Analiza i Predykcja Cen Samochodów Volkswagen Passat

Tomasz Jaskólski 272741

10 lipca 2024

Spis treści

1 Wstęp **2**

2 Zbieranie danych **2**

3 Przetwarzanie wstępne i oczyszczanie danych **3**
3.1 Eliminacja wartości odstających 3
3.2 Korekta danych 4
3.3 Przypisanie brakujących wartości 4

4 Wizualizacja danych po oczyszczeniu **5**
4.1 Rozkłady zmiennych 5
4.2 Wykresy kategorii zmiennych 9

5 Analiza korelacji zmiennych z ceną samochodu **16**

6 Implementacja Modeli Predykcyjnych **18**
6.1 Regresja Liniowa i Grzbietowa 18
6.2 Polynomial Regression 20
6.3 Decision Tree i Random Forest 22
6.4 CatBoost 24
6.5 Porównanie Modeli 25

7 Podsumowanie i wnioski **25**

1 Wstęp

Współczesny rynek samochodowy jest dynamiczny i zróżnicowany, a zrozumienie czynników wpływających na ceny pojazdów jest kluczowe dla konsumentów, czy dealerów. Samochody Volkswagen Passat, ze względu na swoją popularność i renomę stanowią interesujący przedmiot analizy. Wybrano model Passata B8 produkowany w latach 2014-2023, ze względu na dużą liczbę ogłoszeń oraz stosunkowo niewielką ilość odstających ofert.

Inspiracją do przeprowadzenia tego badania była, oprócz szeroko pojętego zainteresowania motoryzacją, chęć lepszego zrozumienia, w jaki sposób różne parametry techniczne i użytkowe, takie jak rok produkcji, przebieg, pojemność silnika czy typ paliwa, wpływają na końcową cenę pojazdu. Dodatkowo, uwzględniono również aspekty takie jak napęd na cztery koła, bez wypadkowość czy faktura VAT.

Analiza przeprowadzona w ramach projektu ma na celu:

- Zbadanie zależności między ceną samochodu a jego parametrami technicznymi i użytkowymi,
- Stworzenie modeli regresyjnych, które umożliwią predykcję cen samochodów na podstawie wybranych parametrów,
- Porównanie skuteczności różnych metod modelowania.

Niniejszy raport przedstawia metodykę badania, wyniki analizy oraz wnioski płynące z przeprowadzonych modeli predykcyjnych, co pozwoli na dokładniejsze zrozumienie mechanizmów kształtujących ceny samochodów na rynku wtórnym.

2 Zbieranie danych

Dane użyte w niniejszym projekcie zostały zebrane z serwisu Otomoto, który jest jednym z najpopularniejszych portali ogłoszeniowych w Polsce dla samochodów używanych. Zbiór danych powstał poprzez jednorazowe załączenie skryptu pobierającego i umieszczającego w bazie danych wszystkie dostępne ogłoszenia dotyczące samochodów Volkswagen Passat B8.

Łącznie zebrano około 2100 rekordów, zawierających szczegółowe informacje o każdym z samochodów, takie jak:

- Cena [zł] *(int)*,
- Rok produkcji *(int)*,
- Przebieg [km] *(int)*,

- Pojemność silnika [cm^3] *(int)*,
- Moc silnika [hp] *(int)*,
- Rodzaj paliwa *(enum('Diesel', 'Petrol'))*,
- Typ skrzyni biegów *(enum('Automatic', 'Manual'))*,
- Informacja o bezwypadkowości *(bool)*,
- Kraj pochodzenia *(string)*,
- Napęd na cztery koła *(bool)*,
- Faktura VAT *(bool)*,
- Typ nadwozia *(enum('Estate', 'Sedan'))*,

Wykorzystując filtry na stronie Otomoto, usunięto ogłoszenia aut uszkodzonych oraz nowych, koncentrując się jedynie na samochodach używanych.[1].

3 Przetwarzanie wstępne i oczyszczanie danych

W ramach projektu przeprowadzono szereg operacji wstępnego przetwarzania oraz oczyszczania danych, mających na celu uzyskanie możliwie najbardziej reprezentatywnego zbioru danych. Nieocenianą pomoc stanowiła tutaj biblioteka ydata-profiling.

3.1 Eliminacja wartości odstających

W celu usunięcia wartości odstających tzw. outlinerów zastosowano następujące kryteria:

- **Cena:** Usunięto samochody z ceną poniżej 30 000 PLN oraz powyżej 160 000 PLN. Samochody z ceną poniżej 30 000 PLN są najczęściej uszkodzone lub wystawione niepoprawnie (np. cena w ogłoszeniu stanowi kwotę cesji leasingu nie wliczając pozostałych do opłacenia rat i wykupu), podczas gdy samochody powyżej 160 000 PLN często reprezentują nowe pojazdy lub wyjątkowo drogie egzemplarze, które nie są reprezentatywne dla rynku wtórnego.

[1] `https://www.otomoto.pl/osobowe/uzywane/volkswagen/passat?search[filter_enum_generation]=gen-b8-2014&search[order]=created_at_first:desc&page=`

- **Rok produkcji:** Wyeliminowano wszystkie pojazdy wyprodukowane przed rokiem 2014, aby uniknąć uwzględnienia Passatów starszej generacji, które mogły być mylnie sklasyfikowane jako nowsze modele.
- **Przebieg:** Samochody z przebiegiem poniżej 10 000 km zostały usunięte, ponieważ najprawdopodobniej są to pojazdy nowe wystawione jako używane. Ponadto, usunięto pojazdy z przebiegiem powyżej 400 000 km, gdyż takie przebiegi są niereprezentatywne dla rynku wtórnego.
- **Pojemność silnika:** Pojemność silnika dla Passata generacji B8 mieści się w zakresie od 1395 cm^3 do 1984 cm^3. Ogłoszenia dotyczące samochodów, których pojemność silnika nie mieściła się w tym zakresie, zostały wyeliminowane. Dodatkowo, skorygowano niepoprawne pojemności silników zgodnie z opisaną poniżej metodą.
- **Moc silnika:** Samochody z mocą poniżej 120 HP oraz powyżej 280 HP zostały usunięte, ponieważ Passaty tej generacji nie były wyposażone w silniki o takich mocach, co wskazuje na błędne wypełnienie ogłoszeń.

Dzięki zastosowaniu wyżej wymienionych kryteriów wyeliminowana łącznie około 60 błędnych ogłoszeń i outlinerów.

3.2 Korekta danych

Wartości pojemności i mocy silnika zostały skorygowane zgodnie z poniższymi zasadami:

- **Pojemność silnika:** Jeżeli pojemność silnika nie znajdowała się w zbiorze możliwych wartości, została dopasowana do najbliższej poprawnej wartości dla danego typu paliwa.
- **Moc silnika:** Jeżeli moc silnika nie znajdowała się w zbiorze możliwych wartości, została dopasowana do najbliższej poprawnej wartości. Różnice te najprawdopodobniej wynikały z niewielkich odchyleń pomiędzy jednostkami mocy HP i PS.

