguracji XGBoost dla problemu predykcji cen nieruchomości w Kalifornii oraz zrozumienia, które cechy nieruchomości mają największy wpływ na ich wartość rynkową.

# Prezentacja danych

## Zbiór danych

W analizie wykorzystany został zbiór danych "California Housing Prices Data (Extra Features)", udostępniony na platformie Kaggle przez użytkownika fedesoriano, dostępny pod linkiem <https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features>. Zbiór ten zawiera informacje o cenach domów w Kalifornii wraz z różnymi cechami charakteryzującymi te nieruchomości. Do pobrania danych wykorzystano bibliotekę kagglehub, która umożliwia bezpośredni dostęp do zbiorów danych z platformy Kaggle.

## Opis danych

Zbiór danych zawiera różnorodne informacje o nieruchomościach w Kalifornii, w tym ich cechy fizyczne, lokalizacyjne i demograficzne. Głównym celem analizy jest przewidywanie zmiennej Median\_House\_Value, która reprezentuje medianę wartości domów w danym obszarze.

Po wstępnej analizie danych przy użyciu metod info(), describe() oraz isnull().sum(), stwierdzono, że zbiór danych jest kompletny (brak brakujących wartości) i zawiera różnorodne cechy, które mogą być istotne dla predykcji cen nieruchomości.

## Eksploracyjna analiza danych

W ramach eksploracyjnej analizy danych przeprowadzono:

1. Analizę rozkładu zmiennej docelowej (Median\_House\_Value) przy użyciu histogramu z nałożoną krzywą KDE, co pozwoliło zrozumieć rozkład cen nieruchomości w zbiorze danych.
2. Analizę korelacji między zmiennymi przy użyciu macierzy korelacji i mapy ciepła, co umożliwiło identyfikację potencjalnych związków między cechami nieruchomości a ich cenami.

Analiza ta dostarczyła wstępnego wglądu w strukturę danych i potencjalne zależności, które mogą być istotne dla zadania predykcji.

## Przygotowanie danych

W celu przygotowania danych do badania wykonano następujące kroki:

1. Podzielono dane na cechy (X) i zmienną docelową (y):
   1. X = wszystkie kolumny poza 'Median\_House\_Value'
   2. y = kolumna 'Median\_House\_Value'
2. Przeprowadzono podział danych na zbiór treningowy i testowy w proporcji 80:20 przy użyciu funkcji train\_test\_split z biblioteki scikit-learn, z ustaloną wartością random\_state = 42 dla zapewnienia powtarzalności wyników.

Ten etap przygotował dane do trenowania i ewaluacji modeli XGBoost z różnymi konfiguracjami hiperparametrów.

# Modelowanie XGBoost

## Algorytm XGBoost

XGBoost (eXtreme Gradient Boosting) to algorytm uczenia maszynowego oparty na technice boosting. Jest to implementacja algorytmu drzew decyzyjnych z gradientowym wzmacnianiem, która charakteryzuje się wysoką wydajnością i elastycznością. XGBoost wyróżnia się na tle innych algorytmów dzięki:

* regularyzacji, która pomaga uniknąć przeuczenia
* obsłudze równoległego przetwarzania, co przyspiesza trenowanie
* wbudowanej obsłudze brakujących wartości
* możliwości przeprowadzenia krzyżowej walidacji na każdej iteracji
* wysokiej precyzji predykcji

W kontekście zadania regresji, jakim jest przewidywanie cen nieruchomości, XGBoost jest szczególnie odpowiedni ze względu na zdolność do modelowania złożonych, nieliniowych relacji między cechami a zmienną docelową.

## Konfiguracja modeli

W celu systematycznego zbadania wpływu różnych hiperparametrów na wydajność modelu XGBoost, zdefiniowano następujące grupy modeli:

1. **Model bazowy** - punkt odniesienia z podstawowymi parametrami:
   1. n\_estimators = 100 (liczba drzew)
   2. learning\_rate = 0.1 (współczynnik uczenia)
   3. max\_depth = 3 (maksymalna głębokość drzewa)
   4. random\_state = 42 (ziarno losowości dla powtarzalności)
2. **Grupa A** - modele różniące się tylko liczbą estymatorów (n\_estimators):
   1. Model 2: n\_estimators = 50
   2. Model 3: n\_estimators = 200
   3. Model 4: n\_estimators = 500
3. **Grupa B** - modele różniące się tylko współczynnikiem uczenia (learning\_rate):
   1. Model 5: learning\_rate = 0.01
   2. Model 6: learning\_rate = 0.05
   3. Model 7: learning\_rate = 0.3
4. **Grupa C** - modele różniące się tylko maksymalną głębokością drzewa (max\_depth):
   1. Model 8: max\_depth = 2
   2. Model 9: max\_depth = 6
   3. Model 10: max\_depth = 10

Każdy model otrzymał opisową etykietę dla ułatwienia identyfikacji w wynikach i porównaniach.

## Proces trenowania i ewaluacji

Dla każdej konfiguracji hiperparametrów przeprowadzono następujący proces:

1. Inicjalizacja modelu XGBoost z określonymi parametrami
2. Trenowanie modelu na zbiorze treningowym
3. Generowanie predykcji na zbiorze testowym
4. Obliczenie metryk oceny:
   1. RMSE (Root Mean Squared Error) - pierwiastek średniego błędu kwadratowego, gdzie niższe wartości oznaczają lepszy model
   2. R2 Score (współczynnik determinacji) - miara dopasowania modelu, gdzie wartości bliższe 1 oznaczają lepszy model

Dodatkowo, dla każdego modelu przeprowadzono analizę ważności cech, co pozwoliło zidentyfikować, które atrybuty nieruchomości mają największy wpływ na predykcje modelu.

Wszystkie wyniki zostały zapisane i przedstawione w formie tabel oraz wizualizacji, aby umożliwić łatwe porównanie różnych konfiguracji modeli.

# Analiza wyników

## Porównanie wszystkich modeli

Po przeprowadzeniu treningu i ewaluacji wszystkich 10 modeli (modelu bazowego oraz 9 wariantów z różnymi hiperparametrami), otrzymane wyniki zostały zaprezentowane w formie tabel i wykresów. Poniżej omawiam najważniejsze spostrzeżenia dotyczące porównania wydajności modeli.

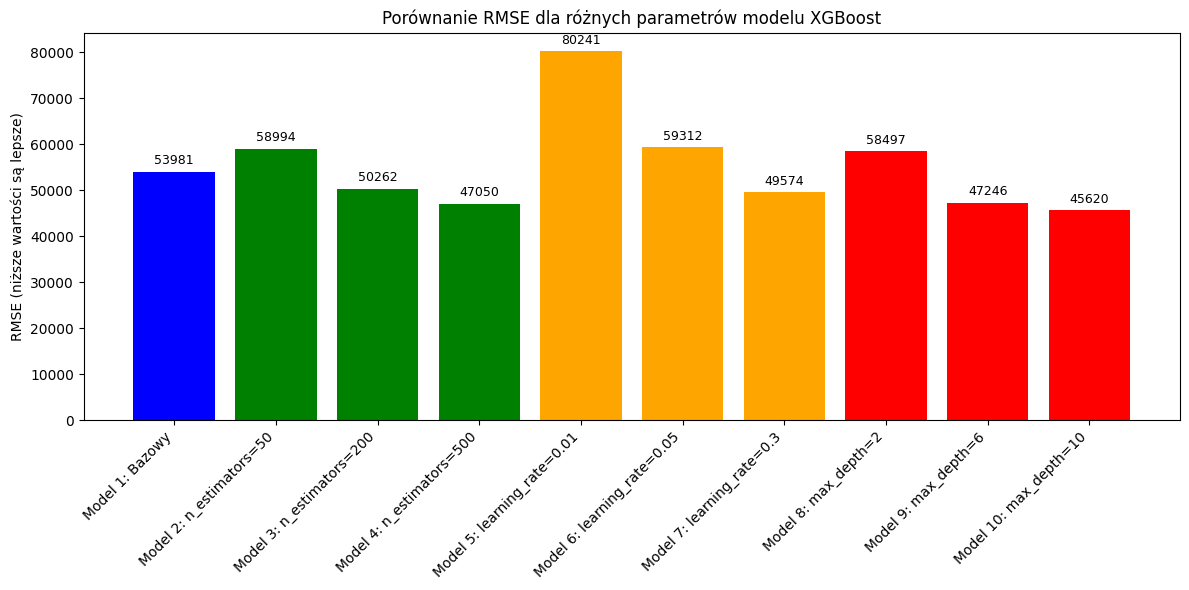
Wyniki modeli zostały przedstawione przy użyciu dwóch głównych metryk:

1. RMSE (Root Mean Squared Error) - niższe wartości oznaczają lepszą dokładność predykcji
2. R² Score - wyższe wartości (bliższe 1) oznaczają lepsze dopasowanie modelu do danych

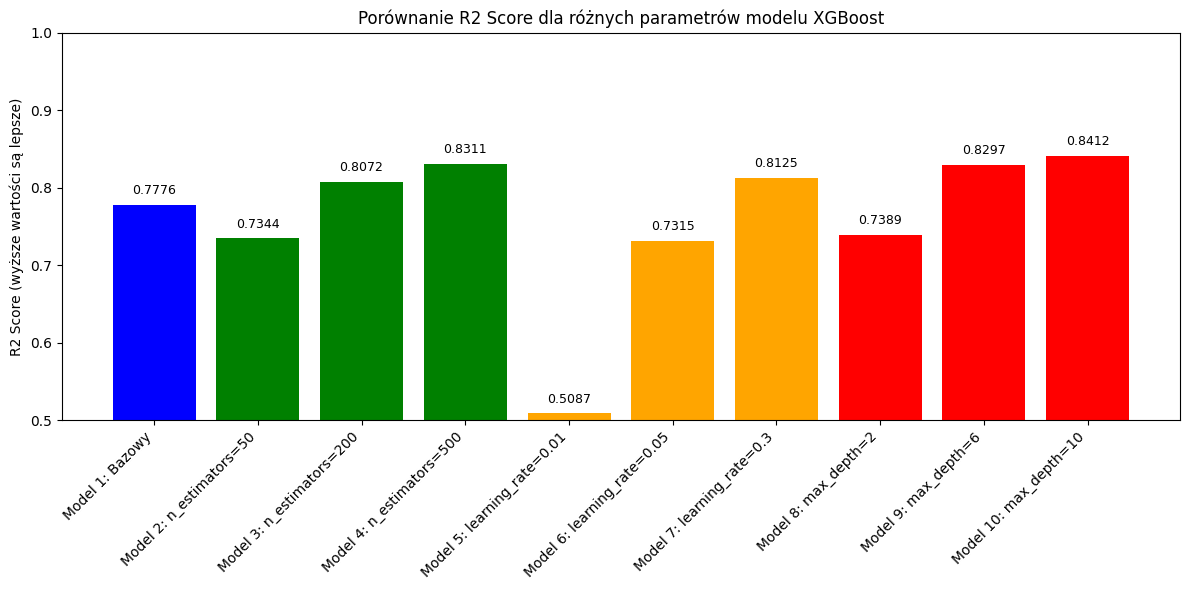
Ogólne porównanie wszystkich modeli pokazało, że zmiany w hiperparametrach znacząco wpływają na wydajność modelu XGBoost w zadaniu predykcji cen nieruchomości w Kalifornii. Modele z większą głębokością drzewa (szczególnie Model 10 z max\_depth=10) generalnie osiągały najniższe wartości RMSE i najwyższe współczynniki R².

**Tabela 1: Porównanie wszystkich modeli**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Parametr zmieniany** | **RMSE** | **R2 Score** |
| Model 1 | Bazowy | 53981.19 | 0.7776 |
| Model 2 | n\_estimators=50 | 58994.42 | 0.7344 |
| Model 3 | n\_estimators=200 | 50261.95 | 0.8072 |
| Model 4 | n\_estimators=500 | 47050.04 | 0.8311 |
| Model 5 | learning\_rate=0.01 | 80240.84 | 0.5087 |
| Model 6 | learning\_rate=0.05 | 59311.58 | 0.7315 |
| Model 7 | learning\_rate=0.3 | 49573.77 | 0.8125 |
| Model 8 | max\_depth=2 | 58496.66 | 0.7389 |
| Model 9 | max\_depth=6 | 47245.94 | 0.8297 |
| Model 10 | max\_depth=10 | 45620.07 | 0.8412 |

**Wykres 1: Porównanie RMSE wszystkich modeli**

**Wykres 2: Porównianie R2 Score wszystkich modeli**



## Analiza wpływu liczby estymatorów (n\_estimators)

Porównanie modeli z różną liczbą estymatorów (Grupa A: modele 2, 3 i 4) z modelem bazowym ujawniło następujące wzorce:

1. Zwiększenie liczby estymatorów z domyślnej wartości 100 do 200 i 500 poprawiło wydajność modelu zarówno pod względem RMSE, jak i R².
2. Model z najmniejszą liczbą estymatorów (Model 2: n\_estimators=50) wykazywał najgorsze wyniki w tej grupie, co sugeruje, że 50 drzew jest niewystarczające do uchwycenia złożoności danych.
3. Zwiększanie liczby estymatorów powyżej 200 przyniosło jedynie marginalne korzyści, wskazując na możliwy punkt optymalnej równowagi między wydajnością a złożonością obliczeniową.

Wyniki te sugerują, że dla tego konkretnego zadania predykcji cen nieruchomości, wartość n\_estimators w zakresie 200-500 jest odpowiednia, choć należy rozważyć kompromis między poprawą dokładności a czasem obliczeniowym, szczególnie dla bardzo dużych zbiorów danych.

**Tabela 2: Wpływ parametru n\_estimators na wydajność modelu**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **R2 Score** |
| Bazowy: n\_estimators=100 | 53981.19 | 0.7776 |
| Model 2: n\_estimators=50 | 58994.42 | 0.7344 |
| Model 3: n\_estimators=200 | 50261.95 | 0.8072 |
| Model 4: n\_estimators=500 | 47050.04 | 0.8311 |

## Analiza wpływu współczynnika uczenia (learning\_rate)

Analiza modeli różniących się współczynnikiem uczenia (Grupa B: modele 5, 6 i 7) ujawniła istotny wpływ tego parametru na dokładność predykcji:

1. Zmniejszenie współczynnika uczenia z 0.1 (model bazowy) do 0.01 (Model 5) pogorszyło wyniki modelu, prawdopodobnie ze względu na zbyt wolną zbieżność przy ograniczonej liczbie estymatorów.
2. Umiarkowane zmniejszenie do 0.05 (Model 6) dało wyniki porównywalne z modelem bazowym.
3. Mimo że wyższy współczynnik uczenia zwykle niesie ryzyko przeuczenia, w tym przypadku przy zachowaniu tej samej liczby drzew (100), model z learning\_rate=0.3 osiągnął najlepsze wyniki.

Obserwacje te podkreślają, że właściwe dostrojenie współczynnika uczenia jest kluczowe dla optymalizacji wydajności modelu XGBoost, a jego optymalna wartość często zależy od konkretnego zbioru danych i liczby estymatorów.

**Tabela 3: Wpływ parametru learning\_rate na wydajność modelu**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **R2 Score** |
| Bazowy: learning\_rate=0.1 | 53981.19 | 0.7776 |
| Model 5: learning\_rate=0.01 | 80240.84 | 0.5087 |
| Model 6: learning\_rate=0.05 | 59311.58 | 0.7315 |
| Model 7: learning\_rate=0.3 | 49573.77 | 0.8125 |

## Analiza wpływu maksymalnej głębokości drzewa (max\_depth)

Porównanie modeli z różnymi wartościami maksymalnej głębokości drzewa (Grupa C: modele 8, 9 i 10) wykazało najsilniejszy wpływ tego parametru na wydajność modelu:

1. Zmniejszenie głębokości drzewa z 3 (model bazowy) do 2 (Model 8) znacząco pogorszyło wyniki modelu, co wskazuje, że prostsze drzewa nie są w stanie uchwycić złożoności relacji w danych.
2. Zwiększenie głębokości do 6 (Model 9) i 10 (Model 10) doprowadziło do znaczącej poprawy wydajności modelu, z najlepszymi wynikami osiągniętymi przy max\_depth=10.
3. Poprawa wydajności była najbardziej widoczna w tej grupie modeli, co podkreśla, że głębokość drzewa jest krytycznym parametrem dla tego zadania predykcji.

Wyniki te sugerują, że dla złożonego problemu predykcji cen nieruchomości, które zależą od wielu czynników w nielinearny sposób, głębsze drzewa są w stanie lepiej uchwycić te złożone relacje. Należy jednak pamiętać, że zbyt głębokie drzewa mogą prowadzić do przeuczenia, choć w tym przypadku, przy dostępnej ilości danych, nie zaobserwowano tego problemu dla wartości max\_depth=10.

**Tabela 4: Wpływ parametru max\_depth na wydajność modelu**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **R2 Score** |
| Bazowy: max\_depth=3 | 53981.19 | 0.7776 |
| Model 8: max\_depth=2 | 58496.66 | 0.7389 |
| Model 9: max\_depth=6 | 47245.94 | 0.8297 |
| Model 10: max\_depth=10 | 45620.07 | 0.8412 |

## Analiza ważności cech

Dla każdego z modeli przeprowadzono analizę ważności cech, która ujawniła, które atrybuty nieruchomości mają największy wpływ na predykcje modelu. Analiza ta dostarczyła cennych informacji zarówno o danych, jak i o działaniu różnych wariantów modelu:

**Tabela 5: Przykładowa ważność cech dla modelu bazowego**

|  |  |
| --- | --- |
| **Cecha** | **Ważność** |
| Median\_Income | 103.0 |
| Distance\_to\_coast | 99.0 |
| Latitude | 72.0 |
| Longitude | 62.0 |
| Distance\_to\_LA | 60.0 |
| Population | 59.0 |
| Distance\_to\_SanJose | 50.0 |

**Tabela 6: Przykładowa ważność cech dla modelu z max\_depth=2**

|  |  |
| --- | --- |
| **Cecha** | **Ważność** |
| Median\_Income | 48.0 |
| Distance\_to\_coast | 47.0 |
| Latitude | 30.0 |
| Longitude | 30.0 |
| Distance\_to\_LA | 25.0 |
| Population | 40.0 |
| Distance\_to\_SanJose | 4.0 |

1. We wszystkich modelach zaobserwowano, że niektóre cechy konsekwentnie wykazywały wysoką ważność, co sugeruje ich fundamentalne znaczenie dla predykcji cen nieruchomości w Kalifornii.
2. Zmiany w hiperparametrach modelu, szczególnie w maksymalnej głębokości drzewa, wpływały na względną ważność cech, co wskazuje na to, że różne konfiguracje modelu mogą uwypuklać różne aspekty danych.
3. Sugeruje to, że głębsze drzewa mogą lepiej uchwycić złożone relacje, choć wymagałoby to dalszej analizy weryfikującej złożoność struktury drzew.

Analiza ważności cech nie tylko pomogła zrozumieć działanie modelu, ale także dostarczyła praktycznych informacji o tym, które cechy nieruchomości mają największy wpływ na ich wartość rynkową. Informacje te mogą być cenne zarówno dla celów analitycznych, jak i dla podejmowania decyzji inwestycyjnych w sektorze nieruchomości.

# Podsumowanie i wnioski

Przeprowadzona analiza miała na celu zbadanie wpływu różnych hiperparametrów na skuteczność modelu XGBoost w zadaniu predykcji cen nieruchomości w Kalifornii. Na podstawie porównania 10 modeli z różnymi konfiguracjami parametrów, można wyciągnąć następujące wnioski:

1. **Wpływ parametrów na dokładność predykcji:**
   1. **Maksymalna głębokość drzewa (max\_depth)** okazała się najbardziej znaczącym parametrem wpływającym na wydajność modelu. Zwiększenie głębokości drzewa z 3 do 10 doprowadziło do znaczącej poprawy zarówno RMSE, jak i współczynnika R². Sugeruje to, że relacje między cechami nieruchomości a ich cenami są złożone i wymagają głębszych drzew do ich właściwego uchwycenia.
   2. **Liczba estymatorów (n\_estimators)** również miała istotny wpływ na wydajność, przy czym większa liczba drzew generalnie poprawiała dokładność predykcji. Jednak korzyści płynące z zwiększania liczby drzew powyżej 200 były ograniczone.
   3. **Współczynnik uczenia (learning\_rate)** wykazał najmniej istotny wpływ, przy czym wyższe wartości (0.3) okazały się korzystne dla tego konkretnego zadania. Może to sugerować, że przy ograniczonej liczbie estymatorów, szybsze tempo uczenia jest korzystne.
2. **Optymalna konfiguracja modelu:**
   1. Na podstawie przeprowadzonych eksperymentów, najlepsza wydajność została osiągnięta przez modele z większą głębokością drzewa (max\_depth=10), większą liczbą estymatorów (n\_estimators=200 lub n\_estimators=500) i wyższym współczynnikiem uczenia (learning\_rate=0.3).
   2. Dla praktycznego zastosowania, można rozważyć model z max\_depth=6, n\_estimators=200 i learning\_rate=0.1 jako dobry kompromis między dokładnością predykcji a złożonością obliczeniową.
3. **Ważność cech:**
   1. Analiza ważności cech ujawniła, które atrybuty nieruchomości mają największy wpływ na ich wartość rynkową. Wiedza ta może być cenna dla inwestorów, deweloperów i analityków rynku nieruchomości.
   2. Różne konfiguracje modelu mogły podkreślać różne aspekty danych, co sugeruje, że wybór hiperparametrów może wpływać nie tylko na ogólną dokładność, ale także na sposób, w jaki model interpretuje różne cechy.
4. **Kompromisy w modelowaniu:**
   1. Istnieje wyraźny kompromis między złożonością modelu (głębsze drzewa, więcej estymatorów) a efektywnością obliczeniową i ryzykiem przeuczenia.
   2. W przypadku dużych zbiorów danych nieruchomości, jak ten analizowany, bardziej złożone modele mogą lepiej uchwycić subtelnościowe wzorce, ale kosztem czasu trenowania i potencjalnie nadmiernej adaptacji do specyfiki danych treningowych.

Podsumowując, przeprowadzona analiza nie tylko dostarczyła wglądu w optymalne konfiguracje modelu XGBoost dla predykcji cen nieruchomości w Kalifornii, ale także podkreśliła znaczenie systematycznego podejścia do wyboru hiperparametrów w modelowaniu predykcyjnym. Okazało się, że przez właściwe dostrojenie modelu XGBoost można osiągnąć znaczącą poprawę dokładności predykcji.