ZASTOSOWANIE REGRESJI LOGISTYCZNEJ DO OKREŚLENIA PRAWDOPODOBIEŃSTWA SPRZEDAŻY ZASOBU MIESZKANIOWEGO

Łukasz MACH

Streszczenie: W artykule przedstawiono proces budowy modelu regresji logistycznej, którego celem jest wspomaganie procesu podejmowania decyzji na rynku mieszkaniowym. Opracowany model regresji logistycznej, będzie definiował prawdopodobieństwo dokonania transakcji na rynku mieszkaniowym (rynek wtórny) oraz będzie wskazywał zmienne statystycznie istotnie wpływające na kształtowanie się popytu. Wartość uzyskanego prawdopodobieństwa, będzie stanowić jedną z podstawowych przesłanek decyzyjnych.

Słowa kluczowe: prognozowanie, regresja logistyczna, podejmowanie decyzji.

1. Wprowadzenie

Poprawnie przygotowany oraz merytorycznie przeprowadzony proces podejmowania decyzji jest kluczowym czynnikiem wpływającym na zmniejszenie luki informacyjnej co implikuje poprawę konkurencyjności przedsiębiorstw. Dobrze przygotowany proces decyzyjny pozwala zminimalizować ryzyko źle podjętych decyzji oraz pozwala wzmocnić pozycję konkurencyjną przedsiębiorstw.

W dobie globalnego kryzysu gospodarczego, przedsiębiorstwa powinny szczególną uwagę przywiązywać do trafnych decyzji, szczególnie w branżach szczególnie narażonych na oddziaływanie kryzysu. Uwzględniając teorię ekonomii, głównym czynnikiem wpływającym na konkurencyjność gospodarek są inwestycje, które w znacznym stopniu są implikowane przez branżę budowlaną (m.in.. łatwość sprzedaży zasobu mieszkaniowego).

Cele niniejszego artykułu, jest opracowanie jakościowego modelu decyzyjnego, którego głównym zadaniem będzie określenie prawdopodobieństwa dokonania transakcji na rynku nieruchomości oraz wskazanie zmiennych istotnie wpływających na dokonanie transakcji. Zdefiniowany model bazuje na regresji logistycznej, która z powodzeniem jest wykorzystywana w procesach decyzyjnych, w których zmienna zależna na charakter dychotomiczny. Wybór modelu bazującego na funkcji logistycznej ma również swoje uzasadnienie, gdyż większości zjawisk społeczno-ekonomicznych, ma parametry zbliżone do przebiegu funkcji logistycznej.

Artykuł składa się z trzech głównych części. W pierwszej opisano możliwość analizy zmiennej zależnej typu dychotomicznego za pomocą regresji logistycznej. Część druga pokazuję przykład praktycznego zastosowania regresji logistycznej do wyznaczania prawdopodobieństwa sukcesu (sprzedaży mieszkania), natomiast część trzecia to wnioski z przeprowadzonych badań. Zastosowanie trójetapowego podejścia opisywanego problemu pozwala na jego kompleksowe rozwiązanie.

2. Analiza danych jakościowych za pomocą regresji logistycznej

W wielu przypadkach procesy decyzyjne bazują na modelach regresji wielorakiej, tzn., takiej, w której analizujemy wpływ kilku zmiennych niezależnych na jedną zmienną zależną typu mierzalnego [2,5]. Autor niniejszej publikacji zajmował się tym zagadnieniem we wcześniejszych pracach badawczych, w których za pomocą modelowania liniowego (regresji wielorakiej), wskazywał na istotne determinanty określające cenę nieruchomości na rynku mieszkaniowym [4].

Natomiast w sytuacji, gdy zmienna zależna jest typu dychotomicznego, powinniśmy zastosować regresją logistyczną. W badaniach ekonomicznych, bardzo popularnym przykładem zastosowania regresji logistycznej jest analiza zdolności spłaty zaciągniętych kredytów bankowych, natomiast w badaniach społecznych możliwość wskazania prawdopodobieństwa np. zakupu nowego produktu przez klienta, z punktu widzenia określonych (statystycznie istotnych) cech dotyczących produktu oraz specyfiki procesu decyzyjnego nabywcy.