3.3 Przypisanie brakujących wartości

W przypadku brakujących wartości dotyczących pochodzenia samochodów, zastosowano przypisanie na podstawie istniejącego rozkładu danych zagranicznych samochodów. Brakujące wartości wynikały z faktu, że sprzedający nie podali tych informacji. Po dokładnej analizie części ogłoszeń stwierdzono, że większość tych

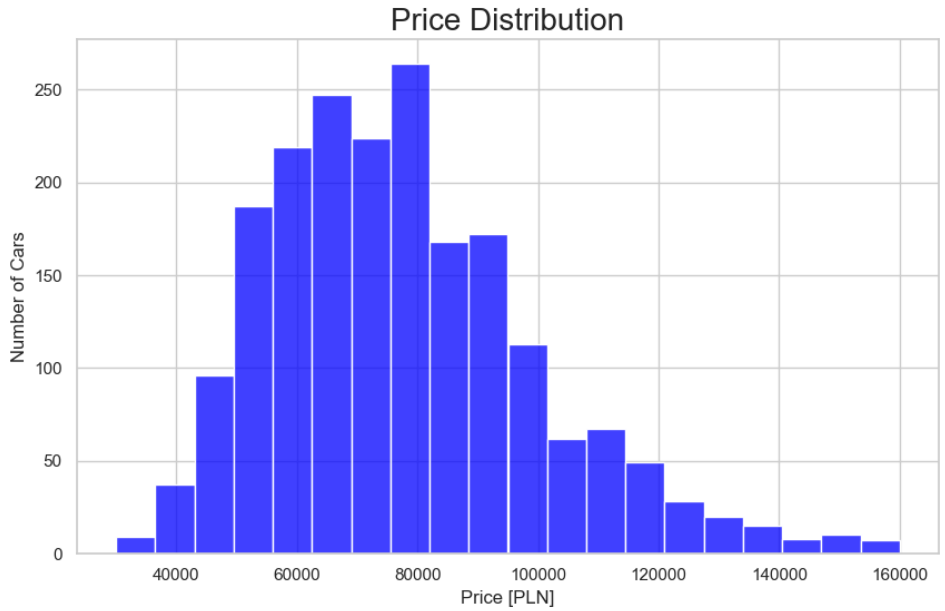
samochodów pochodzi z zagranicy, jednak sprzedający celowo pomijali tę informację, aby ominąć filtrowanie na Otomoto. W związku z tym, brakujące wartości przypisano losowo, wykorzystując wagi na podstawie rozkładu danych zebranych ogłoszeń aut pochodzących z zagraniczny.

4 Wizualizacja danych po oczyszczeniu

W celu lepszego zrozumienia rozkładu i wzajemnych zależności danych po procesie oczyszczania, przeprowadzono szczegółową wizualizację wybranych zmiennych. Poniżej znajdują się opisy wykresów, które zostały uwzględnione w analizie.

4.1 Rozkłady zmiennych

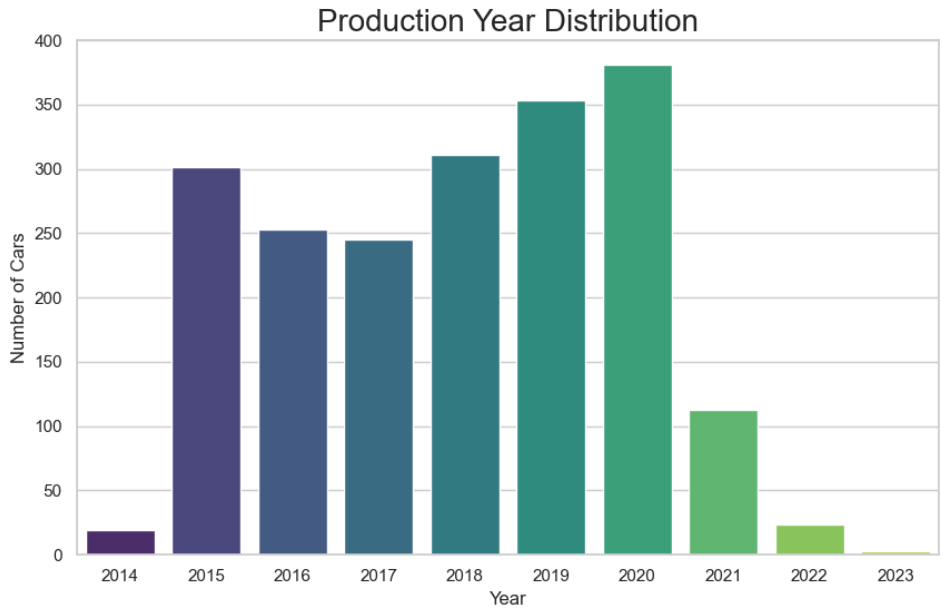
- **Rozkład cen samochodów:** Histogram przedstawiający rozkład cen samochodów w zbiorze danych, z największą koncentracją cen w przedziale 60,000 - 90,000 PLN, co wskazuje na dominującą cenę rynkową.



Rysunek 1: Rozkład cen samochodów

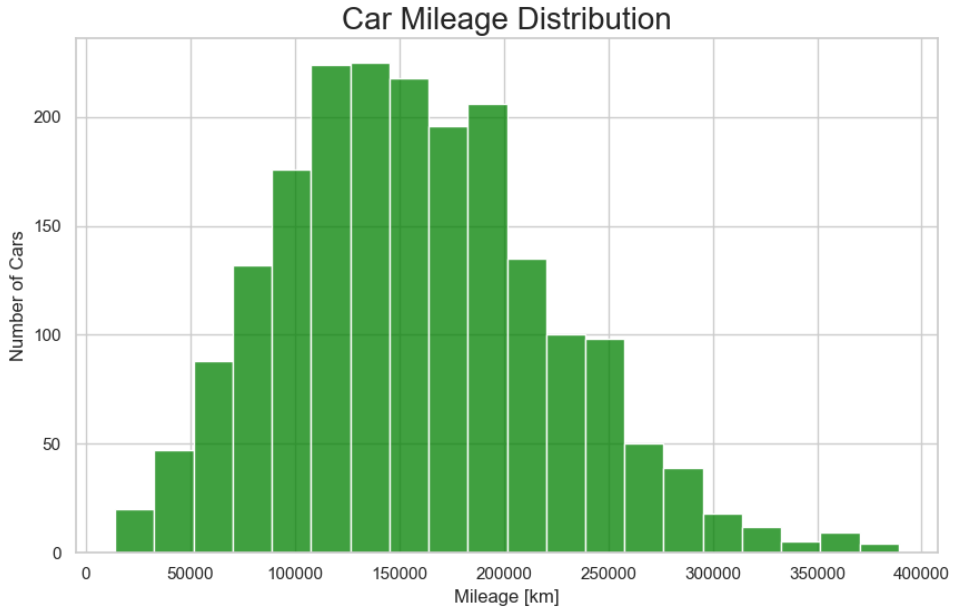
- **Rozkład roczników produkcji:** Histogram prezentujący rozkład lat produkcji samochodów. Wykres słupkowy pokazuje, że większość samochodów

w zbiorze danych pochodzi z lat 2015 - 2020 z najwyższymi wartościami dla 2019 i 2020, co może sugerować, że są to samochody po-leasingowe.



