Wiktoria Galarowicz

e-mail: 177503@student.ue.wroc.pl

ORCID: 0009-0006-8381-3569

Uniwersytet Ekonomiczny we Wrocławiu

Predykcja cen mieszkań w dużych miastach w Polsce z wykorzystaniem uczenia maszynowego

DOI: 10.15611/2024.53.6.05
JEL Classification: R31

Cytuj jako: Galarowicz, W. (2024). Predykcja cen mieszkań w dużych miastach w Polsce z wykorzystaniem uczenia maszynowego. W: A. Grześkowiak, P. Peternek (red.), *Zastosowanie metod ilościowych w ekonomii i finansach* (s. 68-81). Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

Streszczenie: Artykuł analizuje kluczowe zmienne wpływające na kształtowanie cen na rynku nieruchomości w dużych miastach w Polsce. Omówiono w nim pojęcie rynku nieruchomości oraz jego najważniejszych cech, szczegółowo wyjaśniono, jak zmieniały się ceny mieszkań na przestrzeni lat w Polsce. W latach 90. XX wieku ceny notowały wzrost wywołany liberalizacją i łatwym dostępem do kredytów hipotecznych. Jednakże po kryzysie finansowym z 2007 roku rynek doświadczył znacznych wahań. Pandemia COVID-19, w porównaniu z programem Bezpieczny Kredyt 2%, wywarła niewielki wpływ na ceny. Analiza cen mieszkań i kluczowych zmiennych wpływających na ich kształtowanie obejmowała przygotowanie danych, imputowanie brakujących wartości oraz identyfikację odstających obserwacji. Do prognozowania cen wykorzystano algorytmy uczenia maszynowego, Random Forest (las losowy) i XGBoost.

Słowa kluczowe: rynek nieruchomości, ceny mieszkań, uczenie maszynowe, model Random Forest, model XGBoost

1. Wstęp

Artykuł porusza tematykę rynku nieruchomości, który odgrywa kluczową rolę w gospodarce każdego kraju. Jest jednym z najważniejszych sektorów, który wpływa na życie codzienne milionów ludzi. Zrozumienie mechanizmów kształtujących ceny mieszkań w dużych miastach jest istotne zarówno dla deweloperów, inwestorów, jak i nabywców mieszkań. Motywacją do przeprowadzenia niniejszego badania była potrzeba głębszej analizy czynników wpływających na ceny mieszkań oraz ocena skuteczności metod uczenia maszynowego w prognozowaniu tych cen. Dla de-

weloperów i inwestorów wiedza o przyszłych trendach cenowych jest kluczowa w podejmowaniu decyzji inwestycyjnych. Dla osób planujących zakup mieszkania wyniki badania mogą być pomocne w podejmowaniu świadomych decyzji konsumenckich.

Celem teoretycznym artykułu jest zdobycie informacji o zmianach na rynku nieruchomości w latach 1989-2024. Natomiast celem badawczym artykułu jest stworzenie modeli uczenia maszynowego przewidującego ceny mieszkań w dużych miastach w Polsce. Oprócz tego szukana jest również odpowiedź na pytanie badawcze, jakie cechy nieruchomości i jej otoczenia najbardziej wpływają na kształtowanie się cen nieruchomości.

Skupienie się na dużych miastach wynika z ich znaczenia ekonomicznego i demograficznego. Duże miasta charakteryzują się większą dynamiką rynku nieruchomości, co czyni je interesującymi obszarami analizy. Metody uczenia maszynowego, takie jak Random Forest i XGBoost, zostały wybrane ze względu na ich wysoką skuteczność w analizie dużych zbiorów danych oraz zdolność do identyfikacji nieliniowych zależności między zmiennymi.

2. Definicja i cechy rynku nieruchomości

Na przestrzeni lat rynek nieruchomości charakteryzuje się dynamicznymi zmianami. Od 1989 roku (początek transformacji ustrojowej) nieruchomości przestały być dobrami socjalnymi i stały się przedmiotem handlu. Są to specyficzne dobra, które podlegają obrotowi prawami do nich, a nie samymi budynkami czy gruntami (Gołąbeska, 2017). Oznacza to, że transakcje dotyczą transferu różnorodnych praw użytkowania i obowiązków właścicieli. Te prawa obejmują własność, współwłasność, użytkowanie wieczyste gruntu, prawa rzeczowe ograniczone (użytkowanie, służebność), spółdzielcze własnościowe prawo do lokalu, hipotekę oraz prawa zobowiązaniowe jak najem i dzierżawa.

Obrót nieruchomościami można podzielić na dwie kategorie (Gołąbeska, 2017): rynkowe i nierynkowe. Transakcje rynkowe obejmują kupno/sprzedaż, najem/dzierżawę, użytkowanie, podnajem i zamianę, podczas gdy transakcje nierynkowe dotyczą darowizn, spadkobrania, wywłaszczenia, uwłaszczenia, eksmisji i egzekucji. Obroty nierynkowe nie mają na celu osiągnięcia zysku, ale mogą być wynikiem przeniesienia własności w różnych sytuacjach, np. spadek, darowizna, wywłaszczenie czy egzekucja. Transakcje rynkowe zwykle służą zaspokojeniu potrzeb mieszkaniowych lub prowadzeniu działalności gospodarczej. Mogą też być wynikiem inwestycji w celu osiągnięcia dodatkowych dochodów, np. poprzez zysk z transakcji kupna-sprzedaży, dochody z reklam lub dochody czynszowe z najmu nieruchomości. W zależności od rodzaju nieruchomości inwestycja może być krótko- lub długoterminowa oraz służyć zabezpieczeniu finansowemu na przyszłość.

