Raport z przeprowadzonych działań w analizie danych nieruchomości

Wprowadzenie

Raport przedstawia analizę danych dotyczących nieruchomości na rynku nowojorskim. Dane zostały wczytane, przetworzone, a następnie wykorzystane do trenowania kilku modeli regresji w celu przewidywania cen nieruchomości. Zastosowane metody obejmowały regresję liniową, regresję wielomianową (2. i 3. stopnia), regresję Hubera oraz KNN.

Wykorzystane dane

Dane dotyczące nieruchomości zawierały następujące zmienne:

- BROKERTITLE: tytuł brokera
- TYPE: typ nieruchomości
- PRICE: cena nieruchomości (zmienna zależna)
- BEDS: liczba sypialni
- BATH: liczba łazienek
- PROPERTYSQFT: powierzchnia nieruchomości
- ADRESY: różne składowe adresów (zostały zakodowane numerycznie)
- STATE: stan
- LATITUDE, LONGITUDE: współrzędne geograficzne
- density: zagęszczenie zaludnienia w obszarze

Przetwarzanie danych

1. **Czyszczenie danych:**
    - Usunięcie wartości odstających na podstawie wartości interkwartylowych (IQR).
    - Normalizacja danych numerycznych przy użyciu standardowego skalera.
    - Zakodowanie zmiennych tekstowych na numeryczne.
2. **Analiza korelacji:**
    - Obliczono korelacje zmiennych z ceną nieruchomości i usunięto te o najmniejszej korelacji.
    - Wykonano wykresy korelacji.

Metody

3. **Regresja liniowa:**
    - Model trenowany na zmiennych `BEDS`, `BATH`, `PROPERTYSQFT`, `LATITUDE`, `LONGITUDE`, `BROKERTITLE`.
    - Wyniki:
        - Średni błąd bezwzględny (MAE): 782641.36
        - Mediana błędu absolutnego: 467250.50
        - R^2: 0.1527
4. **Regresja Hubera:**
    - Zastosowano model odporny na wartości odstające.
    - Wyniki:
        - MAE: 640458.38
        - Mediana błędu absolutnego: 265348.0
        - R^2: 0.0906
5. **Regresja wielomianowa 2. stopnia:**
    - Model złożony z cech wielomianowych drugiego stopnia.
    - Wyniki:
        - R^2: 0.2473
6. **Regresja wielomianowa 3. stopnia:**
    - Model złożony z cech wielomianowych trzeciego stopnia.
    - Wyniki:
        - R^2: 0.3168
7. **KNN (K-Nearest Neighbors):**
    - Model oparty na sąsiadach najbliższych sąsiadach.
    - Wyniki:
        - R^2: 0.2739

Porównanie modeli

- **Regresja liniowa:** Najniższe wartości R^2 (0.1527), co sugeruje najsłabsze dopasowanie modelu do danych. MAE i mediana błędu wskazują na znaczne różnice między wartościami przewidywanymi a rzeczywistymi.
- **Regresja Hubera:** Wyższy MAE niż regresja liniowa, ale niższa mediana błędu absolutnego, co wskazuje na obecność kilku dużych błędów. Najniższe R^2 (0.0906).
- **Regresja wielomianowa 2. stopnia:** Lepsze dopasowanie niż modele liniowe i Huber, R^2 wynosi 0.2473.

- **Regresja wielomianowa 3. stopnia:** Najlepsze dopasowanie ze wszystkich modeli, R^2 wynosi 0.3168.
- **KNN:** Lepsze dopasowanie niż regresja liniowa i Huber, ale gorsze niż modele wielomianowe, R^2 wynosi 0.2739.

Konkluzja

Modele regresji wielomianowej, zarówno 2., jak i 3. stopnia, osiągnęły najlepsze wyniki pod względem dopasowania do danych (najwyższe wartości R^2). Model regresji liniowej oraz Hubera wykazały najniższe wartości R^2, co sugeruje, że są najmniej efektywne w przewidywaniu cen nieruchomości. KNN osiągnął wyniki lepsze niż modele liniowe, ale gorsze niż wielomianowe. Ogólnie, najlepszym modelem do przewidywania cen nieruchomości na podstawie dostępnych danych okazała się regresja wielomianowa 3. stopnia.