Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Rafał Kaczmarek
Nr albumu: 396945

Michał Szłapa
Nr albumu: 397384

Analiza czynników wyceny aut osobowych

Praca wykonana na zaliczenie przedmiotu
Zastosowania języka Python
prowadzonego przez
Michała Palińskiego oraz Kristófa Gyódiego

Warszawa, maj 2021

Spis treści

Wstęp 2

Dane 3

Metodologia 7

Model 7

Podsumowanie 10

Bibliografia 11

Wstęp

Samochody osobowe są wynalazkiem, który niewątpliwie wpłynął na ułatwienie życia codziennego. Nie są już one dobrem luksusowym. Rynek samochodów osobowych jest dobrze rozwinięty, co objawia się mnogością aut gotowych do zakupu, zarówno tych nowych, jak i używanych. W dobie rosnącej świadomości ludzi w kwestii ochrony środowiska oraz ochrony zdrowia, temat samochodów osobowych jest często poruszany jako czynnik negatywnie wpływający na wspomniane kwestie. Jednakże, użyteczność samochodów jest tak duża, że ludzie nie wyobrażają sobie świata bez tych maszyn, natomiast podejmowane są działania mające na celu zniwelowanie negatywnych efektów użytkowania samochodów takie jak budowa obwodnic oraz, coraz powszechniejsze, stosowanie silników elektrycznych, które są odpowiedzią na problem emisyjności szkodliwych gazów. Oczywiście ma to wpływ na koszty produkcji, a co za tym idzie, także na cenę aut. W literaturze podejmowany jest temat czynników wpływających na cenę samochodów. W artykule Stefana Lessmanna oraz Stefana Voßa[1] weryfikowano modele statystyczne wykorzystane w celu prognozowania cen sprzedaży samochodów używanych. Zostało sprawdzonych wiele modeli, zarówno liniowych jak i nieliniowych, i z pracy badaczy wynika, że losowa regresja lasów jest szczególnie skuteczna w prognozowaniu cen sprzedaży, a metoda najczęściej wykorzystywana w innych pracach - regresja liniowa - nie cechuje się wysoką dokładnością precyzji prognozowania. Dodatkowo, badacze podjęli kwestię przewagi informacyjnej i potwierdzili, że występuje ona wśród sprzedawców samochodów, przez co mogą oni dokładniej prognozować ceny niż agencje badania rynku. W artykule autorstwa Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic oraz Jasmin Kevric[2] została przedstawiona predykcja ceny przy wykorzystaniu technik uczenia maszynowego. W tym celu wykorzystali oni takie techniki jak sztuczna sieć neuronowa, maszyna wektorów nośnych i las losowy. Dane użyte do badania zostały zebrane przy zastosowaniu metod web scrapingu z portalu internetowego autopijaca.ba. Za pomocą wspomnianych danych oszacowano model, w którym do szacowania ceny samochodu z kategorii tanich oraz drogich wykorzystywana jest maszyna wektorów nośnych, a cena samochodów z średniej półki cenowej szacowana jest przy użyciu metody sieci neuronowych. Model w opisanej formie uzyskał dokładność 87,38% na danych testowych.

[1] S.Lessmann, S.Voß *Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy*, International Journal of Forecasting, 2017, Vol 33, Issue 4, s.864-877
[2] E.Gegic et al. *Car Price Prediction using Machine Learning Techniques*, TEM Journal, 2019, Vol 8, Issue 1, s.113-118

W wyborze auta istotnych jest wiele czynników, ale jednym z kluczowych jest cena. W związku z tym postanowiliśmy sprawdzić, jakie czynniki kształtują cenę samochodów osobowych na rynku polskim oraz na podstawie estymowanego modelu przeprowadzić predykcję ceny samochodów osobowych. Zainteresowani stwierdzeniem Stefana Lessmanna oraz Stefana Voßa dotyczącym wykorzystania regresji liniowej do prognozowania opisywanego zjawiska, zdecydowaliśmy się zweryfikować jak ta metoda sprawdzi się w prognozowaniu cen nie tylko używanych samochodów osobowych dostępnych na rynku polskim, ale także nowych. Dane użyte w tym celu zostały zebrane w sposób podobny do tego, który został opisany w wspomnianym artykule autorstwa Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic oraz Jasmin Kevric. Poza wykonaniem prognozy zweryfikowaliśmy dwie hipotezy, z których jedna zakłada że bezwypadkowość wpływa na wzrost ceny samochodu, a druga zakłada, że kolor samochodu wpływa na cenę samochodu.

