Regresja liniowa. Analiza w kontekście zbioru danych Auto-MPG.

Krzysztof Chołys

Uniwersytet Łódzki,
Wydział Fizyki i Informatyki Stosowanej,
Zaawansowane Metody Obliczeniowe
10 stycznia 2024

Spis treści

1 Wstęp

1.1 Cel pracy

Paca skupia się na analizie możliwości zastosowania metod uczenia maszynowego do przewidywania zużycia paliwa przez samochody na podstawie danych z zestawu Auto-MPG. Wykorzystując model regresji liniowej przeprowadzona zostanie estymacja zużycia paliwa (wyrażonego jako litry na 100 km) w zależności od parametrów technicznych pojazdów, takich jak liczba cylindrów, przyspieszenie, czy waga. Analiza ta obejmie etapy obróbki danych (ang. preprocessing), ich normalizacji oraz podziału na zbiory treningowe i testowe. Na koniec oceniona zostanie skuteczność modelu poprzez analizę jego dokładności, co pozwoli na weryfikację jego zdolności do generalizacji. Kluczowym celem pracy jest zbudowanie zaplecza praktycznego dla poznanej dotychczas teorii.

2 Podstawy teoretyczne

2.1 Kluczowe pojęcia

Zmienna zależna - W statystyce i modelowaniu, zmienna zależna jest wynikiem, który jest mierzony i analizowany w celu zrozumienia wpływu zmiennych niezależnych. Konkretnie w przypadku regresji liniowej, jest to wartość, która jest modelowana lub przewidywana na podstawie zmiennych niezależnych.

Zmienna niezależna - Typ zmiennej, który wykorzystuje się w modelu matematycznym, by wpłynąć na zmienną zależną. W regresji liniowej symbolizuje szereg cech, których wartości zmieniają ostateczny wynik zmiennej niezależnej. Przykładowo, przewidując wartość dolara w następnym miesiącu, bazując na danych historycznych, zmiennymi niezależnymi byłyby: stopa inflacji, indeksy giełdowe i stopy procentowe banków.

Zmienna kategoryczna - Typ zmiennej, która przyjmuje wartości wskazujące na przynależność do określonej kategorii lub klasy. Przykładami mogą być biologiczna płeć (mężczyzna, kobieta), rodzaj pojazdu (samochód, motocykl, rower) czy grupa krwi (A, B, AB, 0). W praktyce, modelowanie danych często wymaga by zmienne te zostały zakodowane jako zmienne numeryczne. W analizie statystycznej i uczeniu maszynowym zmienne kategoryczne są ważne, ponieważ pozwalają na analizę i modelowanie danych, które nie są naturalnie liczbowe.

Błąd średniokwadratowy (ang. Mean Squared Error, MSE) - Popularna miara wykorzystywana w ocenie jakości modeli regresyjnych w statystyce i uczeniu maszynowym. MSE jest średnią wartością kwadratów różnic pomiędzy przewidywanymi a rzeczywistymi wartościami zmiennej zależnej. Jest to sposób na kwantyfikację błędu modelu; mniejsza wartość MSE wskazuje na lepsze dopasowanie modelu do danych. MSE jest użyteczne, ponieważ podkreśla większe błędy (poprzez podnoszenie ich do kwadratu) i jednocześnie aprobuje małe (mniejsze niż 1 i większe od 0, wykorzystując ten sam mechanizm). Pozwala to na łatwą interpretację dokładności modelu i dalszą analizę danych.

