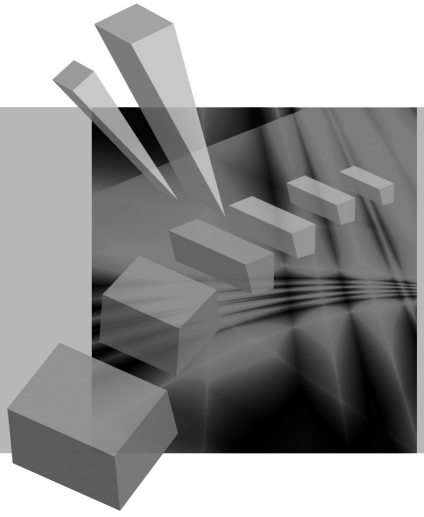
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

278

Taksonomia 20

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 278 • 2013

Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Michał Trzęsiok
Uniwersytet Ekonomiczny w Katowicach

WYCENA RYNKOWEJ WARTOŚCI NIERUCHOMOŚCI Z WYKORZYSTANIEM WYBRANYCH METOD WIELOWYMIAROWEJ ANALIZY STATYSTYCZNEJ

Streszczenie: W artykule przedstawiono próbę wykorzystania wybranych nieparametrycznych metod regresji do analizy cen transakcyjnych mieszkań w Warszawie. Zaprezentowano procedurę symulacyjnego doboru najlepszego, w sensie dokładności predykcji, narzędzia analizy danego zjawiska oraz metodę uzyskania dodatkowej wiedzy o kształtowaniu się cen transakcyjnych mieszkań oraz o istotności wpływu poszczególnych zmiennych objaśniających.

Słowa kluczowe: nieparametryczne metody regresji, wielowymiarowa analiza porównawcza, rynek nieruchomości.

1. Wstęp

Celem artykułu jest próba sprawdzenia możliwości wykorzystania wybranych statystycznych metod uczących się (*machine learning, data mining*) do szacowania wartości rynkowej mieszkań zlokalizowanych w Warszawie. Metody wyceny nieruchomości są nieustannie rozwijane.

Analiza danych dotyczących rynku nieruchomości wiąże się z licznymi problemami i wymaganiami, którym metody analizy statystycznej muszą sprostać. Należą do nich: asymetria rozkładów badanych zmiennych, występowanie zmiennych objaśniających mierzonych na różnych skalach – zarówno słabych, jak i mocnych, występowanie wartości oddalonych, czy ogólnie – błędy pomiaru w danych. Nieparametryczne metody regresji zostały stworzone jako narzędzia analizy, które mogą być stosowane również w przypadku tego typu problemów [Trzęsiok, Trzęsiok 2009]. Brak jednak jednoznacznych wskazań, którą z metod należy wykorzystać do danego problemu. Do wyboru metody najczęściej stosuje się podejście symulacyjne – buduje się wiele modeli i wybiera się model o najlepszych zdolnościach predykcyjnych.

Podstawową wadą statystycznych metod uczących się jest ich bardzo ograniczona interpretowalność [Guyon i in. 2006; Trzęsiok 2010]. W artykule wskazano najlepsze, w sensie dokładności predykcji, narzędzia analizy oraz wykorzystano dodatkową procedurę pozwalającą na interpretację zbudowanego modelu – pozyskanie

dodatkowej wiedzy o kształtowaniu się cen transakcyjnych mieszkań oraz o istotności wpływu poszczególnych zmiennych objaśniających.

2. Procedura badawcza

Zbiór danych poddany analizie

Zbiór danych rzeczywistych *mieszkania* został utworzony na podstawie danych o zrealizowanych transakcjach sprzedaży mieszkań, udostępnianych przez serwis internetowy `www.oferty.net`. Z kolei dane tego serwisu pochodzą z 16 warszawskich biur obrotu nieruchomościami współpracujących z serwisem `oferty.net`[1]. Dane dotyczą transakcji sprzedaży mieszkań zrealizowanych od 2007 do 2010 r.