Rynek nieruchomości jest miejscem, gdzie odbywają się transakcje kupna, sprzedaży, dzierżawy czy wynajmu, lecz jego definicja jest różna w zależności od kontekstu. Jak podaje E. Gołąbeska (2007) można go postrzegać jako interakcje między ludźmi lub instytucjami zajmującymi się nieruchomościami, wszystkie transakcje nieruchomościami, forum do zawierania umów oraz zestaw układów, gdzie ustalane są ceny i prawa własności lub wymianę prawa własności nieruchomości na inne aktywa. Niemniej, najbardziej kompletną definicją wydaje się być stwierdzenie, że rynek nieruchomości to ogół warunków, w których dokonuje się przekazywania praw do nieruchomości i zawierania umów dotyczących ich użytkowania (Gołąbeska, 2007).

Trzema głównymi czynnikami, które kształtują rynek nieruchomości (w tym również rynek mieszkaniowy) są: popyt, podaż i cena. Popyt na nieruchomości odzwierciedla siłę nabywczą osób poszukujących lokali do kupna lub wynajmu. Można wyróżnić dwa rodzaje popytu: potencjalny i efektywny (Gołąbeska, 2017). Popyt potencjalny związany jest z ludzkimi potrzebami i aspiracjami, podczas gdy popyt efektywny opiera się na dostępnych środkach finansowych. Istotną cechą popytu na nieruchomości jest ich unikalność i brak substytutów, ponieważ jak twierdzi E. Kucharska-Stasiak (2000), potrzeba posiadania własnego mieszkania lub domu nie może być zastąpiona innym dobrem. Podaż odnosi się do liczby dostępnych na rynku nieruchomości, jakie mają do zaoferowania producenci w określonym czasie i za określoną cenę. Wzrost cen nieruchomości skutkuje zwykle wzrostem podaży, ponieważ wyższe ceny zachęcają producentów do szybkiego zarobku poprzez sprzedaż lub wynajem (Gołąbeska, 2017). Podaż nieruchomości obejmuje zarówno istniejące nieruchomości na rynku wtórnym, jak i nowe inwestycje budowlane na rynku pierwotnym. Obejmuje to także różne formy transakcji, takie jak sprzedaż, wynajem, dzierżawa czy najem lokali. Deweloperzy i właściciele nieruchomości są głównymi dostawcami na rynku nieruchomości, dostarczając zarówno nowych, jak i zmodernizowanych mieszkań dostosowanych do bieżących trendów oraz oczekiwań klientów. Natomiast cena jest relacją pomiędzy popytem a podażą.

Rynek nieruchomości znacznie różni się od innych rynków, a jego stałymi cechami, jak podaje M.J. Bryx (2006) są, m.in: mała elastyczność popytu i podaży, niedoskonałość, lokalny charakter, niejednolitość, niska efektywność, niepowtarzalność i duży zakres interwencji państwa. Mała elastyczność popytu i podaży ukazuje się poprzez małą reakcję na zmiany cen. Niedoskonałość rynku objawia się w nieprawidłowym funkcjonowaniu mechanizmu ustalania ceny równowagi mimo swobodnego popytu i podaży. To wynika częściowo z niemożności zmiany lokalizacji nieruchomości, ograniczonej podaży oraz długotrwałego procesu inwestycyjnego. Lokalny charakter rynku ogranicza konkurencję, ponieważ duże znaczenie ma lokalizacja nieruchomości. Niejednolitość wynika z różnorodności rodzajów nieruchomości (czego przejawem jest podział na rynek międzynarodowy, krajowy, regionalny i lokalny). Natomiast niską efektywność potwierdza problem z właściwą wyceną nieruchomości, a informacje, które posiada inwestor, mogą nie być wystarczające,

aby opracować strategie handlowe. Dodatkowo występuje duży interwencjonizm państwa poprzez wysokość podatków, politykę czynszową i regionalną oraz ochronę zabytków. Rynek nieruchomości jest wyjątkowy nie tylko ze względu na unikatowość poszczególnych nieruchomości, ale także ze względu na ich liczbę, łączną powierzchnię i wartość. Niepowtarzalność tego rynku wynika z różnorodności popytu i podaży, która jest determinowana przez strukturę społeczeństwa i jego zamożność. Silne przywiązanie do lokalizacji sprawia, że ludzie chętniej pozostają w danym regionie, co zwiększa popyt na lokalnym rynku. Te cechy są niemal stałe niezależnie od lokalizacji, czasu czy otoczenia prawno-ekonomicznego, co wynika z charakteru samej nieruchomości jako towaru.

3. Zmiany na rynku mieszkaniowym w ostatnich trzydziestu latach w Polsce

Na wzrost i spadki cen mieszkań wpływają takie czynniki, jak rozwój gospodarczy, zmiany demograficzne i czynniki polityczne. Początek lat 90. XX wieku był okresem intensywnego rozwoju rynku nieruchomości w Polsce (Gołąbeska, 2017). Upadek komunizmu wywołał społeczną euforię i zwiększył optymizm co do przyszłości. W tym czasie ceny mieszkań dynamicznie rosły, ponieważ ludzie oczekiwali lepszego życia i chcieli posiadać własne mieszkania. Liberalizacja gospodarki, wprowadzenie wolnego rynku oraz łatwy dostęp do kredytów hipotecznych spowodowały zwiększenie popytu na mieszkania, co doprowadziło do szybkiego wzrostu cen na rynku pierwotnym i wtórnym.

