Predykcja wartości samochodów używanych przy pomocy metod uczenia maszynowego

Patryk Śliwiński

Plan prezentacji

Wstęp

Opis zbioru danych

Opis zastosowanych metod

Przeprowadzone eksperymenty

Porównanie modeli

Wstęp

Cel projektu: Opracowanie modelu, który na podstawie parametrów samochodu będzie przewidywał cenę używanego samochodu.

Zastosowanie biznesowe: Wsparcie osób sprzedających i kupujących samochody w ustaleniu atrakcyjnej ceny.

Opis zbioru danych

Cechy:

- Numeryczne: rok produkcji, przebieg , spalanie, pojemność silnika
- Kategoryczne: model, rodzaj skrzyni biegów, rodzaj silnika

Zmienna przewidywana: **cena**

Opis zbioru danych

Rozmiar zbioru: 32092 rekordy

Badane marki:

- Audi
- Skoda
- Volkswagen

Opis zastosowanych metod

1. Przetwarzanie wstępne

- Kodowanie One-Hot dla danych kategorycznych (OneHotEncoder)
- Standaryzacja danych numerycznych (StandardScaler)

Opis zastosowanych metod

2. Podział danych:

- Zbiór treningowy: 80%
- Zbiór testowy: 20%

Opis zastosowanych metod

3. Wykorzystane modele:

- RandomForestRegressor
- GradientBoostingRegresor
- MLPRegressor
- LinearRegression

Opis zastosowanych metod

4. Optymalizacja hiperparametrów:

Metoda: GridSearchCV z 5-krotną walidacją krzyżową.

Metryka: RMSE

Opis zastosowanych metod

5. Testowanie modeli:

Metryka: RMSE

Przeprowadzone eksperymenty

Dwa podejścia

1

Losowy podział

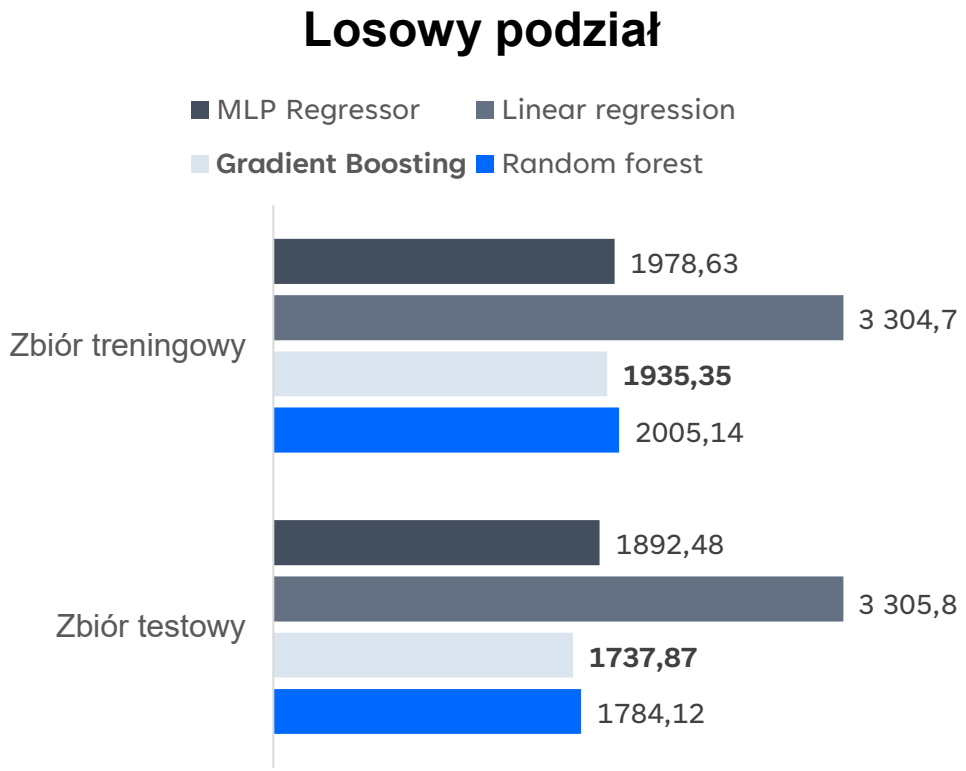
Losowy podział na dane treningowe i testowe