Podkreślić należy, że wykorzystane w analizie wartości zmiennej zależnej to ceny *transakcyjne*. Ze względu na występowanie dużych różnic między cenami ofertowymi mieszkań a cenami transakcyjnymi jakość danych, a co za tym idzie – również jakość budowanych modeli statystycznych z wykorzystaniem danych transakcyjnych, jest znacznie wyższa.

Podstawowe charakterystyki zbioru *mieszkania* przedstawiono w tab. 1.

Tabela 1. Charakterystyki zbioru danych *mieszkania*

Liczebność zbioru	Liczba zmiennych objaśniających		
	ilorazowych	porządkowych	nominalnych
990	5	1	2

Źródło: opracowanie własne.

Zmienne objaśniające (pierwotne) opisujące mieszkania zawarte w zbiorze uczącym to:

X_1 – powierzchnia użytkowa mieszkania (m^2),

X_2 – lokalizacja (nazwa dzielnicy Warszawy),

X_3 – odległość mieszkania od centrum (km),

X_4 – liczba pokoi,

X_5 – kondygnacja (1 – parter, 2 – mieszkanie na pierwszym piętrze itd.),

[1] Biura nieruchomości zasilające bazę danych serwisu `oferty.net`, to: *4 ŚCIANY Sp. z o.o.*, *AD DRĄGOWSKI*, *ADAMPOLSKI Nieruchomości s.c.*, *Akces Nieruchomości S.C.*, *Bracia Strzelczyk Sp. z o.o.*, *CENTURY 21 Nieruchomości WS*, *Emmerson SA, Warszawa oddział Atrium Centrum*, *Habitats Real Estate Sp. z o.o.*, *Maxon Nieruchomości Sp. z o.o.*, *MW Nieruchomości*, *Neodom Agencja Nieruchomości Oddział I Warszawa*, *NeSUS Nieruchomości*, *Open Home Sp. z o.o.*, *PRIMO Biuro Nieruchomości*, *Unia Nieruchomości Sp. z o.o.*, *UNIWERS Nieruchomości Doradztwo Prawne* oraz *WARSZAWIAK Nieruchomości.*

X_6 – rok oddania do użytkowania,

X_7 – typ własności (mieszkanie: spółdzielcze, własnościowe, hipoteczne, spółdzielczo-własnościowe)[2],

X_8 – stan mieszkania (5 – bardzo dobry, 4 – dobry, 3 – do wykończenia, 2 – do remontu)[3].

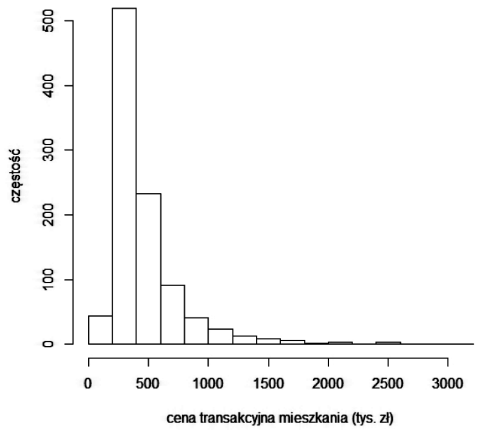
Zmienną zależną (Y) była cena transakcyjna mieszkania (tys. zł).

Ze względu na silną asymetrię prawostronną zmiennej zależnej (por. tab. 2 i rys. 1) oraz wobec występowania zmiennych nominalnych zbiór danych pierwotnych, przed przystąpieniem do analizy, przekształcono, zastępując zmienne

Tabela 2. Struktura badanych mieszkań ze względu na cenę transakcyjną

Średnia $\bar{y}$	Współczynnik zmienności $V_s(y)$	Współczynnik asymetrii $\lambda_3(y)$
460 tys. zł	68%	3,2
Min	Me	Max
160 tys. zł	366 tys. zł	2 770 tys. zł

Źródło: opracowanie własne.