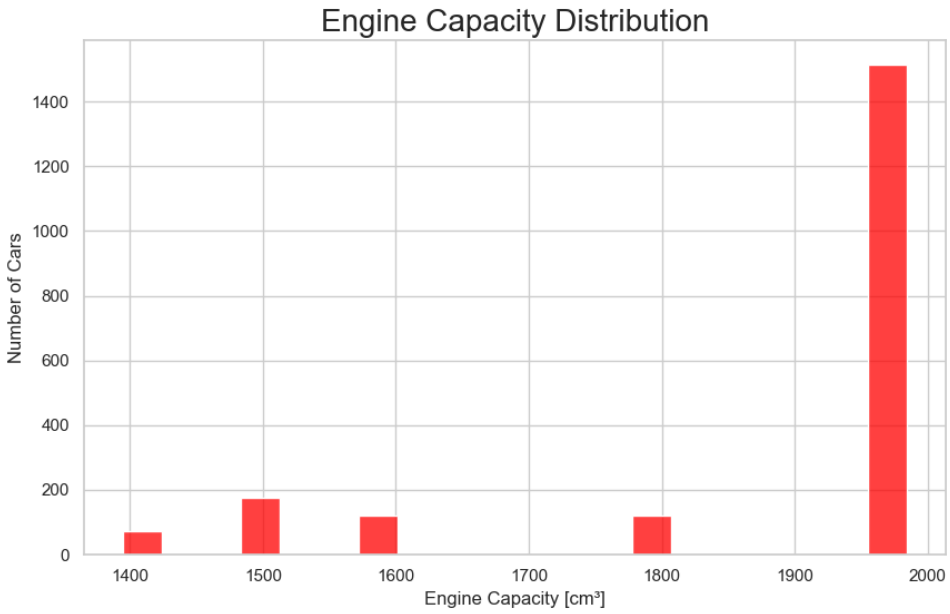
Rysunek 2: Rozkład roczników produkcji

- **Rozkład przebiegu samochodów:** Histogram ukazujący rozkład przebiegu samochodów. Wykres ten wskazuje, że największa liczba pojazdów ma przebieg w zakresie 100,000 - 200,000 km, co jest typowe dla używanych samochodów tej klasy.



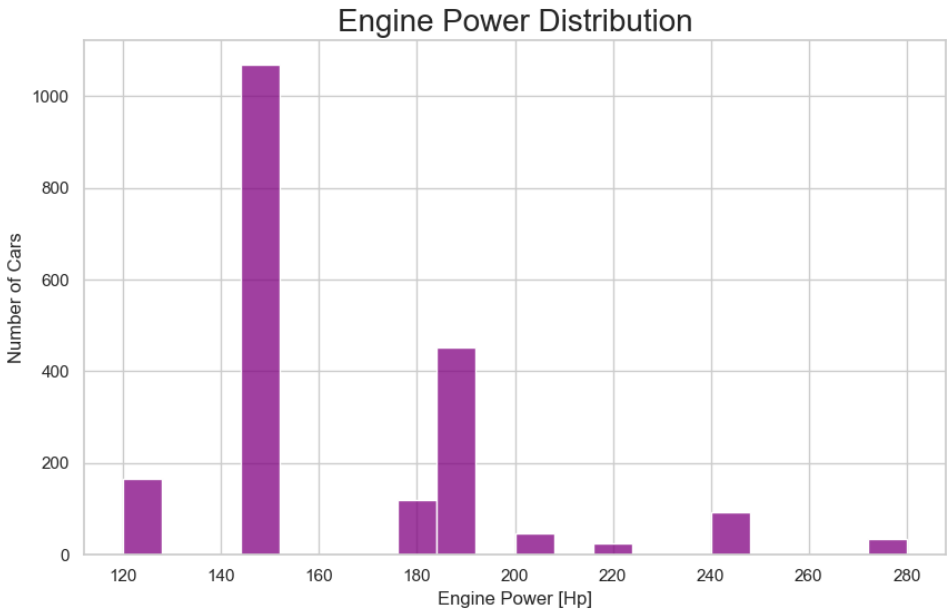
Rysunek 3: Rozkład przebiegu samochodów

- **Rozkład pojemności silnika:** Histogram przedstawiający rozkład pojemności silnika samochodów. Ukazuje on dominację pojemności zbliżonych do 2000 cm^3 co wskazuje na dominację silników 2.0 TDI oraz 2.0 TFSI.



Rysunek 4: Rozkład pojemności silnika

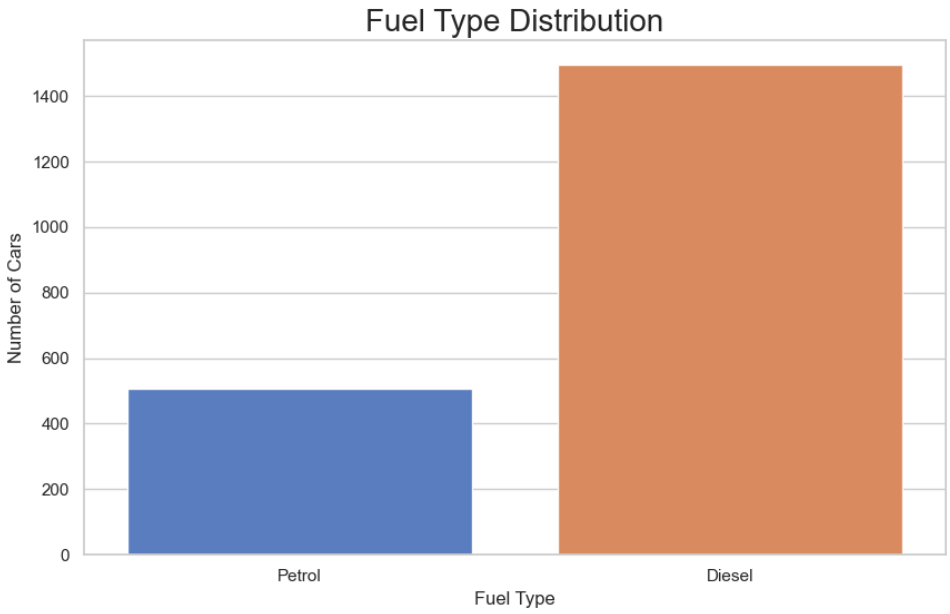
- **Rozkład mocy silnika:** Histogram ilustrujący rozkład mocy silnika w samochodach. Ten wykres pokazuje, że najczęściej spotykane moce to około 150 HP, co odpowiada typowym konfiguracjom dla tego modelu.