W latach 2000-2005 przeważała korzystna sytuacja dla deweloperów, gdyż na rynku nieruchomości mieszkaniowych panowała hossa. Deweloperzy przejęli rolę spółdzielni mieszkaniowych budujących mieszkania. Wyjątkiem był 2002 rok i początek 2003 roku, kiedy nastąpiło spowolnienie i recesja. Ten stan nie utrzymał się długo. Ożywienie na rynku nieruchomości było w znacznym stopniu wynikiem przystąpienia Polski do Unii Europejskiej oraz wzmożonego zainteresowania możliwością zaciągania kredytów hipotecznych. W roku 2004 polski sektor bankowy doświadczył boomu hipotecznego, który był rezultatem znacznego zainteresowania społeczeństwa właśnie tą formą finansowania nieruchomości. Wzrost zadłużenia polskich gospodarstw domowych był widoczny w zwiększonej liczbie udzielanych kredytów klientom indywidualnym. Wzrost zadłużenia hipotecznego przekształcił się w Polsce w boom ekonomiczny w sektorze bankowym, co doprowadziło do powstania silnej asymetrii w systemie finansowania rynku mieszkaniowego. Deweloperzy preferowali tanie źródła finansowania, takie jak przedpłaty klientów, a ryzyko kredytowe w ich działalności było wyższe, co wymagało odpowiednich zabezpieczeń. Boom hipoteczny spowodował duży wzrost cen nieruchomości mieszkaniowych po 2005 roku, co z kolei wywołało reakcję ze strony podażowej. W okresie wzrostu cen, czyli w latach 2007-2008, marże deweloperów sięgały poziomu 50-60% (Kucharska-

-Stasiak i in., 2012). Wzrost cen mieszkań nie był uzasadniony wzrostem dochodów gospodarstw domowych. Dodatkowo, czynniki makroekonomiczne, których spadające wartości przyczyniły się do kryzysu finansowego z 2007 roku, negatywnie wpłynęły na rynek nieruchomości. W drugiej połowie 2008 roku oddano do użytku mieszkania, których budowę rozpoczęto wcześniej, co spowodowało nadmierną podaż i spadek cen nieruchomości. W związku z tym deweloperzy ograniczyli nowe inwestycje i zaczęli proponować klientom upusty lub dostosowywali nieruchomości do ich potrzeb (np. przez przekształcanie mieszkań na mniejsze). Doprowadziło to do poprawy sytuacji dopiero w połowie 2009 roku, a w 2010 rynek ustabilizował się, choć nadal przeważała podaż nad popytem. W 2012 roku wprowadzono zmiany prawne, takie jak zakończenie programu „Rodzina na swoim" (Majorek, 2013) i ustawę deweloperską, co skutkowało wzrostem liczby dostępnych nieruchomości mieszkaniowych i rozpoczęciem nowych inwestycji przez deweloperów, a w 2013 roku rynek mieszkań zbliżył się do stanu równowagi, a jego dalszy rozwój zależał głównie od sytuacji makroekonomicznej. Obserwowano wtedy poprawę wskaźników sytuacji mieszkaniowej: wzrost zasobu mieszkaniowego, spadek liczby osób w mieszkaniu i zwiększenie średniej powierzchni użytkowej mieszkania na osobę. W 2015 roku odnotowano najwyższy poziom sprzedaży od 2007. Większość transakcji była finansowana przez nabywców ze środków własnych. Liczba transakcji rosła przy jednoczesnym braku wzrostu cen nieruchomości. Wzrost popytu spowodowany niskimi stopami procentowymi skłonił do większych inwestycji gotówkowych przez osoby posiadające kapitał, które szukały bardziej opłacalnych alternatyw dla nisko oprocentowanych lokat. Nieruchomości coraz częściej postrzegano nie tylko jako miejsce do zamieszkania, ale także jako atrakcyjną formę inwestycji kapitału.

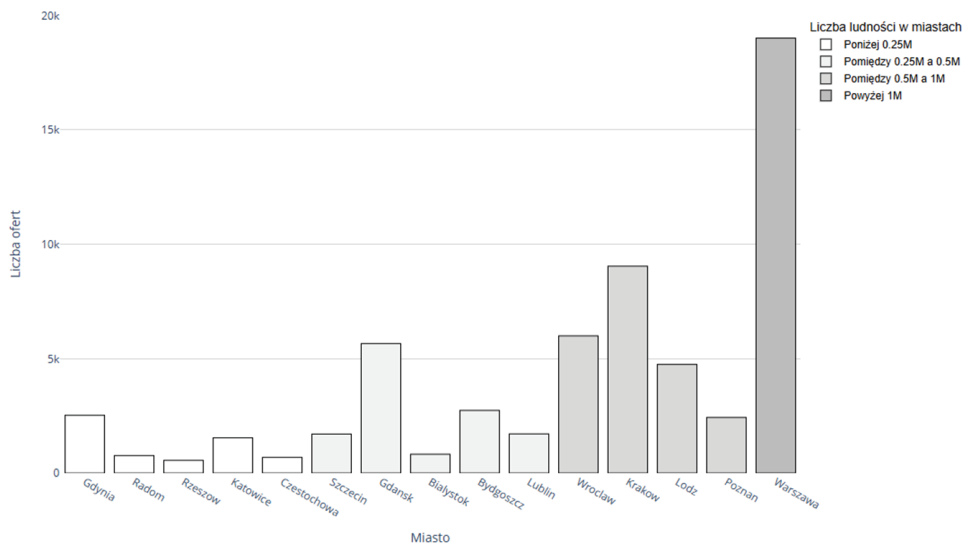
Jak podają B. Samorek i M. Cichocki (2023) zarówno rynek pierwotny, jak i wtórny, wykazały się wysoką odpornością na międzynarodowy kryzys spowodowany pandemią. W 2020 roku nie zaobserwowano istotnego wpływu pandemii na ceny nieruchomości, a dynamika zmian cen pozostała na podobnym poziomie jak w poprzednich latach. Pomimo okresowych fluktuacji rynek nieruchomości pozostawał stosunkowo stabilny, choć pandemia miała wpływ na pewne okresowe zmiany w gospodarce kraju. Końcem roku 2020, mimo zaostrzenia przepisów, ceny nieruchomości w wielu miastach powróciły do poziomu z początku roku, a zainteresowanie zakupem i sprzedażą nieruchomości wzrosło.

W lipcu 2023 roku można było zacząć składanie wniosku o Bezpieczny Kredyt 2%. Biuro Prasowe FIN-HUB (b.d.) podaje, że kredytobiorcy mogli liczyć na rządowe dopłaty do rat kredytu przez dziesięć lat. Spowodowało to spadek miesięcznych rat do 40%. Zainteresowanie tym programem miało wyraźny wpływ na ceny mieszkań. Wprowadzenie kredytu hipotecznego z rządową dopłatą zwykle powoduje wzrost cen mieszkań zarówno na rynku pierwotnym, jak i wtórnym. Można było zauważyć wyraźną tendencję podnoszenia cen nieruchomości do poziomu 700-800 tys. zł, czyli do ustalonych przez rząd limitów dopłat.

4. Analiza cen mieszkań w dużych miastach w Polsce

4.1. Opis zbioru danych

Przeprowadzono analizę, której celem było stworzenie modeli prognozujących ceny mieszkań w dużych miastach w Polsce oraz porównanie skuteczności tych modeli. Do analizy wybrano dwa algorytmy uczenia maszynowego: las losowy oraz XGBoost. Dodatkowo postanowiono zbadać, jakie czynniki najbardziej wpływają na ceny mieszkań, co pozwoli na lepsze zrozumienie rynku nieruchomości.



Rys. 1. Liczba dostępnych ofert mieszkaniowych w dużych miastach w Polsce

Źródło: opracowanie własne.

Wykorzystane w analizie dane pochodzą ze strony internetowej Kaggle.com (Jamroz, 2024) i dotyczą ceny oraz właściwości nieruchomości w piętnastu największych miastach w Polsce (Warszawa, Kraków, Wrocław, Łódź, Poznań, Gdańsk, Szczecin, Bydgoszcz, Lublin, Katowice, Białystok, Częstochowa, Rzeszów, Radom, Gdynia). Odnoszą się one do okresu sierpień 2023-styczeń 2024. Ich zbiór zawierał następujące zmienne:

- *city* – nazwa miasta, w którym znajduje się nieruchomość;
- *squareMeters* – powierzchnia mieszkania w metrach kwadratowych;
- *rooms* – liczba pokoi w mieszkaniu;
- *floor* – piętro, na którym znajduje się mieszkanie;
- *floorCount* – liczba pięter w budynku;
- *centreDistance* – odległość od centrum miasta w kilometrach;

- *poiCount* – liczba punktów usługowych w promieniu 500 metrów od mieszkania;
- *[poiName]Distance* – odległość do najbliższego punktu usługowego (szkoły, kliniki, urzędy pocztowe, przedszkola, restauracje, uczelnie, apteki);
- *ownership* – rodzaj własności nieruchomości;
- *has[features]* – informacja czy nieruchomość posiada określone cechy (miejsce parkingowe, balkon, ochronę, pomieszczenie gospodarcze);
- *price* – cena oferty w PLN.

Wśród analizowanych danych było ponad 40 000 duplikatów; po ich usunięciu pozostało około 60 000 rekordów. Zdecydowana większość ofert pochodzi z Warszawy (rys. 1). Warto zauważyć, że im większa liczba ludności w mieście, tym więcej ofert mieszkań.

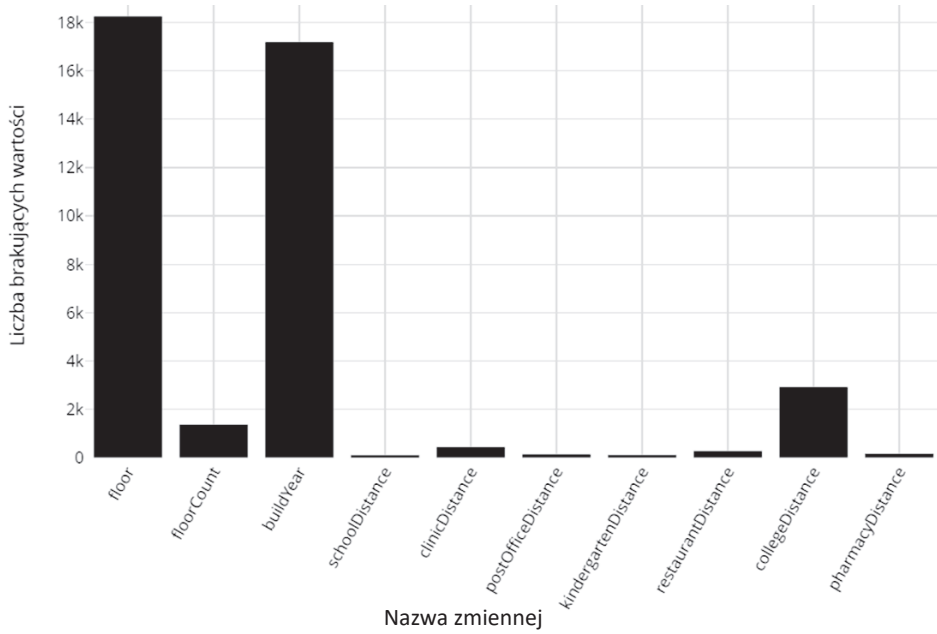
W związku z faktem, że zmienną zależną jest cena mieszkania, zdecydowano się na obliczenie współczynnika korelacji liniowej Pearsona (Sadowski, 1976). Silną zależność dodatnią zauważono dla zmiennych: liczba metrów kwadratowych mieszkania i liczba pokoi w mieszkaniu. Pozostałe zmienne charakteryzują się bardzo słabą korelacją, co wskazuje na niezasadność budowy klasycznych modeli regresji, przez co zdecydowano się na wykorzystanie algorytmów uczenia maszynowego.