Autor kontynuując wcześniejsze badania dotyczące rynku mieszkaniowego (dotyczące zastosowania regresji wielorakiej), w niniejszej publikacji zastosował regresję logistyczną do określenia prawdopodobieństwa sprzedaży nieruchomości mieszkaniowej, z punktu widzenia czasu oczekiwania nieruchomości na sprzedaż (prawdopodobieństwo sprzedaży nieruchomości w czasie krótszym od średniego czasu oczekiwania).

Zaletą regresji logistycznej jest to, że interpretacja wyników jest bardzo podobna do metod stosowanych w regresji klasycznej. Jednakże, należy również zaznaczyć, że regresja logistyczna w porównaniu do regresji wielorakiej jest bardziej skomplikowaną obliczeniowo, oraz wyliczenie wartości i sporządzenie wykresów reszt często nie wnosi nic nowego do modelu [6].

Logistyczny model regresji dla zmiennej dychotomicznej wyrażony jest wzorem 1 [3,5,6].

$$P(Y = 1/ x_1, x_2, ..., x_k) = \frac{e^{a_0 + \sum_{i=1}^{k} a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^{k} a_i x_i}} \tag{1}$$

gdzie

$a_i, i = 0,1,2,...,k$ - parametry strukturalne modelu w regresji logistycznej,

$x_1, x_2, ..., x_k$ – zmienne niezależne, które mogą być zarówno ilościowe jak i jakościowe.

Do oszacowania współczynników regresji, nie możemy użyć popularnej metody najmniejszych kwadratów, gdyż wymaga ona założenia o stałości wariancji. Z tego względu do estymacji parametrów w regresji logistycznej wykorzystuje się metodę największej wiarygodności (MNW) [3,5,6].

W regresji logistycznej, prócz interpretacji współczynników regresji, dochodzi jeszcze jeden parametr tj. iloraz szans. Jest to stosunek prawdopodobieństwa, że jakieś zdarzenie wystąpi do prawdopodobieństwa, że ten przypadek się nie pojawi. Dla określonego przykładu A, możemy to zapisać w postaci wzoru 2 [6].

$$S(A) = \frac{P(A)}{1 - P(A)} \tag{2}$$

Interpretacja szans wyrażonych wzorem 2 jest przydatna do wyjaśnienia oszacowanego modelu logistycznego. Możemy udowodnić, że gdy wybrana zmienna niezależna wzrośnie o jednostkę, to iloraz szans zmieni się $\exp(a_i)$ razy. Jeśli $\exp(a_i) > 1$ to należy się spodziewać wzrostu ilorazu szans, natomiast gdy $\exp(a_i) < 1$ to jego spadku. W przypadku, gdy zmienna niezależna jest zmienną zero-jedynkową, to $\exp(a_i)$ oznacza, ile razy wzrasta iloraz dla wartości zmiennej zależnej równej jedności [3].

W kolejnym kroku procesu budowy modelu logistycznego powinniśmy określić miary dopasowania modelu oraz dokonać weryfikacji poprawności modelu (testowanie modelu). Miary dopasowania modelu możemy wyrazić za pomocą wartości pseudo-R^2 McFaddena, jak również za pomocą miar R-kwadrat Cragga-Uhlera, Nagelkerke czy Coxa-Snella. W uwagi na nieco inne podejście obliczeniowe, należy zachować szczególną uwagę przy porównywaniu modeli logistycznych, w których zostały określone miary dopasowania różnych autorów [6].

