Marcin Dugiełło

e-mail: marcin.dugiello@gmail.com

ORCID: 0009-0002-3645-3442

Uniwersytet Ekonomiczny we Wrocławiu

Porównanie jakości modeli prognozowania na podstawie cen transakcyjnych nieruchomości mieszkalnych we Wrocławiu

DOI: 10.15611/2024.80.2.04
JEL Classification: I12, I14

Cytuj jako: Dugiełło, M. (2024). Porównanie jakości modeli prognozowania na podstawie cen transakcyjnych nieruchomości mieszkalnych we Wrocławiu. W: H. Dudycz (red.), *Informatyka w biznesie* (s. 48-62). Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

Streszczenie: Badanie koncentruje się na porównaniu skuteczności metod uczenia maszynowego, takich jak lasy losowe oraz regresja liniowa w kontekście wyceny nieruchomości mieszkalnych na podstawie danych transakcyjnych. Głównym celem artykułu jest ocena zdolności tych metod do precyzyjnego prognozowania cen nieruchomości. Artykuł rozpoczyna się od teoretycznego wprowadzenia do wyceny nieruchomości, omawiając definicje i istniejące podejścia w Polsce. Następnie przedstawione są założenia badania empirycznego zawierające szczegółowy opis zbioru danych oraz używane zmienne. Analiza skupia się na rynku mieszkalnym we Wrocławiu w latach 2014-2023. Zastosowano metodę GridSearchCV do optymalizacji parametrów modeli predykcyjnych. Wnioski płynące z badania pokazują potencjał modeli nieliniowych w estymacji cen nieruchomości i jednocześnie podkreślają znaczenie odpowiedniej kalibracji hiperparametrów przed przystąpieniem do estymacji.

Słowa kluczowe: nieruchomości, wycena nieruchomości, modele prognozowania, uczenie maszynowe, liniowa regresja, las losowy, ANOVA, hiperparametry, GridSearchCV

1. Wstęp

Zakup nieruchomości jest jedną z kluczowych decyzji finansowych w życiu, mającą wpływ nie tylko na nas samych, ale i na naszych bliskich. Oszacowanie wartości nieruchomości mieszkalnej to proces wymagający precyzji i zaawansowanej analizy, co stanowi wyzwanie zarówno dla specjalistów, jak i dla przeciętnych osób.

Współczesne technologie oferują narzędzia mogące znacząco wspomóc ten proces. Zaawansowane metody, takie jak uczenie maszynowe oraz podstawowe modele regresji liniowej wykorzystujące ogromne zbiory danych transakcyjnych,

mogą być nieocenioną pomocą w dokładniejszym określaniu wartości nieruchomości. Dzięki nim możliwe jest uzyskanie precyzyjnych wycen, co ma istotne znaczenie dla podejmowania decyzji inwestycyjnych.

Uczenie maszynowe, dzięki swojej zdolności do analizowania dużej liczby danych i wykrywania subtelnych wzorców, oraz modele regresyjne, prognozujące na podstawie historycznych danych, mogą znacząco zwiększyć świadomość potencjalnych nabywców i sprzedawców. To z kolei przyczynia się do poprawy przejrzystości oraz efektywności wyceny na rynku nieruchomości.

Celem artykułu jest sprawdzenie potencjału modeli predykcyjnych w praktycznej wycenie nieruchomości na podstawie rzeczywistych danych oraz porównanie skuteczności liniowej metody regresyjnej z metodą uczenia maszynowego.

W kolejnej części artykułu skoncentrowano się na teorii wyceny nieruchomości, przedstawiono jej definicje, podziały oraz istniejące podejścia do niej w Polsce. W następnej opisano założenia badania empirycznego, w tym modele prognozowania oraz inne narzędzia analizy danych, a potem omówiono szczegółowo kluczowe zmienne i zbiór danych, który posłużył do przeprowadzenia badania. W dalszych opisano przeprowadzoną analizę rynku mieszkalnego we Wrocławiu, bazując na transakcjach nieruchomości z lat 2014-2023. Procedura GridSearchCV została wykorzystana do optymalizacji hiperparametrów modeli prognozujących. Na zakończenie porównano modele przy użyciu wskaźników oceny predykcji regresyjnej oraz histogramów błędów bezwzględnych, prezentujących efektywność obu podejść w kontekście wyceny nieruchomości.

2. Teoria wyceny nieruchomości

Definicja nieruchomości najczęściej cytowana w polskich publikacjach naukowych i literaturze fachowej znajduje swoje umocowanie w ustawie z dnia 23 kwietnia 1964 r. w Kodeksie Cywilnym, w artykule 46 paragrafie 1. Zgodnie z tą definicją: „Nieruchomościami są części powierzchni ziemskiej stanowiące odrębny przedmiot własności (grunty), jak również budynki trwale związane z gruntem lub części takich budynków, jeżeli na mocy przepisów szczególnych stanowią odrębny od gruntu przedmiot własności". Definicja ta podkreśla, że nieruchomość jest przede wszystkim określonym fragmentem ziemi, co czyni ją trwałą i niemożliwą do usunięcia. Jest to definicja klarowna i zwięzła, co ułatwia jej stosowanie w praktyce oraz w terminologii naukowej i prawnej.

Nieruchomości można podzielić na trzy główne kategorie (Ustawa z dnia 23 kwietnia 1964 r.) – gruntowe, budynkowe i lokalowe. Każda z nich ma unikalne cechy i wymaga specyficznego podejścia do wyceny. Grunty to tereny, które zgodnie z przepisami prawnymi obejmują zarówno powierzchnię, jak i przestrzeń nad i pod nią. Kluczowym elementem w wycenie nieruchomości jest więc precyzyjne wyznaczenie granic, co wymaga pomiarów geodezyjnych i procedur prawnych.