Rys. 1. Struktura badanych mieszkań ze względu na cenę transakcyjną

Źródło: opracowanie własne.

[2] Kategorie tej zmiennej nie są zgodne z obowiązującym stanem prawnym. Taką klasyfikacją posługują się jednak rozważane warszawskie biura obrotu nieruchomościami. Autor artykułu dysponuje takim zbiorem danych, a przedstawione kategorie zmiennych wynikają z analizowanego materiału statystycznego.

[3] Kategorie tej zmiennej nie zostały wyjaśnione przez publikujący dane serwis `oferty.net`. Pozostawiono jednak oryginalny zestaw zmiennych wraz z nazwami kategorii, mimo braku wyjaśnień używanych nazw kategorii, gdyż głównym celem analizy było sprawdzenie adekwatności wybranych metod WAS do szacowania wartości rynkowej nieruchomości, nie zaś analiza jakości danych prezentowanych przez portal `oferty.net`.

nominalne zmiennymi sztucznymi, usuwając obserwacje, w których występowały braki wartości, oraz przeskalowując zmienną zależną – dzieląc jej wartości przez powierzchnię użytkową mieszkania. W ten sposób w analizowanym zbiorze danych zmienną zależną była cena jednego metra kwadratowego mieszkania.

Charakterystyki przetworzonego zbioru danych poddanego dalszej analizie przedstawiono w tab. 3.

Tabela 3. Charakterystyki przetworzonego zbioru danych *mieszkania* poddanego dalszej analizie

Liczebność zbioru	Liczba zmiennych objaśniających		
	ilorazowych	porządkowych	sztucznych
747	5	1	22

Źródło: opracowanie własne.

Zadanie polegało na znalezieniu zależności pozwalającej wyznaczać wartość rynkową 1 m^2 mieszkania na podstawie przedstawionych cech objaśniających. Cena nieruchomości jest zmienną mierzoną na skali mocnej (ilorazowej), więc przedstawione zadanie jest zadaniem *analizy regresji*.

Wybrane metody wielowymiarowej analizy statystycznej jako potencjalne narzędzia analizy

Do utworzenia modelu wyceny wykorzystano następujące nieparametryczne metody regresji:

BAGGING – metoda łączenia równoległego drzew regresyjnych,
RFOREST – zagregowane drzewa regresyjne Breimana,
MART – addytywna metoda drzew regresyjnych MART,
PPR – metoda rzutowania,
NNET – sieci neuronowe,
SVM – metoda wektorów nośnych,
FLEXMIX – metoda regresji wykorzystująca mieszanki rozkładów.

Wszystkie obliczenia przeprowadzone zostały z wykorzystaniem programu statystycznego **R** z dołączonymi bibliotekami oraz autorskimi procedurami programu **R**. Wykorzystane nazwy metod regresji odpowiadają nazwom funkcji programu **R**. Szczegółowy opis wykorzystanych metod można znaleźć w pracach: [Gatnar 2001; Gatnar 2008; Hastie i in. 2001; Walesiak, Gatnar (red.) 2009].

Kryterium wyboru metody

Zbudowano wiele modeli wyceny mieszkań, wykorzystując wszystkie rozważane metody regresji i dobierając za każdym razem optymalnie parametry danej metody[4]. Ostatecznie wybrano tę metodę, która zbudowała model o największej *do-*

[4] Mnogość zbudowanych modeli wyceny wynika z ich budowania siedmioma różnymi metodami, przy czym każdej pojedynczej metodzie odpowiada kilka modeli dla różnych kombinacji jej wewnętrznych parametrów.

kładności predykcji, zaś dokładność predykcji oszacowano, obliczając wartość błędu średniokwadratowego metodą *b*–CV (MSE_{CV}), tj. metodą sprawdzania krzyżowego z podziałem zbioru uczącego na $b = 10$ części. W ten sposób ocena zdolności predykcyjnych modeli była realizowana na obserwacjach nieuczestniczących w budowie modelu. Losowy podział zbioru na $b = 10$ części został zrealizowany na wstępie analizy, aby każda z rozpatrywanych metod budowała modele na takim samym zestawie zbiorów uczących.