Procedura testowania wyników estymacji w modelowaniu logistycznym jest dokonywana za pomocą wartości statystyki testu ilorazu wiarygodności. Hipoteza zerowa dla tak zdefiniowanych założeń mówi, że wszystkie parametry modelu, bez wyrazu wolnego są równe zeru. Statystyka testu jest wyrażona jest wzorem nr 3 [3,5,6].

$$2(\ln L_{MP} - \ln L_{MZ}) \tag{3}$$

gdzie:

$\ln L_{MP}$ - logarytm funkcji wiarygodności dla modelu pełnego;

$\ln L_{MZ}$ - logarytm funkcji wiarygodności dla modelu zredukowanego;

oraz z założenia ma rozkład λ^2 z liczbą stopni swobody równą liczbie zmiennych objaśniających modelu pełnego.

Należy również zaznaczyć, że dobór zmiennych do modelu jest poddany tym samym wymaganiom, co w klasycznym modelu liniowym.

W końcowym etapie budowy modelu logistycznego można obliczyć prognozę ex-post wartości zmiennej zależnej dla każdej obserwacji. Przyjmuje się, że trafność prognozy ex-post wygonie jest przedstawić za pomocą tablicy trafności. Obliczając prognozę ex-post obowiązują dwie zasady, tj. zasada standardowa stosowana przy próbie zbilansowanej oraz zasada optymalnej wartości granicznej [3].

Założenia poprawności analizy regresji logistycznej [6]:

- wybór próby należy przeprowadzić w sposób losowy,
- w procesie przygotowania danych usunąć dane odstające
- zastosować odpowiednie kodowanie, tzn., zmiennej zależnej wartość jeden przypisać dl przypadku nas interesującego,
- włączyć do modelu wszystkie zmienne istotne statystycznie,
- wyłączyć z modelu wszystkie nieistotne zmienne,
- pamiętać o liniowej zależność transformacji logitowej od zmiennych niezależnych,

- unikać zjawiska współliniowości,
- pamiętać o dużej liczebności próby, tj. powinno być spełnione założenie $N > 10(k+1)$, gdzie k jest liczbą parametrów. Jednakże, za minimalną liczebność powinno się określić na 100 - elementów.

3. Zastosowanie regresji logistycznej w praktyce – studium przypadku

Przy budowie modelu regresji logistycznej w pierwszej kolejności sprawdzono czy występują obserwacje odstające. Proces eliminacji obserwacji odstających przeprowadzono, dla każdej zmiennej mierzalnej. W tym celu na bazie wykresu plot-box oraz obliczonych miar położenia (mediana, kwartyl1, kwartyl3) odrzucono wszystkie obserwacje podejrzane o odstające (za kryterium klasyfikacji zmiennej jako odstającej przyjęto +/- 1,5 IQR).

Następnie zdefiniowano cel badania, którym jest zbudowanie modelu logistycznego, sprawdzającego prawdopodobieństwo sprzedaży mieszkania w czasie krótszym niż średnia czasu oczekiwania na sprzedaż ze wszystkich ofert wystawionych do sprzedaży.

Dla rekordów wykorzystanych w niniejszym badaniu średni czas oczekiwania nieruchomości na sprzedaż wynosi 88,64 dni. Wejściowy model regresji logistycznej został zapisany w postaci formalnej we wzorze nr 4.

$$(Y = 1/ x_1, x_2, ..., x_{10}) = \frac{e^{a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + a_{10}x_{10}}}{1 + e^{a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + a_{10}x_{10}}} \tag{4}$$

gdzie:

x_1 - logarytm naturalny ceny mieszkania;

x_2 - powierzchnia mieszkania;

x_3 - liczba pokoi w mieszkaniu;

x_4 - piętro, na którym znajduje się mieszkanie;

x_5 - kuchnia w mieszkaniu;

x_6 - ocena położenia mieszkania;

x_7 - standard wykończenia mieszkania;

x_8 - liczba kondygnacji w budynku;

x_9 - technologia budowy;

x_{10} - lokalizacja mieszkania w budynku;

$a_0, a_1, a_2, ..., a_{10}$ – parametry strukturalne modelu.

Po zbudowaniu formalnego modelu regresji logistycznej, przeprowadzono estymację jego parametrów strukturalnych. W tabeli 1 przedstawiono wyniki estymacji z zastosowaniem metody Quasi-Newtona.

