Sprawozdanie z Projektu - Ekonometria

Jakub Anczyk

2023-05-06

```
knitr::opts_chunk$set(include = FALSE)

PackageNames <- c("regclass", "tidyverse", "stargazer", "magrittr", "moments", "stringr", "dplyr",  "ca

for (i in PackageNames){
  if(!require(i, character.only = T)){
    install.packages(i, dependencies = T)
    require(i, character.only = T)
  }
}
```

```
## Ładowanie wymaganego pakietu: regclass

## Ładowanie wymaganego pakietu: bestglm

## Ładowanie wymaganego pakietu: leaps

## Ładowanie wymaganego pakietu: VGAM

## Ładowanie wymaganego pakietu: stats4

## Ładowanie wymaganego pakietu: splines

## Ładowanie wymaganego pakietu: rpart

## Ładowanie wymaganego pakietu: randomForest

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.

## Ładowanie wymaganego pakietu: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine()  masks randomForest::combine()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x ggplot2::margin() masks randomForest::margin()
## Ładowanie wymaganego pakietu: stargazer
##
##
## Please cite as:
##
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
##
##
## Ładowanie wymaganego pakietu: magrittr
##
##
## Dołączanie pakietu: 'magrittr'
##
##
## Następujący obiekt został zakryty z 'package:purrr':
##
##     set_names
##
##
## Następujący obiekt został zakryty z 'package:tidyr':
##
##     extract
##
##
## Ładowanie wymaganego pakietu: moments
##
## Ładowanie wymaganego pakietu: caret
##
## Ładowanie wymaganego pakietu: lattice
##
##
## Dołączanie pakietu: 'lattice'
##
##
## Następujący obiekt został zakryty z 'package:regclass':
##
##     qq
##
##
##
## Dołączanie pakietu: 'caret'
```

```
##
##
## Następujący obiekt został zakryty z 'package:purrr':
##
##     lift
##
##
## Następujący obiekt został zakryty z 'package:VGAM':
##
##     predictors
##
##
## Ładowanie wymaganego pakietu: lmtest
##
## Ładowanie wymaganego pakietu: zoo
##
##
## Dołączanie pakietu: 'zoo'
##
##
## Następujące obiekty zostały zakryte z 'package:base':
##
##     as.Date, as.Date.numeric
##
##
##
## Dołączanie pakietu: 'lmtest'
##
##
## Następujący obiekt został zakryty z 'package:VGAM':
##
##     lrtest
##
##
## Ładowanie wymaganego pakietu: corrplot
##
## corrplot 0.92 loaded
```

Przewidywanie cen nieruchomości w Miami

Zestaw danych wykorzystany do projektu pochodzi z Kaggle i zawiera podstawowe parametry 13 932 domów jednorodzinnych sprzedanych w Miami w 2016 roku: https://www.kaggle.com/datasets/deepcontractor/miami-housing-dataset.

Celem projektu jest zbudowanie modelu pozwalającego przewidzieć cenę rynkową nieruchomości na podstawie jej parametrów.

Budowa modelu

W pierwszym kroku wybieram wszystkie dane z podanego zestawu danych i tworzę zestaw danych wyłącznie numerycznych. Kod nie zawiera kolejnych kroków czyszczenia i dostosowywania danych, ponieważ dane są już gotowe do analizy.

Poniżej znajduje się krótkie podsumowanie dostępnych zmiennych:

Następnie zaczynam budowę modelu ekonometrycznego od zbudowania macierzy korelacji, macierzy R0 oraz R:

W kolejnym kroku buduję macierz zawierającą zestawienie wszystkich możliwych kombinacji zestawień zmiennych.

Następnie obliczam całkowitą pojemność informacji każdej uzyskanej w ten sposób kombinacji.

Finalnie sortuję uzyskane w ten sposób dane i wybieram kombinację o największej możliwej pojemności informacyjnej.

Teraz mozemy podsumować uzyskany w ten sposób model.

Jak widać, wszystkie wybrane zmienne są bardzo istotne dla modelu, osiągając p-value poniżej 2e-16, tj. mniejszą niż 0.0000000000000002. Zmienne te to kolejno:

TOT_LVG_AREA (total living area) - to wartość odpowiadająca całkowitej powierzchni użytkowej lokalu w stopach kwadratowych.

Znak przy tym współczynniku jest dodatni, co oznacza, że wzrost o jedną jednostkę (stopę kwadratową) oznacza wzrost ceny obiektu o $200.70 (ceteris paribus).

SPEC_FEAT_VAL (total value of special features) - podana w dolarach wartość wszystkich dodatkowych cech lokalu, takich jak baseny, jaccuzi, szklarnie przydomowe etc.,

Znak przy tym współczynniku także jest dodatni, co oznacza, że wzrost o jedną jednostkę, tj. wzrost wartości znajdujących się w obrębie budynku obiektów o jednego dolara oznacza wzrost ceny samego obiektu o $3.61 (ceteris paribus).

SUBCNTR_DI (subcenter distance) - odległość od najbliższego subcentrum, tj. centrum działalności handlowej lub usługowej. Wraz ze wzrostem odległości o jedną stopę, cena nieruchomości maleje o $2.70 (ceteris paribus).

STRUCTURE_QUALITY - jakość/klasa (strukturalna) budynku. Zgodnie z modelem przeskok o jedną klasę jakości w górę oznacza wzrost średniej ceny obiektu aż o $95340.00 (ceteris paribus).

Diagnostyka

Test normalności reszt Test normalności reszt pokazuje, że statystyka R-kwadrat jest niewystarczająca, aby pozytywnie ocenić model.

Aby poprawić model, próbuję najpierw przekształcić go za pomocą logarytmu.

Jest to rzeczywiście wystarczające aby uzyskać wartość R-kwadrat powyżej 70%, jednak podniesienie ceny jeszcze to czwartej potęgi pomoże uzyskać R-kwadrat bliższe 75%. Wciąż jednak nawet takie przekształcenie nie jest wystarczające by pozbyć się tzw. “grubych ogonów”.

Taki rozkład oznacza dużo wyższe niż w rozkładzie normalnym prawdopodobieństwo wystąpienia skrajnie wysokich (gruby prawy ogon) lub niskich (gruby lewy ogon) wartości.

Wykres wartości teoretycznych i reszt Jak widać na wykresie wartości teoretycznych i reszt, obserwacje koncentrują się wyraźnie wokół zera, choć wykres jest zdecydowanie niesymetryczny i ma rożną wariancję reszt dla różnych wartości teoretycznych.

Oznacza to, że błąd jest różny w zależności od wartości przewidzianej przez model.

Test normalności reszt - Test Shapiro-Wilka Wynik testu Shapiro-Wilka pokazuje, że są podstawy do odrzucenia hipotezy zerowej o normalności reszt, tj. wartości resztowe nie mają rozkładu normalnego.

Heterosdedastyczność - Test Breuscha-Pagana Bazując na bardzo małym p-value, istnieją podstawy do odrzucenia hipotezy zerowej o braku heteroskedastyczności. Oznacza to, że wariancja reszt w istocie jest różna a model niedopasowany.

Heterosdedastyczność - Test Durbina-Watsona Bazując na bardzo małym p-value dla tego testu, istnieją także podstawy do odrzucenia hipotezy zerowej o braku autokorelacji. Oznacza to, że zmienne są między sobą istotnie skorelowane, co także oznacza słabe dopasowanie modelu.

Podsumowanie

Ostatecznie model po przeprowadzeniu diagnostyki zostaje odrzucony z powodu braku spełnienia podstawowych założeń, między innymi o normalności reszt i nieheteroskedastyczności.