Grunty dzielą się na działki niezabudowane (np. rolne, leśne, wodne, kopalne i nieużytki) oraz zabudowane (Malec i Stachura, 2006), które obejmują budynki

i inne stałe obiekty. Struktury zwane naniesieniami, oddzielone od gruntu, klasyfikowane są jako nieruchomości budynkowe. Ten podział jest istotny, gdy różne podmioty posiadają grunt i budynki, jak to ma miejsce w użytkowaniu wieczystym. W takich przypadkach Skarb Państwa jest właścicielem gruntu, a osoby prywatne budynków. W praktyce jednak budynek i grunt są nierozłączne, co sprawia, że pojęcie nieruchomości budynkowej funkcjonuje głównie w podstawie prawnej.

Indywidualnie wydzielone jednostki w budynkach, mające przeznaczenie mieszkalne lub użytkowe, klasyfikowane są jako nieruchomości lokalowe. Samodzielność lokalu zapewnia trwałe oddzielenie go ścianami, co może obejmować także przynależne pomieszczenia, takie jak garaże czy komórki, nawet jeśli nie są bezpośrednio połączone z lokalem. Każdy lokal musi posiadać własną księgę wieczystą, która potwierdza jego niezależność (Kucharska-Stasiak, 2006). W niniejszym badaniu zbiór danych będzie wyłącznie zawierał nieruchomości określone jako lokalowe.

Wycena nieruchomości obejmuje różnorodne podejścia, które pozwalają na dokładne określenie jej wartości w zależności od dostępnych danych i specyficznych uwarunkowań rynkowych (Cymerman i Cymerman, 2024).

Podejście porównawcze polega na wycenie nieruchomości poprzez porównanie cen transakcyjnych nieruchomości o podobnych cechach i lokalizacji. Zakłada się, że ceny, po jakich sprzedawane były te nieruchomości, odzwierciedlają rynkową wartość. Metoda ta jest powszechnie stosowana ze względu na swoją prostotę i oparcie na rzeczywistych danych rynkowych, co czyni ją jednym z najbardziej przystępnych i wiarygodnych sposobów oceny wartości nieruchomości.

Kolejnym podejściem jest metoda dochodowa, która opiera się na ocenie wartości nieruchomości przez analizę przewidywanych przepływów pieniężnych, które są generowane na bieżąco lub mają potencjał do generowania ich w przyszłości. Podejście to wymaga prognozowania przyszłych dochodów oraz uwzględnienia sposobu wykorzystania nieruchomości w celu zapewnienia maksymalnych korzyści finansowych. Jest szczególnie użyteczne w przypadku nieruchomości komercyjnych, gdzie generowany dochód stanowi główny wskaźnik wartości. Podejście kosztowe umożliwia uzyskanie wartości odtworzeniowej nieruchomości poprzez określenie kosztu nabycia gruntu, kosztów odtworzenia jego części składowych oraz ich zużycia. Jest stosowane przy wycenie nieruchomości, które nie są przedmiotem obrotu rynkowego (Korenik i Zakrzewska-Półtorak, 2021). Metoda ta jest używana głównie dla nieruchomości specjalistycznych lub unikalnych, gdzie brakuje odpowiednich danych porównawczych, co czyni inne podejścia mniej praktycznymi.

W sytuacjach gdy specyficzne uwarunkowania uniemożliwiają zastosowanie podejścia porównawczego lub dochodowego, stosuje się podejście mieszane. Łączy ono elementy podejść porównawczego i dochodowego, by precyzyjnie określić wartość nieruchomości. W jego ramach stosuje się trzy metody: pozostałościową, kosztów likwidacji i metodę wskaźników szacunkowych gruntu. Takie połączenie pozwala na bardziej elastyczną i kompleksową ocenę wartości nieruchomości, uwzględniając różnorodne aspekty jej wyceny.

Pełna wartość i trafność wyceny nieruchomości są potwierdzone tylko wtedy, gdy wszystkie etapy procesu są przeprowadzone prawidłowo z dokładnym zebraniem danych i odpowiednim przypisaniem wagi czynnikom wpływającym na wartość. W artykule zaprezentowano podejście porównawcze, które opiera się na analizie statystycznej rynku nieruchomości. Estymacja wartości opiera się na modelach ekonometrycznych – regresji wielorakiej, analizy trendów i sztucznych sieciach neuronowych (Korenik i Zakrzewska-Półtorak, 2021). Dzięki temu minimalizuje się subiektywizm, bazując wyłącznie na danych z transakcji rynkowych. Takie podejście zwiększa precyzję wyceny oraz pozwala na obiektywne prognozowanie zmian wartości nieruchomości w dynamicznym otoczeniu rynkowym.

3. Założenia realizacji i narzędzia badawcze

Celem przeprowadzonego badania było ocenienie i porównanie efektywności dwóch modeli prognozowania wyceny nieruchomości – modelu opartego na regresji liniowej oraz modelu wykorzystującego las losowy. Ich efektywność zmierzono przy użyciu takich wskaźników jak współczynnik determinacji (R^2), błąd średniokwadratowy (MSE) oraz średni bezwzględny błąd procentowy (MAPE). Analizę przeprowadzono na podstawie danych dotyczących cen transakcyjnych mieszkań we Wrocławiu obejmujących okres od 2014 do 2023 roku. Badanie przeprowadzono zgodnie z następującymi etapami:

1. wybór odpowiednich zmiennych ze zbioru danych do analizy;
2. uzupełnienie brakujących danych i czyszczenie zbioru danych;
3. wizualizacja rynku mieszkaniowego we Wrocławiu;
4. optymalizacja hiperparametrów modeli w celu uzyskania jak najlepszej precyzji predykcji;
5. predykcja wyceny nieruchomości przy użyciu zoptymalizowanych modeli;
6. ocena efektywności modeli za pomocą miar jakości predykcji.

W badaniu zastosowano las losowy jako model uczenia maszynowego. Jest to metoda łącząca wyniki wielu drzew decyzyjnych, co poprawia dokładność predykcji. Za jej pomocą tworzy się wiele drzew na podstawie losowych podzbiorów danych i cech, a końcowy wynik jest uśrednieniem lub głosowaniem większościowym (Biau i Scornet, 2016). Lasy losowe mają kilka kluczowych zalet: są mniej podatne na przeuczenie, mogą być stosowane do regresji i klasyfikacji oraz są odporne na brakujące dane i szum.

Drugim modelem, który został przetestowany, jest regresja liniowa – intuicyjny model szeroko stosowany w różnych dziedzinach nauki i przemysłu. W literaturze regresja liniowa jest często opisywana jako jeden z elementów podejścia porównawczego w wycenie nieruchomości (Korenik i Zakrzewska-Półtorak, 2021). Opierając się na analizie statystycznej i ekonometrycznej, zalicza się ją do metod analizy statystycznej rynku, co pozwala na mniej subiektywne szacowanie wartości nieruchomości.

Pozostałe narzędzia zastosowane w badaniu, które przyczyniły się do poprawy estymacji modeli, to Welch-ANOVA, GridSearchCV oraz sprawdzian krzyżowy.

4. Wybór zmiennych do badania

W badaniu wykorzystano dane z Systemu Analiz i Monitorowania Rynku Obrotu Nieruchomościami (AMRON), który jest jedyną w Polsce międzybankową, wystandaryzowaną bazą danych o nieruchomościach, ich cenach i wartościach (Amron, b.d.). Stanowi ona istotne źródło informacji na temat rynku nieruchomości. Badanie opiera się na danych dotyczących transakcji na rynku miejskim we Wrocławiu przeprowadzonych w latach 2013-2023.

Początkowy zbiór danych składał się z 45 319 wierszy i 25 zmiennych, przy czym każdy wiersz reprezentował pojedynczą transakcję. Niektóre zmienne zostały usunięte z powodu braku różnorodności wewnątrz zmiennej, co sugeruje niską wartość informacyjną. Inne zmienne zostały wyeliminowane, ponieważ brakowało jednolitych kategorii, co komplikowałoby proces standaryzacji i wymagało zastosowania dodatkowych narzędzi do przekształcenia ich w użyteczne informacje. Oprócz tego część zmiennych wykorzystano jako filtry lub użyto ich do utworzenia nowej zmiennej o nazwie „cena zaktualizowana", która lepiej odzwierciedlała aktualne wartości rynkowe.

Po przeprowadzeniu filtracji, selekcji i przetworzeniu danych pierwotny zbiór danych został zredukowany do 23 394 wierszy zawierających 8 zmiennych objaśniających oraz jedną zmienną objaśnianą.

Zmienna „cena zaktualizowana" służy jako zmienna zależna w analizie i reprezentuje obecną wartość rynkową każdego z transakcyjnych lokali mieszkalnych, zaktualizowaną na dzień 01.01.2024. Badanie obejmuje wszystkie transakcje dotyczące nieruchomości lokalowych przeprowadzone na terenie Wrocławia od 01.01.2014 do 31.12.2023. Średnia dynamika wzrostu cen w ciągu ostatnich 10 lat wyniosła 203%. W poniższej tabeli 1 przedstawiono opis zmiennych, które zostały dostosowane do dalszej analizy.

Tabela 1. Zmienne wykorzystane w badaniu

Oznaczenie zmiennej	Nazwa zmiennej	Opis zmiennej	Typ zmiennej
1	2	3	4
Y	Cena zaktualizowana	Cena transakcyjna zaktualizowana na dzień 01.01.2024.	Zmienna ilościowa zmiennoprzecinkowa
X1	Źródło informacji	Zmienna określająca, z jakiego typu umowy pochodzą dane transakcji.	Zmienna kategoryczna
X2	Dzielnica	Lokalizacja nieruchomości według dzielnicy.	Zmienna kategoryczna

Tabela 1. (cd)

1	2	3	4
X3	Rodzaj budynku	Zmienna określająca typ budynku, w którym znajduje się nieruchomość.	Zmienna kategoryczna
X4	Rok budowy	Rok budowy budynku, w którym znajduje się nieruchomość.	Zmienna ilościowa całkowita
X5	Powierzchnia użytkowa	Zmienna określająca powierzchnię podmiotu.	Zmienna ilościowa zmiennoprzecinkowa
X6	Piętro	Zmienna określająca, na którym piętrze znajduje się nieruchomość.	Zmienna ilościowa całkowita
X7	Liczba pokoi	Liczba pokoi w nieruchomości.	Zmienna ilościowa całkowita
X8	Liczba pięter w budynku	Liczba pięter budynku, w którym znajduje się podmiot.	Zmienna ilościowa całkowita

Źródło: opracowanie na podstawie własnych badań.

W analizie wybrano zmienną przedstawiającą całkowitą cenę lokalu, a nie cenę za metr kwadratowy. Dzięki temu można ustalić bezpośrednią cenę transakcyjną, która jest łatwiejsza do zobrazowania dla potencjalnych użytkowników przedstawionych modeli wyceny.