Tab. 1. Wstępne wyniki estymacji modelu regresji logistycznej

	Stała	X1	X2.	X3	X4	X5	X6	X7	X8	X9	X10
Ocena	4,21	-0,09	-0,09	1,65	-0,37	0,48	-1,03	-0,65	0,10	-0,08	-0,15
Błąd standardowy	14,08	1,24	0,03	0,77	0,13	0,73	0,57	0,56	0,13	0,69	0,56
t(110)	0,30	-0,07	-2,77	2,15	-2,78	0,65	-1,82	-1,16	0,81	-0,11	-0,27
Poziom p	0,77	0,94	0,01	0,03	0,01	0,52	0,07	0,25	0,42	0,91	0,79

Analizując przedstawione wyniki zmiennymi istotnie wpływającymi na zmienną Y są x_2, x_3, x_4, x_6. Po kolejnym odrzuceniu zmiennych nieistotnych statystycznie, ostateczna postać modelu regresji logistycznej przyjęła postać wyrażoną wzorem 5.

$$P(Y) = \frac{e^{3,67-0,10x_2+1,73x_3-0,30x_4-1,18x_6}}{1+e^{3,67-0,10x_2+1,73x_3-0,30x_4-1,18x}} \tag{5}$$

gdzie:

x_2 - powierzchnia mieszkania;

x_3 - liczba pokoi w mieszkaniu;

x_4 - piętro, na którym znajduje się mieszkanie;

x_6 - ocena położenia mieszkania.

Natomiast podsumowanie procesu estymacji przedstawiono w tabeli 2.

Tab. 2. Wyniki procesu estymacji dla zmiennych istotnych

	Stała	X2	X3	X4	X6
Ocena	3,67	-0,10	1,73	-0,30	-1,18
Błąd standardowy	0,95	0,03	0,62	0,10	0,55
t(110)	3,88	-3,48	2,80	-2,92	-2,16
Poziom p	0,00	0,00	0,01	0,00	0,03
-95%CL	1,79	-0,16	0,50	-0,50	-2,26
+95%CL	5,55	-0,04	2,96	-0,10	-0,10
Chi-kwadrat Walda	15,02	12,13	7,82	8,51	4,68
Poziom p	0,00	0,00	0,01	0,00	0,03
Iloraz szans z jend.	39,35	0,90	5,65	0,74	0,31
-95%CL	6,02	0,85	1,66	0,61	0,10
+95%CL	257,36	0,96	19,27	0,91	0,91
Iloraz szans zakr.		0,00	5753,88	0,04	0,31
-95%CL		0,00	12,46	0,00	0,10

+95%CL		0,00	2656223,00	0,35	0,91

Wartość statystyki p dla całego modelu przyjęła wartość 0,0000502, co świadczy o istotności modelu w porównaniu do modelu tylko z wyrazem wolnym, co potwierdza cel badania, że zbudowany model wnosi coś nowego. Ponadto, należy poddać interpretacji, tzw. logarytm wiarygodności, który jest miarą dopasowania całego modelu. Logarytm ten obliczany jest za pomocą statystyki $-2\log$ z maksimum wiarygodności zbudowanego modelu i modelu tylko zawierającym wyraz wolny.

W zbudowanym modelu wartości te wynoszą odpowiednio 122,3 oraz 147,3. Duża różnica pomiędzy tymi statystykami, ma rozkład zbliżony do chi-kwadrat. Statystyka ta to pierwszy krok weryfikacji istotności modelu. Na bazie powyższych wartości obliczono pseudo R^2 McFaddena i wyniósł on 0,17.

Dokonując interpretacji otrzymanych wyników możemy wnioskować, że:

- każdy dodatkowy pokój w mieszkaniu zwiększa 5,65 razy prawdopodobieństwo sprzedaży nieruchomości;
- mamy 35% szans na sprzedaż mieszkania, jeśli zwiększymy jego powierzchnię o 10 m^2, natomiast 12% szans na sprzedaż, gdy powierzchnię zwiększymy o 20 m^2 ,w stosunku do powierzchni bazowej;
- w tabeli 3 widzimy, że mieszkanie na 3 piętrze w stosunku do mieszkania na 1 piętrze ma o połowę mniejszą szansę na sprzedaż, natomiast jeszcze mniejsze szanse na sprzedaż (0,30) ma mieszkanie znajdujące się na 5 piętrze (w porównaniu do mieszkania znajdującego się na 1 piętrze).

Tab. 3. Iloraz szans na sprzedaż mieszkania w zależności od liczby piętra

Piętro	1	2	3	4	5	6	7	8	9	10
Wielkość zmniejszenia		0,74	0,55	0,41	0,30	0,22	0,16	0,12	0,09	0,07

W tabeli 4 przedstawiono poprawnie i niepoprawnie zakwalifikowane przypadki dla wyliczonego modelu. Obliczono również iloraz szans jako stosunek iloczynu poprawnie zaklasyfikowanych przypadków do iloczynu niepoprawnie zakwalifikowanych i wynosi on 4,76. Wartość większa od jedności oznacza, że ta klasyfikacja jest lepsza od tej, którą zostałaby przeprowadzona przez przypadek.

Tab. 4. Tablica trafności

	Przewidywane 0	Przewidywane 1.	Procent poprawności
0,000000	14,00	25,00	35,90
1,000000	8,00	68,00	89,47

4. Podsumowanie

Przeprowadzony proces badawczy potwierdza przydatność zbudowanego modelu logistycznego, w procesie podejmowania decyzji na rynku mieszkaniowym. Zbudowany model może być stosowany, jako narzędzie wspomagające podjęcie trafnej decyzji inwestycyjnych przez deweloperów jak i indywidualnych sprzedających. Przyjmując za kryterium czas sprzedaży mieszkania (liczbę dni potrzebną na sprzedaż mieszkania),

możemy przy określonych parametrach wartościujących mieszkanie wskazać prawdopodobieństwo powodzenia transakcji oraz możemy przeprowadzić wielowymiarową interpretację parametrów. W przedstawionym artykule przedstawiono szczegółową analizę modelu regresji logistycznej, którego celem jest określenie prawdopodobieństwa sprzedaży nieruchomości w czasie krótszym od średniego czasu oczekiwania na sprzedaż.

Autor niniejszej publikacji podjął również próbę budowy modeli regresji logistycznej, w których celem było określenie prawdopodobieństwa sprzedaży nieruchomości w czasie krótszym niż jeden miesiąc oraz w czasie krótszym niż jeden tydzień. Niestety na bazie posiadanych danych w obydwu modelach wszystkie zmienne niezależne okazały się nieistotne statystycznie.

W dalszych etapach badawczych, próbując poprawić jakość zbudowanego modelu regresji logistycznej, będzie podjęta próba poprawy jakości danych wejściowych użytych do budowy modelu (np. poprzez zwiększenie rekordów wykorzystywanych do analizy).

Literatura

1. Churrchill, G.A. Badania marketingowe, Podstawy metodologiczne, PWN, Warszawa, 2002.
2. Dittmann P., Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie, Oficyna Ekonomiczna, Kraków 2004.
3. Gruszczyński M., Kuszewski T., Podgórska M. (red.), Ekonometria i badania operacyjne, PWN, Warszawa, 2009.
4. Mach Ł., Econometric model structure as a support tool in real property market parameters featuring, The 19th International DAAAM SYMPOSIUM "Intelligent Manufacturing & Automation: Focus on Next Generation of Intelligent Systems and Solutions", 22-25th October 2008, Vienna.
5. Maddala G.S., Ekonometria, Wydawnictwo PWN, Warszawa, 2008.
6. Stanisz A., Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny. T2. Modele liniowe i nieliniowe, Statsoft, Kraków, 2007.

Dr inż. Łukasz MACH
Politechnika Opolska
Wydział Zarządzania i Inżynierii Produkcji
45-370 Opole, ul. Ozimska 75
telefon: 774234031
e-mail: l.mach@po.opole.pl

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Artykuł współfinansowany przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY