

WARSZTATY MINI PW

Projekt OLX nr 2



Plan prezentacji

O1

EDA oraz wybrane metodologie

O2

.....

Otrzymane wyniki

O3

Wnioski



EDA

Analiza Eksploracyjna Danych

- Rozbicie kolumn tekstowych tj. params, media_types itp.
- Wykorzystanie danych tylko z rynku wtórnego.
- Zastąpienie pustych napisów brakami danych.
- Sprawdzenie i poprawienie wartości odstających.
- Sprawdzenie i poprawienie błędów we wpisanych informacjach.
- Usunięcie kolumn prawie w całości wybrakowanych.
- Zamiana dat na timestamp.
- Zastąpienie braków danych w kolumnach kategorycznych najczęściej występującą wartością.
- Zastąpienie braków danych w kolumnach numerycznych średnią.
- Zbadanie korelacji.
- Standaryzacja danych do modeli liniowych.
- Skorzystanie z pakietu zawierającego ogólnodostępne mapy, aby znaleźć wszystkie ważne punkty w mieście takie jak parki, szkoły, przystanki, centra handlowe i markety. Na tej podstawie obliczono odległość w metrach do najbliższych punktów.
- Analiza wykresów zawierających informacji o dacie wystawienia ogłoszenia oraz rozmieszczenia ogłoszeń na mapie miasta z podziałem na ceny.

Wybrane metody:

- Lasso
- Ridge
- Regresja liniowa
- Lasy losowe
- XGBoost
- AutoGluon



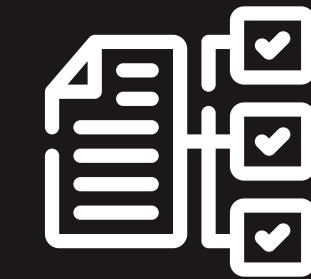
Użyto również standardowe
pakietы do przetwarzania danych.

Użyte pakiety:

- scikit-learn
- folium
- branca
- osmnx
- geopy
- autogluon
- scikit-optimize
- xgboost
- transformers



Metody znajdują się w
osobnych plikach.



Instalacja potrzebnych pakietów
znajduje się w notatnikach.

WYNIKI



	Model bazowy	Model bazowy z danymi geograficznymi	Model bazowy z cechami tekstowymi	Model bazowy z cechami tekstowymi BERT	Model pełny
Lasso i regresja liniowa	1007.51	1008.82			
Ridge i regresja liniowa	1007.51	1008.82			
RandomForest	723.92	757.65	815.83	987.73	824.52
XGBoost	720.15	736.08	675.06	750.66	670.57
Autogluon			LightGBMXT 271.76		WeightedEnsemble 798.76

Tabela: Wyniki **RMSE** dla wszystkich utworzonych modeli.

Wnioski

- Najniższy błąd RMSE (670.57) dla metody ręcznej osiągnięto dla modelu XGBoost z pełnym zestawem cech (bazowy + geograficzne + tekstowe). Wskazuje to, że im bardziej złożone i zróżnicowane cechy są uwzględnione w modelu, tym lepsze są jego wyniki.
- Modele liniowe mogą mieć ograniczenia w uchwyceniu nieliniowych zależności w danych nieruchomości.
- Uwzględnienie cech tekstowych może poprawić jakość predykcji, zwłaszcza dla modeli nieliniowych, takich jak XGBoost. Modele te mogą lepiej analizować szczegóły oferty, takie jak unikalne opisy czy cechy nieruchomości, co wpływa na dokładniejsze przewidywanie cen.
- Dane geograficzne jak i tekstowe poprawiają wyniki. Wymaga to jednak dokładnej analizy i wydobycia odpowiednich informacji. Na podstawie modelu z XGBoost widać, że dodanie cech geograficznych pogarsza wynik, a dodanie cech z tekstu polepsza. Jednak gdy to razem połączymy otrzymujemy najlepszy wynik dla tego modelu.
- Ze względu na złożoność danych AutoGluon radzi sobie najlepiej. Uwzględnia metody tekstowe, ale nie ma dostępu do map zawierających dodatkowe informacje o położeniu mieszkań. Jest to jednak metoda, która obliczeniowo plasuje się na ostatnim miejscu.





**Dziękujemy
za uwagę!**