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 opisujących domy.
* **Zbiór testowy** liczył **1459 przykładów** (bez etykiety SalePrice).

# Modele

W projekcie porównano działanie czterech modeli:

1. **Regresja Grzbietowa (Ridge Regression)** – Model regularyzowany metodą L2, zapobiegający nadmiernemu dopasowaniu.
2. **Gradient Boosting Regressor** – Model wykorzystujący algorytm boosting do poprawy jakości predykcji.
3. **XGBoost Regressor** – Ulepszona wersja boostingu, zoptymalizowana pod kątem szybkości i dokładności.
4. **Sieć Neuronowa (TensorFlow/Keras)** – Model głębokiego uczenia składający się z kilku warstw neuronów.

# Ewaluacja

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. Model Regresji Grzbietowej (Ridge Regression) uzyskał najwyższy współczynnik determinacji R² (0.917342) oraz najniższą wartość błędu RMSE (25179.66), co oznacza, że dobrze dopasowuje się do danych i zapewnia stabilność predykcji. Regularizacja L2 ogranicza wpływ ekstremalnych wartości, co poprawia uogólnienie modelu.
2. Model Gradient Boosting Regressor osiągnął najlepszy wynik w metryce RMSLE (0.136689), co wskazuje na lepsze radzenie sobie z wartościami odstającymi i różnicami w skali predykcji. Jego R² (0.904258) i RMSE (27099.36) są nieco gorsze niż w przypadku Ridge Regression, co sugeruje, że model może być bardziej wrażliwy na różnorodność danych.
3. Model XGBoost Regressor uzyskał R² na poziomie 0.910411 oraz RMSE wynoszące 26214.12. Jego wynik RMSLE (0.143371) jest wyższy niż dla Gradient Boosting, co sugeruje, że model mógł mieć trudności z poprawnym odwzorowaniem niektórych wartości odstających. XGBoost może wymagać dalszego tuningu hiperparametrów, np. regularyzacji lub optymalizacji głębokości drzew.
4. Model Sieci Neuronowej (TensorFlow/Keras) uzyskał wyniki porównywalne do XGBoost, jednak jego RMSLE (0.144571) jest najwyższe spośród wszystkich modeli. Może to wynikać z niewystarczającej liczby danych treningowych lub potrzeby lepszego doboru hiperparametrów, takich jak liczba warstw czy funkcje aktywacji. Mimo to model dobrze generalizuje dane, osiągając R² na poziomie 0.908552.