2.2 Wzory

Postać ogólna regresji liniowej:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{1}$$

gdzie:

- y_i – wartość zmiennej zależnej dla i-tej obserwacji
- β_0 – wyraz wolny (przecięcie z osią Y)
- β_j – współczynnik przy j-tej zmiennej niezależnej
- x_{ij} – wartość j-tej zmiennej niezależnej dla i-tej obserwacji
- ϵ_i – błąd dla i-tej obserwacji

Metoda najmniejszych kwadratów:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2}$$

gdzie:

- $\hat{\beta}$ – wektor estymowanych współczynników regresji
- X – macierz zawierająca wartości zmiennych niezależnych wraz z kolumną jedności dla wyrazu wolnego
- X^T – transpozycja macierzy X
- $(X^T X)^{-1}$ – odwrotność macierzy $X^T X$
- y – wektor obserwowanych wartości zmiennej zależnej

3 Metodologia

3.1 Język i biblioteki

Program został napisany w języku Python 3.11. Wykorzystane zostały biblioteki:

- numpy – obliczenia numeryczne
- pandas – przetwarzanie danych
- sklearn – przetwarzanie danych i budowanie modelu regresji liniowej
- matplotlib – wizualizacja danych
- seaborn – wizualizacja danych

3.2 Struktura programu

Program zaprojektowany do analizy danych ze zbioru Auto-MPG składa się z kilku kluczowych elementów, które współpracują, aby umożliwić efektywną analizę i modelowanie danych. Projekt bazowany jest na paradygmacie obiektowym.

1. **Wczytywanie danych:** Klasa `CarsUtils` zawiera metodę `read_cars_from_file`, która wczytuje dane z pliku. Przetwarza ona surowe dane, konwertując je na strukturę DataFrame z biblioteki pandas.

2. **Przetwarzanie danych:** Program zawiera metody do przekształcania danych, takie jak `convert_mpg_to_litres_per_100km`, która konwertuje zużycie paliwa z mil na galon paliwa (MPG) na litry na 100 km, oraz `normalise_data`, służącą do normalizacji wybranych kolumn danych. Ponadto zmienna kategoryczna `origin` została skonwertowana do wartości numerycznych $1, 2, 3$, odpowiadających kolejno: Stanom Zjednoczonym Ameryki, Europie i Japonii. W procesie wykorzystano kodowanie typu one-hot (ang. one-hot encoding).

3. **Analiza i wizualizacja:** Dostępne są różnorodne metody do wizualizacji danych, w tym wykresy punktowe, histogramy i mapy korelacji. Te narzędzia wizualizacji pomagają zrozumieć relacje między różnymi cechami w danych. Znajdujące się w dalszej części pracy wykresy zostały wygenerowane z użyciem tych właśnie metod.

4. **Modelowanie:** Używając klasy `LinearRegression` z biblioteki `sklearn`, program buduje i trenuje model regresji liniowej na przetworzonych danych. Model ten jest następnie wykorzystywany do przewidywania i oceny na podstawie zbioru testowego.

5. **Ocena modelu:** Program ocenia dokładność modelu, korzystając z takich metryk jak średni błąd kwadratowy (MSE) i prezentuje wyniki.

4 Wyniki

4.1 Podstawowe obserwacje

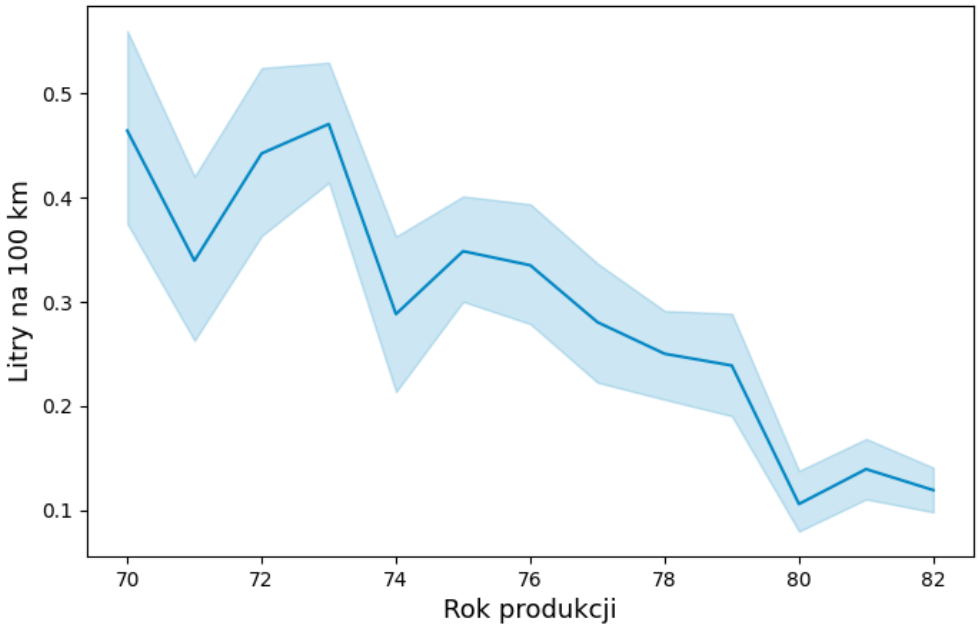
Po inicjalnej analizie danych, ze względu na ich szeroką rozbieżność numeryczną (1), konieczna jest ich normalizacja.

