Temat: System Inteligentnej Wyceny Pojazdów Używanych z wykorzystaniem metod Machine Learning

Autorzy projektu:

1. Filip Porzucek
2. Ryszard Redelbach
3. Wiktor Maliszewski
4. Filip Makuch

1. **Wstęp i Cel Biznesowy**

Rynek samochodów używanych charakteryzuje się dużą dynamiką oraz subiektywnością wycen. Proces ręcznego ustalania wartości pojazdu jest podatny na błędy ludzkie oraz manipulacje. Celem niniejszego projektu było zaprojektowanie i zaimplementowanie systemu opartego na algorytmach uczenia maszynowego (Machine Learning), który pozwala na automatyczną i obiektywną estymację ceny rynkowej samochodu na podstawie jego parametrów technicznych.

Głównym założeniem biznesowym projektu jest stworzenie narzędzia, które mogłoby wspierać portale ogłoszeniowe (sugerowanie ceny sprzedającemu). Projekt zrealizowano w oparciu o zbiór danych Australian Vehicle Prices, zawierający pierwotnie ponad 16 000 rekordów.

2. **Inżynieria Danych (Data Engineering)**

Kluczowym wyzwaniem projektu była niska jakość surowych danych, która uniemożliwiała bezpośrednie zastosowanie modeli predykcyjnych. Proces przygotowania danych (ETL) podzielono na trzy etapy.

2.1**. Parsowanie i Standaryzacja (Parsing)**

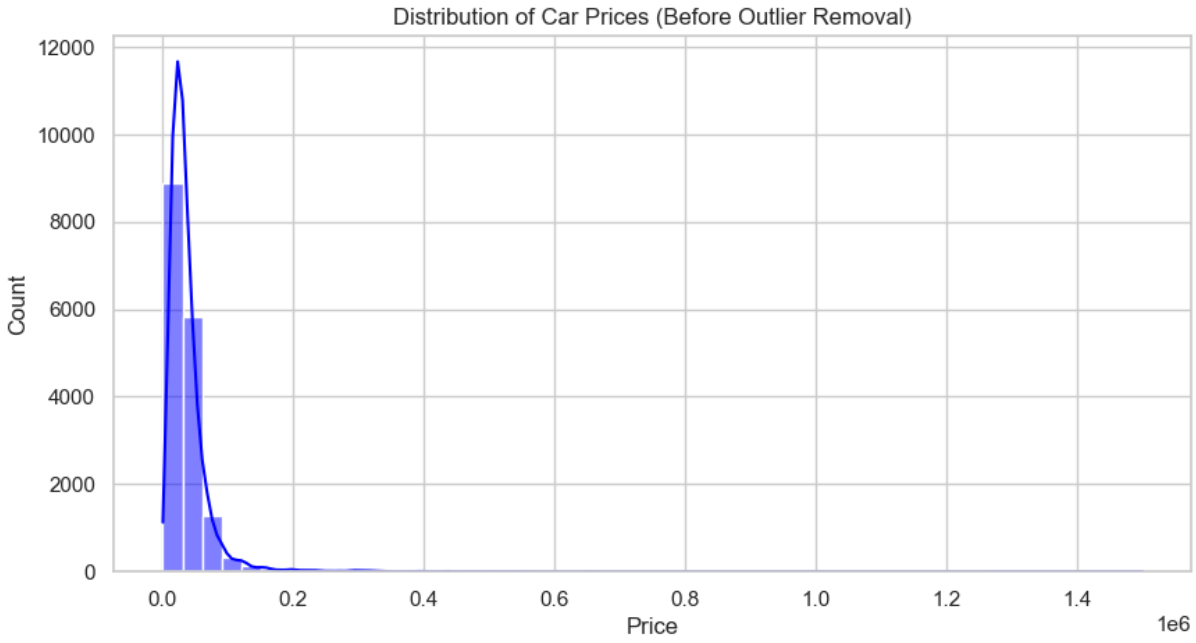
Większość kluczowych atrybutów znajdowała się w formacie tekstowym nieustrukturyzowanym. Zastosowano wyrażenia regularne (Regex) oraz dedykowane funkcje parsujące w celu ekstrakcji informacji numerycznych:

