Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Kamil Matuszelański
Nr albumu: 387078

DETERMINANTY OCEN RESTAURACJI W SERWISIE ZOMATO

Praca zaliczeniowa
na przedmiot Ekonometria

Praca wykonana pod kierunkiem
Rafała Walaska

Warszawa, styczeń 2019

SPIS TREŚCI

ROZDZIAŁ I. Wstęp 3
ROZDZIAŁ II. Wstępna analiza danych 4
ROZDZIAŁ III. Estymacja modelu 7
Pierwszy model 7
Model na znaczących zmiennych 7
Ostateczny model 8
Testy diagnostyczne 9
Interpretacja parametrów 10
BIBLIOGRAFIA 11

Wstęp

W pracy tej postaram się określić determinanty ocen wystawianych restauracjom na portalu zomato.pl. Portal ten powstał w 2008 roku w Indiach, obecnie działa w 24 krajach i miesięcznie korzysta z niego 190 mln unikalnych użytkowników. W Polsce serwis pojawił w 2014 roku, po wykupieniu polskiej strony gastronauci.pl. Najważniejszą funkcjonalnością strony jest możliwość wyszukiwania restauracji z danego obszaru, a także udostępnianie oraz przeglądanie opinii innych użytkowników. Każdy użytkownik publikujący słowną recenzję wystawia również ocenę restauracji na 5-stopniowej skali. Jest to istotny czynnik determinujący zainteresowanie użytkowników portalu, dlatego też informacja o tym co wpływa na ocenę restauracji może być bardzo cenna dla restauratorów.

Temat predykcji oceny biznesów nie jest wyczerpująco opisany w literaturze. Większość badań opiera się na metodach analizy sentymentu wykonywanych na cząstkowych recenzjach internautów w różnych serwisach społecznościowych. Poza tym wiele badań jest ukierunkowanych na utworzenie silnika rekomendacyjnego. W pracy "Applications of Machine Learning to Predict Yelp Ratings" autorstwa Carbona, Fujii i Veerina analizowany był zbiór danych udostępniony przez firmę Yelp. Znalazły się w nim dane dotyczące 42 tysięcy biznesów (głównie punktów gastronomicznych). W pracy analizowano możliwość przewidzenia średniej oceny restauracji na podstawie zmiennych dostępnych w zbiorze. Znalazły się w nich między innymi lokalizacja, ilość recenzji, zakres cenowy, a także kilka zmiennych zerojedynkowych (na przykład czy dana restauracja umożliwia dostawę lub czy została otagowana przez użytkowników jako dobra na śniadanie). Poza oryginalnymi zmiennymi zostały też dodane zmienne wyznaczające sentyment recenzji tekstowych napisanych przez użytkowników, a także zmienna powstała przez klastrowanie przestrzenne metodą k-średnich [1] (stworzono 15 klastrów). Do predykcji oceny wykorzystano zarówno metody klasyfikacji jak i regresji (ocena była z przedziału 1-5 ze skokami co 0,5 , do algorytmów klasyfikacji utworzono sztuczną zmienną przyjmującą wartość 1 jeżeli ocena była wyższa od 3). Najlepsze rezultaty uzyskano za pomocą algorytmów maszyny wektorów wspierających[2] i regresji logistycznej, a najistotniejszymi zmiennymi okazały się sentyment recenzji, współrzędne geograficzne, ilość recenzji a także rozmiar klastra. W pracy Hu, L., Sun, A., i Liu, Y. pt. "Your neighbors affect your ratings" testowano hipotezę o tym że istnieje zależność pomiędzy ocenami sąsiednich restauracji. Okazało się że istnieje słaba pozytywna korelacja (ok 0.1) pomiędzy oceną danej restauracji a średnią oceną od 1 do 10 sąsiadów. Dalsza część badania używała tej obserwacji do stworzenia silnika rekomendacyjnego na podstawie analizy tekstów recenzji.

[1] Metoda ta polega na przypisaniu obserwacji które znajdują się blisko siebie do jednej grupy.