3. Wyniki analizy

W tabeli 4 przedstawiono obliczone na zbiorze *mieszkania* wartości błędów MSE_{CV} dla każdej z rozpatrywanych metod.

Tabela 4. Wartości błędów średniokwadratowych obliczonych metodą sprawdzania krzyżowego dla zbioru *mieszkania* dla różnych metod regresji

Metoda regresji	Błąd MSE_{CV}	Współczynnik determinacji R^2
RFOREST	1,989	0,924
MART	1,997	0,648
BAGGING	2,146	0,762
SVM	2,532	0,685
PPR	2,563	0,448
NNET	2,634	0,434
FLEXMIX	3,081	0,291

Źródło: opracowanie własne.

Najlepsze własności predykcyjne w przypadku zbioru danych *mieszkania* ma model zbudowany metodą zagregowanych drzew regresyjnych Breimana (*random forest*). W związku z tym model ten został wykorzystany w dalszej analizie. Warto jednak zwrócić uwagę na niewielkie różnice w zdolnościach predykcyjnych, występujące między najlepszymi trzema z rozważanych modeli. Oznacza to, że dla danego zbioru danych alternatywnie można wykorzystać do zbudowania modelu wyceny nieruchomości metodę MART lub BAGGING. Na szczególną uwagę zasługuje fakt, że owe trzy najwyżej sklasyfikowane metody to metody wykorzystujące podejście wielomodelowe (budują modele zagregowane), gdzie pojedyncze modele to drzewa regresyjne. Potwierdza to liczne wyniki wskazujące na dobre własności modeli drzew regresyjnych oraz zasadność stosowania technik łączenia modeli [Gatnar 2008].

Dodatkowo w tab. 4 przedstawiono wartości współczynnika determinacji. Z porównania stopnia dopasowania modeli do danych ze zbioru uczącego wynika, że model zbudowany metodą RFOREST jest nie tylko najlepszy pod względem dokładności predykcji, ale również najlepiej dopasowany do danych.

W kolejnym etapie analizy dokonano klasyfikacji zmiennych objaśniających na zmienne istotne i redundantne. W tym celu zastosowano uniwersalną iteracyjną procedurę eliminacji zmiennych pojedynczo, która w każdym kroku usuwa jedną zmienną – tę, której usunięcie powoduje poprawę bądź najmniejsze pogorszenie zdolności predykcyjnej modelu zbudowanego na zmniejszonym zestawie zmiennych [Trzęsiok 2009; Guyon i in. 2006]. Wyniki działania procedury identyfikacji zmiennych nieistotnych przedstawiono w tab. 5.

Tabela 5. Wynik działania procedury eliminacji zmiennych na zbiorze *mieszkania*

Numer iteracji	Numer usuniętej zmiennej	Błąd klasyfikacji MSE_{CV}	Błąd standardowy pomiaru błędu MSE_{CV}
		2,008	0,480
1	1	1,984	0,520
2	4	2,084	0,515
3	2	2,099	0,533
4	3	2,176	0,609
5	5	2,256	0,661
6	6	2,420	0,808
7	**7**	2,522	0,879
8	**8**		

Źródło: opracowanie własne.

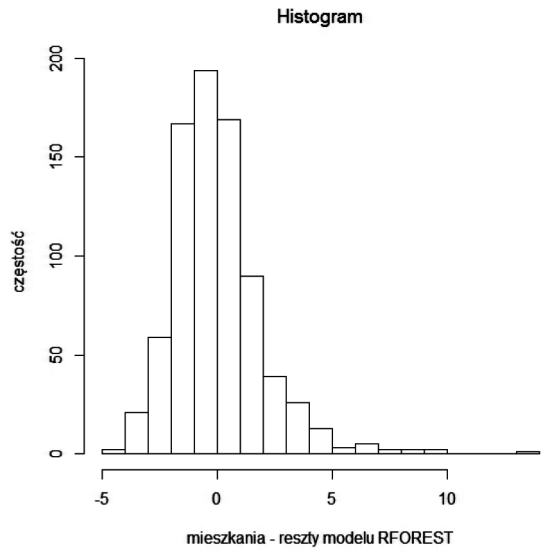
Po pierwszej iteracji (po usunięciu pierwszej zmiennej) uzyskano najmniejszy błąd MSE_{CV} = 1,984. Uwzględnienie odpowiadającego mu błędu standardowego pomiaru (równego 0,52) pozwoliło na wskazanie jako najlepszego modelu zbudowanego w 6. iteracji (błąd tego modelu jest nie większy niż minimalny błąd MSE_{CV} powiększony o odpowiadający mu błąd pomiaru). Oznacza to, że zmienne zidentyfikowane w pierwszych sześciu iteracjach zostają usunięte z dalszej analizy jako nieistotne. Jedynie dwie zmienne zostały więc zidentyfikowane jako istotne. Model zbudowany z wykorzystaniem zestawu tylko tych dwóch zmiennych ma, z dokładnością do błędu pomiaru, taką samą zdolność predykcyjną jak najlepszy ze zbudowanych modeli, czyli również jak model zbudowany na zestawie zawierającym komplet zmiennych objaśniających. Największy wpływ na zmienną zależną ma zmienna nr 8 opisująca *stan mieszkania*. Druga ze zmiennych istotnych to X_7 – *typ własności*. Ranking uwzględniający wszystkie zmienne przedstawiono w tab. 6.

Tabela 6. Ranking zmiennych w modelu wyceny nieruchomości zbudowanym metodą RFOREST na zbiorze *mieszkania*

Pozycja w rankingu	Numer zmiennej	Nazwa zmiennej
1	**8**	Stan mieszkania
2	**7**	Typ własności
3	6	Rok oddania do użytkowania
4	5	Kondygnacja
5	3	Odległość od centrum
6	2	Dzielnica
7	4	Liczba pokoi
8	1	Powierzchnia użytkowa

Źródło: opracowanie własne.

Najmniej istotna okazała się zmienna X_1 wskazująca powierzchnię użytkową mieszkania, co może się wydawać niepokojące, lecz należy pamiętać, że wskutek przeskalowania zmiennej zależnej (podzielenie ceny transakcyjnej mieszkania przez jego powierzchnię użytkową) zależność Y od X_1 została wyeliminowana.



Rys. 2. Struktura reszt modelu RFOREST

Źródło: opracowanie własne.

W kolejnym kroku analizy zbudowano model, wykorzystując metodę RFOREST oraz zbiór danych, w którym pozostawiono wyłącznie zmienne X_8 i X_7, zidentyfiko-

wane jako istotne. Wykorzystując podział zbioru uczącego na 10 części z metody sprawdzania krzyżowego, ponownie budowano modele na 9 spośród 10 części, wyznaczano predykcję dla obserwacji z wyodrębnionej jednej części i obliczono w ten sposób reszty modelu. Czyniąc tak dla każdej z 10 części, otrzymano zbiór reszt modelu, który poddano analizie. Strukturę reszt modelu przedstawiono na rys. 2 oraz podsumowano statystykami pozycyjnymi w tab. 7.

Tabela 7. Kwartyle reszt bezwzględnych modelu

Q_1	*Me*	Q_3
0,5 tys. zł	1,1 tys. zł	1,9 tys. zł

Źródło: opracowanie własne.