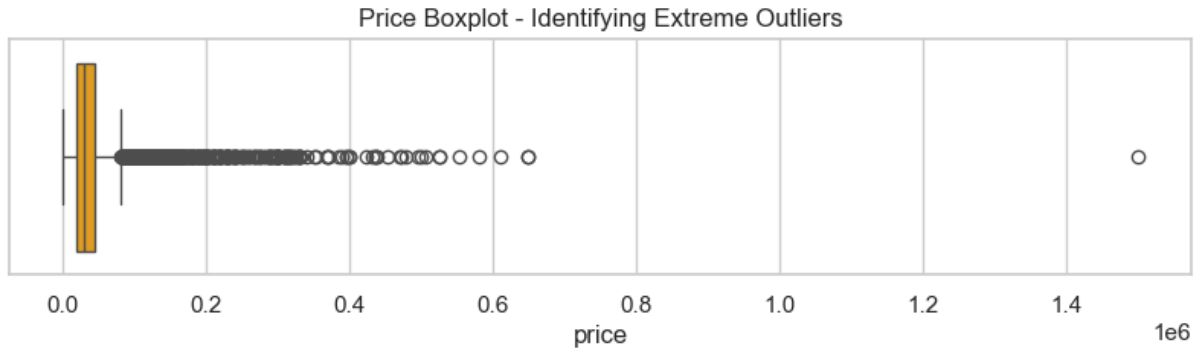
- **Pojemność silnika**: Wyodrębniono wartość w litrach (np. z ciągu "**4 cyl, 2.2 L**" uzyskano wartość numeryczną **2.2**).
- **Zużycie paliwa**: Przekonwertowano wartości tekstowe (c na

  float oraz wyciagnieta wartość numeryczna 8.7.
- Lokalizacja: Rozdzielono kolumnę adresu na miasto (City) oraz stan/region (State), co pozwoliło na lepszą generalizację geograficzną.

2.2. Analiza Exploracyjna (EDA) i Czyszczenie Danych

Przeprowadzono szczegółową analizę rozkładu zmiennej celu (Price). Wstępna wizualizacja wykazała silną prawoskośność rozkładu oraz obecność wartości odstających (outlierów), które mogłyby negatywnie wpłynąć na stabilność modelu.

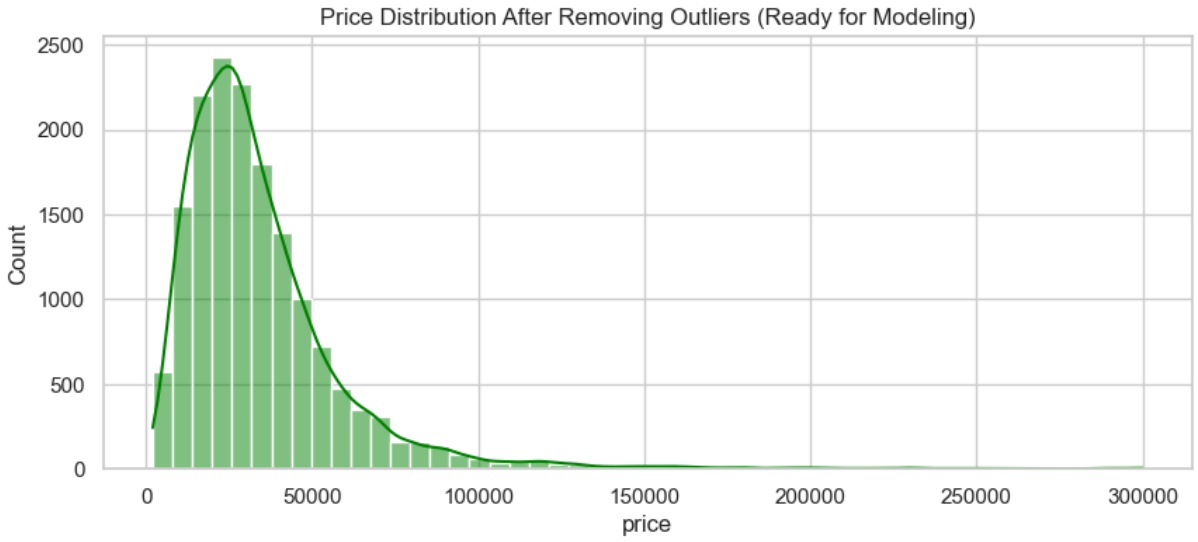




Na podstawie analizy wizualnej podjęto decyzję o zastosowaniu tzw. twardych odcięć (Hard Cutoffs). Ze zbioru treningowego usunięto:

- Pojazdy o cenie powyżej 300 000 AUD (pojazdy luksusowe i kolekcjonerskie, stanowiące margines rynku).
- Pojazdy o przebiegu powyżej 400 000 km (wysokie zaszumienie danych).
- Rekordy z błędnymi cenami (np. poniżej 2000 AUD).