[2] Jest to metoda klasyfikacji która w najprostszym przypadku polega na znalezieniu prostej która najlepiej oddziela od siebie grupy obserwacji.

Wstępna analiza danych

Analizę oraz estymację modelu wykonałem w środowisku R. Do pobrania danych ze strony zomato.pl posłużył mi pakiet rvest, a dalsze analizy wykonywałem używając pakietów lmtest a także tidyverse.

Model został opracowany na podstawie danych o restauracjach w Warszawie dostępnych w serwisie Zomato. Pierwotnie do bazy należało 2273 obserwacji.

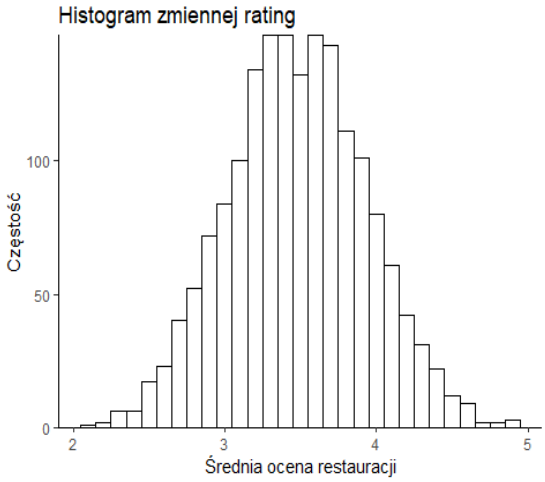
Do zmiennych w bazie należą:

- **Średnia ocena** (zmienna zależna)
- Średnia cena dla dwóch osób
- Liczba ocen wystawionych przez użytkowników portalu Zomato dla danej restauracji
- Serwowane rodzaje kuchni
- Dzielnica Warszawy w której znajduje się dana restauracja
- Koordynaty geograficzne

A także nazwa restauracji oraz jej numer identyfikacyjny, które nie będą użyte w estymacji modelu.

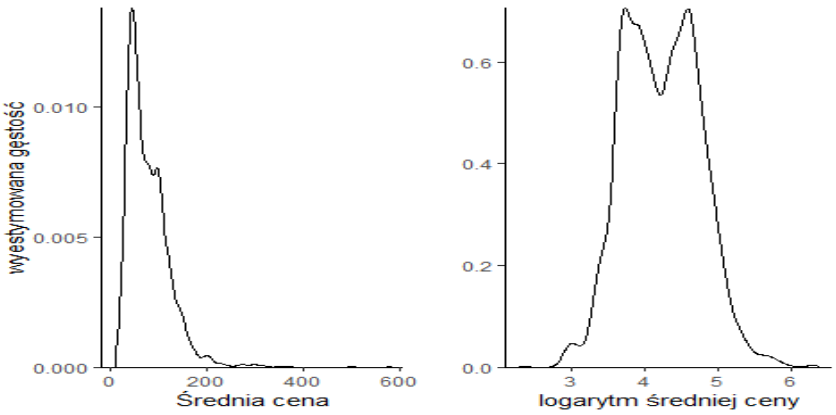
Recenzje punktowe są wystawiane jeżeli dana restauracja otrzymała chociaż 1 głos. Ponadto w 8 obserwacjach w zmiennej cena występują braki danych. Z tego powodu usunąłem z bazy część obserwacji, więc ostatecznie pozostały 1722 restauracje do modelu.

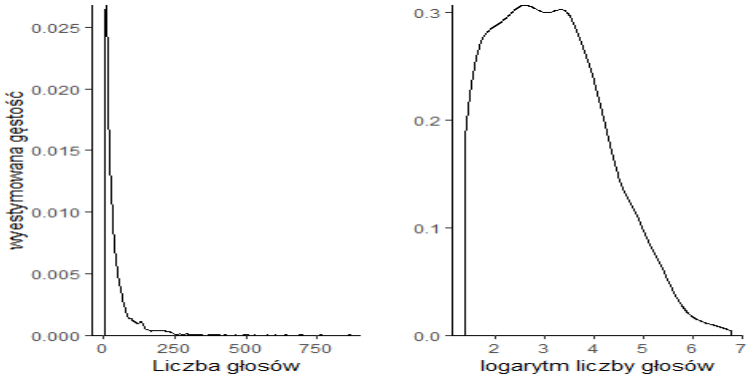
Poniżej opisałem najważniejsze cechy zmiennych z perspektywy modelowania:



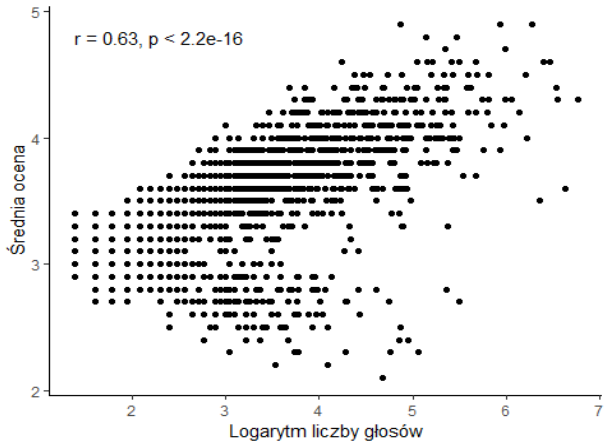
- Średnia ocena czyli zmienna niezależna ma rozkład bliski rozkładowi normalnemu. Z perspektywy modelu istotne są 2 kwestie. Po pierwsze, 544 restauracje nie mają żadnej wystawionej oceny. Dlatego zdecydowałem się je usunąć, co zmniejszyło ilość obserwacji do 1729. Po drugie, ocena może przyjmować wartości od 2 do 5, ze skokiem co 0.1, dlatego też należy zaznaczyć że oszacowania powyżej 5 i poniżej 2 wynikające z modelu są błędne. Być może właściwsze byłoby użycie algorytmów klasyfikacji (np. wieloklasowej regresji logistycznej).

- Zarówno średnia cena jak i liczba głosów mają bardzo skośne rozkłady. Na odpowiednich wykresach przedstawiłem ich gęstości wyestymowane metodą KDE, a także ich logarytmy. Można przypuszczać że użycie tych zmiennych ale zlogarytmowanych może dać lepsze rezultaty.

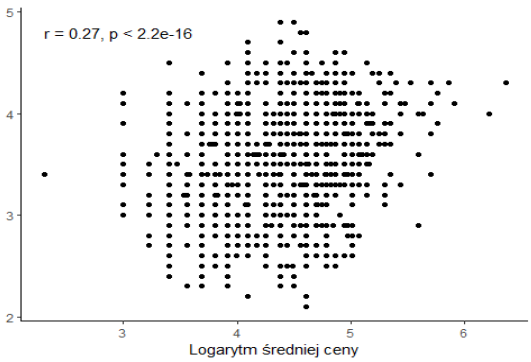




- Występuje silna pozytywna korelacja pomiędzy średnią oceną a logarytmem liczby głosów, co jest zgodne z przewidywaniami (Jeżeli restauracja jest dobra to może liczyć na zwiększenie zainteresowania nią, choćby ze względu na marketing szeptany). Ta zmienna będzie prawdopodobnie bardzo istotna w modelu. To co może powodować problemy przy modelowaniu to fakt że zależność nie przypomina liniowej, a bardziej funkcję y^2=x. Można to interpretować w taki sposób że chęć wyrażenia swojej opinii o restauracji na portalu jest największa z jednej strony kiedy jesteśmy bardzo zawiedzeni obsługą, a z drugiej kiedy restauracja była ponadprzeciętnie dobra.