Rysunek 5: Rozkład mocy silnika

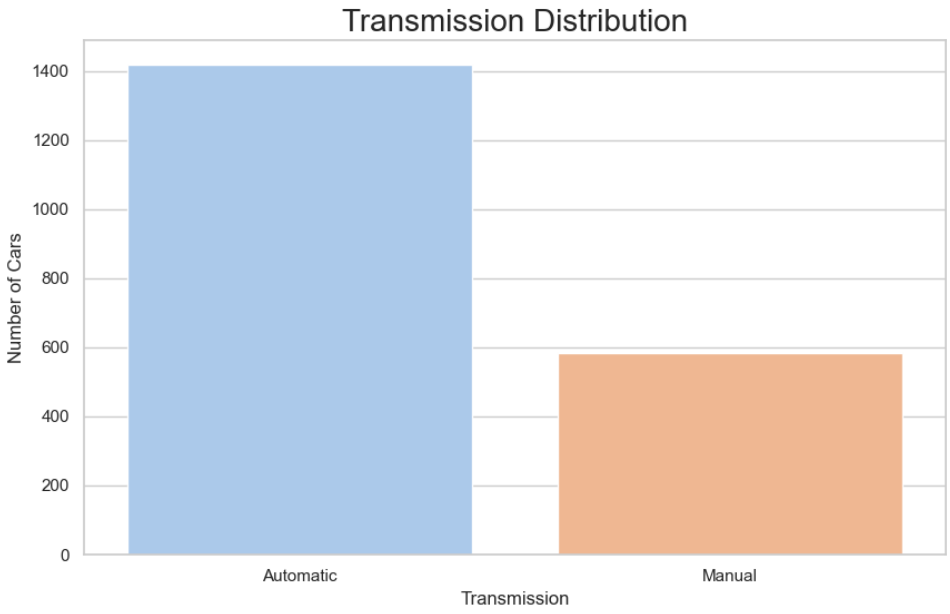
4.2 Wykresy kategorii zmiennych

- **Rozkład rodzajów paliwa:** Wykres słupkowy prezentujący rozkład rodzajów paliwa używanych w samochodach (Diesel, Benzyna). Wykres wskazuje istotnie wyższą popularność silników na ropę, co odzwierciedla preferencje użytkowników w Polsce.



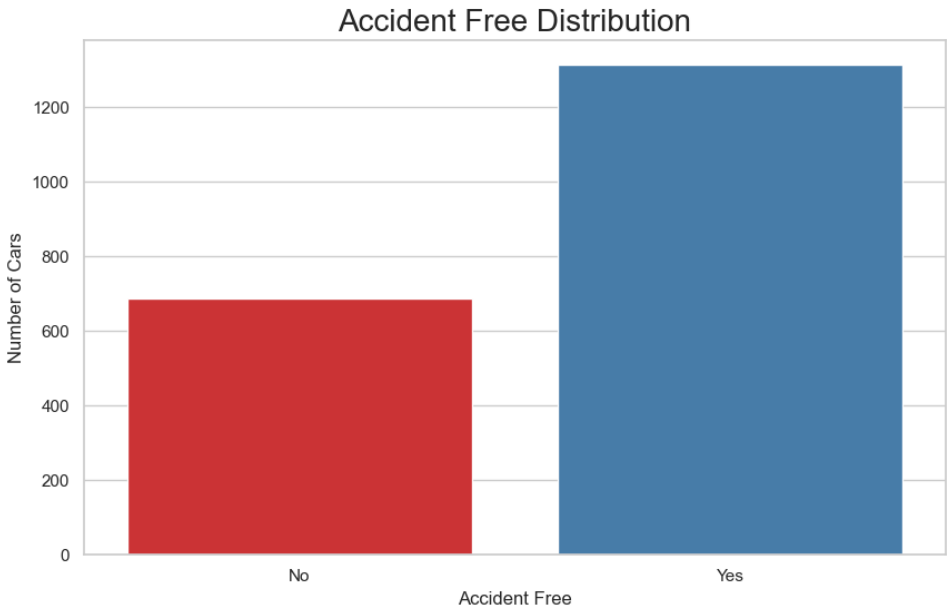
Rysunek 6: Rozkład rodzajów paliwa

- **Rozkład typów skrzyni biegów:** Wykres słupkowy przedstawiający rozkład typów skrzyni biegów (Automatyczna, Manualna). Wykres pokazuje, że skrzynie automatycznie są popularniejsze niż tradycyjne.



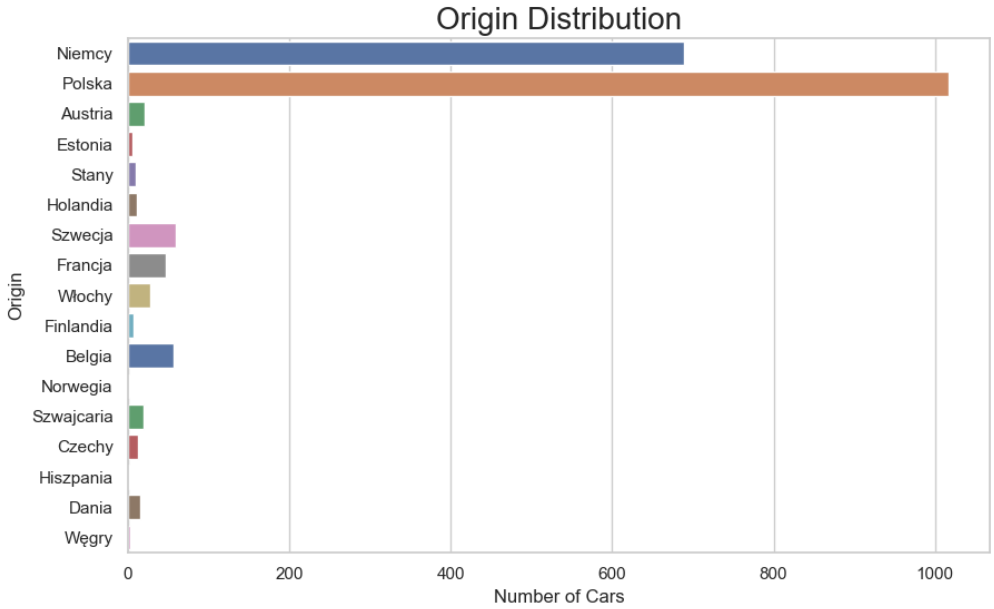
Rysunek 7: Rozkład typów skrzyni biegów

- **Rozkład bez wypadkowości:** Wykres słupkowy ukazujący rozkład informacji o bez wypadkowości samochodów. Pokazane jest, że większość samochodów w zbiorze danych jest bezwypadkowa, co może świadczyć o dobrym stanie technicznym oferowanych pojazdów lub (prawdopodobniej) o przemilczaniu negatywnych informacji przez sprzedających.