Dzięki tym zabiegom uzyskano zbiór danych reprezentujący typowy rynek konsumencki, co znacząco poprawiło jakość późniejszego modelowania.



2.3. Transformacja Cech (Feature Engineering)

W celu przygotowania danych dla algorytmów uczenia maszynowego zastosowano Scikit-Learn Pipeline, co zapobiegło wyciekowi danych (Data Leakage) między zbiorem treningowym a testowym:

- **Zmienne numeryczne:** Braki danych uzupełniono medianą, a następnie zastosowano skalowanie (RobustScaler) w celu zrównoważenia wpływu zmiennych o różnych rzędach wielkości (np. Rok vs Przebieg).
- **Zmienne kategoryczne:** Zastosowano kodowanie gorąco-jedynkowe (One-Hot Encoding) dla marek, typów paliwa i nadwozia, traktując braki danych jako osobną kategorię informacyjną.
- **Zmienna celu:** Zastosowano transformację logarytmiczną ceny (LogPrice), co pozwoliło na zbliżenie rozkładu zmiennej celu do rozkładu normalnego.

3. Metodyka Badań i Modelowanie

W celu weryfikacji postawionej hipotezy oraz znalezienia optymalnego rozwiązania, przeprowadziliśmy serię eksperymentów z wykorzystaniem trzech różnych klas algorytmów: modelu liniowego (Baseline), zespołowego modelu drzewiastego (Ensemble) oraz głębokiej sieci neuronowej (Deep Learning).

3.1. Procedura Walidacji Zbiór danych podzielono na część treningową (80%) oraz testową (20%) z wykorzystaniem ziarna losowości (random_state=42) w celu zapewnienia powtarzalności wyników.

Ze względu na specyfikę zmiennej celu (ceny pojazdów), trening odbywał się na wartościach zlogarytmizowanych, co zredukowało wpływ rzędów wielkości na proces

optymalizacji wag modeli. Wyniki końcowe (predykcje) były poddawane transformacji odwrotnej (np.expm1), aby obliczyć błędy w rzeczywistej walucie (AUD).

3.2. Badane Modele

- **Model 1: Regresja Liniowa (Linear Regression)** Zastosowana jako punkt odniesienia (Baseline). Jest to model prosty, interpretowalny, ale zakłada liniową zależność między cechami a ceną, co w przypadku utraty wartości pojazdu w czasie może być niewystarczające.
- **Model 2: Random Forest Regressor (Las Losowy)** Algorytm oparty na uczeniu zespołowym (bagging), składający się z wielu drzew decyzyjnych. Jest naturalnie odporny na wartości odstające i potrafi modelować nieliniowe zależności.
    - *Optymalizacja:* Zastosowano metodę RandomizedSearchCV (60 iteracji) w celu doboru hiperparametrów.
    - *Wynik strojenia:* Najlepsze rezultaty osiągnięto dla parametrów: n_estimators=200 oraz max_depth=20. Ograniczenie głębokości drzewa pozwoliło uniknąć przeuczenia (overfittingu).
- **Model 3: Głęboka Sieć Neuronowa (PyTorch)** Zaprojektowano sieć typu Feed-Forward (MLP) z trzema warstwami ukrytymi i funkcją aktywacji ReLU. Celem eksperymentu było sprawdzenie, czy techniki Deep Learningu, skuteczne w rozpoznawaniu obrazów, poradzą sobie lepiej z danymi tabelarycznymi niż klasyczne algorytmy ML.

4. Wyniki Eksperymentu

Ewaluację modeli przeprowadzono w oparciu o dwie kluczowe metryki:

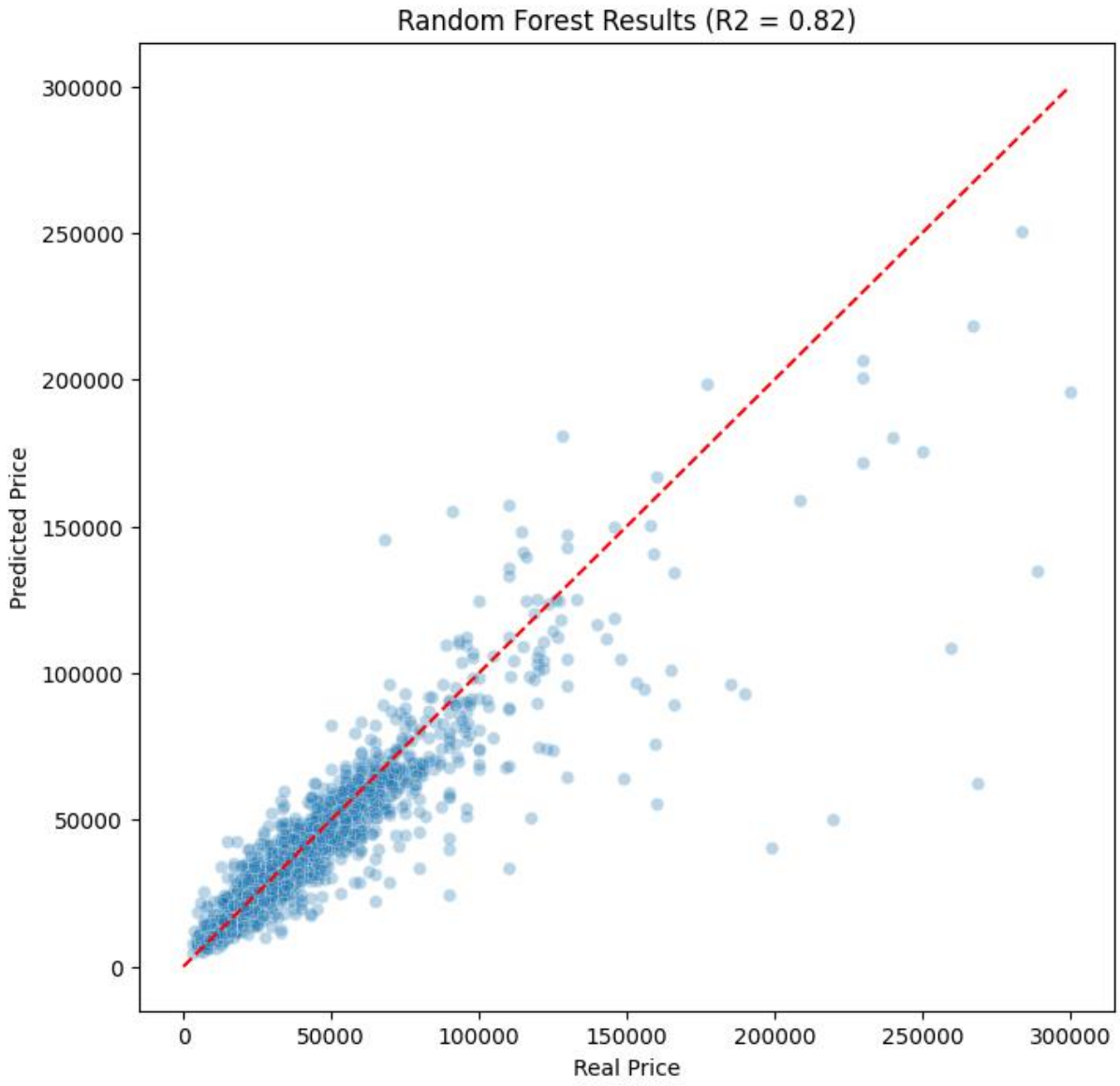
1. **R2 Score (Współczynnik determinacji):** Określa, jaki procent zmienności ceny jest wyjaśniany przez model (im bliżej 1.0, tym lepiej).
2. **MAE (Mean Absolute Error):** Średni błąd bezwzględny wyrażony w walucie (AUD), który jest łatwy do interpretacji biznesowej.