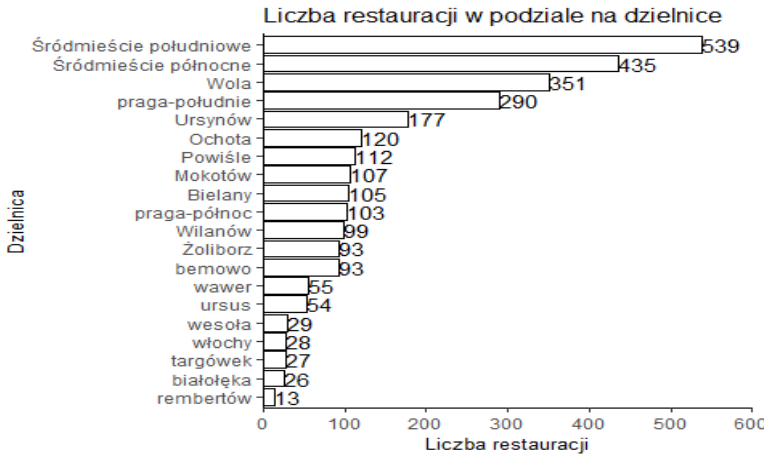
- Występuje także niewielka pozytywna korelacja pomiędzy zmienną niezależną a logarytmem ceny. Może to oznaczać że właściciele dobrych restauracji mogą narzucać większe marże, a i tak zostanie to zrekompensowane przez jakość usługi.



- Poniższa tabela przedstawia najczęstsze rodzaje kuchni. Uwagę zwraca fakt że już 10. pod względem popularności kategoria zawiera zaledwie 3% restauracji. Aby nie wprowadzać zmiennych które mogą wpływać na nadmierne dopasowanie modelu (w przypadku gdy dana kategoria zawiera zbyt mało obserwacji wnioskowanie o ogólnej jakości restauracji z danej kategorii mija się z celem), stworzyłem nową zmienną kuchnia, która przyjmuje wartość ‘inne’ jeżeli wcześniejsza wartość występowała rzadziej niż w 2% restauracji. Taką wartość ostatecznie przyjmuje ok. 15% obserwacji

Kuchnia	Liczba restaruacji	Procent restauracji
Polska	263	15.21 %
Kawiarnia	163	9.43 %
Włoska	150	8.68 %
Pizza	134	7.75 %
Sushi	94	5.44 %
Europejska	82	4.74 %
Burgery	56	3.24 %
Przekąski	56	3.24 %
Wietnamska	53	3.07 %
Chińska	52	3.01 %

- Na poniższym wykresie można zauważyć że duża część restauracji znajduje się w ścisłym centrum Warszawy. Ponadto wyróżniającymi się dzielnicami są Wola, Praga-Południe oraz Ursynów.



Estymacja modelu

Pierwszy model

Najpierw do wyestymowania modelu użyłem wszystkich zmiennych. Od razu rzuciła się w oczy duża liczba nieznaczących zmiennych, które wynikają z dużej ilości poziomów zmiennych kuchnia i dzielnica. W rozkodowanej zmiennej dzielnica wszystkie zmienne zerojedynkowe poza tą oznaczającą Śródmieście Północne są nieznaczące z p-value>0.1.Natomiast znaczące kuchnie to Japońska, Europejska oraz Inne. Zmienne wyznaczające koordynaty geograficzne również nie okazały się znaczące. Co więcej współczynnik R^2 nie jest zadowalający.

Statystyka	Wartość
R^2	0.3425
Skorygowane R^2	0.3277
RSS	0.3736
p-value (z testu F na łączną istotność zmiennych)	0.0000

Model na znaczących zmiennych

Ten model wyestymowałem używając wyłącznie znaczących zmiennych ze wstępnego modelu. Co więcej przekształciłem zmienne cena i ilość głosów za pomocą logartmu, licząc że poprawi to jakość dopasowania.

Parametry tego modelu opisane są w poniższej tabeli. Zwraca uwagę fakt polepszenia wskaźnika R^2 do 40%, co jest bliskie rezultatowi otrzymanemu przez Hu, L., Sun, A., i Liu, Y. przy użyciu regresji liniowej. W modelu tym pomimo dużej wytłumaczalności pojawiają się poważne problemy co do założeń. Najważniejszym z nich jest to że test RESET z p-value<0.0001 odrzuca poprawność formy funkcyjnej. Oznacza to że wnioskowanie na podstawie tego modelu będzie błędne, co więcej przeprowadzanie dalszych testów diagnostycznych mija się z celem. Aby rozwiązać ten problem do kolejnego modelu włączyłem interakcje pomiędzy zmiennymi.

zmienna	Wyestymowana wartość	Błąd std.	Statystyka t	p.value
Wyraz wolny	2.5542	0.0737	34.6348	0.0000
liczba_glosow_log	0.2423	0.0080	30.1717	0.0000
cena_log	0.0322	0.0183	1.7612	0.0784
kuchnia_japonska	0.2591	0.0718	3.6104	0.0003
kuchnia_europejska	0.1598	0.0408	3.9141	0.0001
kuchnia_desery	0.1985	0.0572	3.4704	0.0005
kuchnia_inne	0.1024	0.0203	5.0347	0.0000
kuchnia_kebab	-0.1620	0.0501	-3.2358	0.0012
kuchnia_miedzynarodowa	0.2485	0.0500	4.9757	0.0000

Statystyka	Wartość
R^2	0.4269000
Skorygowane R^2	0.4243000
RSS	0.3458000
p-value (z testu F na łączną istotność zmiennych)	0.0000000
p-value testu RESET	0.0000006

Ostateczny model

Ten model wyestymowałem używając interakcji między zlogarytmowaną ceną, zlogarytmowaą liczbą głosów, a także ze zmiennymi zerojedynkowymi powstałymi z rozkodowanej zmiennej kuchnia (ostatecznie wpływ na poprawienie dopasowania miały kuchnie: międzynarodowa, europejska, kebab oraz japońska). W tym modelu wszystkie zmienne są istotne z p-value<0.01, ponadto test F z p-value<0.0001 wskazuje na łączną istotność wszystkich zmiennych. Wyjaśnione jest 40% wariancji (mierzonej współczynnikiem R^2).

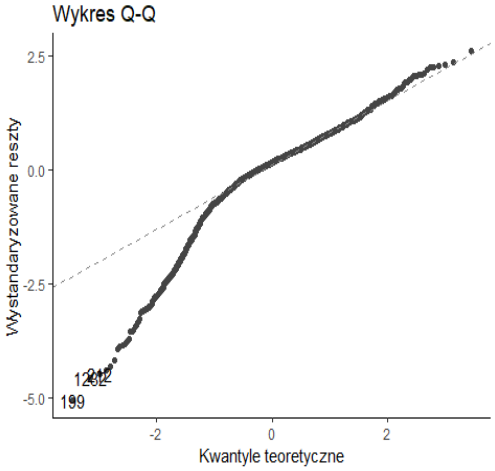
zmienna	Wyestymowana wartość	Błąd std.	Statystyka t	p-value
Wyraz wolny	2.8221	0.0223	126.5461	0.0000
liczba_glosow_log*cena_log	0.0491	0.0016	31.1856	0.0000
liczba_glosow_log*cena_log*kuchnia_miedzynarodowa	0.0122	0.0034	3.5334	0.0004
liczba_glosow_log*cena_log*kuchnia_europejska	0.0055	0.0023	2.3456	0.0191
liczba_glosow_log*cena_log*kuchnia_kebab	-0.0173	0.0050	-3.4973	0.0005
liczba_glosow_log*cena_log*kuchnia_japonska	0.0145	0.0045	3.2182	0.0013

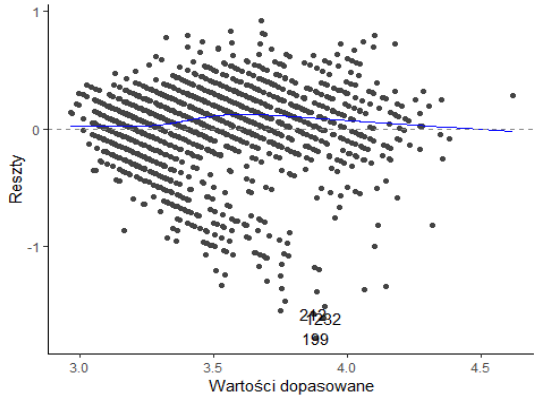
Statystyka	Wartość
R^2	0.4014000
Skorygowane R^2	0.3997000
RSS	0.3531000
p-value (z testu F na łączną istotność zmiennych)	0.0000000
p-value testu RESET	0.3179421

Testy diagnostyczne

Na modelu wykonałem również podstawowe testy diagnostyczne.