	count	mean	std	min	25%	50%	75%	max
cylinders	313.0	5.482428	1.700446	3.0	4.0	4.0	8.0	8.0
displacement	313.0	195.517572	103.766567	70.0	105.0	151.0	302.0	455.0
horsepower	313.0	104.594249	38.283669	46.0	76.0	95.0	129.0	230.0
weight	313.0	2986.124601	841.133957	1613.0	2234.0	2855.0	3645.0	5140.0
accel	313.0	15.544089	2.817864	8.0	13.5	15.5	17.3	24.8
model_year	313.0	76.207668	3.630136	70.0	73.0	76.0	79.0	82.0
USA	313.0	0.645367	0.479168	0.0	0.0	1.0	1.0	1.0
Europe	313.0	0.153355	0.360906	0.0	0.0	0.0	0.0	1.0
Japan	313.0	0.201278	0.401597	0.0	0.0	0.0	0.0	1.0

Rysunek 1: Ogólne statystyki uczącego zestawu danych

Średnie wartości cech zestawu zaczynają się na 0.15, a kończą na 2986. Porównywanie tak różnych liczb byłoby ciężkie do usystematyzowania. W celu ułatwienia późniejszej analizy i interpretacji zastosowano normalizację przedstawionych powyżej (1) wartości.

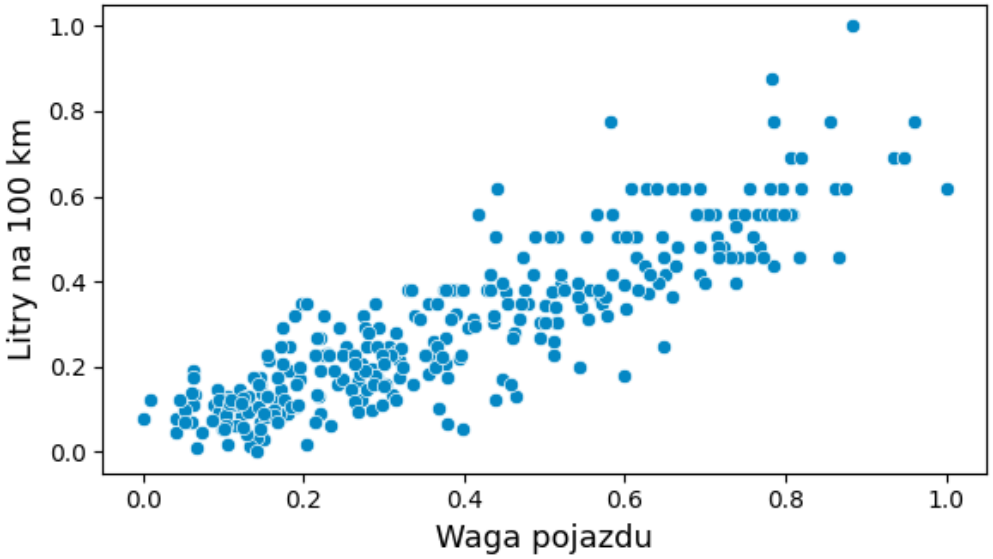
Jednym z pierwszych kroków po wczytaniu i obróbce danych była wizualizacja kilku cech w zestawieniu ze spalaniem pojazdu na 100 kilometrów wyrażonym w litrach.