Tabela 1. Zestawienie wyników końcowych na zbiorze testowym.

Model	**R2 Score**	**MAE (Średni błąd)**	**Wnioski**
Regresja Liniowa	0.76	~6,500 AUD	Solidny wynik bazowy. Model poprawnie identyfikuje ogólne trendy,

Model	R2 Score	MAE (Średni błąd)	Wnioski
			ale traci precyzję przy autach droższych (nieliniowość).
Random Forest (Tuned)	**0.82**	**~5,180 AUD**	**Najlepszy model.** Wysoka precyzja i stabilność. Błąd na poziomie 5 tys. AUD jest akceptowalny biznesowo przy średnich cenach aut.
Sieć Neuronowa	~0.20	>10,000 AUD	Model nieefektywny dla tego zbioru danych. Rzadka macierz cech (powstała po One-Hot Encodingu) utrudniła sieci znalezienie wzorców.

4.1. Analiza Błędów i Wizualizacja Najwyższą skuteczność osiągnął model Random Forest. Jak pokazuje wykres "Predykcja vs Rzeczywistość" punkty dla tego modelu układają się najbliżej idealnej linii diagonalnej, szczególnie w segmencie aut popularnych (do 50,000 AUD).

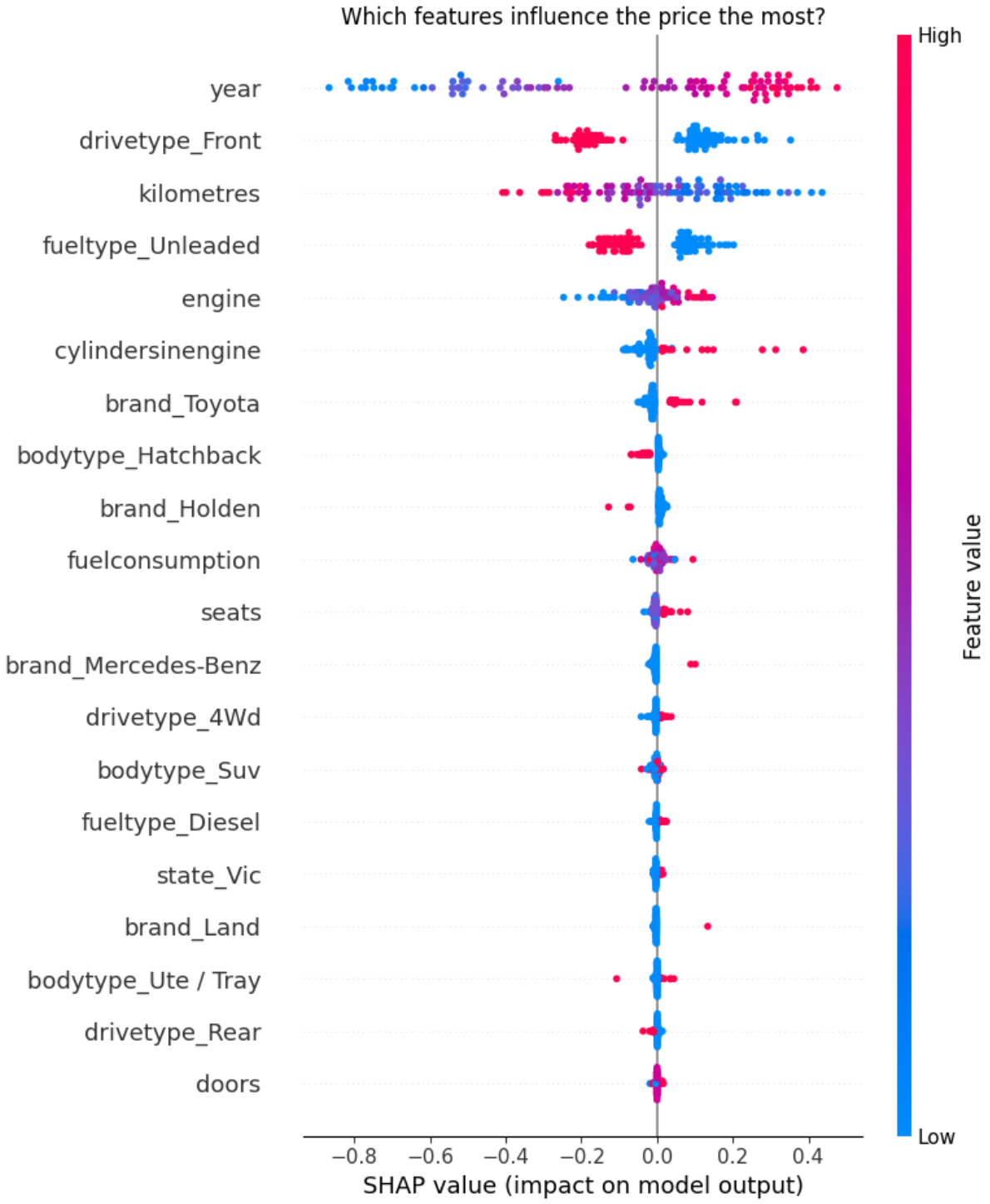


5. Wyjaśnialność Modelu (XAI)

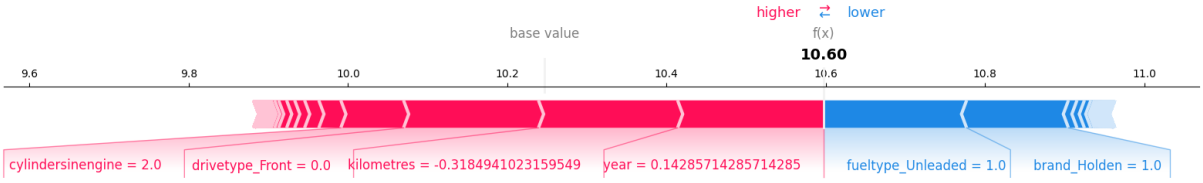
Aby zweryfikować, czy zwycięski model (Random Forest) "rozumie" logikę rynku, a nie tylko dopasowuje liczby w sposób losowy, zastosowano analizę **SHAP (SHapley Additive exPlanations)**. Pozwala ona określić wpływ poszczególnych cech na finalną wycenę.

5.1. Ranking Ważności Cech Analiza wykresu SHAP Summary Plot wykazała, że kluczowymi czynnikami cenotwórczymi są:

1. **Rok Produkcji (Year):** Najsilniejsza cecha. Nowsze roczniki drastycznie podnoszą wycenę.
2. **Przebieg (Kilometres):** Silna korelacja ujemna – wyższy przebieg obniża wartość pojazdu.