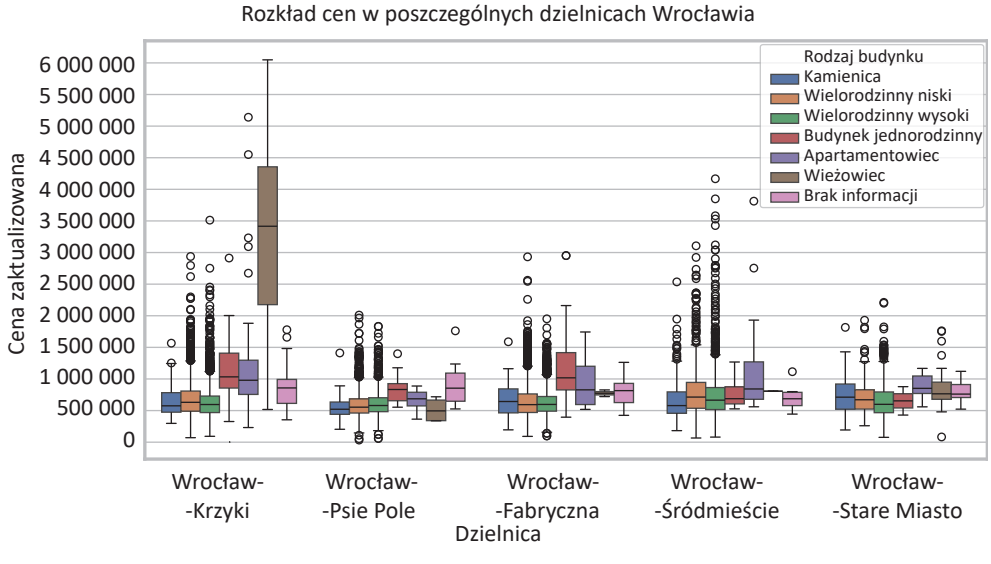
5. Analiza danych rynku mieszkalnego

Do przetwarzania danych wykorzystano środowisko Visual Studio Code. Zbiór danych został wczytany i przekształcony w strukturę znaną jako ramka danych (*data frame*) przy użyciu bibliotek Pandas i Numpy. Umożliwiło to zlokalizowanie brakujących danych. Braki stwierdzono w następujących zmiennych: „rodzaju budynku" (99 brakujących wartości), „piętrze" (43 brakujące wartości), „liczbie pięter w budynku" (85 brakujących wartości) oraz „liczbie pokoi" (jedna brakująca wartość). Z uwagi na to, że nawet pojedyncza brakująca wartość w zmiennych może zakłócić estymację cen nieruchomości, przeprowadzono proces uzupełnienia tych braków. Dla zmiennych numerycznych zastosowano metodę uzupełnienia medianą wartości z danej kolumny, natomiast dla zmiennych kategorycznych przyjęto strategię zastąpienia brakujących danych najczęściej występującą kategorią.

W kolejnym kroku podzielono zmienne na dwa główne typy: dane kategoryczne i numeryczne. Dane numeryczne zostały dodatkowo sklasyfikowane jako ciągłe lub dyskretne, co pozwoliło na dokładniejsze analizowanie zmiennych i identyfikację obszarów wymagających poprawy. W procesie tworzenia modeli uczenia maszynowego kluczowe jest identyfikowanie wartości ekstremalnych, które mogą znacząco wpłynąć na jakość predykcji (Sabourin, 2021) szczególnie w przypadku

modelowania wyceny nieruchomości, gdzie model przeznaczony do estymacji typowego mieszkania może wykazać większe błędy, jeśli nie zostaną odpowiednio dostosowane wartości graniczne. Analiza wykresów pudełkowych dla wszystkich numerycznych zmiennych objaśniających wskazała, że wartości graniczne powinny być ustalone dla zmiennych, takich jak „rok budowy", „powierzchnia użytkowa", „piętro", „liczba pokoi" i „liczba pięter w budynku".

Nastepnym etapem badania była wizualizacja kategorycznych zmiennych objaśniających. Na rysunku 1, przedstawiającym dane przed ustaleniem wartości granicznych na wykresie pudełkowym, wyraźnie wyróżniała się kategoria „wieżowce" dla zmiennej „rodzaj budynku" w dzielnicy Wrocław-Krzyki. Szczególnie odstający budynek – Sky Tower, podkreślał swoją unikalność w kontekście dzielnicy i całego miasta. Z uwagi na ograniczoną liczbę nieruchomości przypisanych do kategorii „wieżowiec" (zaledwie 168) oraz „brak informacji" podjęto decyzję o ich wykluczeniu z dalszej analizy. Po ustaleniu wartości granicznych dla numerycznych zmiennych objaśniających i eliminacji kategorii „wieżowiec" z analizy osiągnięto bardziej zrównoważony rozkład cen, co jest przedstawione na rysunku 2.

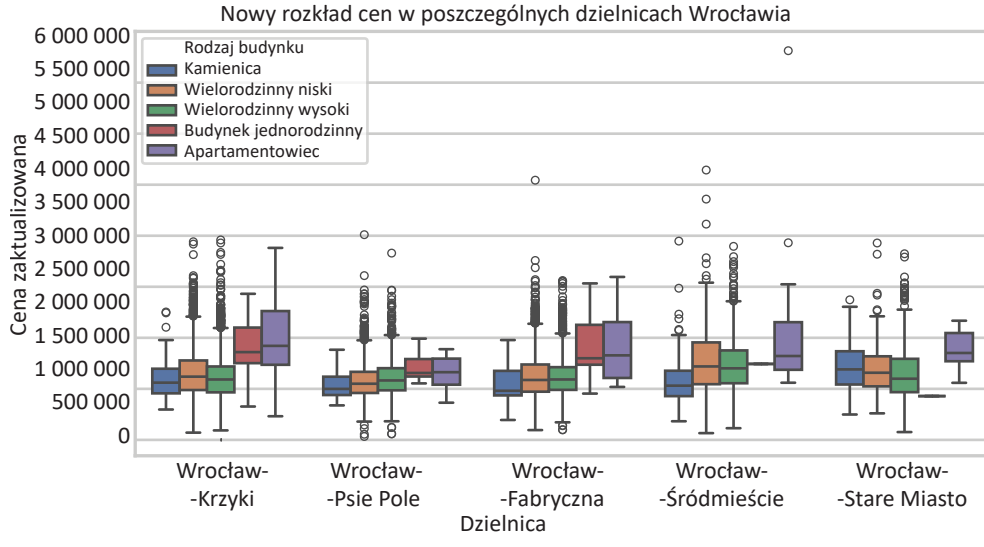


Rys. 1. Rozkład cen nieruchomości w poszczególnych dzielnicach Wrocławia

Źródło: opracowanie na podstawie badań własnych.

Ostatni etap analizy obejmował zastosowanie analizy wariancji, znanej jako ANOVA, będącej techniką modelowania liniowego umożliwiającą ocenę związków między zmiennymi. ANOVA jest wykorzystywana do badania różnic w przewidywanych średnich wartościach między różnymi kategoriami jednej zmiennej lub kombinacjami kategorii dwóch zmiennych (Shaw i Mitchell-Olds, 1993). Dzięki swojej

prostocie implementacji i skuteczności w porównywaniu danych pomiędzy grupami jest często stosowana w analizach statystycznych. W badaniu wykorzystano bibliotekę Pingouin, rozszerzenie statystyczne Pythona. Skupiono się na porównaniu danych w zbiorze, wykorzystując test Levene'a do sprawdzania heterogeniczności zbioru danych. Odrzucenie hipotezy zerowej w tym teście skutkowało zastosowaniem metody Welch-ANOVA (Pingouin, b.d.), która zakłada nierówność wariancji. Zmienne kategoryczne, takie jak „źródło informacji", „dzielnica" i „rodzaj budynku" zostały zbadane, a analiza potwierdziła ich istotny wpływ na kształtowanie się cen nieruchomości. Wyniki wskazują, że te kategorie powinny być uwzględniane przy modelowaniu cen nieruchomości ze względu na ich znaczący wpływ na cenę zaktualizowaną.



Rys. 2. Rozkład cen nieruchomości w poszczególnych dzielnicach Wrocławia po wyznaczeniu wartości granicznych

Źródło: opracowanie na podstawie badań własnych.

Po przeprowadzeniu wszelkich niezbędnych przekształceń zbiór danych przygotowany do estymacji wartości nieruchomości składał się z 19 126 wierszy reprezentujących rzeczywiste transakcje rynkowe.

6. Dostrajanie hiperparametrów

Hiperparametry są kluczowymi współczynnikami w modelach uczenia maszynowego, które nie są uczące się bezpośrednio przez model podczas treningu, ale muszą być ustalane przez użytkownika. Ich odpowiednia kalibracja jest niezbędna, by uzyskać optymalną jakość i dokładność modeli predykcyjnych (Serrano, 2022). Proces ten,

znany jako dostrajanie hiperparametrów, wymaga eksploracji różnych scenariuszy ustawień dla każdego parametru.

W ramach badania rozróżniono dwie główne grupy hiperparametrów: te związane z inżynierią danych oraz te specyficzne dla modeli uczenia maszynowego. Analiza tych parametrów opisana została szczegółowo w tabeli 2, która prezentuje funkcje, opisy działania oraz hiperparametry wchodzące w skład poszczególnych funkcji.

W przypadku modeli uczenia maszynowego, takich jak lasy losowe, istotne hiperparametry obejmują maksymalną liczbę rozgałęzień modelu oraz kryteria oceny jakości podziału. Proces znajdowania optymalnych parametrów dla tych modeli został przeprowadzony z użyciem narzędzia GridSearchCV z biblioteki scikit-learn, co pozwoliło na systematyczne przeszukiwanie przestrzeni parametrów przy użyciu walidacji krzyżowej (Scikit-Learn, b.d.). W analizie rozważano różne ustawienia hiperparametrów, sprawdzając 96 możliwych konfiguracji dla regresji liniowej i 384 dla lasów losowych. Algorytm miał za zadanie znaleźć takie ustawienie hiperparametrów, które zapewni najlepsze wyniki pod względem współczynnika determinacji (Chicco i in., 2021), jednocześnie minimalizując standardowe odchylenie wyników.

Cały proces poszukiwania i optymalizacji hiperparametrów został zilustrowany na diagramach (zob. rys. 3 i 4), które przedstawiają ocenę różnych konfiguracji hiperparametrów i ich wpływ na jakość predykcji. Ten etap uwidacznia, jak istotne jest odpowiednie dostosowanie hiperparametrów, aby zwiększyć skuteczność modeli predykcyjnych w dziedzinie uczenia maszynowego.

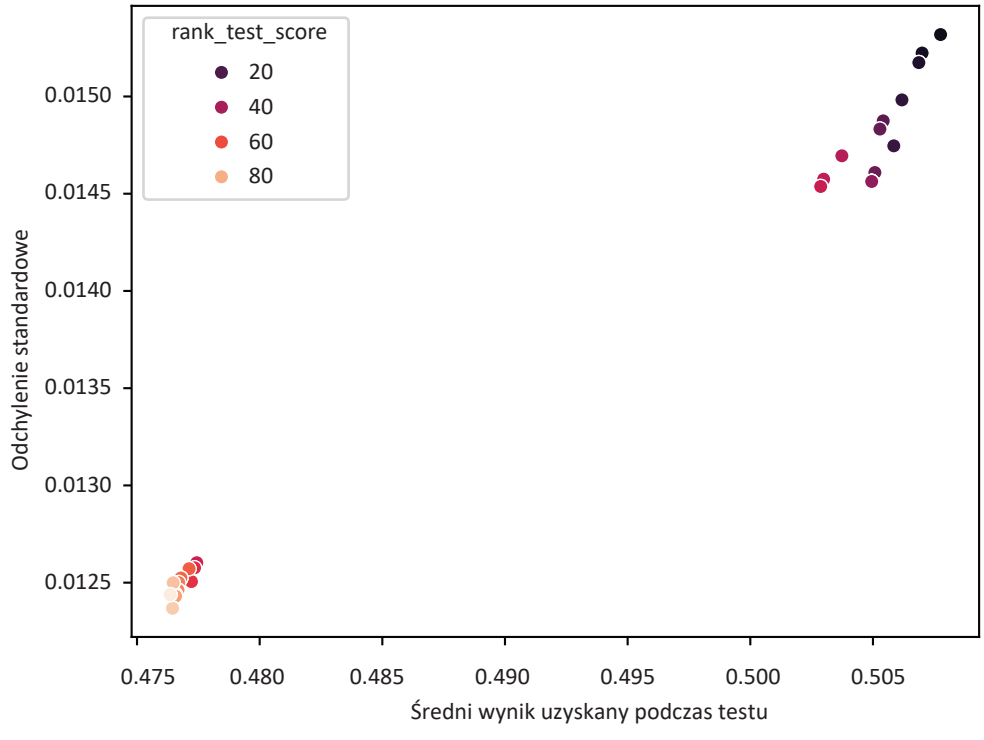
Tabela 2. Zestawienie hiperparametrów analizowanych za pomocą metody GridSearchCV

Nazwa funkcji	Opis działania	Hiperparametry w funkcji
Funkcje związane z inżynierią danych		
1	2	3
Uzupełnienie brakujących numerycznych danych	Zastępuje brakujące dane numeryczne.	• Wybór metody imputacji: średnia czy mediana.
Uzupełnienie brakujących kategorycznych danych	Zastępuje brakujące dane kategoryczne.	• Wybór metody imputacji: odrzucenie brakujących danych lub uzupełnienie o najczęściej występującą kategorię.
Kodowanie rzadkich etykiet	Grupuje rzadkie kategorie w nową kategorię o nazwie „Rare”.	• Wybór minimalnej częstotliwości, aby obserwacja była uznana za częstą: 0,01 czy 0,05. • Wybór pożądanej liczby przedziałów: 5 lub 10.
Kodowanie porządkowe	Przekształca zmienne kategoryczne na zmienne liczbowe.	• Wybór preferowanej metody kodowania: porządkowa czy arbitralna.
Dyskretyzacja	Dzieli zmienne numeryczne ciągłe na przedziały o równej częstotliwości.	• Wybór pożądanej liczby przedziałów: 10, 50 czy 100.

Tabela 2. (cd.)

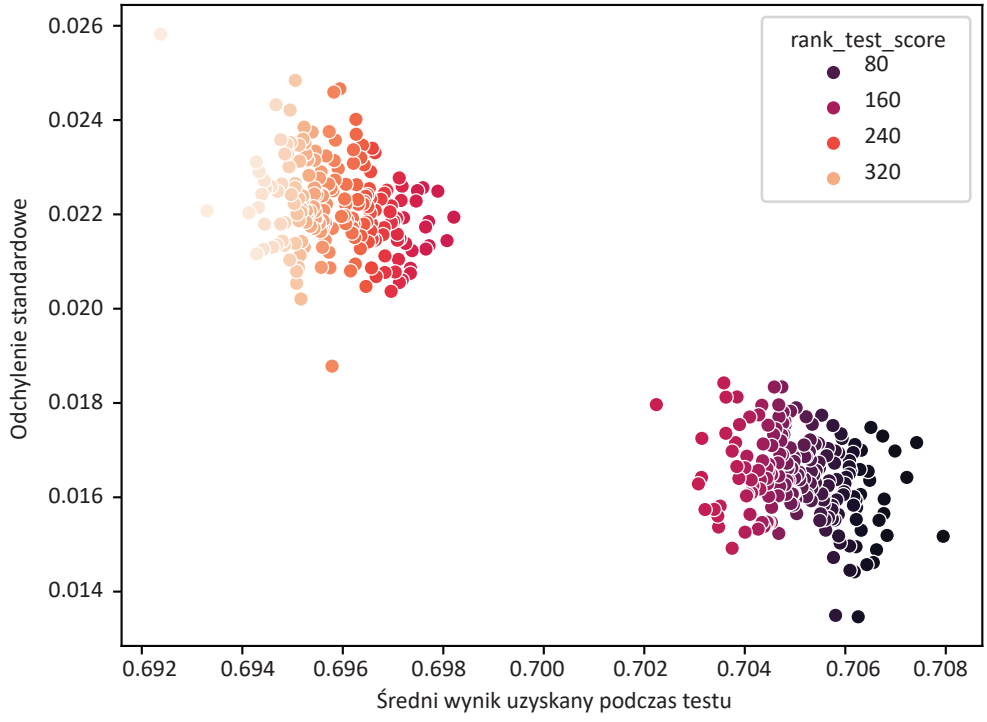
1	2	3
Modele uczenia maszynowego		
Regresja liniowa	Liniowy model uczenia maszynowego, który stosuje regresję liniową.	• Brak
Las losowy	Model, który dokonuje prognoz poprzez podział danych na coraz mniejsze grupy na podstawie określonych kryteriów.	• Wybór maksymalnej liczby rozdziałów modelu: 100 czy bez ograniczeń. • Wybór oceny jakości podziału: współczynnik determinacji czy błąd średniokwadratowy.

Źródło: opracowanie na podstawie badań własnych.



Rys. 3. Procedura dostrajania hiperparametrów w modelu regresji liniowej

Źródło: opracowanie na podstawie badań własnych.



Rys. 4. Procedura dostrajania hiperparametrów w modelu lasów losowych

Źródło: opracowanie na podstawie badań własnych.

Na rysunkach 3 i 4 widoczne są różnice pomiędzy zastosowanymi procedurami dostrajania hiperparametrów. W przypadku lasów losowych algorytm koncentrował się na uzyskaniu jak najlepszego wyniku przy minimalnym odchyleniu standardowym. Tymczasem w regresji liniowej obserwuje się odmienny scenariusz, gdzie odchylenie standardowe jest coraz większe. Dodatkowo, wyniki pokazują różnice w średnim wyniku współczynnika determinacji – dla lasów losowych jest on znacznie wyższy. Wskazuje to na potencjalnie mniejsze błędy przy estymacji cen nieruchomości przy użyciu tego modelu.

7. Porównanie zastosowanych modeli predykcji

Modele predykcyjne w badaniu zostały poddane szczegółowej ocenie przy użyciu specjalnie opracowanej metody, co umożliwiło kompleksową analizę ich jakości. Cały proces rozpoczął się od podziału danych na wiele podzbiorów, które zostały następnie poddane walidacji krzyżowej, umożliwiając uśrednienie wyników z różnych scenariuszy i zmniejszenie ryzyka wynikającego z jednorazowego podziału danych na nieregularny zbiór testowy i treningowy. Analiza obejmowała przygotowa-

nie danych, utworzenie listy metryk służących do oceny predykcji, konfigurację walidacji krzyżowej, bezpośrednią procedurę oceny modelu, a na koniec zebranie i zapisanie wyników (zob. tab. 3).

Tabela 3. Wyniki oceny prognozy

Wskaźnik oceny	Modele predykcji cen nieruchomości	
	Liniowa regresja	Las losowy
R^2	51%	71%
MSE	23 696 919 851,17 zł²	14 043 462 762,58 zł²
RMSE	153 938,04 zł	118 505,12 zł
MAPE	19%	13%

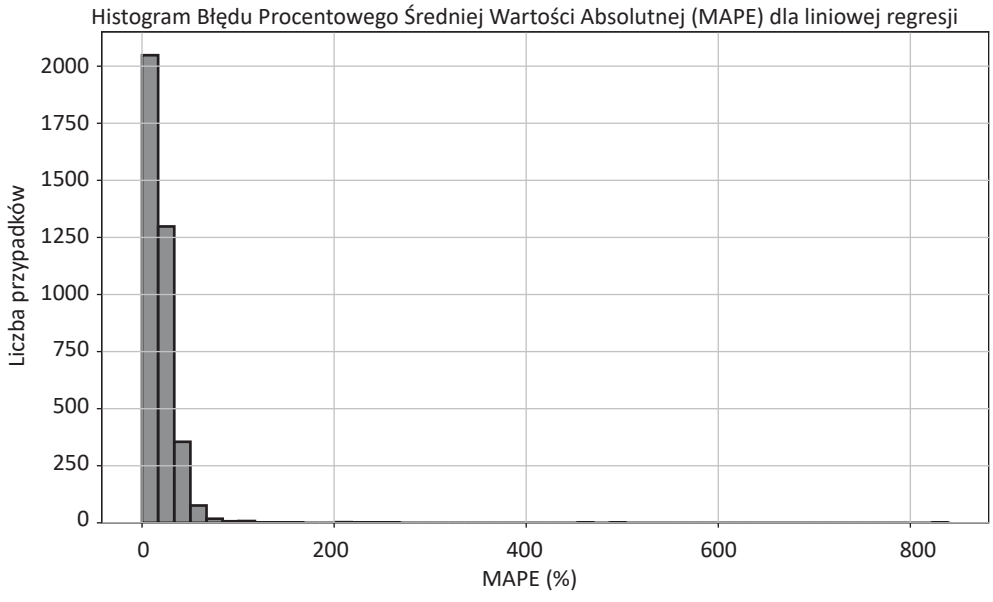
Źródło: opracowanie na podstawie badań własnych.

Analiza wyników predykcji cen nieruchomości pokazuje, że model lasu losowego jest znacznie skuteczniejszy w estymacji wartości nieruchomości niż regresja liniowa. Jednakże należy zauważyć, że las losowy jest również bardziej czasochłonny i wymaga większych zasobów obliczeniowych. W każdym z analizowanych wskaźników – współczynniku determinacji, średnim błędzie kwadratowym, pierwiastku średniego błędu kwadratowego oraz średnim błędzie procentowym – las losowy osiągnął lepsze wyniki. Te rezultaty sugerują, że w kontekście analizowanych danych, które mogą charakteryzować się większą nieliniowością lub zróżnicowaniem, lasy losowe okazują się być bardziej efektywnym narzędziem do estymowania cen nieruchomości niż regresja liniowa.

Na rysunkach 5 i 6 przedstawiono histogramy rozkładu błędu procentowego średniej wartości bezwzględnej dla obu modeli predykcyjnych.

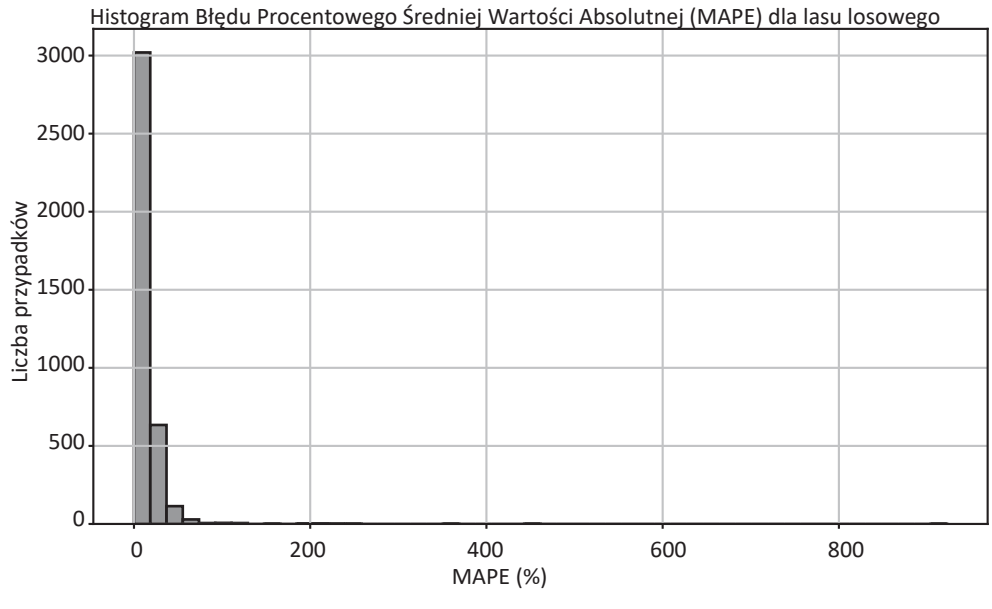
Na podstawie przedstawionych histogramów błędu procentowego średniej wartości absolutnej (MAPE) dla dwóch modeli predykcyjnych – regresji liniowej i lasu losowego – można zaobserwować znaczące różnice w dokładności tych metod. Histogram dla regresji liniowej pokazuje szerszy rozkład błędów z większą liczbą obserwacji o wyższym MAPE, co wskazuje na mniej spójne i często gorsze wyniki predykcji. Histogram dla lasu losowego pokazuje znacznie węższy zakres błędów, gdzie większość wartości koncentruje się w niższych przedziałach MAPE. Wynika z tego, że las losowy jest bardziej skuteczny w minimalizowaniu błędów i zapewnia bardziej niezawodne oraz dokładne oszacowania cen nieruchomości. Spośród możliwych 3826 estymacji ponad 3000 mieści się w najniższym przedziale błędu MAPE.

Te wyniki potwierdzają wyższą efektywność lasów losowych w kontekście modelowania cen nieruchomości, co jest zgodne z wcześniejszą analizą wskazującą na ich lepsze ogólne wyniki w porównaniu do regresji liniowej.



Rys. 5. Histogram błędu bezwzględnego w procentach dla liniowej regresji

Źródło: opracowanie na podstawie badań własnych.



Rys. 6. Histogram błędu bezwzględnego w procentach dla lasu losowego

Źródło: opracowanie na podstawie badań własnych.

8. Zakończenie

Na podstawie przeprowadzonego badania można wnioskować, że metody prognozowania, a zwłaszcza uczenie maszynowe, wykazują znaczący potencjał w dziedzinie wyceny nieruchomości, opierając się na rzeczywistych transakcjach. Las losowy, osiągając R^2 na poziomie około 71%, potwierdza swoją zdolność do przewidywania zmienności cen nieruchomości na podstawie dostępnych danych. Ponadto średni błąd bezwzględny rzadko przekracza 13%, co świadczy o precyzji prognoz tego modelu.

Oba analizowane modele oparto na metodzie GridSearchCV do optymalizacji parametrów, co zaowocowało uzyskaniem najlepszych możliwych wyników. Pomimo gorszych rezultatów regresji liniowej należy zauważyć, że jej efektywność może być ograniczona w przypadku szerokich obszarów, takich jak całe miasto. Warto podkreślić, że rynek nieruchomości ma charakter lokalny, co może wymagać rozwoju modeli dedykowanych do estymacji cen na poziomie mniejszych obszarów np. poszczególnych osiedli. Skuteczność takich modeli zależy od dostępu do aktualnych cen transakcyjnych oraz starannego oczyszczania danych z nadzwyczajnych odchyleń.

Wyniki przeprowadzonych badań mają istotne implikacje zarówno dla nauki, jak i praktyki. Z naukowego punktu widzenia poszerzają one wiedzę na temat zastosowania zaawansowanych metod uczenia maszynowego w wycenie nieruchomości. Z praktycznego punktu widzenia, badania dostarczają narzędzi, które mogą znacząco usprawnić ten proces, minimalizując subiektywizm. Ujednolicenie skali oceny standardu nieruchomości mogłoby dodatkowo poprawić jakość predykcji, ułatwiając porównywalność wyników. Takie podejście zwiększa precyzję prognoz i pozwala na dokładniejszą ocenę wartości nieruchomości w kontekście lokalnych rynków. Ponadto wzbogacenie analiz o nowe zmienne i nowoczesne techniki regresji mogłoby jeszcze bardziej podnieść trafność i dokładność wycen.

Literatura

Amron. (b.d.). *O systemie*. Amron.pl. Pobrano z https://www.amron.pl/strona.php?tytul=o-systemie

Biau, G., i Scornet, E. (2016). A Random Forest Guided Tour. *Test*, (25), 197-227.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Coefficient of Determination R-squared is More IIformative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Computer Science*, (623). https://doi.org/10.7717/peerj-cs.623

Cymerman, R. i Cymerman, J. (2024). *Wycena nieruchomości*. Wydawnictwo Politechniki Koszalińskiej.

Korenik, S. i Zakrzewska-Półtorak, A. (2021). *Nieruchomości i ich wycena*. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

Kucharska-Stasiak, E. (2006). *Nieruchomości w gospodarce rynkowej.* Wydawnictwo Naukowe PWN.

Malec, T. i Stachura E. (2006) *Nieruchomości – proces inwestycyjny.* Wydawnictwo Wyższej Szkoły Bankowości i Finansów.

Pingouin. (b.d.). *Guideline*s. Pingouin-stats.org. Pobrano z https://pingouin-stats.org/build/html/guidelines.html

Sabourin, A. (2021). *Extreme Value Theory and Machine Learning*. Doctoral dissertation, Institut Polytechnique de Paris.

Scikit-Learn. (b.d.). *GridSearchCV*. Scikitlearn.org. Pobrano z https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Serrano, L. G. (2022). *Machine Learning*. Manning Publications Co., 71-72.

Shaw, R. G., i Mitchell-Olds, T. (1993). ANOVA for Unbalanced Data: An Overview. *Ecology*, *74*(6), 1638--1645.

Ustawa z dnia 23 kwietnia 1964 r. – Kodeks cywilny (Dz. U. 20240.1061 t.j., art. 46)

Comparison of the Quality of Forecasting Models Based on Transaction Prices of Residential Real Estate in Wroclaw

Abstract: The study focuses on comparing the effectiveness of machine learning methods, such as random forests, and linear regression in the context of residential property valuation based on transactional data. The main objective of the research is to evaluate the ability of these methods to accurately forecast property prices. The article begins with a theoretical introduction to property valuation, discussing definitions and existing approaches in Poland. Subsequently, the empirical study assumptions are presented, including a detailed description of the dataset and variables used. The analysis focuses on the residential market in Wroclaw from 2014 to 2023. The GridSearchCV method was employed to optimise the parameters of predictive models. The findings of the study demonstrate significant potential in nonlinear models for property price estimation, emphasising the importance of proper hyperparameter calibration before estimation begins.

Keywords: real estate, property valuation, forecasting models, machine learning, linear regression, random forest, ANOVA, hyperparameters, GridSearchCV