Rysunek 2: Wykres zależności roku produkcji pojazdu od spalania paliwa

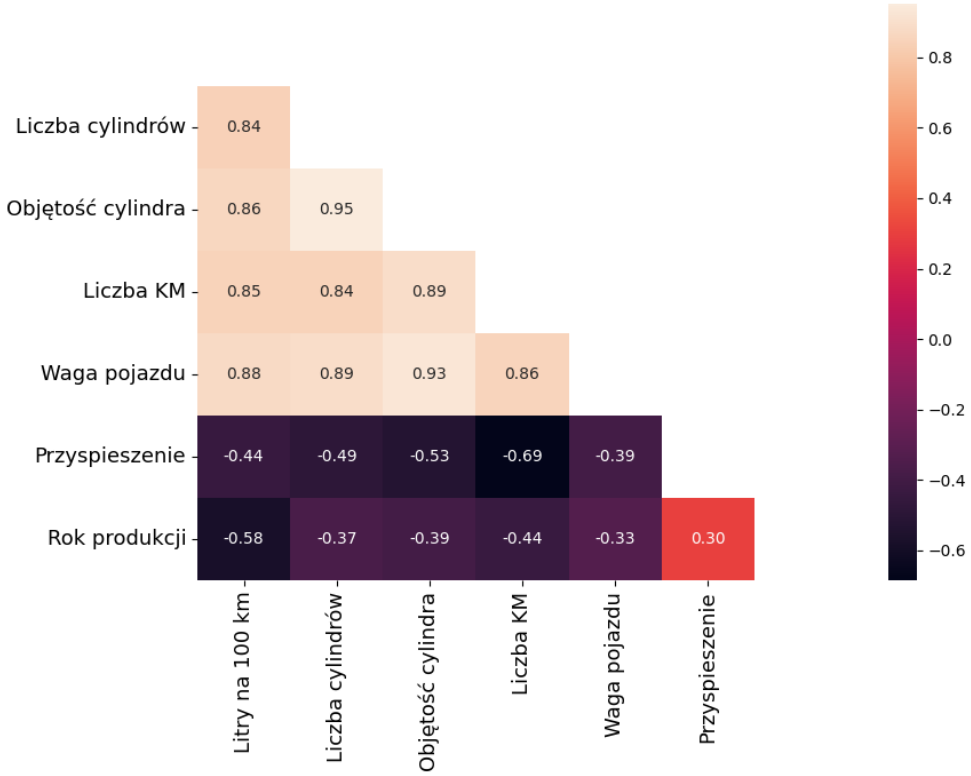
Na wykresie przedstawiono opisaną zależność z uwzględnieniem rozproszenia danych. Poza łatwo zauważalną umiarkowaną negatywną korelacją między współczynnikami, ciekawe jest zmniejszające się odchylenie standardowe danych. Potencjalnie może to wynikać z małej ilości dostępnych danych.

Analogicznie analizować można każdą z cech w zestawieniu ze sobą. Najwyższy współczynnik Pearsona w relacji ze spalanym paliwem osiąga waga pojazdu (3). Z danych punktowych można odczytać dodatni współczynnik kierunkowy, który jasno zaznacza obecność prostej. Dane znajdują się w wyraźnie bliskim otoczeniu linii, co świadczy o silnej, wysokiej korelacji danych.



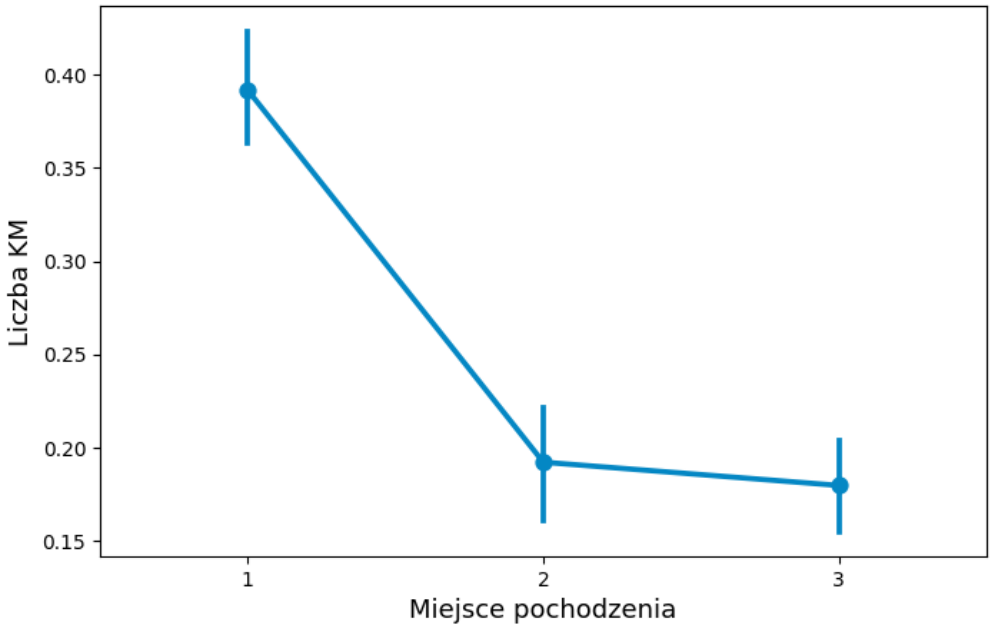
Rysunek 3: Wykres zależności wagi pojazdu od spalania paliwa

Problemem występującym przy tego typu analizie cech jedna po drugiej, jest relatywnie niska ilość informacji w porównaniu do czasu poświęconego na jej wydobycie. Dużo lepszym i bardziej pełnym zbiorem informacji jest wykorzystanie mapy współczynników Pearsona cech względem siebie nawzajem. Rozwiązanie to zaprezentowano na wykresie poniżej (4)



Rysunek 4: Wykres zależności współczynników Pearsona cech względem siebie

Analiza takiego wykresu pozwala zauważyć, że spalanie jednoznacznie rośnie wraz ze wzrostem wagi pojazdu, liczby koni mechanicznych, liczby cylindrów i objętości cylindra. Ułatwia to tym samym zauważenie i badanie zależności pomiędzy cechami, które wpływają na siebie wzajemnie, a w ostateczności, na ilość spalanego paliwa. Gdyby wzbogacić ten zestaw informacji o fakt, że wg danych w Stanach Zjednoczonych produkuje się pojazdy właśnie z dużą wagą, liczbą koni (5) i liczbą cylindrów, można by dojść do łatwego wniosku - samochody wyprodukowane w Stanach Zjednoczonych będą spalały więcej paliwa w porównaniu do pojazdów z Europy czy Japonii.

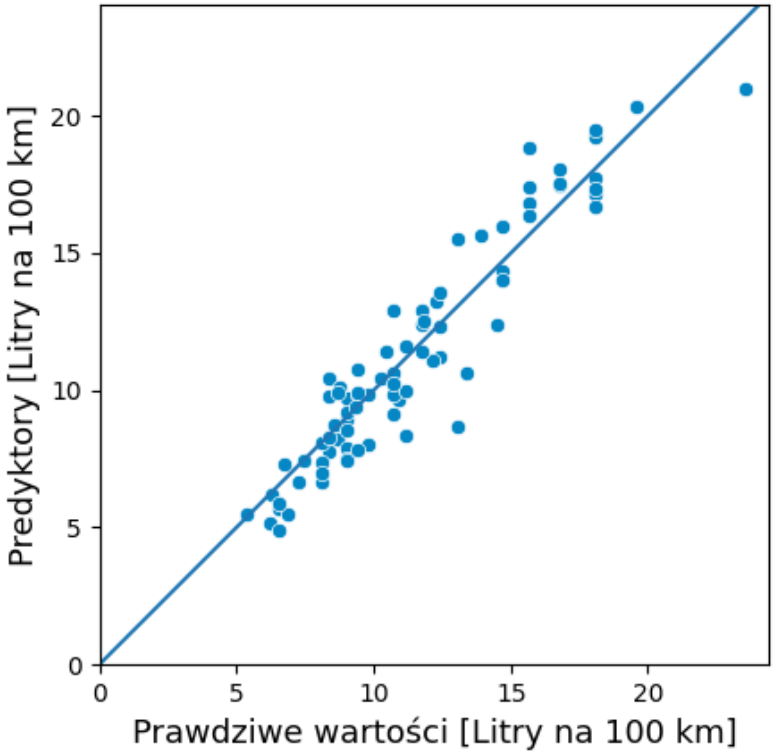


Rysunek 5: Wykres zależności miejsca pochodzenia pojazdu i liczby koni mechanicznych. 1 – Stany Zjednoczone, 2 – Europa, 3 – Japonia

Powyższa relacja (5) zachodzi w podobnej skali dla innych parametrów: wagi, liczby cylindrów i objętości cylindra. Poza faktem, że samochody pochodzące ze Stanów Zjednoczonych spalają najwięcej paliwa, warto zauważyć, że samochody z Japonii plasują się jednocześnie na drugim końcu spektrum. Europa zajmuje tym samym drugie miejsce.

4.2 Porównanie wyników różnych metod regresji liniowej

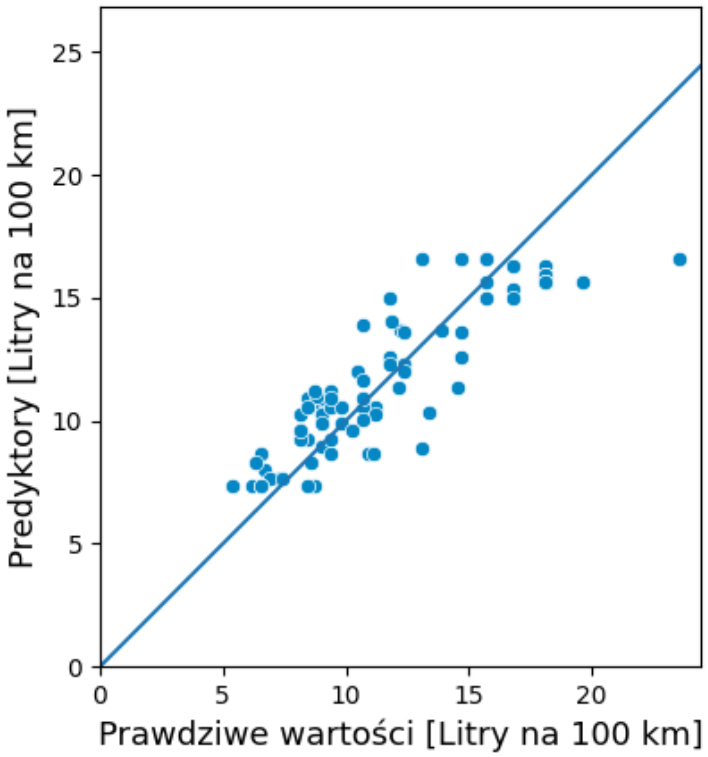
Najbardziej oczywistym, najprostszym i jak się okaże, najbardziej efektywnym rozwiązaniem jest metoda `LinearRegression()` bazująca na metodzie najmniejszych kwadratów.



Rysunek 6: Wykres zależności wartości prawdziwych od wartości przewidywanych testowego zestawu danych. Regresja liniowa, metoda najmniejszych kwadratów.

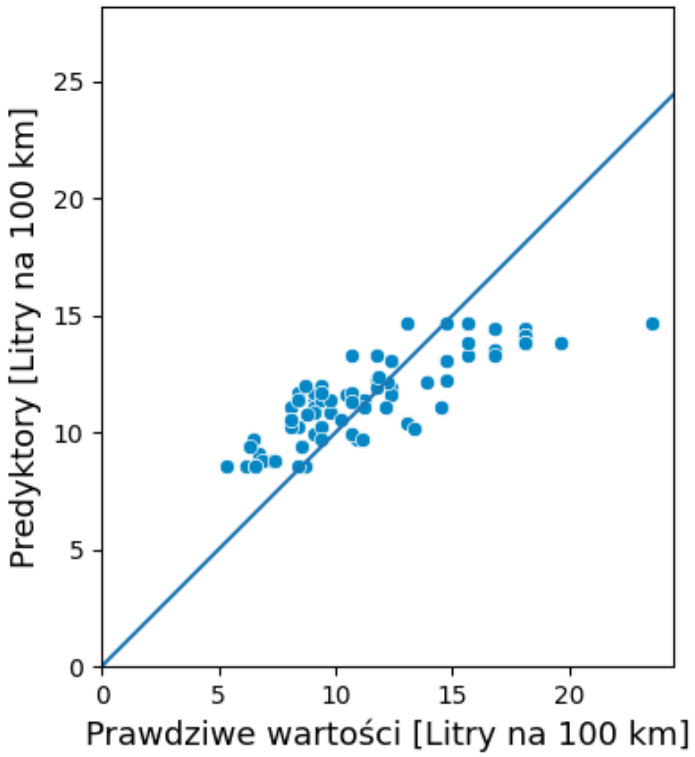
Dane rozkładają się wzdłuż linii z wartością błędu średniokwadratowego równą 1.74.

Jedną z innych przykładowych typów regresji liniowej jest `Lasso()`. Charakterystyczne jest dla niej zbieganie się wartości do wybranego punktu centralnego. Efekty zastosowania tego rodzaju regresji widać na wykresach (7) i (8).



Rysunek 7: Wykres zależności wartości prawdziwych od wartości przewidywanych testowego zestawu danych. Regresja liniowa, metoda LASSO. Parametr $\alpha = 1$.

W porównaniu do wykresu bazującego na metodzie najmniejszych kwadratów (6), dane wyraźnie koncentrują się bliżej środka. Zwiększenie wartości parametru alfa, który jest odpowiedzialny za regularyzację L1 (dodanie do funkcji celu kary proporcjonalnej do wartości bezwzględnej współczynników regresji) będzie skutkowało jeszcze silniejszym nagromadzeniem punktów danych wokół centrum wykresu. Naturalnie, metoda ta może być dobrze wykorzystana do przeprowadzenia regresji na zbiorze danych, które występują w relatywnie wysokim rozrzedzeniu.



Rysunek 8: Wykres zależności wartości prawdziwych od wartości przewidywanych testowego zestawu danych. Regresja liniowa, metoda LASSO. Parametr $\alpha = 3$.

Wykres analogiczny do poprzedniego (7) z większą wartością współczynniku alfa. Efektem jest jeszcze wyższa koncentracja danych wokół centrum wykresu.