Tabela 8. Kwartyle ceny transakcyjnej 1 metra kwadratowego mieszkania

Q_1	Me	Q_3
7,4 tys. zł	8,5 tys. zł	9,8 tys. zł

Źródło: opracowanie własne.

Nie należy do roli statystyka oceniać, czy obliczona mediana reszt modelu równa 1,1 tys. zł, czyli informacja o środkowej wartości odchyleń wartości prognozowanej przez model od wartości rzeczywistej, to dużo czy mało. Dla porządku należy jednak zestawić informacje dotyczące reszt modelu ze statystykami pozycyjnymi zmiennej zależnej (por. tab. 8).

Wydaje się uprawniony wniosek, że metody zagregowanych drzew regresyjnych można skutecznie wykorzystywać do budowania modeli wyceny nieruchomości. Metody te nie wymagają kodowania numerycznego zmiennych nominalnych. Zbudowany model charakteryzuje się bardzo dobrym dopasowaniem ($R^2 = 0{,}924$), lecz tylko stosunkowo dobrą dokładnością predykcji (o czym świadczą kwartale reszt modelu przedstawione w tab. 7). Informacje zawarte w analizowanym zbiorze danych nie pozwoliły jednak na zbudowanie modelu, który mógłby służyć jako *automatyczne* narzędzie wyceny nieruchomości. Uzyskanie dobrej wyceny wartości rynkowej mieszkania wymaga więc nie tylko wykorzystania zbudowanego modelu, ale również dodatkowej ingerencji eksperta.

4. Podsumowanie

W przypadku analizowanego zbioru danych z rynku nieruchomości zbudowane modele statystyczne, wykorzystujące zagregowane modele drzew regresyjnych, należy traktować jako narzędzia wspomagające proces podejmowania decyzji, a końcowa wycena powinna uwzględniać dodatkowe informacje (o ile są dostępne) oraz opinie eksperta. Zastosowanie dodatkowych procedur wspomagających interpretowanie wyników modelowania metodami regresji nieparametrycznej pozwala na uproszczenie modelu bez zmniejszania jego zdolności predykcyjnych oraz na pozyskanie dodatkowej wiedzy o wpływie poszczególnych zmiennych diagnostycznych na cenę nieruchomości.

Literatura

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, „Biblioteka Ekonometryczna", Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Guyon I., Gunn S., Nikravesh M., Zadeh L. (red.) (2006), *Feature Extraction, Foundations and Applications*, Springer.

Hastie T., Tibshirani R., Friedman J. (2001), *The Elements of Statistical Learning*, Springer Verlag, N.Y.

Trzęsiok J., Trzęsiok M. (2009), *Nieparametryczne metody regresji*, [w:] M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu* ***R***, Wydawnictwo Naukowe PWN, Warszawa, s. 156-192.

Trzęsiok M. (2009), *Problem doboru zmiennych do modelu dyskryminacyjnego budowanego metodą wektorów nośnych,* [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 16, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu nr 47, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, s. 214-222.

Trzęsiok M. (2010), *Wyodrębnianie reguł klasyfikacyjnych z modelu dyskryminacyjnego budowanego metodą wektorów nośnych,* [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 17, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu nr 107, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, s. 316-324.

Walesiak M., Gatnar E. (red.) (2009), *Statystyczna analiza danych z wykorzystaniem programu* **R**, Wydawnictwo Naukowe PWN, Warszawa.

REAL ESTATE MARKET VALUE ESTIMATION BASED ON MULTIVARIATE STATISTICAL ANALYSIS

Summary: The paper presents a nonparametric regression approach to the market value estimation problem for flats in Warsaw. We present a procedure for choosing the best model in terms of predictive ability, but also show how to fix the machine learning' lack of interpretation by extracting knowledge about variable importance.

Keywords: nonparametric regression, multivariate statistical analysis, real estate market.