Dane

Dane pozyskano ze strony internetowej otomoto.pl, przy pomocy pakietu *scrapy*. Zbiór danych zawiera 3324 obserwacji, które dotyczą wystawionych na sprzedaż samochodów osobowych. W jego skład wchodzą poniższe zmienne zmiennych:

- Price – cena wystawionego pojazdu osobowego w PLN
- Number – numer identyfikacyjny oferty
- Brand – producent pojazdu
- Year – rok produkcji wystawionego pojazdu
- Capacity – pojemność skokowa pojazdu w cm^3
- Aso – zmienna zerojedynkowa o tym czy pojazd podlega autoryzowanej stacji obsługi
- Color – kolor wystawionego pojazdu
- Condition – zmienna zerojedynkowa 1 – auto używane 0 – w p.p
- First_owner – zmienna zerojedynkowa 1 – pierwszy właściciel 0 – w p.p
- Fuel_type – rodzaj paliwa: 'Diesel', 'Benzyna', 'Benzyna+LPG', 'Benzyna+CNG', 'Hybryda', 'Elektryczny'
- Horse_power – liczba koni mechanicznych
- Mileage – przebieg w kilometrach
- No_accidents – zmienna zerojedynkowa 1 – brak szkody 0 - szkoda
- Number_of_doors – liczba drzwi w pojeździe: 2,3,4,5

- Transmission - rodzaj skrzyni biegów: 'Manualna', 'Automatyczna'
- Type – rodzaj auta : 'Kombi', 'Kompakt', 'Sedan', 'Auta miejskie', 'Coupe', 'SUV', 'Minivan', 'Auta małe', 'Kabriolet'

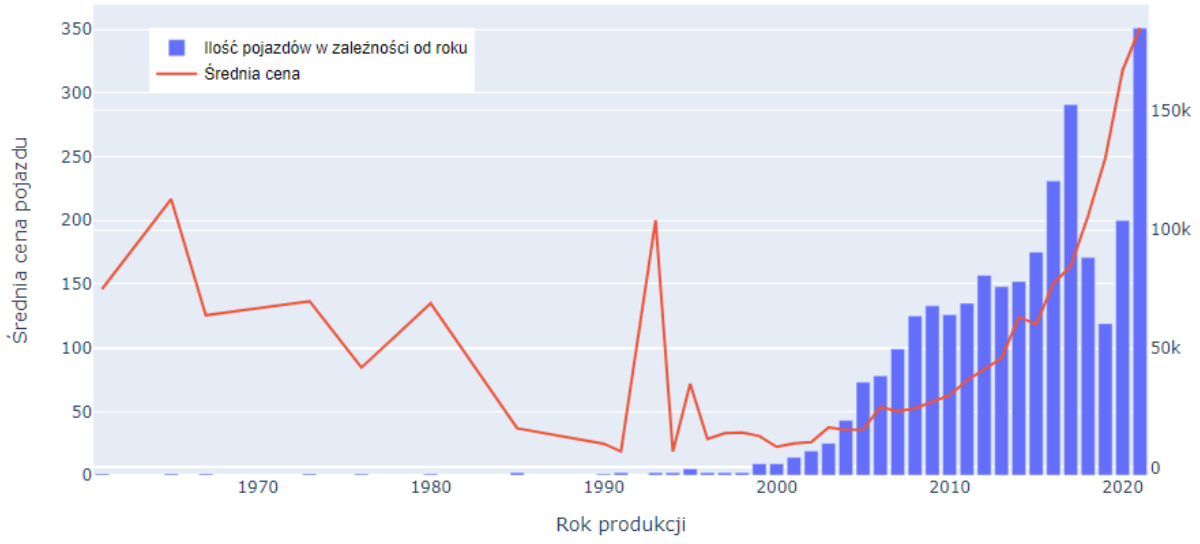
W celu wykorzystania tych danych w modelu stworzono zmienną Year_grouped, która opisywała rok produkcji pogrupowany na przedziały < 2011; 2011-2015; 2016-2018; 2019-2021. Dodatkowo zmieniono definicję zmiennej fuel_type, która opisuje 3 rodzaje paliwa - 'Diesel', 'Benzyna' (w tym 'Benzyna+LPG', 'Benzyna+CNG'), ‘Inne’ (w tym 'Hybryda', 'Elektryczny'). Ze zbioru usunięto obserwacje dla marek samochodów, których liczba była mniejsza niż 50. Pozostałe działania na danych opierały się na oczyszczeniu danych, w tym ujednoliceniu jednostek, usunięciu wartości pustych oraz niepoprawnych.

Tabela 1. Podstawowe statystyki opisowe dla zmiennych liczbowych

	price	capacity	Horse_power	mileage
count	3324.00	3313.00	3323.00	3321.00
mean	79174.57	19335.28	164.73	119847.55
std	82780.34	7263.32	75.93	95245.93
min	1500.00	5053.00	4.00	1.00
25%	26875.00	14993.00	115.00	37000.00
50%	50000.00	19683.00	150.00	114160.00
75%	97900.00	19973.00	190.00	183152.00
max	689900.00	66003.00	659.00	1111111.00

Źródło: Opracowanie własne na podstawie danych

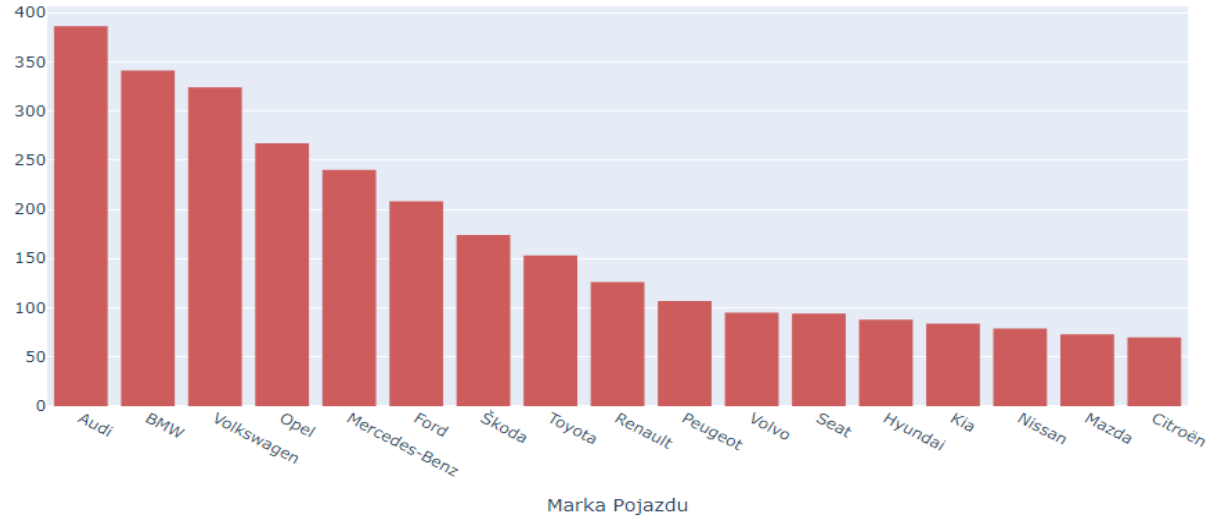
Wykres 1. Liczba pojazdów w zależności od roku produkcji wraz ze średnią ceną



Źródło: Opracowanie własne na podstawie danych

Najwięcej pojazdów spośród zbioru danych wyprodukowano w roku 2021. Dla aut z tego rocznika średnia cena jest najwyższa, co zdaje się być intuicyjne. Nowszy pojazd jest prawdopodobnie lepszy niż ten wyprodukowany kilka lat temu. Co ciekawe, średnia cena od roku 1970 do 2000 mocno waha się. Przyczyną prawdopodobnie jest to, że obserwacji z tego roku jest niewiele. Dodatkowo mogą to być auta kolekcjonerskie, których cena może być znacząco wyższa.

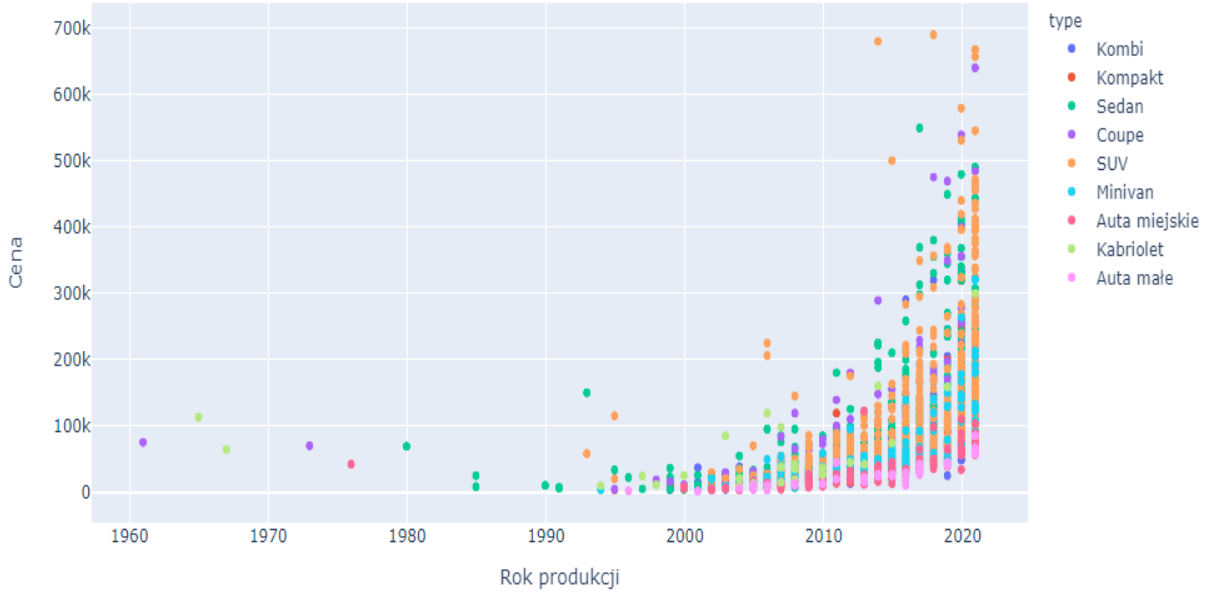
Wykres 2. Liczba wystawionych pojazdów w zależności od marki.



Źródło: Opracowanie własne na podstawie danych

Najwięcej pojazdów dostępnych na aukcji internetowej w otomoto to samochody marki Audi, BMW oraz Volkswagen. Analizując wykres zdaje się, że auta najpopularniejszych marek, wystawiane są na sprzedaż najczęściej.

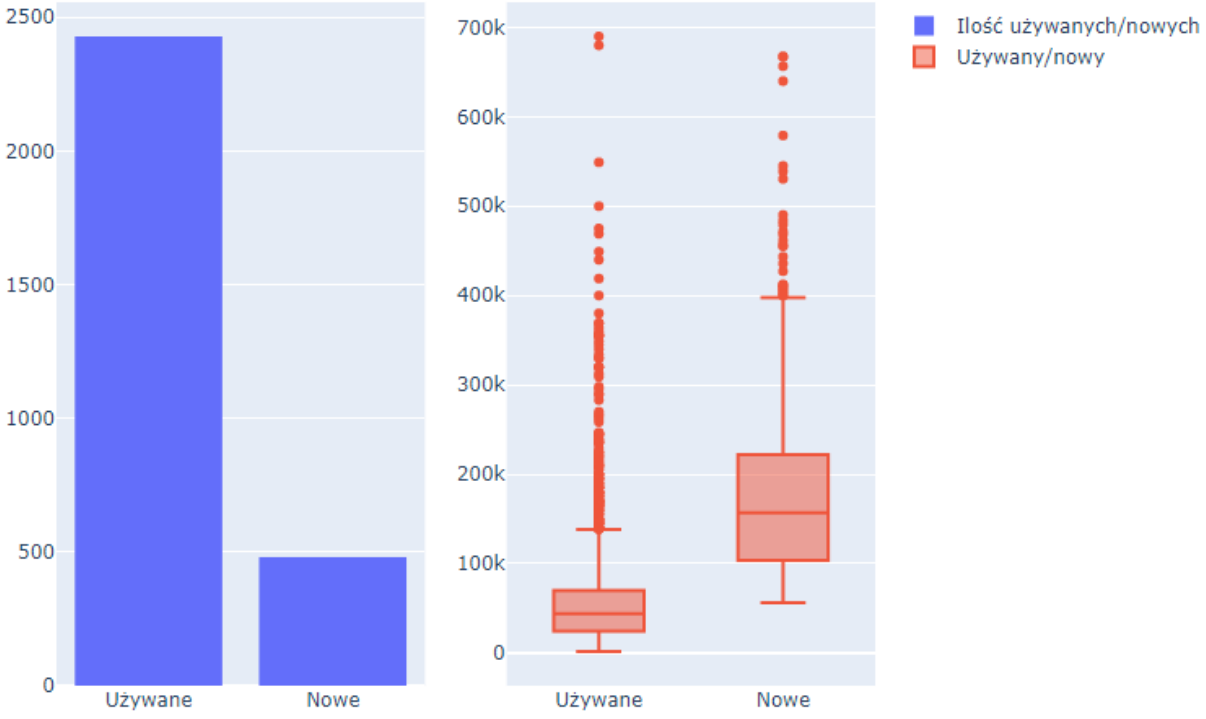
Wykres 3. Cena pojazdu w zależności od roku produkcji i typu samochodu



Źródło: Opracowanie własne na podstawie danych

Na powyższym wykresie przedstawiono scatter plot cen i roku produkcji. Z wykresu można wywnioskować, że najdroższym typem pojazdu jest SUV. Najczęściej tego typu pojazd jest najdroższy dla danego roku produkcji.

Wykres 4. Status auta w zależności od ceny



Źródło: Opracowanie własne na podstawie danych

Powyżej zaprezentowano prosty histogram oraz wykres pudełkowy zmiennej dotyczącej statusu pojazdu. Większość wystawionych pojazdu to samochody używane, co wydaje się intuicyjne. Cena używanych aut najczęściej znajduje się w przedziale 20-60 tys., zaś nowych w przedziale od 100 tys. do 200 tys. Co ciekawe najdroższy całym zbiorze danych jest samochód używany.

Metodologia

W celach sprawdzenia co wpływa na cenę aut oraz na jej predykcję posłużono się Metodą Najmniejszych Kwadratów (OLS). Ważna jest odpowiednia diagnostyka modelu, a przede wszystkim spełnienie założeń KMRL: wariancja składnika losowego jest stała, liniowa forma funkcyjna, zerowa wartość oczekiwana błędu losowego, nielosowość zmiennych, zerowa wariancja pomiędzy dwoma dowolnymi błędami losowymi. W tej metodzie istotne jest to że estymator nazywamy estymatorem nieobciążonym, jeżeli jego wartość oczekiwana jest równa wartości szacowanego parametru. Nieobciążoność estymatora daje najlepsze możliwe oszacowanie parametrów modelu.[3] Zbiór danych podzielono na zestaw treningowy i testowy. Utworzone zostanie kilka modeli, z których wybrany zostanie ten najbardziej optymalny. Wraz z kolejnymi etapami modelu będzie dokonywana weryfikacja zmiennych, tak aby otrzymać najbardziej optymalny model. Zmienne nieistotne bądź te zaburzające regresje m.in. przez duże wskaźniki współliniowości, będą usuwane z modelu.

Model

Celem tej pracy jest zbadanie wpływu poszczególnych czynników na cenę samochodów osobowych oferowanych na rynku polskim na próbie treningowej oraz wykonanie predykcji cen na próbie testowej. Aby tego dokonać zamieniono zmienne jakościowe na zmienne dyskretne oraz przeprowadzono standaryzację zmiennych ciągłych. Dodatkowo podzielono zbiór danych na próbę treningową oraz próbę testową w stosunku 7:3. Wykorzystano model regresji liniowej estymowany przy użyciu Metody Najmniejszych Kwadratów. Wybór tej metody był podyktowany charakterystyką danych, które zawierają wiele zmiennych jakościowych, podczas gdy zmienną zależną jest zmienna ciągła. Dodatkowo, chcieliśmy zweryfikować stwierdzenie Stefana Lessmanna oraz Stefana Voßa dotyczącym wykorzystania

[3] J. Mycielski, Ekonometria, Uniwersytet Warszawski. Wydział Nauk Ekonomicznych, Warszawa, 2010

regresji liniowej, które wskazywało na to, że regresja liniowa nie cechuje się wysoką dokładnością precyzji prognozowania w tym przypadku.

Pierwotny model powstał po wstępnej selekcji zmiennych z danych zebranych przy użyciu web scrapingu. Opisany jest poniższym równaniem:

$$Price = \beta_0 + \beta_1 \text{capacity} + \beta_2 \text{horse\_power} + \beta_3 \text{mileage} + \beta_4 \text{brand} + \beta_5 \text{color} + \beta_6 \text{condition} + \beta_7 \text{fuel\_type} + \beta_8 \text{transmission} + \beta_9 \text{type} + \beta_{10} \text{year\_grouped} + \beta_{11} \text{aso} + \beta_{12} \text{first\_owner} + \beta_{13} \text{no\_accidents} + \beta_{14} \text{number\_of\_doors} + \varepsilon_i$$

Następne etapy pracy nad modelem obejmowały działania mające na celu uzyskanie jak najlepszego modelu biorąc pod uwagę istotność zmiennych, statystyki VIF, AIC, BIC oraz wyniki testów weryfikujących poprawność modelu. W wyniku tych działań powstał model końcowy opisany poniższym równaniem:

$$Price = \beta_0 + \beta_1 \text{horse\_power} + \beta_2 \text{mileage} + \beta_3 \text{brand} + \beta_4 \text{fuel\_type} + \beta_5 \text{year\_grouped} + \beta_6 \text{no\_accidents} + \varepsilon_i$$

Dokładne wyniki regresji dla powyższego modelu zostały przedstawione na rysunku 1. P-value związane ze statystyką F jest równe 0 co oznacza, że zmienne objaśniające są łącznie istotne na przyjętym poziomie istotności 0,05. R^2 równe 0.782 wskazuje na to, że model wyjaśnia 78,2% zmienności zmiennej zależnej. Z przeprowadzonej analizy wynika że na cenę auta silnie wpływają takie czynniki jak moc silnika, przebieg, rok produkcji a także informacje na temat czy samochód jest zasilany olejem napędowym oraz czy jest bezwypadkowy. Spośród wymienionych czynników, jedynie wyższa wartość przebiegu samochodu wpływa negatywnie na cenę. Dodatkowo można zauważyć, że im młodszy rocznik auta, tym większy wpływ ma on na cenę. Informacja na temat producenta samochodu miała także znaczenie na cenę, ale w niektórych przypadkach zmienne były nieistotne statystycznie.

W pracy postawiono dwie hipotezy badawcze. Pierwsza z nich zakłada, że informacja o bezwypadkowości samochodu wpływa na wzrost ceny tego samochodu. Z modelu wynika, że nie ma podstaw do odrzucenia tej hipotezy, ponieważ zmienna *no_accidents=1* jest istotna oraz wpływa pozytywnie na zmienną zależną. Druga hipoteza zakłada, że kolor samochodu ma statystycznie istotny wpływ na cenę samochodu. W modelu zmienna opisująca kolor samochodu okazała się nieistotna, co oznacza, że należy odrzucić tę hipotezę.

Przy pomocy opisywanego modelu przeprowadzono predykcję ceny na próbie testowej. R^2 predykcji jest równy 80% co oznacza dobre dopasowanie predykcji ceny przy użyciu modelu estymowanego na próbie treningowej, do ceny przedstawionej w próbie testowej. Wykres 5 przedstawia porównanie ceny uzyskanej w predykcji (y_pred) do ceny w próbie testowej (y_test). Na podstawie wykresu można zauważyć, że ceny prognozowane mają tendencje do zawyżenia cen w porównaniu do stanu rzeczywistego.

Rys 1. Wyniki regresji końcowego modelu

OLS Regression Results

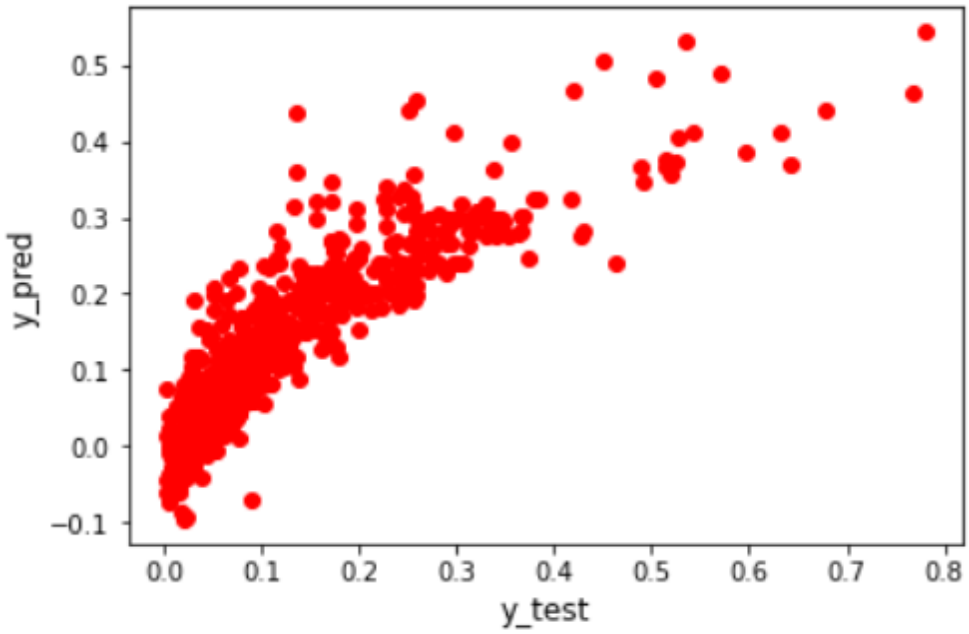
Dep. Variable:	price	R-squared:	0.782
Model:	OLS	Adj. R-squared:	0.779
Method:	Least Squares	F-statistic:	286.5
Date:	Sun, 16 May 2021	Prob (F-statistic):	0.00
Time:	23:22:29	Log-Likelihood:	2953.3
No. Observations:	2025	AIC:	-5855.
Df Residuals:	1999	BIC:	-5709.
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0130	0.008	-1.678	0.094	-0.028	0.002
horse_power	0.4958	0.012	40.173	0.000	0.472	0.520
mileage	-0.1855	0.013	-13.900	0.000	-0.212	-0.159
BMW	-0.0016	0.005	-0.303	0.762	-0.012	0.008
Citroën	-0.0128	0.009	-1.475	0.140	-0.030	0.004
Ford	-0.0301	0.006	-4.937	0.000	-0.042	-0.018
Hyundai	-0.0043	0.009	-0.483	0.629	-0.021	0.013
Kia	-0.0203	0.009	-2.378	0.017	-0.037	-0.004
Mazda	-0.0151	0.009	-1.680	0.093	-0.033	0.003
Mercedes-Benz	0.0262	0.006	4.672	0.000	0.015	0.037
Nissan	-0.0273	0.009	-3.076	0.002	-0.045	-0.010
Opel	-0.0195	0.006	-3.270	0.001	-0.031	-0.008
Peugeot	-0.0245	0.008	-3.177	0.002	-0.040	-0.009
Renault	-0.0233	0.007	-3.174	0.002	-0.038	-0.009
Seat	-0.0237	0.008	-2.985	0.003	-0.039	-0.008
Toyota	-0.0137	0.007	-1.917	0.055	-0.028	0.000
Volkswagen	-0.0085	0.005	-1.560	0.119	-0.019	0.002
Volvo	-0.0086	0.008	-1.066	0.287	-0.024	0.007
Škoda	-0.0329	0.007	-4.978	0.000	-0.046	-0.020
Benzyna+Gaz	0.0147	0.008	1.905	0.057	-0.000	0.030
Diesel	0.0355	0.003	12.092	0.000	0.030	0.041
Inny	0.0449	0.010	4.668	0.000	0.026	0.064
2011-2015	0.0096	0.004	2.435	0.015	0.002	0.017
2016-2018	0.0380	0.005	8.046	0.000	0.029	0.047
2019-2021	0.1127	0.006	17.793	0.000	0.100	0.125
no_accidents=1	0.0152	0.003	5.320	0.000	0.010	0.021

Omnibus:	1168.473	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27848.238
Skew:	2.243	Prob(JB):	0.00
Kurtosis:	20.605	Cond. No.	22.2

Źródło: Opracowanie własne na podstawie danych

Wykres 5. Porównanie oszacowań predykcji (y_pred) do oszacowań ze zbioru testowego.



Źródło: Opracowanie własne na podstawie danych

Podsumowanie

W niniejszej pracy przestawiono regresję liniową estymowaną przy użyciu Metody Najmniejszych Kwadratów, a następnie wyniki z tej regresji posłużyły do przeprowadzenia predykcji. W opisywanym modelu zmienną zależną była cena samochodu osobowego, a zmiennymi niezależnymi były różne cechy, które są często wymieniane podczas sprzedaży samochodu. Dane do tego badania pochodziły ze strony otomoto.pl i zostały zebrane przy użyciu metod web scrapingu. Wyniki przeprowadzonego badania wskazują, że na cenę auta silnie wpływają takie czynniki jak moc silnika, przebieg, rok produkcji a także informacje na temat czy samochód jest zasilany olejem napędowym oraz czy jest bezwypadkowy. Hipoteza zakładająca, że kolor samochodu ma statystycznie istotny wpływ na cenę samochodu została odrzucona, natomiast z modelu wynikło, że nie ma podstaw do odrzucenia hipotezy zakładającej, że informacja o bezwypadkowości samochodu wpływa na wzrost jego ceny. Przeprowadzona predykcja ceny na próbie testowej wskazała, że R^2 równe 80%, co oznacza dobre dopasowanie ceny uzyskanej za pomocą predykcji do ceny rzeczywistej. Odnosząc się do stwierdzenia Stefana Lessmanna oraz Stefana Voßa dotyczącego wykorzystania regresji liniowej do predykcji takiego zjawiska, wyrażamy przekonanie, że opisywana przez nas metoda nie jest najlepszą metodą do predykcji ceny aut osobowych, co można potwierdzić

wynikiem predykcji uzyskanej w pracy *Car Price Prediction using Machine Learning Techniques*[4], ale biorąc pod uwagę, że w predykcji R^2 było równe 80%, nie można wnioskować, że regresja liniowa nie cechuje się wysoką dokładnością precyzji prognozowania. Niniejsze badanie ukazuje, że należy rozważyć regresję liniową podczas wykonywania predykcji cen aut osobowych, ale w celu uzyskania lepszych wyników proponujemy wykorzystanie bardziej zaawansowanych metod.

Bibliografia

Gegic E. et al. *Car Price Prediction using Machine Learning Techniques*, TEM Journal, 2019, Vol 8, Issue 1

Lessmann S., Voß S. *Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy*, International Journal of Forecasting, 2017, Vol 33, Issue 4

Mycielski J., *Ekonometria*, Uniwersytet Warszawski. Wydział Nauk Ekonomicznych, Warszawa, 2010

[4] E.Gegic et al. op. cit.