Raport z projektu

06-DUMALI0 Semestr zimowy 2024/2025

# Cel projektu

Celem projektu było stworzenie modelu, który przewiduje ceny domów na podstawie danych dotyczących cech nieruchomości i jej otoczenia. Projekt oparto na danych z konkursu **House Prices - Advanced Regression Techniques** na platformie Kaggle. Modele zostały ocenione pod kątem zdolności do przewidywania zmiennej docelowej SalePrice.

# Dane

Dane pochodzą z konkursu **House Prices - Advanced Regression Techniques** ([link do konkursu](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview)).

* **Zbiór uczący** zawierał **1460 przykładów** i **79 cech**, a **Zbiór testowy** liczył **1459 przykładów**

# Modele

1. **Regresja Grzbietowa (Ridge Regression)** – wykorzystuje regularyzację **L2 (alpha=10)**, co ogranicza wpływ mniej istotnych cech i zapobiega przeuczeniu. Trenowanie odbyło się na logarytmicznie przekształconych cenach.
2. **Gradient Boosting Regressor** – wykorzystuje **100 estymatorów** i współczynnik uczenia **0.1**, stopniowo redukując błędy predykcji poprzez iteracyjne wzmacnianie słabych modeli. Trening przeprowadzono na logarytmicznie przekształconych cenach
3. **XGBoost Regressor** – wykorzystuje **100 estymatorów** i współczynnik uczenia **0.05**, co pozwala na stopniowe dopasowywanie do danych i redukcję błędów. Trening przeprowadzono na logarytmicznie przekształconych wartościach cen.
4. **Sieć Neuronowa (TensorFlow/Keras)** – Sieć neuronowa składa się z trzech warstw ukrytych (64, 32, 16 neuronów) z aktywacją **ReLU**, regularizacją **L2**, **BatchNormalization** i **Dropout (0.2)**, a warstwa wyjściowa zwraca wartość przewidywanej ceny. Model wykorzystuje optymalizator **Adam (learning\_rate=0.02)**, funkcję straty **MSE**, oraz mechanizmy **EarlyStopping** i **ReduceLROnPlateau** dla stabilnego trenowania.

# Ewaluacja Kaggle

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSLE (dane A)** | **RMSLE (dane B)** |
| Regresja Grzbietowa (Ridge) | [0.13826](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) | [0.15472](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) |
| Gradient Boosting Regressor | [0.13252](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) | [0.14761](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) |
| XGBoost Regressor | [0.14496](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) | [0.15777](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) |
| Sieć Neuronowa (TensorFlow) | [0.13864](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) | [0.14222](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions) |

# Ewaluacja (zbiór testowy wydzielony z treningowego)

Do ewaluacji wykorzystano metryki **R² Score**, **RMSE** (Root Mean Squared Error) oraz **RMSLE** (Root Mean Squared Logarithmic Error), które dobrze radzą sobie z porównywaniem różnic w wartościach predykcji względem rzeczywistej ceny.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R²** | RMSE | **RMSLE** |
| Regresja Grzbietowa (Ridge) | 0.917342 | 25179.657608 | 0.137838 |
| Gradient Boosting Regressor | 0.904258 | 27099.362774 | 0.136689 |
| XGBoost Regressor | 0.910411 | 26214.116530 | 0.143371 |
| Sieć Neuronowa (TensorFlow) | 0.908552 | 26484.633820 | 0.144571 |

# Wnioski

1. **Regresja Grzbietowa (Ridge Regression)**: Ten model osiągnął najlepszy wynik R² (0.917342) i najniższe RMSE (25179.66), co sugeruje, że skutecznie radzi sobie z uogólnianiem. Regularizacja L2 ogranicza wpływ cech o mniejszym znaczeniu, co mogło zredukować nadmierne dopasowanie. W porównaniu do modeli boostingu, Ridge okazał się mniej wrażliwy na drobne odstępstwa w danych, co przełożyło się na bardziej stabilne wyniki.
2. **Gradient Boosting Regressor**: Osiągnął najniższy RMSLE (0.136689), co oznacza, że lepiej radzi sobie z wartościami odstającymi i różnicami w skali. Wynik R² (0.904258) oraz wyższe RMSE (27099.36) w porównaniu z Ridge sugerują jednak, że model ten bardziej skupił się na lokalnych strukturach danych, a mniej na całościowym dopasowaniu. W porównaniu z XGBoost, Gradient Boosting okazał się bardziej stabilny, co mogło wynikać z prostszego podejścia do regularyzacji.
3. **XGBoost Regressor**: Wyniki XGBoost są zbliżone do Gradient Boosting, ale wyższy RMSLE (0.143371) i niższe R² (0.910411) wskazują, że model miał większe trudności z radzeniem sobie z wartościami odstającymi. XGBoost często wymaga bardziej precyzyjnego tuningu, np. dopasowania głębokości drzew czy parametrów regularyzacji. W porównaniu do Ridge, model ten lepiej radzi sobie z bardziej złożonymi zależnościami, ale w tym przypadku mogło to prowadzić do lekkiego przeuczenia.
4. **Sieć Neuronowa (TensorFlow/Keras)**: Mimo podobnych wyników do XGBoost, najwyższy RMSLE (0.144571) oraz nieco niższy R² (0.908552) sugerują, że model miał trudności z generalizacją. W porównaniu z Ridge, sieć neuronowa była bardziej podatna na ograniczoną ilość danych, co jest typowym wyzwaniem w przypadku głębokiego uczenia. Możliwe, że lepsze wyniki można by uzyskać poprzez zastosowanie rozbudowanego preprocessingu danych lub większej liczby przykładów treningowych.