Metoda	**Współczynniki**	**Przecięcie**	**MSE**
LSM	[0.4, -2.8, 5.4, 8.9, 1.3, -0.3, 0.5, -0.4, -0.1]	29.04	1.74
Ridge ($\alpha = 1$)	[0.4, 0.0, 4.0, 6.5, 1.1, -0.3, 0.4, -0.3, -0.2]	29.23	1.78
Ridge ($\alpha = 3$)	[0.6, 0.9, 3.0, 4.6, 0.7, -0.3, 0.4, -0.2, -0.2]	29.30	1.96
Lasso ($\alpha = 1$)	[1.4, 0.0, 0.0, 0.0, -0.0, -0.3, 0.0, -0.0, -0.0]	28.35	3.54
Lasso ($\alpha = 3$)	[0.7, 0.0, 0.0, 0.0, -0.0, -0.3, 0.0, -0.0, -0.0]	29.20	6.10
ElasticNet ($\alpha = 1$)	[1.3, 0.0, 0.0, 0.0, -0.0, -0.4, 0.0, -0.0, -0.0]	31.66	3.73
ElasticNet ($\alpha = 3$)	[0.8, 0.0, 0.0, 0.0, -0.0, -0.3, 0.0, -0.0, -0.0]	33.46	5.58

Tabela 1: Tabela porównawcza różnych metod regresji liniowej

Z zestawionych danych wynika jasno, że do zbioru Auto-MPG najlepiej sprawdziła się podstawowa metoda regresji liniowej bazującej na metodzie najmniejszych kwadratów. W przypadku innego zestawu danych, bądź nawet inaczej rozproszonych danych, użycie pozostałych metod regresji byłoby podstawne i zależnie od sytuacji, optymalne.

Współczynnik	**Wartość**
Liczba cylindrów	0.4
Pojemność cylindrów	-2.8
Liczba koni mechanicznych	5.4
Waga	8.9
Przyspieszenie	1.3
Rok produkcji	-0.3
Stany Zjednoczone	0.5
Europa	-0.4
Japonia	-0.1

Tabela 2: Tabela współczynników dla metody regresji liniowej LSM

Poziom dokładności metod regresji ewaluowany błędem średniokwadratowym (MSE) jednoznacznie określił podstawową regresję liniową bazowaną na metodzie najmniejszych kwadratów (LSM) jako najbardziej precyzyjną. Analizując wartości współczynników uzyskiwanych dzięki jej zastosowaniu, nie jest zaskakujące, że kluczowymi aspektami wpływającym na spalanie samochodu jest jego waga i liczba koni mechanicznych. Są to dwie cechy pojazdu, które rosną proporcjonalnie do ilości zużytego paliwa. Co ciekawe, najistotniejszą cechą po drugiej stronie spektrum okazała się pojemność cylindrów. I to nie sama cecha jest tu interesująca, ale jej skala – współczynnik -2.8 to trochę ponad 31% wartości współczynnika wagi, który utrzymuje się na poziomie 8.9. Niespodziewane wartości ukazały się również przy etykietach pochodzenia pojazdu – przypomnijmy, że samochody z najmniejszą liczbą koni mechanicznych pochodziły z Japonii (5). Tymczasem to Europa plasuje się na pierwszym miejscu pod względem ekonomicznych samochodów.

5 Podsumowanie

5.1 Co udało się zbadać?

W ramach pracy udało się skutecznie zbudować i zastosować model regresji liniowej do analizy i przewidywania zużycia paliwa na podstawie danych ze zbioru Auto-MPG. Przeprowadzona została obróbka i normalizacja danych. Pozwoliło to na wydobywanie użytecznych informacji i trendów. Udało się napisać model, który z zadowalającą dokładnością przewiduje zużycie paliwa, bazując na takich parametrach jak pojemność silnika, waga pojazdu, czy rok produkcji.

Praca ta pozwoliła na zrozumienie wpływu różnych cech technicznych pojazdów na ich efektywność paliwową. Co więcej, eksperymentowanie z różnymi metodami regresji, takimi jak Ridge, Lasso, czy ElasticNet, dało możliwość porównania ich skuteczności i wpływu na wyniki modelu. Podsumowując, projekt doprowadził do ciekawych wniosków dotyczących zastosowania uczenia maszynowego w analizie danych oraz pokazał potencjał modelowania w przewidywaniu parametrów pojazdów.

6 Bibliografia

[1] J. Frost. *Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models*. Statistics By Jim Publishing, 2020. ISBN: 9781735431185.

[2] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.