3. **Pojemność Silnika:** Większe jednostki napędowe korelują z wyższą ceną.


Which features influence the price the most?
year
drivetype_Front
kilometres
fueltype_Unleaded
engine
cylindersinengine
brand_Toyota
bodytype_Hatchback
brand_Holden
fuelconsumption
seats
brand_Mercedes-Benz
drivetype_4Wd
bodytype_Suv
fueltype_Diesel
state_Vic
brand_Land
bodytype_Ute / Tray
drivetype_Rear
doors
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Low
Feature value

5.2. Analiza Lokalna (Force Plot) Dla przykładowej predykcji model poprawnie zidentyfikował, że wysoki rok produkcji (np. 2022) oraz niski przebieg były głównymi czynnikami, które podniosły wycenę powyżej średniej rynkowej.



6. Podsumowanie i Wnioski

Projekt zakończył się sukcesem, osiągając wysoką predykcyjność (R2 > 0.82) przy użyciu algorytmu Random Forest. Przeprowadzone eksperymenty pozwoliły na sformułowanie następujących wniosków:

1. **Jakość Danych:** Kluczem do sukcesu była zaawansowana inżynieria danych (Parsing, Cleaning) oraz usunięcie wartości odstających, a nie sam wybór modelu.
2. **Dobór Algorytmu:** Dla danych tabelarycznych o, relatywnie niewielkiej objętości (~15 tys. rekordów), klasyczne metody zespołowe (Random Forest) okazały się znacznie skuteczniejsze i szybsze w trenowaniu niż Głębokie Sieci Neuronowe (R2 0.82 vs 0.20).
3. **Zastosowanie Biznesowe:** Osiągnięty średni błąd (MAE) na poziomie ok. 5000 AUD pozwala na praktyczne zastosowanie modelu jako narzędzia wspomagającego wycenę w serwisach ogłoszeniowych.