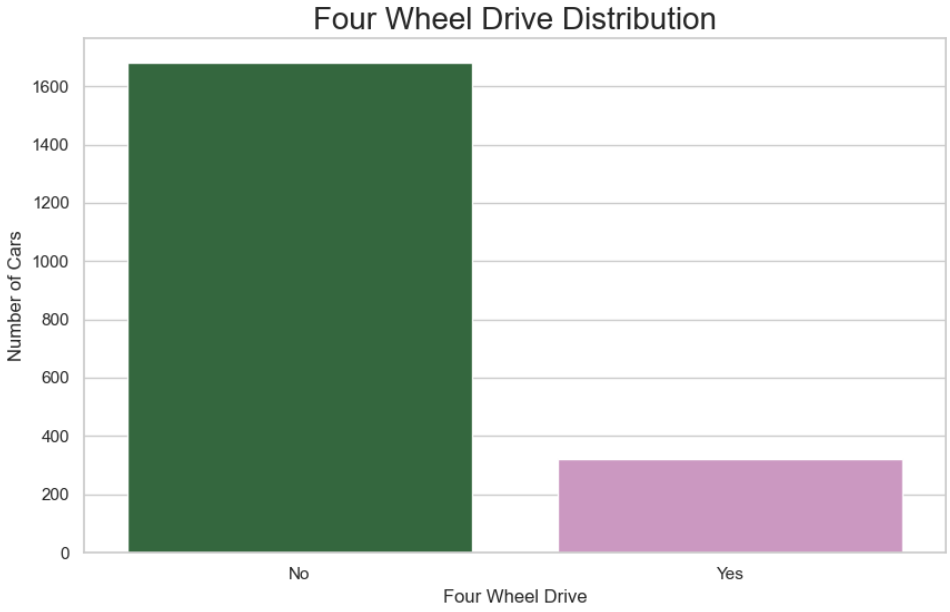
Rysunek 8: Rozkład bez wypadkowości

- **Rozkład pochodzenia samochodów:** Wykres słupkowy przedstawiający pochodzenie samochodów. Na wykresie widać, że większość samochodów pochodzi z Polski i Niemiec, w dalszej kolejności z reszty krajów Europy, a auta z reszty świata stanowią margines.



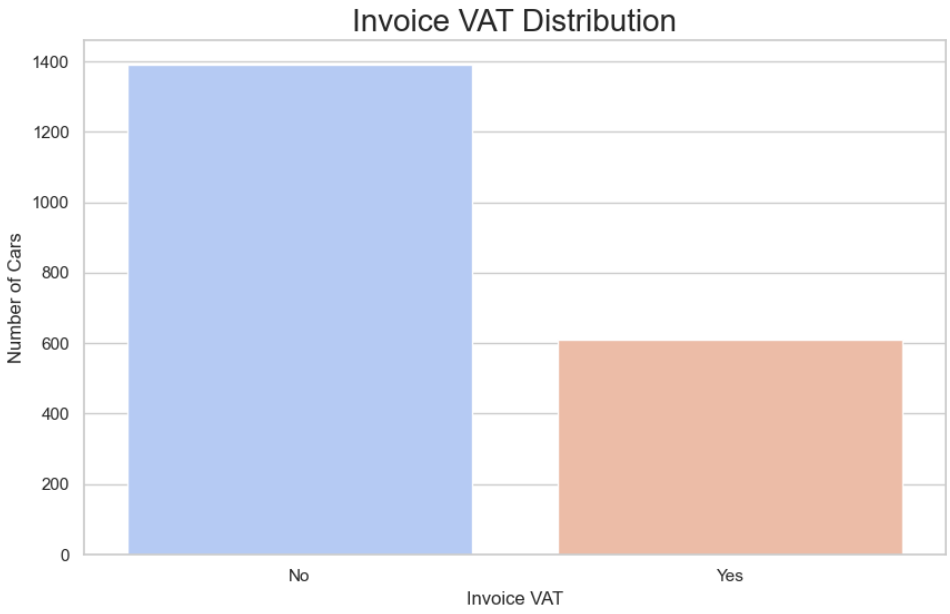
Rysunek 9: Rozkład pochodzenia samochodów

- **Rozkład napędu na cztery koła:** Wykres słupkowy ilustrujący rozkład informacji o napędzie na cztery koła.



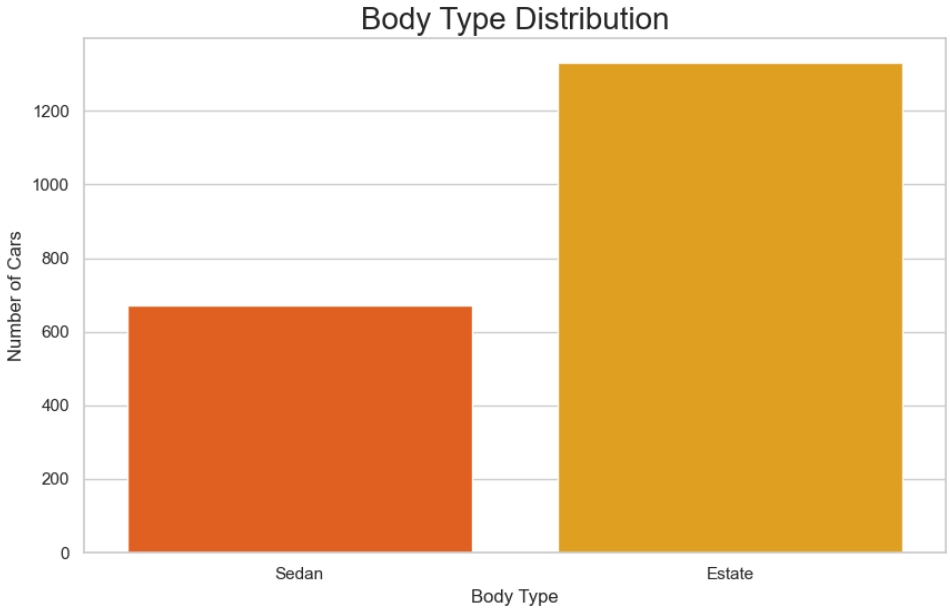
Rysunek 10: Rozkład napędu na cztery koła

- **Rozkład faktur VAT:** Wykres słupkowy prezentujący rozkład informacji o fakturach VAT. Widać, że większość samochodów nie jest sprzedawana z fakturą VAT.



Rysunek 11: Rozkład faktur VAT

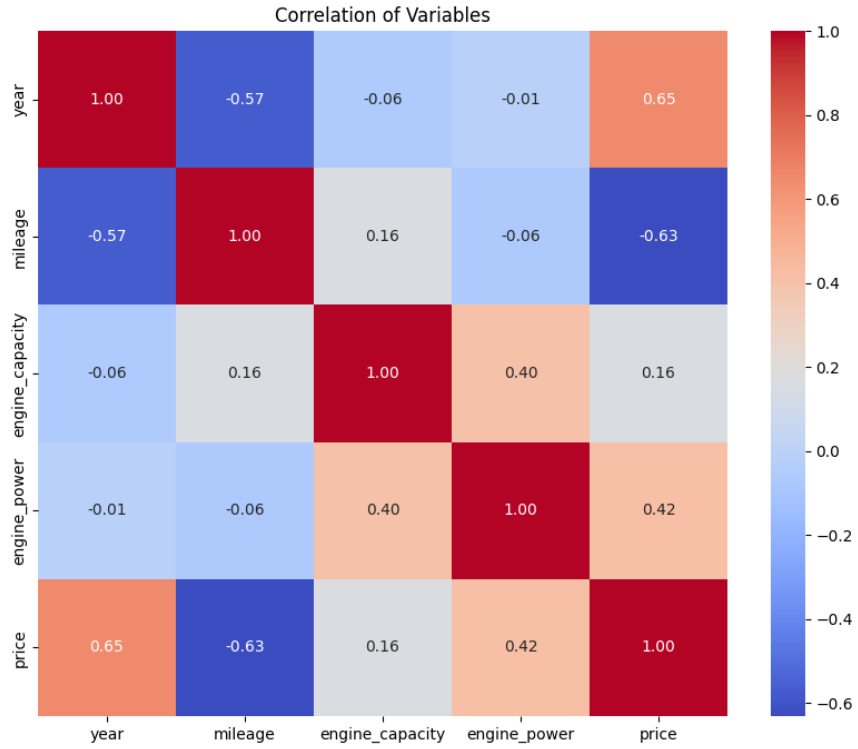
- **Rozkład typów nadwozia:** Wykres słupkowy ukazujący rozkład typów nadwozia (sedan, estate). Widać dominacją nadwozia typu Kombi (Estate) nad Sedanem.



Rysunek 12: Rozkład typów nadwozia

5 Analiza korelacji zmiennych z ceną samochodu

- **Macierz korelacji zmiennych:** Macierz korelacji prezentująca wzajemne zależności pomiędzy zmiennymi numerycznymi (rok, przebieg, pojemność silnika, moc silnika, cena).



Rysunek 13: Macierz korelacji zmiennych

- **Rok produkcji:** Współczynnik korelacji między ceną samochodu a rokiem produkcji wynosi 0.65, co wskazuje na silną dodatnią korelację. Oznacza to, że nowsze modele Passata B8 są generalnie droższe, co jest zgodne z oczekiwaniami.
- **Przebieg:** Korelacja między ceną a przebiegiem samochodu wynosi -0.63, co wskazuje na silną ujemną korelację. Wyższy przebieg samochodu wiąże się z niższą ceną, co jest logiczne, ponieważ większy przebieg często oznacza większe zużycie pojazdu.
- **Pojemność silnika:** Korelacja między pojemnością silnika a ceną wynosi 0.16. Jest to słaba dodatnia korelacja, co sugeruje, że większa pojemność silnika może nieznacznie wpływać na wzrost ceny.

- **Moc silnika:** Korelacja między mocą silnika a ceną wynosi 0.42, co oznacza umiarkowaną dodatnią korelację. Samochody o większej mocy silnika są zazwyczaj droższe.

Podsumowując, cena samochodów Volkswagen Passat B8 jest najbardziej zależna od jego rocznika i przebiegu, podczas gdy moc silnika ma umiarkowany wpływ na wartość pojazdu.

6 Implementacja Modeli Predykcyjnych

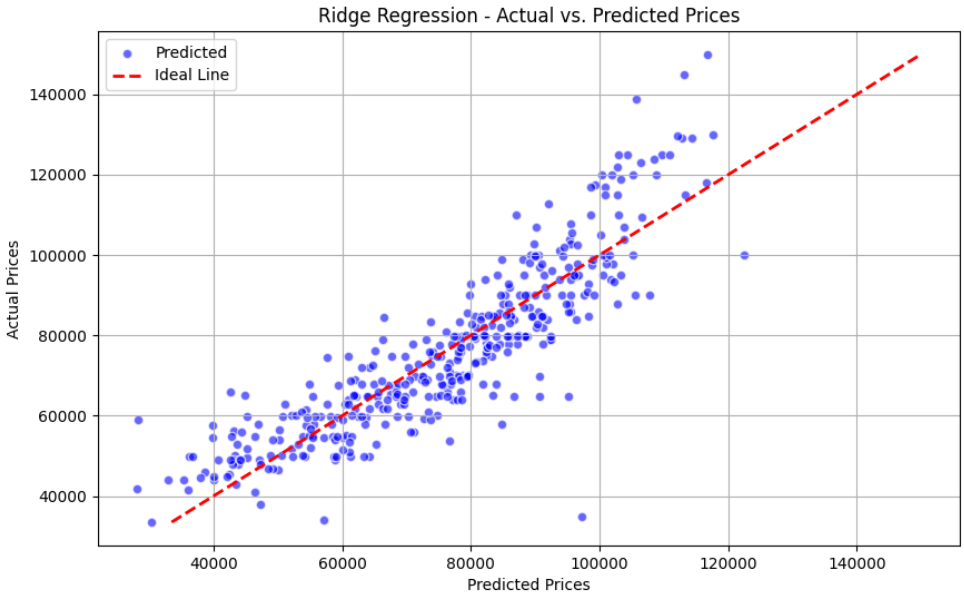
6.1 Regresja Liniowa i Grzbietowa

Regresja Liniowa: Regresja liniowa jest jednym z najprostszych i najczęściej stosowanych modeli regresyjnych, używanym do określenia liniowej zależności pomiędzy zmiennymi niezależnymi a zmienną zależną, w tym przypadku ceną samochodu.

Rysunek 14: Regresja Liniowa

Regresja Grzbietowa (Ridge Regression): Regresja grzbietowa jest techniką regresji liniowej, która dodaje karę do sumy kwadratów współczynników re-

gresji, co pomaga w redukcji nadmiernego dopasowania modelu do danych treningowych.



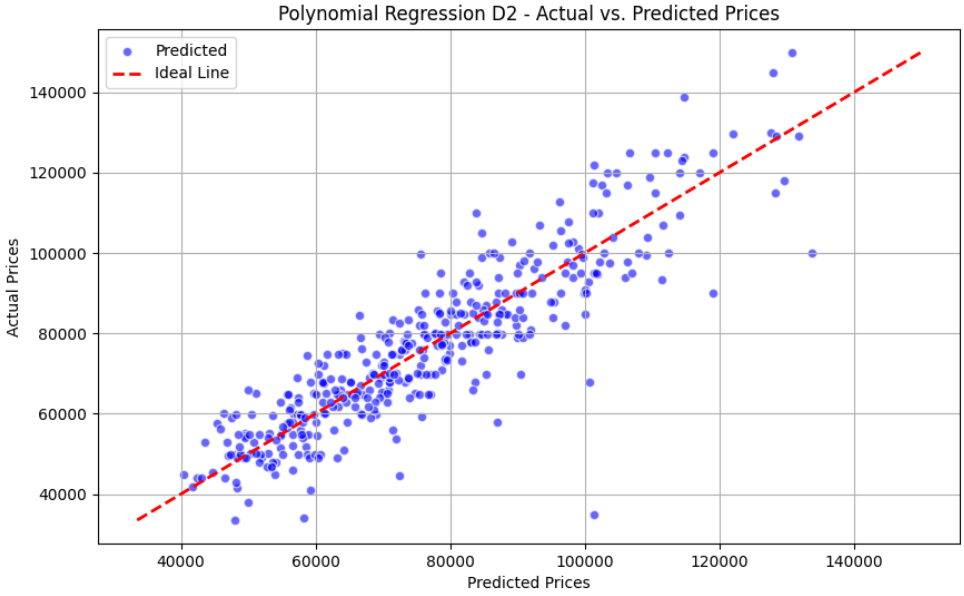
Rysunek 15: Regresja Grzbietowa

Model	**Metryka**	**Wartość**
Regresja Liniowa	MSE Train	135,012,103.46
	R2 Train	0.7349
	MSE Test	99,822,916.19
	R2 Test	0.7689
Regresja Grzbietowa	MSE Train	135,027,313.87
	R2 Train	0.7348
	MSE Test	99,740,396.74
	R2 Test	0.7691

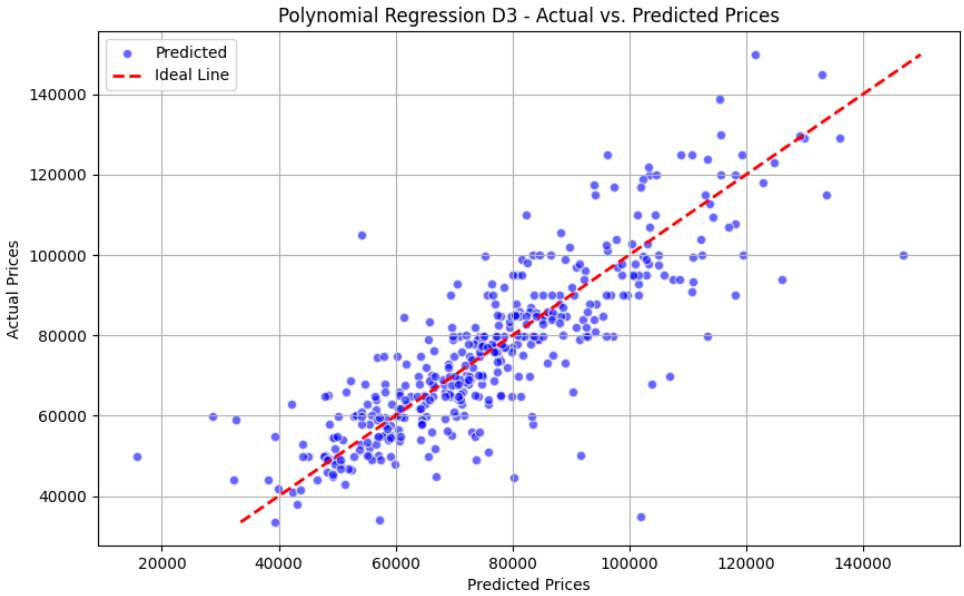
Wyniki modeli regresji liniowej i grzbietowej są bardzo zbliżone, co pokazuje, że oba modele radzą sobie w sposób porównywalny. Modele zostały użyte bez dodatkowych parametrów.

6.2 Polynomial Regression

W ramach analizy wykorzystano również modele regresji wielomianowej drugiego i trzeciego stopnia, aby sprawdzić, czy bardziej złożone relacje między zmiennymi niezależnymi a ceną samochodu poprawią wyniki predykcji.



Rysunek 16: Polynomial Regression D2



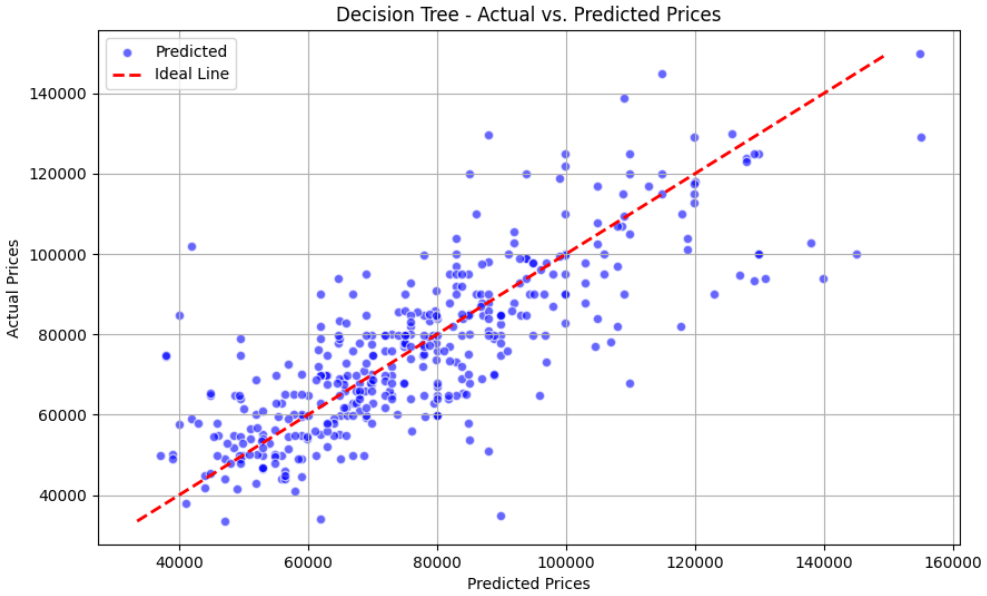
Rysunek 17: Polynomial Regression D3

Model	**Metryka**	**Wartość**
Polynomial Regression D2	MSE Train	95,736,877.48
	R2 Train	0.8120
	MSE Test	86,798,396.99
	R2 Test	0.7991
Polynomial Regression D3	MSE Train	69,109,499.96
	R2 Train	0.8643
	MSE Test	138,314,832.00
	R2 Test	0.6798

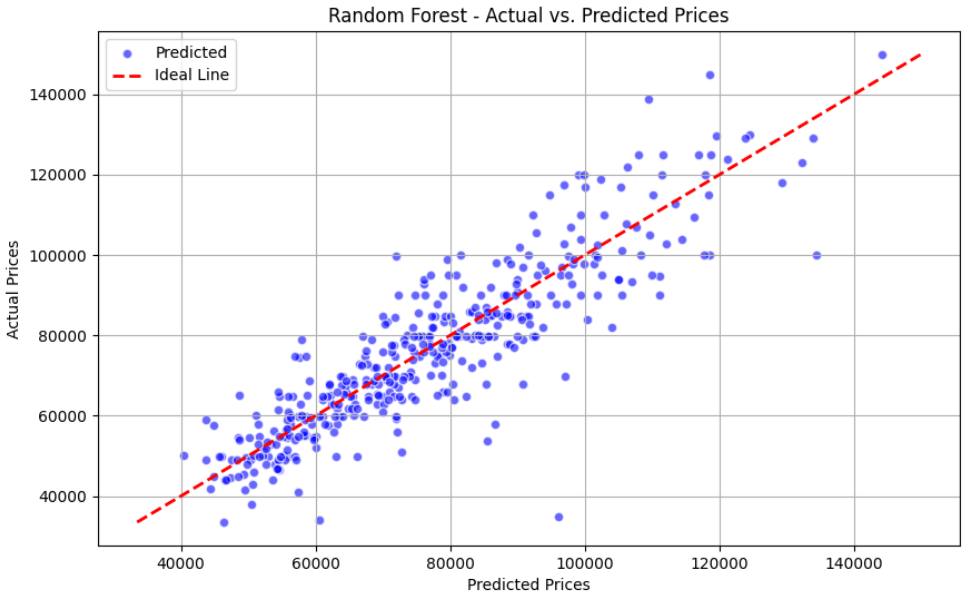
Model wielomianowy drugiego stopnia pokazuje poprawę zarówno w wynikach treningowych, jak i testowych w porównaniu do regresji liniowej i regresji grzbietowej. Model wielomianowy trzeciego stopnia pokazuje dalszą poprawę w wynikach treningowych, jednak wyniki testowe wskazują na overfitting. Testy wielomianów wyższych stopni wykazały dalszą tendencję do popadania w overfitting dla zbioru treningowego i zwracania niedokładnych wyników dla zbioru testowego. Modele zostały użyte z Ridge Regression, gdyż LinearRegression dawał fatalne wyniki (zerowy współczynnik R2).

6.3 Decision Tree i Random Forest

Drzewa decyzyjne i Random Forest to popularne modele stosowane do przewidywania zmiennych ciągłych. Drzewa decyzyjne tworzą model poprzez iteracyjne dzielenie danych na podstawie wartości cech, natomiast lasy losowe łączą wiele drzew decyzyjnych w celu uzyskania bardziej stabilnych i dokładnych predykcji.



Rysunek 18: Decision Tree



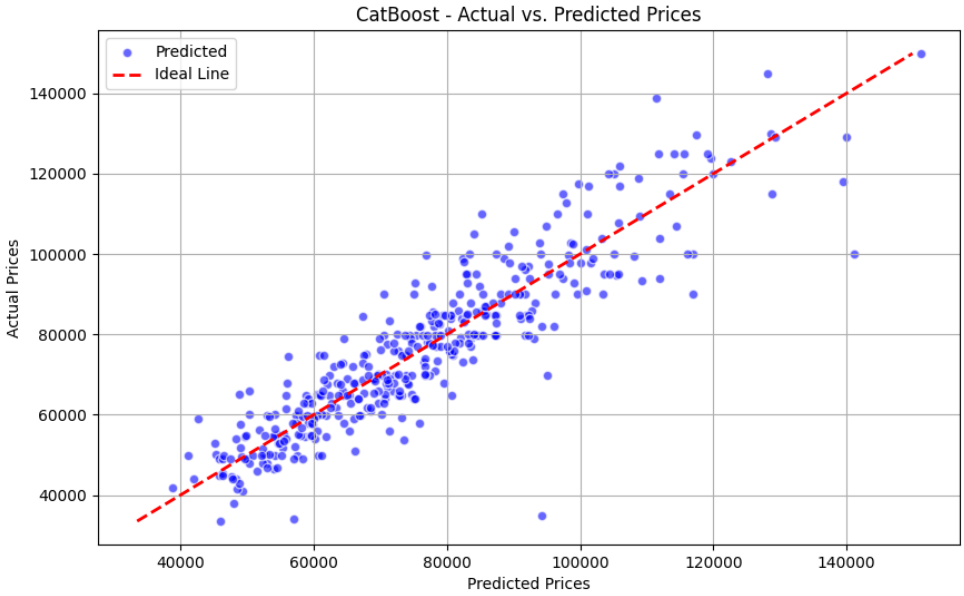
Rysunek 19: Random Forest

Model	**Metryka**	**Wartość**
Decision Tree	MSE Train	24,425.36
	R2 Train	1.0000
	MSE Test	193,259,266.72
	R2 Test	0.5526
Random Forest	MSE Train	19,226,129.47
	R2 Train	0.9622
	MSE Test	91,842,258.86
	R2 Test	0.7874

Model drzewa decyzyjnego doskonale dopasowuje się do danych treningowych, jednak jego zdolność generalizacji na danych testowych jest ograniczona. Model Random Forest osiągnął znacznie lepsze wyniki na danych testowych w porównaniu do drzewa decyzyjnego, co wskazuje na jego zdolność do lepszego uogólniania i przewidywania cen samochodów. Modele zostały użyte bez dodatkowych parametrów (poza seed).

6.4 CatBoost

CatBoost to nowoczesny i zaawansowany model gradientowego wzmocnienia drzew decyzyjnych, który został dodany do analizy dla kontekstu i porównania z innymi modelami. Ten model charakteryzuje się wysoką wydajnością i złożonością, co sprawia, że jest wyjątkowo skuteczny w różnych zadaniach regresji i klasyfikacji.



Rysunek 20: CatBoost

Metryka	**Wartość**
MSE Train	40,025,000
R2 Train	0.9214
MSE Test	77,977,000
R2 Test	0.8195

Model CatBoost osiągnął bardzo dobre wyniki zarówno na danych treningowych, jak i testowych, przewyższając inne modele pod względem dokładności predykcji. Jego zaawansowana architektura pozwala na lepsze dopasowanie do danych, co skutkuje wyższą dokładnością i mniejszym błędem predykcji. Model został użyty bez dodatkowych parametrów (poza seed).

6.5 Porównanie Modeli

Model	MSE Train	R2 Train	MSE Test	R2 Test
Regresja Liniowa	135,012,103.46	0.7349	99,822,916.19	0.7689
Regresja Grzbietowa	135,027,313.87	0.7348	99,740,396.74	0.7691
Polynomial Regression D2	95,736,877.48	0.8120	86,798,396.99	0.7991
Polynomial Regression D3	69,109,499.96	0.8643	138,314,832.00	0.6798
Decision Tree	24,425.36	1.0000	193,259,266.72	0.5526
Random Forest	19,226,129.47	0.9622	91,842,258.86	0.7874
CatBoost	40,024,891.04	0.9214	77,976,584.89	0.8195

Tabela 1: Porównanie wyników różnych modeli predykcyjnych

7 Podsumowanie i wnioski

Przeprowadzona analiza wykazała, że regresja liniowa dostarcza solidnych wyników w przewidywaniu cen samochodów Volkswagen Passat B8. Model regresji wielomianowej drugiego stopnia oferuje zadowalające rezultaty, podczas gdy model CatBoost, jako zaawansowane narzędzie, zapewnia jeszcze precyzyjniejsze prognozy.

Należy podkreślić, że dokładność uzyskanych modeli jest znaczna, szczególnie biorąc pod uwagę czynnik ludzki, który wpływa na wystawienie ogłoszenia i ustalenie ceny pojazdu. W sytuacjach, gdzie cena samochodu jest subiektywnie ustalana przez sprzedającego, uzyskane wyniki modelowania można uznać za co najmniej dobre.

Podsumowując, przeprowadzone testy zakończyły się sukcesem. Modele regresji wykazały zdolność do skutecznego przewidywania cen samochodów.