Raport końcowy z modelu regresji liniowej

Kacper Tymków ADS II Stopień 2025

# Proces z zbierania danych.

Dane zostały pobrane ze strony: <https://www.autosphere.be>

W dniu pobierania ogłoszeń było ich 1110 i tyle też udało się pobrać.

# Zastosowane transformacje zmiennych

Proces przygotowania danych do trenowania modelu obejmował.

* Zmianę nagłówków z polskich na angielskie
* Usunięcie kolumny ‘Wersja’
* Pominięcie kolumn ‘ID’, ‘Opis’, ‘Link’
* Sprawdzenie typu danych:
  + ID object
  + Brand object
  + Model object
  + Fuel object
  + Mileage int64
  + Year int64
  + Transmission object
  + Price int64
  + Bodywork object
  + CO2 float64
  + Color object
  + Euro Standard object
  + Doors int64
  + HP int64
  + Description object
  + Link object
  + dtype: object
* Sprawdzenie wartości NA
  + ID False
  + Brand False
  + Model False
  + Fuel False
  + Mileage False
  + Year False
  + Transmission False
  + Price False
  + Bodywork False
  + CO2 True
  + Color False
  + Euro Standard False
  + Doors False
  + HP False
  + Description False
  + Link False
* ‘CO2’ posiadało 144 puste komórki. Zostały one uzupełnione Medianą
* ‘Standard Euro’ posiadał wartości NaN, obserwacje te zostały wykluczone z trenowania modelu
* Dokonano rekodowania zmiennych.
  + Nadwozie:
* 'Berline': 'Limousine',
* 'Stadswagen': 'City',
* 'Coupé': 'Coupé',
* 'Bedrijfsvoertuig': 'Other',
* 'Andere': 'Other',
* 'Monovolume': 'Van',
* 'SAV': 'SUV',
* 'Break': 'Other',
* 'Cabriolet': 'Cabriolet'
* ‘Kolor’ -> Jedynie zmiana języka z oryginalnego na angielski
* ‘Skrzynia’ -> Jedynie zmiana języka z oryginalnego na angielski
* Podstawowe statystyki:

A screenshot of a computer screen

AI-generated content may be incorrect.

* Rozkłady zmiennych

A graph with numbers and a bar

AI-generated content may be incorrect.

Rysunek 1: Rozkład kolumny 'Cena'

A graph with numbers and a bar

AI-generated content may be incorrect.

Rysunek 2: Rozkład kolumny 'Przebieg'

A graph with numbers and a bar

AI-generated content may be incorrect.

Rysunek 3: Rozkład kolumny 'CO2'

A graph of a graph with numbers

AI-generated content may be incorrect.

Rysunek 4: Rozkład kolumny 'Rok'

A graph with blue bars

AI-generated content may be incorrect.

Rysunek 5: Rozkład kolumny 'HP'

# Historia budowy modelu regresji

Do finalnego zbudowania modelu regresji wykonany próbne modele za pomocą biblioteki *statsmodels*

Każda zmienna została przetestowana pod kątem osiąganego R2 w celu dobrania najlepszych zmiennych do utworzenia ostatecznego modelu regresji.

Wyniki testów:

|  |  |
| --- | --- |
| Kolumna | Osiągnięte R2 |
| Marka | 0.328 |
| Model | 0.826 |
| Paliwo | 0.197 |
| Przebieg | 0.224 |
| Rok | 0.225 |
| Skrzynia biegów | 0.247 |
| Nadwozie | 0.085 |
| CO2 | 0.009 |
| Kolor | 0.030 |
| Euro standard | 0.006 |
| Drzwi | 0.010 |
| Moc | 0.749 |

Pełny model ze wszystkimi zmiennymi osiągnął R2 - 0.941

Część zmiennych osiągnęła bardzo niskie R2 (zmienne oznaczone kolorem czerwonym) z tego względu zostały wyeliminowane z modelu regresji.

Wyniki modeli regresji usunięciu kolumn oznaczonych na czerwono- 0.939

W dalszej kolejności zrezygnowano ze zmiennych ‘Marka’ i ‘Model’. Decyzja ta została podjęta z 4 powodów.

* Bardzo duża liczba potencjalnych opcji nadmiernie komplikujących model regresji
* Potencjalne ograniczenie użyteczności modelu regresji w sytuacji szacowania ceny pojazdu nie będącego na liście
* Wysoki R2 osiągany przez pozostałe zmienne ~0.85
* Model dalej był istotny statystycznie

W dalszej kolejności przeprowadzono test VIF, który dał wyniki:

| **Feature** | **VIF** | **Wniosek** |
| --- | --- | --- |
| **const** | ~2 940 906 | Intercept – naturalnie może mieć bardzo wysoki VIF, nie interpretujemy tego |
| **Mileage** | **3.89** | W normie, brak silnej współliniowości |
| **Year** | **3.84** | W normie, nie jest silnie skorelowana z innymi zmiennymi |
| **HP** | **1.60** | Bardzo dobrze |
| **Fuel\_Electric** | **1.96** | Brak problemów |
| **Fuel\_Hybride** | **1.56** | Brak problemów |
| **Fuel\_LPG** | **1.85** | Brak problemów |
| **Transmission\_Manual** | **1.54** | Brak problemów |

Wniosek: **Żadna zmienna nie przekracza poziomu VIF = 5**, więc nie ma istotnej współliniowości pomiędzy zmiennymi niezależnymi.

**Analizy reszt:**

* Liniowość,
* Homoscedastyczność (stała wariancja reszt),
* Normalność rozkładu reszt,
* Brak utokorelacji reszt.

A graph with blue dots and a red line

AI-generated content may be incorrect.A graph of a normal distribution

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

Statystyka Durbin-Watsona: 1.7702

W dalszej kolejności przetestowano interakcje:

* X\_inter['Year\_HP'] = X['Year'] \* X['HP'] -> 0.865
* X\_inter['Mileage\_HP'] = X['Mileage'] \* X['HP'] -> 0.879
* X\_inter['Fuel\_Electric\_HP'] = X['Fuel\_Electric'] \* X['HP'] -> 0.856
* X\_inter['Fuel\_LPG\_Mileage'] = X['Fuel\_LPG'] \* X['Mileage'] -> 0.859
* X\_inter['Transmission\_Manual\_HP'] = X['Transmission\_Manual'] \* X['HP'] -> 0.858

Widzimy, że interakcje „Year \* HP” i „Mileage \* HP” dają wyraźną poprawę modelu (odpowiednio 0.865 i 0.879 wobec bazowego 0.856)

# Końcowy model regresji użyty w aplikacji

A screenshot of a computer screen

AI-generated content may be incorrect.

Y = -3629424.4660 - 0.0760 \* Mileage + 1796.8177 \* Year + 233.1252 \* HP - 11198.4543 \* Fuel\_Elektrisch - 3389.1961 \* Fuel\_Gas - 1171.1920 \* Fuel\_Hybride - 4388.0221 \* Fuel\_LPG - 1776.4866 \* Transmission\_Manual