4.2. Przygotowanie danych – imputacja braków danych i identyfikacja obserwacji odstających

Przed przystąpieniem do analizy właściwej konieczne było wykonanie szeregu operacji przygotowawczych. Proces ten obejmował, m.in. imputację braków danych oraz identyfikację i obsługę wartości odstających.

Dziesięć zmiennych posiadało niepełne informacje (rys. 2). Najwięcej braków mają zmienne określające rok budowy (około 17 000 braków) i numer piętra, na którym znajduje się mieszkanie (około 18 000 braków). Pozostałe charakteryzują się stosunkowo niewielką liczbą braków. Przez wzgląd na to, że każda z tych zmiennych jest numeryczna, postanowiono najpierw zestandaryzować dane, a następnie imputować braki z wykorzystaniem metody *k*-najbliższych sąsiadów (KNN). Polega ona na poszukiwaniu najbardziej podobnych przypadków ze znanymi wartościami atrybutów, które są używane do estymacji brakujących danych (Malarvizhi i Thanamani, 2012). Do uzupełnienia danych wybrano dziewiętnaście najbardziej podobnych ofert.

Kolejnym etapem analizy była identyfikacja obserwacji odstających. Wykorzystano w tym celu metodę Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Jest to metoda, która identyfikuje punkty jako odstające, jeśli znajdują się w obszarach o niskiej gęstości (Westerlund, 2023). Następnie do automatycznej identyfikacji punktów łamania w krzywej odległości od najbliższych dziewiętnastu sąsiadów zastosowano algorytm KNEED. To podejście wykazało, iż około 14% danych jest obserwacjami odstającymi. Zdecydowano się na ich usunięcie z dalszej analizy.



Rys. 2. Liczba brakujących wartości w poszczególnych kolumnach

Źródło: opracowanie własne.

4.3. Modele uczenia maszynowego

Po podziale zbioru danych na zbiór testowy (20% ofert) i uczący (80% ofert) przystąpiono do modelowania. Pierwszym wykorzystanym modelem był las losowy, który opiera się na zasadzie tworzenia wielu drzew decyzyjnych (Jaiswal i Samikannu, 2017). Do optymalizacji hiperparametrów:

- *n_estimators* – liczba drzew;
- *max_depth* – maksymalna głębokość każdego drzewa;
- *max features* – maksymalna liczba zmiennych brana pod uwagę w modelu;
- *min_samples_split* – minimalna liczba próbek wymagana do podziału węzła;

zdecydowano się wykorzystać bibliotekę Hyperopt (Bergstra i in., 2015). Celem tego procesu było minimalizowanie błędu średniokwadratowego (RMSE), będącego miarą dopasowania modelu do danych (Chicco i in., 2021). Podczas procesu doboru hiperparametrów wygenerowano 500 modeli Random Forest i wybrano najlepszy, czyli o najmniejszym błędzie RMSE.

Drugim wykonanym modelem uczenia maszynowego był XGBoost (Ramraj i in., 2016). Podobnie jak w przypadku modelu Random Forest, również tutaj wykorzy-

stano bibliotekę Hyperopt do doboru optymalnych hiperparametrów. Proces strojenia hiperparametrów obejmował przetestowanie 500 różnych wartości:

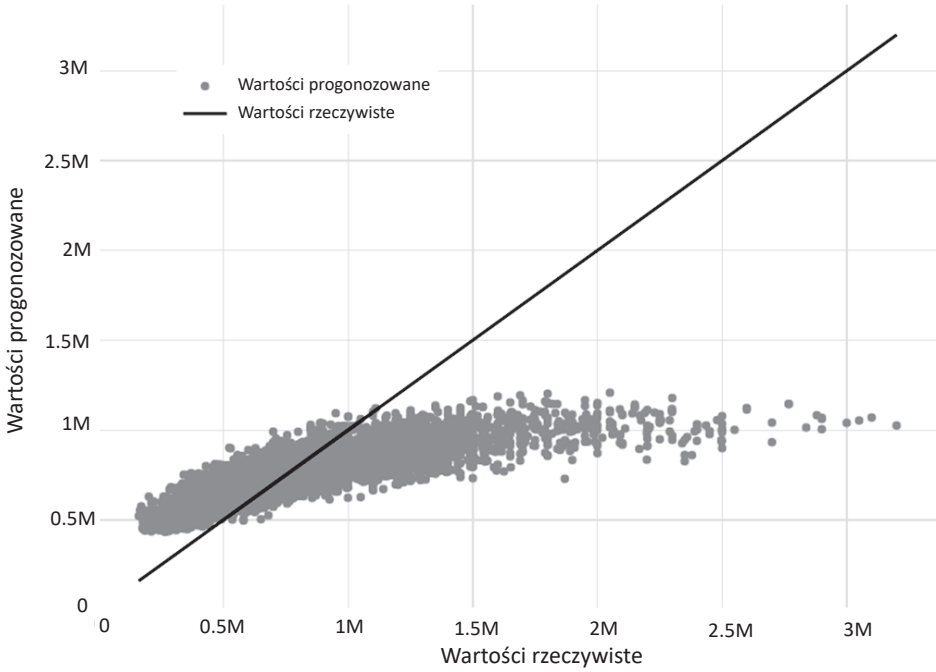
- *n_estimators* – liczba drzew decyzyjnych budowanych w sekwencji;
- *max_depth* – maksymalna głębokość każdego drzewa;
- *learning_rate* – szybkość, z jaką model się uczy;
- *gamma* – parametr, który kontroluje podział węzła podczas budowy drzewa;
- *min_child_weight* – minimalna suma wag obserwacji w danym węźle, aby ten węzeł mógł być podzielony;
- *subsample* – procent przypadków ze zbioru treningowego, który jest używany do trenowania każdego drzewa,
- *colsample_bytree* – kontroluje, ile zmiennych użyto do treningu każdego drzewa w modelu;
- *reg_alpha* – parametr regularyzacji L1 (Lasso). Dodanie składnika regularyzacji L1 pomaga modelowi bardziej skupić się na istotnych cechach;
- *reg_lambda* – parametr regularyzacji L2 (Ridge). Dodanie składnika regularyzacji L2 pomaga kontrolować wielkość wag, zapobiegając im nadmiernemu wzrostowi;

