Raport z projektu: przewidywanie cen ofertowych sprzedaży mieszkań

06-DUMAUI0 2021/SL

Kamil Szostak

Cel projektu

Celem projektu było stworzenie modelu, który przewiduje cenę ofertową sprzedaży mieszkania w Krakowie na podstawie powierzchni, liczby pokoi, piętra, roku budowy, odległości od centrum oraz liczby atrakcji w okolicy.

Dane

Dane pochodzą ze zbioru *Apartment Prices in Poland* (link: https://www.kaggle.com/datasets/krzysztofjamroz/apartment-prices-in-poland).

Skrypt *scraper.py* wyodrębnia dane dotyczące miasta Krakowa za listopad i grudzień 2023 oraz za styczeń i luty 2024 z odpowiednich plików w podfolderze *archive*, dokonuje normalizacji, wyboru wskazanych kolumn, usunięcia pustych wartości, usunięcia duplikatów i połączenia rekordów w jeden *dataframe*, który zapisuje do pliku dane.csv. Zdecydowałem się na wybór Krakowa, gdyż ceny sprzedaży w tym mieście są względnie stabilne (https://sonarhome.pl/statystyki-rynkowe), co umożliwiło połączenie ogłoszeń z czterech miesięcy w celu zwiększenia wielkości zbioru uczącego.

Ostatecznie uzyskano 3847 przykłady, które zarówno dla regresji, jak i dla sieci neuronowej dzielą się na zbiór uczący (3077 przykłady) i zbiór testowy (770 przykłady).

Modele

W projekcie porównano działanie 3 modeli:

- Regresji liniowej (ang. *linear regression*) – pierwsza część w pliku *regression (linear and Random Forest Tree).py*. Jako optymalizatora użyto domyślną w bibliotece *SkLearn* metodę najmniejszych kwadratów.
- Regresji losowego lasu decyzyjnego (ang. *Random Forest Tree regression*) – druga część w pliku *regression (linear and Random Forest Tree).py*. Przyjęto domyślną w bibliotece *SkLearn* liczbę drzew: 100.
- Sieci neuronowej (ang. *neural network*) – plik *neural network.py*. Stworzono trzy warstwy (w tym dwie ukryte z funkcją aktywacji ReLU po 32 neurony obie). Model używa optymalizatora *adam* oraz funkcji straty MSE. Metryką monitorowaną podczas uczenia modelu jest MAE, uczenie następuje w ciągu 1500 epok. Do implementacji skorzystano z biblioteki *Keras*.

Ewaluacja

Do ewaluacji wykorzystano metryki *pierwiastka z błędu średniokwadratowego* (*ang. root-mean-square error,* RMSE) oraz *średniego błędu bezwzględnego* (*ang. mean absolute error,* MAE). Wyniki ewaluacji przedstawia poniższa tabelka:

Model	RMSE	MAE
Regresja liniowa	207 778,63 zł	142 173,87 zł
Regresja losowego lasu decyzyjnego	153 476,51 zł	91 117,76 zł
Sieć neuronowa	190 817,19 zł	125 959,22 zł

Wnioski

Najlepsze wyniki, zarówno pod względem RMSE, jak i MAE uzyskano przy pomocy regresji losowego lasu decyzyjnego. Gorsze metryki uzyskano dla sieci neuronowej, a najgorsze dla modelu regresji liniowej.