- **Poprawność formy funkcyjnej**
  Test RESET sprawdzający poprawność formy funkcyjnej z p-value= 0.32 wskazuje na poprawność formy funkcyjnej.
- **Normalność reszt**
  Wykonałem testy Shapiro-Wilka oraz Jarque-Bera, aby sprawdzić założenie o normalności składników losowych. Oba odrzucają hipotezę o normalności. Z wykresu Q-Q można zauważyć że występuje duża skośność lewostronna (współczynnik skośności na poziomie -1.28). Oznacza to, że wartości dopasowane są w wielu przypadkach przeszacowane w stosunku do prawdziwej wartości. Jednocześnie nie wynika to z dopasowania wartości większych niż 5, co byłoby błędnym oszacowaniem (ponieważ rating może przyjmować wartości 2-5). Ponieważ jednak ilość obserwacji jest duża, brak normalności składnika losowego nie jest tak dużym problemem.





- **Heteroskedastyczność**
  Test Breuscha-Pagana z p-value<0.0001 wskazuje na występowanie heteroskedastyczności. To samo można odczytać z wykresu reszt i wartości dopasowanych. Co więcej, można podejrzewać że heteroskedastyczność występuje z powodu skomplikowanej zależności liczby głosów od oceny (Po wyrzuceniu zmiennej liczna głosów z modelu wariancja składnika losowego jest homoskedastyczna, ale wskaźnik R^2 spada z ok 40% do ok. 10%, dlatego zdecydowałem się na model z heteroskedastycznością ale lepszym dopasowaniem).
- **Współliniowość**
  Obliczyłem wartości współczynnika VIF dla wszystkich zmiennych z modelu. Wszystkie z nich mają wartość w przybliżeniu 1, co oznacza że w modelu nie występuje współliniowość.

Zmienna	VIF
liczba_glosow_log * cena_log	1.0599
liczba_glosow_log * cena_log * kuchnia_miedzynarodowa	1.0076
liczba_glosow_log * cena_log * kuchnia_europejska	1.0453
liczba_glosow_log * cena_log * kuchnia_kebab	1.0074
liczba_glosow_log * cena_log * kuchnia_japonska	1.0096

Interpretacja parametrów

- Jeżeli restauracja serwuje kuchnie inne niż poniżej wymienione, zmiana ceny o 1 spowoduje zmianę oceny o +0.0000049. Taka sama zmiana pojawi się jeżeli liczba głosów zmieni się o 1.
- Jeżeli serwuje daną kuchnię to rating przy zmianie ceny o 1 zmieni się o dodatkowe:
  - +0.0000012 jeżeli kuchnia międzynarodowa
  - +0.0000005 jeżeli kuchnia europejska
  - -0.0000017 jeżeli kuchnia kebab
  - +0.0000015 jeżeli kuchnia japońska
- Jeżeli serwuje daną kuchnię to rating przy zmianie liczby głosów o 1 zmieni się o dodatkowe:
  - +0.0000012 jeżeli kuchnia międzynarodowa
  - +0.0000006 jeżeli kuchnia europejska
  - -0.0000017 jeżeli kuchnia kebab
  - +0.0000015 jeżeli kuchnia japońska

BIBLIOGRAFIA

Hu, L., Sun, A., & Liu, Y. (2014, July). Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 345-354). ACM.

Carbon, K., Fujii, K., & Veerina, P. (2014). Applications of Machine Learning to Predict Yelp Ratings.

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org**.**

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.5. https://CRAN.R-project.org/package=dplyr

Hadley Wickham (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest