Prognozowanie Cen Nieruchomości z Wykorzystaniem Zaawansowanych Modeli Regresji

Kompleksowa Analiza Wyceny Nieruchomości na Podstawie Zestawu Danych "kc_house"

Łukasz Syguła
02.08.2024

Cel projektu:

Opis: Rozwój i ocena różnych modeli regresji do przewidywania cen nieruchomości.

Etapy:

- Przygotowanie danych
- Tworzenie modeli
- Ewaluacja modeli
- Optymalizacja i strojenie
- Raportowanie wyników

Zbiór danych: "kc_house"

Opis danych:

- Dane sprzedaży domów w King County, Washington (USA) z lat 2014-2015.

Kluczowe zmienne: Cena, liczba sypialni, liczba łazienek, powierzchnia mieszkalna, stan, rok budowy, lokalizacja.

Analiza Eksploracyjna (EDA)

Czyszczenie danych:

- Wypełnienie brakujących wartości w `sqft_above`.
- Usunięcie outlierów w kolumnach `bedrooms` i `bathrooms`.

Wizualizacje:

- Histogram cen nieruchomości.
- Boxplot cen nieruchomości.
- Dystrybucja liczby sypialni i łazienek.

Wypełnienie brakujących wartości w `sqft_above`

Filling the two missing values in the 'sqft_above' column with the median to maintain robustness against outliers.

In [8]:
```
median_sqft_above = df['sqft_above'].median()
```

In [9]:
```
df['sqft_above'].fillna(median_sqft_above, inplace=True)
```

In [10]:
```
df.drop(columns='id', inplace=True)
```

In [11]:
```
df
```

Out[11]:

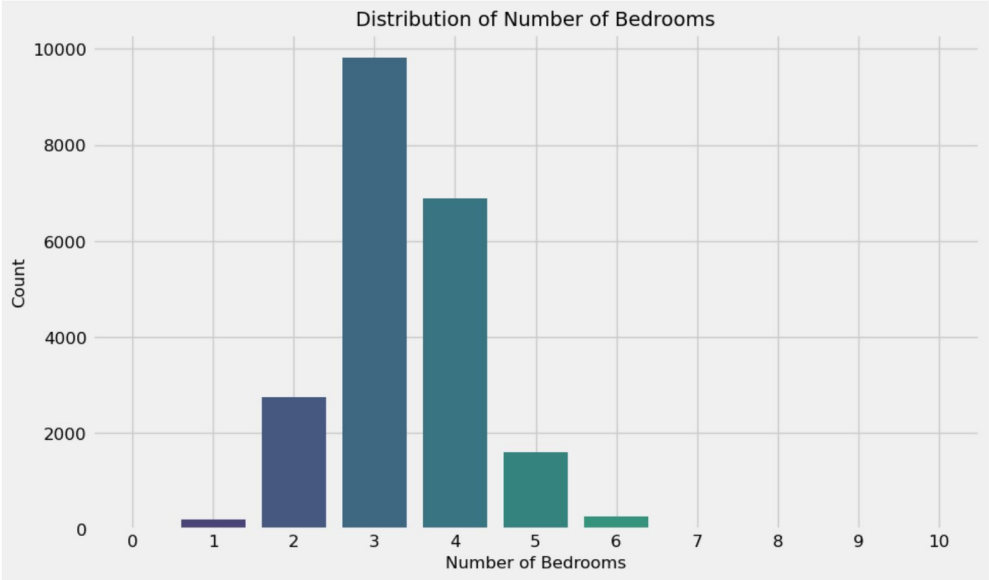
	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqf
0	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	3	7	
1	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	3	7	
2	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	3	6	
3	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	5	7	
4	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	3	8	

Usunięcie outlierów w kolumnie `bedrooms`

Removing outlier values from the 'bedrooms' column to enhance data quality and model accuracy.

In [20]:
```
df = df[df['bedrooms'] <= 10]
```

In [21]:
```
plt.figure(figsize=(10, 6), dpi=100)
sns.countplot(x='bedrooms', data=df, palette='viridis')
plt.title('Distribution of Number of Bedrooms')
plt.xlabel('Number of Bedrooms')
plt.ylabel('Count')
plt.grid(True)
plt.show()
```

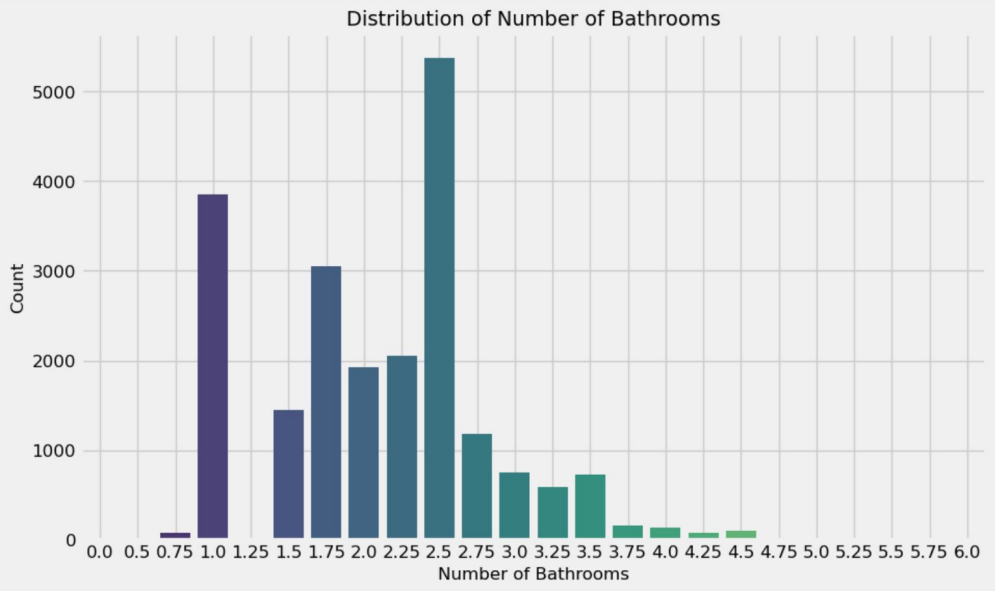


Usunięcie outlierów w kolumnie `bathrooms`

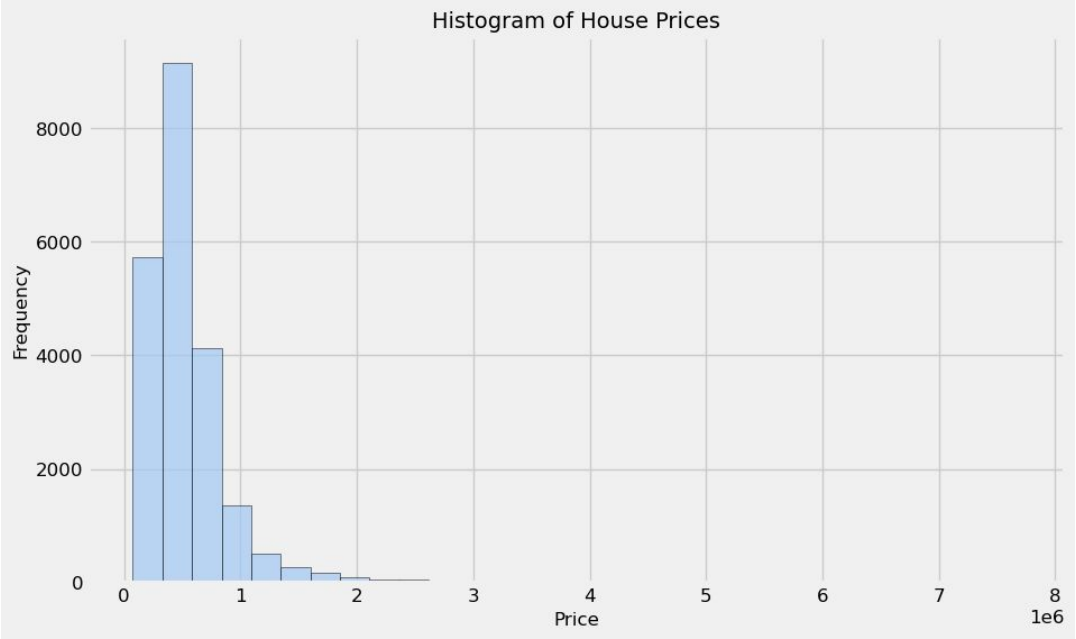
Removing records with 'bathrooms' values greater than 6 to ensure data quality and improve model performance.

In [22]:
```
df = df[df['bathrooms'] <= 6]
```

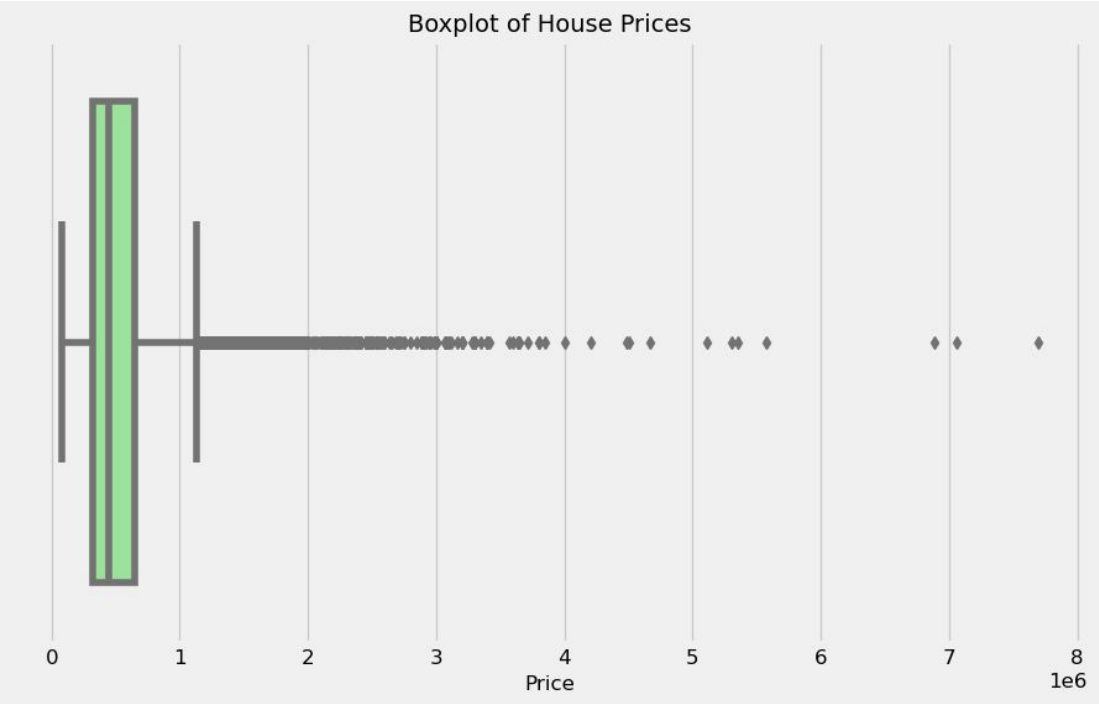
In [23]:
```
plt.figure(figsize=(10, 6), dpi=100)
sns.countplot(x='bathrooms', data=df, palette='viridis')
plt.title('Distribution of Number of Bathrooms')
plt.xlabel('Number of Bathrooms')
plt.ylabel('Count')
plt.grid(True)
plt.show()
```



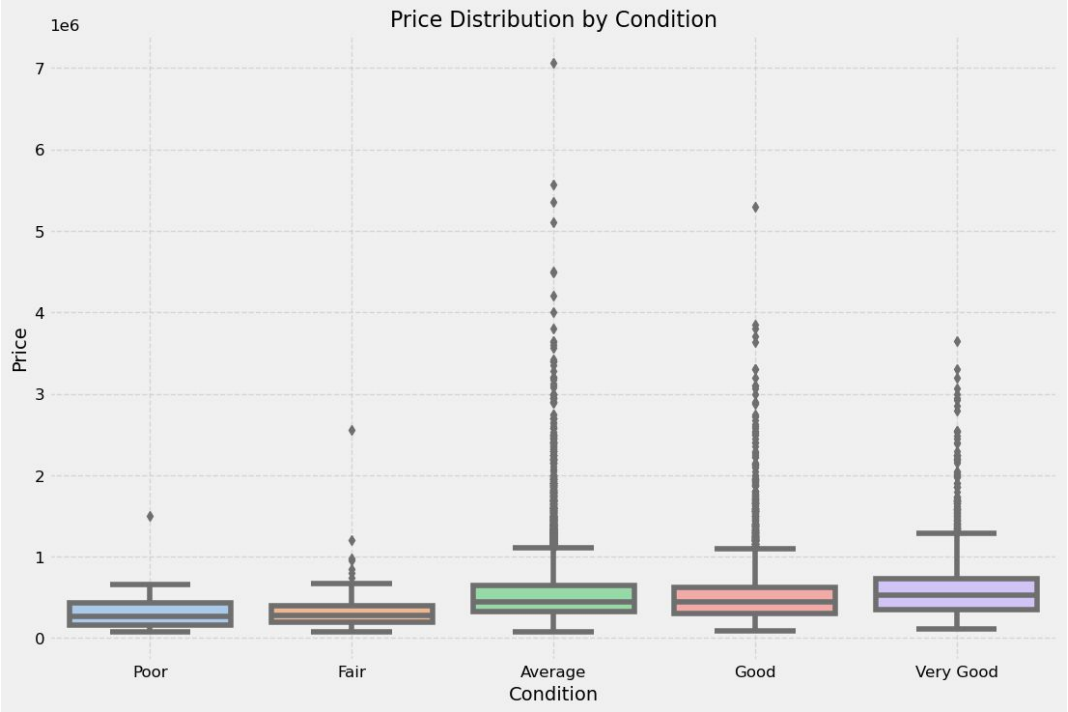
Histogram przedstawia rozkład cen nieruchomości, ukazując, że większość domów ma niższe ceny, z mniejszą liczbą nieruchomości w wyższych przedziałach cenowych.



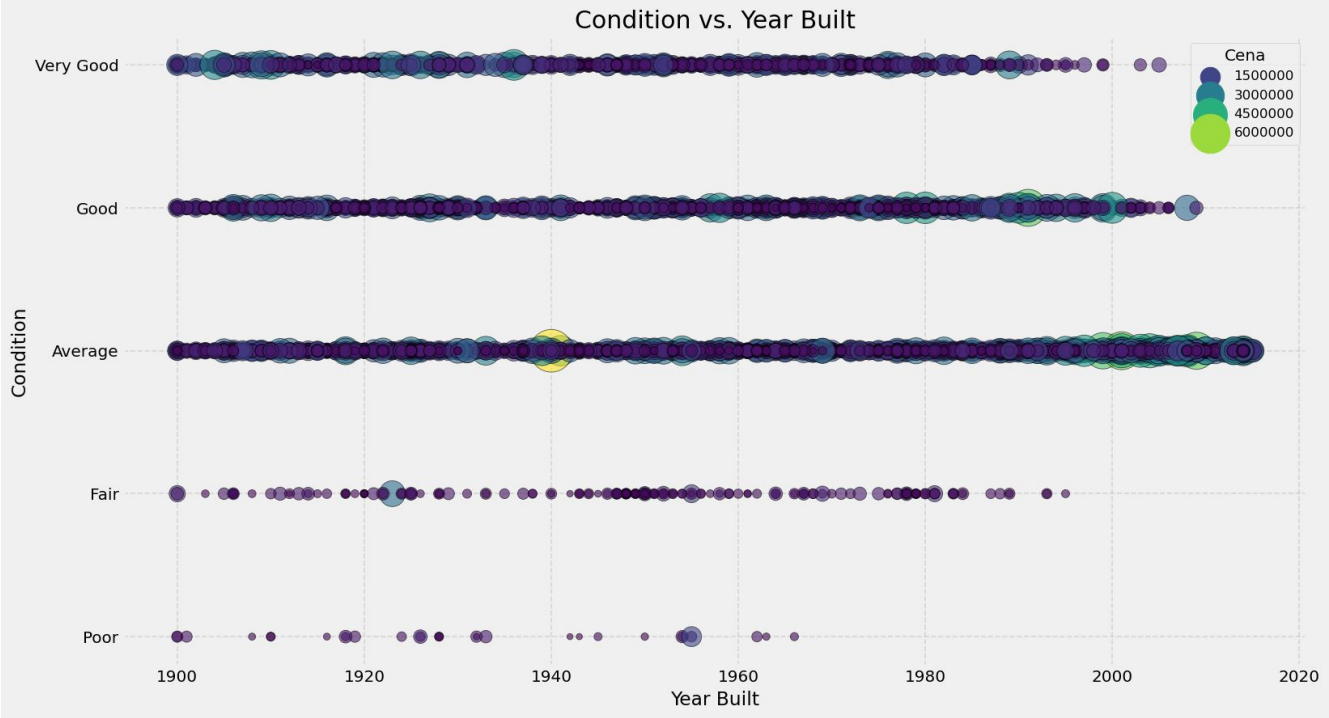
Boxplot ukazuje rozkład cen nieruchomości, wskazując na obecność wartości odstających oraz medianę cen domów.



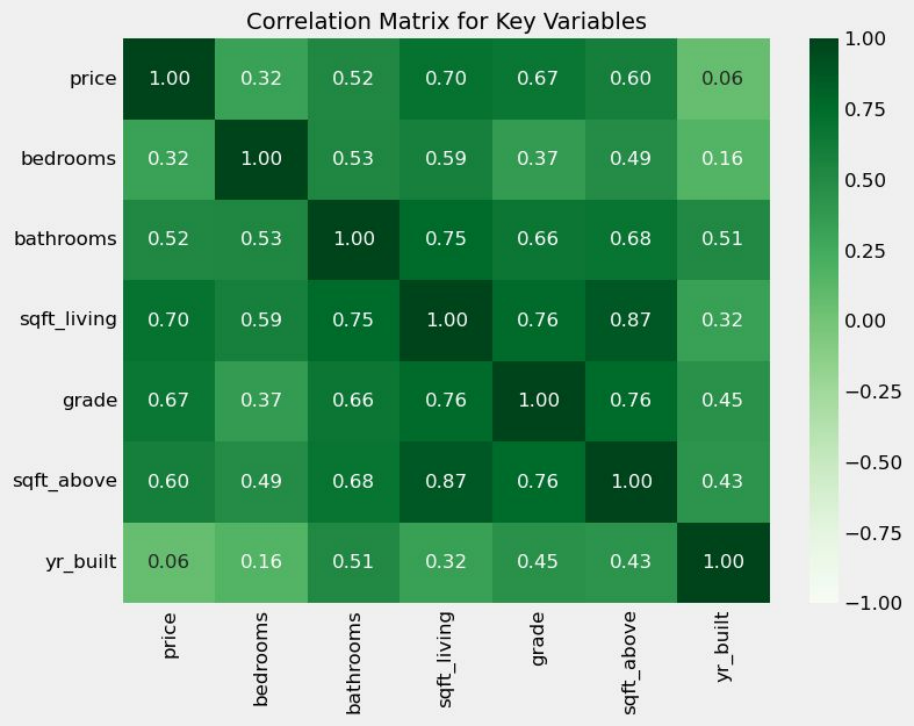
Boxplot przedstawia zależność między stanem nieruchomości a jej ceną. Wykres pokazuje, że lepszy stan techniczny przekłada się na wyższe ceny sprzedaży.



Wykres punktowy pokazuje zależność między rokiem budowy a stanem nieruchomości, z rozmiarem i kolorem punktów reprezentującymi ceny. Widać, że nowsze budynki częściej mają lepszy stan techniczny, co koreluje z wyższymi cenami.



Macierz korelacji pokazuje, które cechy nieruchomości, takie jak liczba pokoi czy rok budowy, mają największy wpływ na cenę.



Modele regresyjne - Implementacja

Przeprowadzone modele:

- Regresja liniowa
- Ridge
- Lasso
- Regresja wielomianowa
- Drzewo decyzyjne
- Las losowy
- Gradient Boosting

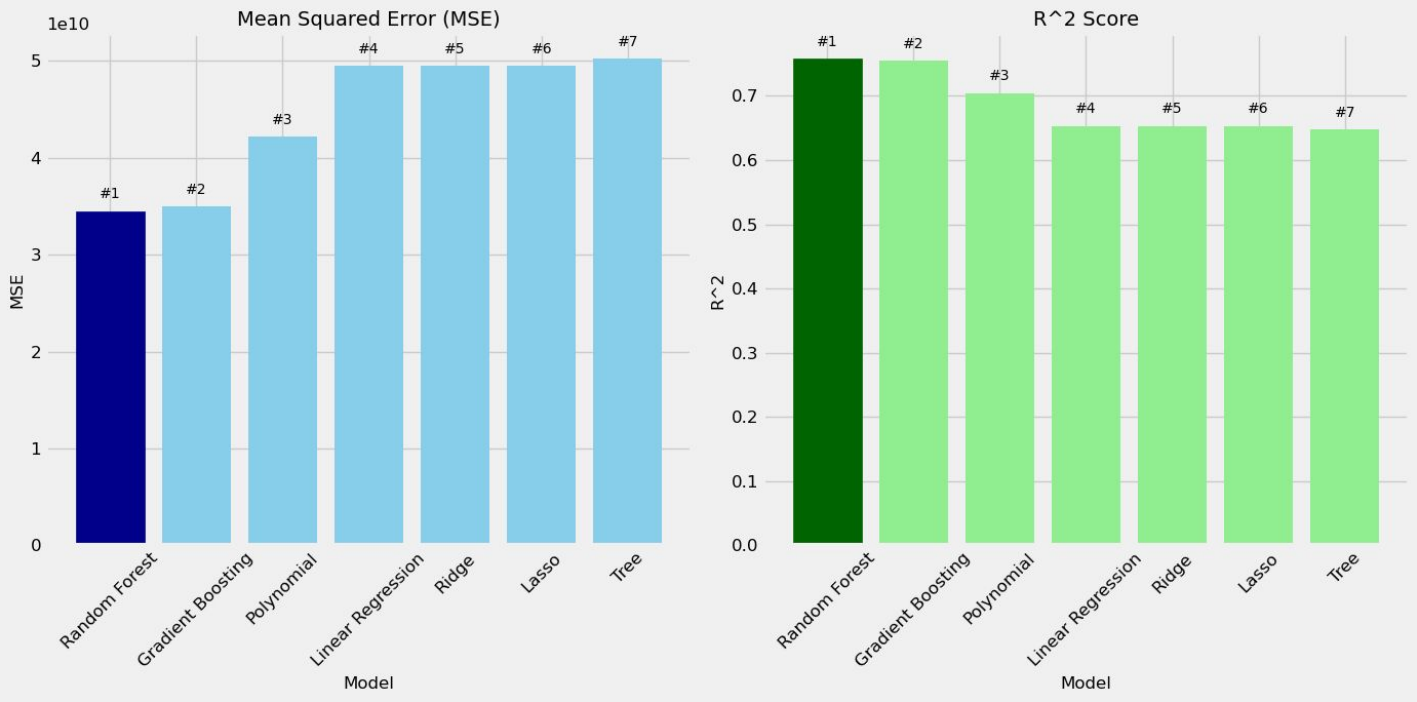
Proces: Podział danych na zestawy treningowy i testowy, normalizacja danych.

Wyniki Modeli - MSE i R^2

Porównanie wyników:

- Wykres słupkowy MSE i R^2 dla wszystkich modeli.
- Podkreślenie najlepszych modeli: Random Forest i Gradient Boosting.

Ten wykres przedstawia porównanie modeli regresyjnych na podstawie ich wartości błędu średniokwadratowego (MSE) oraz współczynnika determinacji (R^2). Modele są uszeregowane od najlepszych do najgorszych w obu kategoriach, co ułatwia identyfikację najbardziej efektywnych algorytmów w przewidywaniu cen nieruchomości.



Wnioski

Podsumowanie wyników:

- Random Forest i Gradient Boosting uzyskały najlepsze wyniki.
- Regresja wielomianowa pokazała znaczny wzrost dokładności w stosunku do regresji liniowej.

Modele regularizowane (Ridge i Lasso): Brak istotnej poprawy.

Rekomendacje i Dalsze Kroki

Wybór modelu: Na podstawie analizy danych, rekomenduję wykorzystanie modeli Random Forest lub Gradient Boosting do przewidywania cen nieruchomości. Oba modele uzyskały najlepsze wyniki w zakresie dokładności przewidywań i są najbardziej skuteczne w tym przypadku.
Dalsze kroki:

1. **Optymalizacja Modeli:** Przeprowadzenie dalszej optymalizacji hiperparametrów dla modeli Random Forest i Gradient Boosting, aby maksymalnie poprawić ich wydajność.
2. **Testowanie na Większych Zbiorach Danych:** Rozszerzenie testów na większe zbiory danych, aby sprawdzić, czy modele zachowują wysoką jakość prognoz w różnych kontekstach.
3. **Walidacja Krzyżowa:** Wykonanie bardziej zaawansowanej walidacji krzyżowej w celu oceny stabilności i generalizacji modeli na różnych podzbiorach danych.
4. **Analiza Cech:** Dokonanie pogłębionej analizy znaczenia cech, aby zrozumieć, które zmienne mają największy wpływ na ceny nieruchomości i ewentualnie przeprowadzenie dalszych prac nad ich selekcją.
5. **Udoskonalenie Modelu:** Rozważenie wykorzystania zaawansowanych technik, takich jak kombinacje modeli (ensemble learning) lub zastosowanie głębokiego uczenia, jeśli dostępne są większe zbiory danych i zasoby obliczeniowe.