aby znaleźć optymalny zestaw dla modelu XGBoost. Spośród tych 500 modeli również wybrano ten, który osiągnął najmniejszą wartość błędu średniokwadratowego (RMSE).

4.4. Porównanie i ocena modeli uczenia maszynowego

Najlepszy model Random Forest osiągnął następujące wartości błędów: MAE wyniosło 113 032,47, co oznacza, że średnio przewidywana cena mieszkania różni się o około 113 032,47 złotych od rzeczywistej ceny; MAPE na poziomie 0,17 wskazuje, że średni względny błąd predykcji wynosi około 17%, co sugeruje, że model przewiduje ceny mieszkań z dokładnością na poziomie 83%; RMSE wyniosło 167 930,67; współczynnik determinacji osiągnął wartość 0,774, co sugeruje, że model stosunkowo dobrze dopasowuje się do danych.

W celu jego dodatkowej oceny stworzono wykres pokazujący porównanie wartości rzeczywistych i prognozowanych (rys. 3). Idealnym dopasowaniem byłoby, gdyby punkty znajdywały się na prostej. Model dobrze przewiduje ceny mieszkań nie droższych niż 1,25 mln. W przypadku droższych mieszkań można zauważyć występowanie rozbieżności między wartościami przewidywanymi a rzeczywistością.

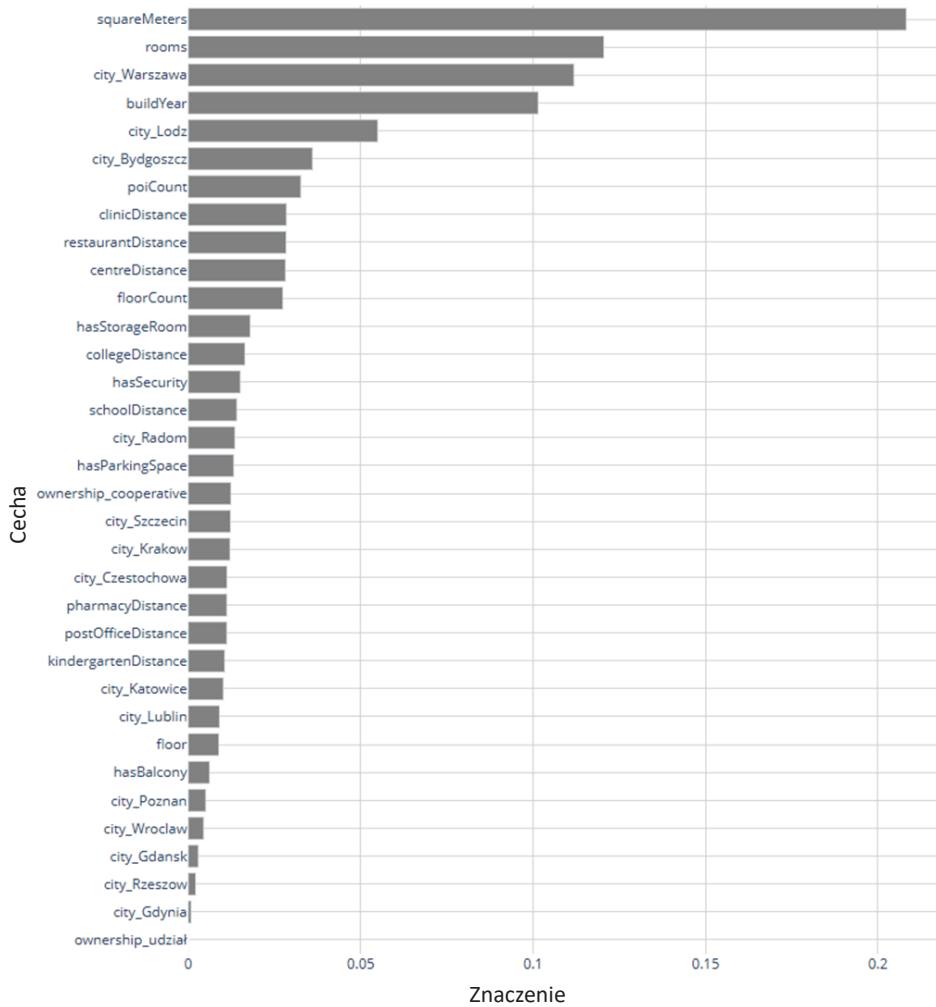


Rys. 3. Porównanie przewidywanych wartości modelu Random Forest z rzeczywistymi cenami mieszkań

Źródło: opracowanie własne.

Na predykcje modelu ceny mieszkań największy wpływ miała zmienna dotycząca wielkości mieszkania (rys. 4): powierzchnia w metrach kwadratowych (20%). Duże znaczenie miała również liczba pokoi w mieszkaniu, informacja, czy mieszkanie znajduje się w Warszawie, oraz rok budowy (po około 10-12%).

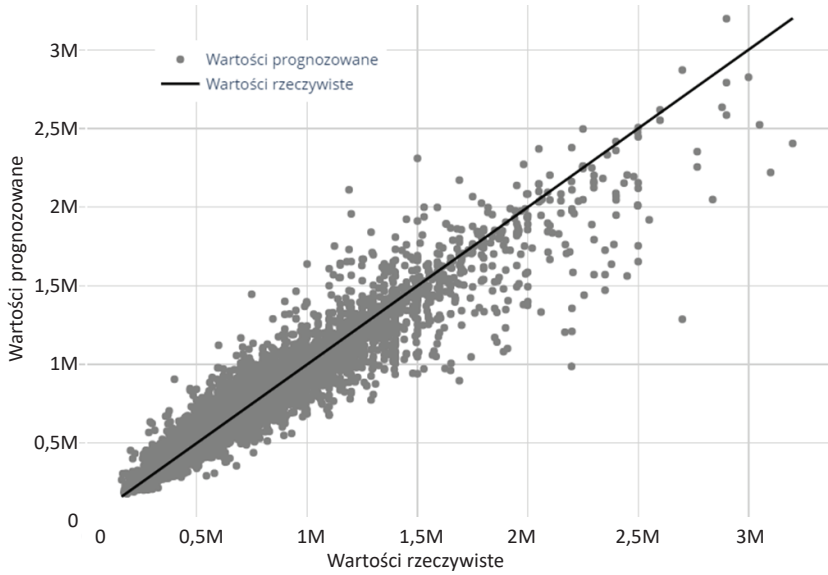
W przypadku modelu XGBoost, najlepszy z modeli osiągnął średni błąd bezwzględny (MAE) około 62 967,83, co oznacza, że przewidywane ceny mieszkań różnią się średnio o około 62 967 złotych od rzeczywistych wartości. Wartość błędu względnego (MAPE) wynosząca 0,09 wskazuje, że model przewiduje ceny z dużą dokładnością, błąd wynosi zaledwie 9%. Dodatkowo niska wartość RMSE (103 939,56) sugeruje, że model ma niską wariancję i dobrze dopasowuje się do danych treningowych. Warto również zauważyć wysoki wynik współczynnika determinacji na poziomie 0,913. Jest to znak, że model XGBoost dobrze odwzorowuje zależności między zmiennymi a cenami mieszkań, co potwierdza jego wysoką skuteczność w przewidywaniu cen mieszkań na podstawie analizowanych danych. Na rysunku 5 widać, że model dobrze przewiduje ceny mieszkań tańszych i gorszych, a punkty rozproszone są stosunkowo blisko prostej.



Rys. 4. Ważność zmiennych w modelu Random Forest

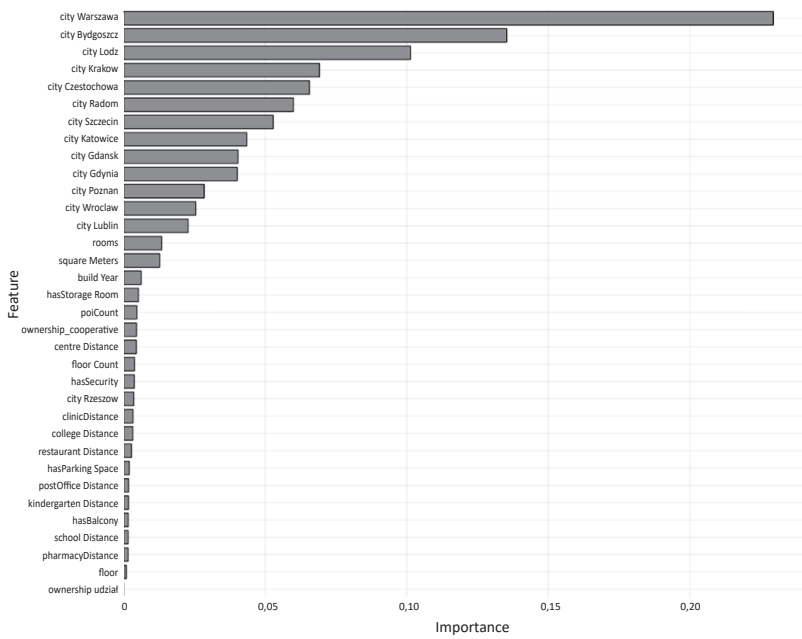
Źródło: opracowanie własne.

Największy wpływ na predykcje modelu miały zmienne dotyczące miasta, w którym znajduje się mieszkanie (rys. 6), w szczególności, czy mieszkanie znajduje się w Warszawie (23%), Bydgoszczy (13,5%) i Łodzi (10 %). W znacznie mniejszym stopniu na prognozy modelu wpływała wielkość mieszkania, czyli powierzchnia w metrach kwadratowych i liczba pokoi (po około 2%).



Rys. 5. Porównanie przewidywanych wartości modelu XGBoost z rzeczywistymi cenami mieszkań

Źródło: opracowanie własne.



Rys. 6. Ważność cech w modelu XGBoost

Źródło: opracowanie własne.

Podsumowując, model XGBoost w porównaniu z Random Forest jest zdecydowanie lepiej dopasowany do danych, czyli jego przewidywania są dużo bardziej precyzyjne. Oznacza to, że na cenę mieszkań najbardziej wpływa miejscowość, w której się ono znajduje.

5. Zakończenie

Ewolucja polskiego rynku nieruchomości od lat 90. XX wieku do współczesności wyraża złożoność procesów społecznych, ekonomicznych i politycznych, które kształtują tę dziedzinę. W obliczu przemian demograficznych, zmieniających się preferencji mieszkańców oraz dynamicznych trendów rynkowych kluczowym wyzwaniem jest wycena mieszkań.

Przeprowadzone badanie dostarczyło istotnych wniosków dotyczących zastosowania metod uczenia maszynowego w prognozowaniu cen mieszkań w dużych miastach. Porównanie modeli Random Forest i XGBoost wykazało, że model XGBoost osiąga mniejsze błędy prognoz (RMSE na poziomie około 104 000 złotych) i wyższą wartość współczynnika determinacji (0,91). Model Random Forest, chociaż wykazał się RMSE wynoszącym około 168 000 złotych, dobrze prognozował ceny dla segmentu mieszkań o wartości poniżej miliona złotych, mając trudności z przewidywaniem cen wyższych, czyli mieszkań dużych lub luksusowych.

Istotnymi czynnikami wpływającymi na predykcje ceny mieszkań okazała się lokalizacja, zwłaszcza miasto, w którym znajduje się nieruchomość. Oprócz tego przeprowadzone wizualizacje potwierdziły dobrą jakość predykcji modelu XGBoost, który skutecznie przewidywał ceny mieszkań zarówno w niższym, jak i wyższym przedziale cenowym.

Literatura

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., i Cox, D. D. (2015). Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. *Computational Science & Discovery*, *8*(1). https://doi.org/10.1088/1749-4699/8/1/014008

Bryx, M. (2006). *Rynek nieruchomości system i funkcjonowanie*. Wydawnictwo Poltex.

Chicco, D., Warrens, M. J., i Jurman, G. (2021). The Coefficient of Determination R-squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *Peerj Computer Science*, (7). https://doi.org/10.7717/peerj-cs.623

FIN-HUB. (b.d.). *Bezpieczny Kredyt 2%*. Pobrano z https://finhub.pl/slownik/bezpieczny-kredyt-2

Gołąbeska, E. (2007). *Rynek nieruchomości i jego podmioty*. Wydawnictwo Wyższej Szkoły Finansów i Zarządzania w Białymstoku.

Gołąbeska, E. (2017). Współczesne trendy na rynku nieruchomości mieszkaniowych. W: E. Broniewicz (red.), *Gospodarowanie przestrzenią w warunkach rozwoju zrównoważonego*, (s. 85-106). Oficyna Wydawnicza Politechniki Białostockiej.

Jaiswal, J. K., i Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. *Conference: 2017 World Congress on Computing and Communication Technologies (WCCCT)*, 65-68. https://doi.org/10.1109/WCCCT.2016.25

Jamroz, K. (2024). *Apartment Prices in Poland*. Pobrano z https://www.kaggle.com/datasets/krzysztofjamroz/apartment-prices-in-poland?resource=download

Kucharska-Stasiak, E. (2000). *Nieruchomość a rynek*. Wydawnictwo Naukowe PWN.

Kucharska-Stasiak, E., Załęczna M. i Żelazowski K. (2012). *Wpływ procesu integracji Polski z Unią Europejską na rozwój rynków nieruchomości*. Wydawnictwo Uniwersytetu Łódzkiego.

Majorek, M. (2013). Rządowy program „Rodzina na swoim". Społeczne, ekonomiczne i medialne reperkusje wdrażania. *Państwo i Społeczeństwo*, *13*(3), 11-30.

Malarvizhi, R., i Thanamani, A. S. (2012). K-nearest Neighbor in Missing Data Imputation. *International Journal of Engineering Research and Development*, *5*(1), 5-7.

Ramraj, S., Uzir, N., Sunil, R., i Banerjee, S. (2016). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. *International Journal of Control Theory and Applications*, *9*(40), 651-662.

Sadowski, W. (1976). *Mała Encyklopedia Statystyki*. Państwowe Wydawnictwo Ekonomiczne Warszawa.

Samorek, B. i Cichocki, M. (2023). *Polski rynek nieruchomości mieszkaniowych*. https://doi.org/10.51733/opm.2023.03

Westerlund, O. (2023). *Cluster Analysis for Outlier Detection. A Case Study of Applying Unsupervised Machine Learning on Diesel Engine Data*. University of Turku. Department of Computing, Faculty of Technology. Pobrano z https://www.utupub.fi/bitstream/handle/10024/174378/Westerlund_Otto_opinnayte.pdf?sequence=1&isAllowed=y

Predicting Housing Prices in Major Cities in Poland Using Machine Learning

Abstract: The article provides an in-depth analysis of the real estate market in major cities in Poland, with a primary focus on predicting housing prices. It discusses the concept of the real estate market and its key features, explaining in detail how housing prices have changed over the years in Poland. In the 1990s, housing prices saw an increase due to liberalisation and easy access to mortgage loans. However, after the financial crisis of 2007, the market experienced significant fluctuations. Compared to the 2% Safe Mortgage programme, the COVID-19 pandemic had a minor impact on prices. Analysing housing prices, the article describes data preparation, imputing missing values, and identifying outliers. Machine learning algorithms, such as Random Forest and XGBoost, were used to forecast prices. These models demonstrated high effectiveness (especially XGBoost). The article analyses key variables influencing price formation in the real estate market in major cities in Poland.

Keywords: real estate market, housing prices, machine learning, Random Forest model, XGBoost model