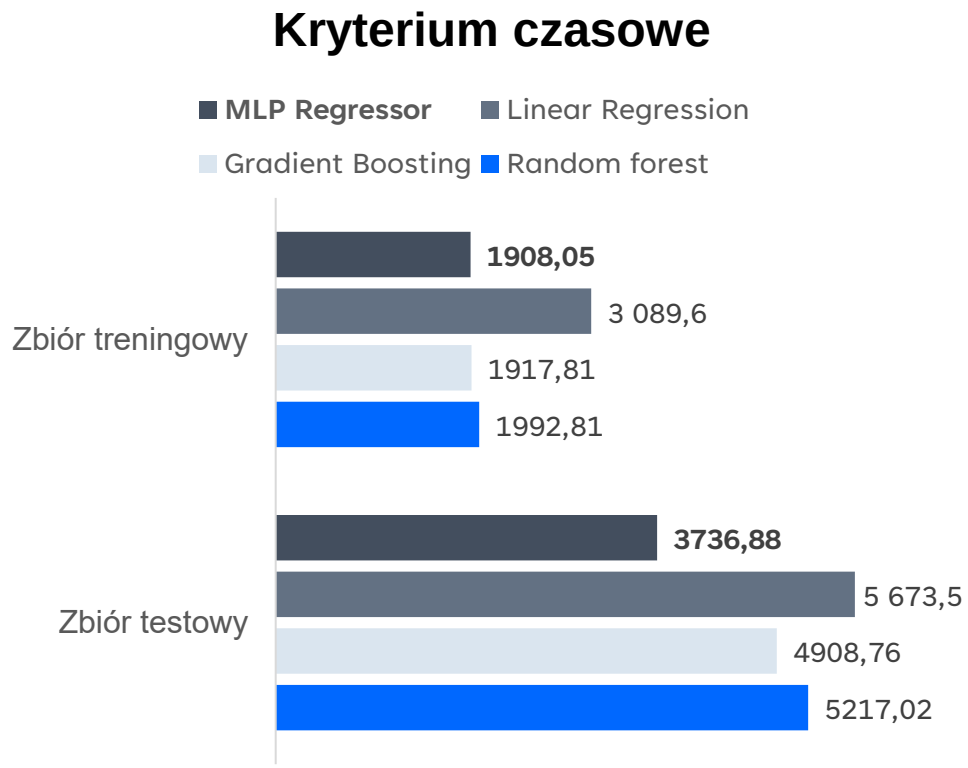
2

Kryterium czasowe

Dane treningowe obejmują pojazdy z roku 2018 i starsze.
Dane testowe dotyczą pojazdów z roku 2019 i nowszych

Ocena modeli - RMSE





Wyniki eksperymentów próba 1 – podział losowy

	RMSE w zbiorze treningowym	Najlepsze parametry	Czas treningu	RMSE w zbiorze testowym
Random Forest	2005.14	{'max_depth': 20}	16.8s	1784.12
Gradient Boosting	**1935.35**	{'learning_rate': 0.1, 'max_depth': 10}	76.1s	**1737.87**
MLP Regressor	1978.63	{'alpha': 0.01, 'hidden_layer_sizes': (50, 50, 50)}	213.9s	1892.48
Linear Regression	3304.67	Standardowe	0.3s	3305.84

Wyniki eksperymentów próba 2 – kryterium czasowe

	RMSE w zbiorze treningowym	Najlepsze parametry	Czas treningu	RMSE w zbiorze testowym
Random Forest	1992.81	{'max_depth': 20}	13.3s	5217.02
Gradient Boosting	1917.81	{'learning_rate': 0.1, 'max_depth': 10}	46.1s	4908.76
MLP Regressor	**1908.05**	{'alpha': 0.0001, 'hidden_layer_sizes': (50, 50)}	327.5s	**3736.88**
Linear Regression	3089.60	Standardowe	0.2s	5673.49

Najlepszy model

Do predykcji danych z zakresu

GradientBoostingRegresor

Do predykcji danych spoza zakresu

MLPRegressor

Dziękuje

Patryk Śliwiński