

7. BHL - Zadanie Artificial Intelligence

Nowy Jork, jako centrum jednej z najludniejszych aglomeracji na świecie, stanowi świetne studium przypadku funkcjonowania ogromnych metropolii. Dane dotyczące posiadania posesji w Nowym Jorku z roku 2021 mogą umożliwić pogłębienie zrozumienia procesu kształtowania się cen nieruchomości w wielkich, przeludnionych miastach. Przeprowadź analizę dostarczonego przez nas zbioru danych i znajdź zastosowanie dla pozyskanej podczas tego procesu wiedzy.

Etapy zadania:

1. Czyszczenie danych i redukcja wymiarowości (10 pkt)

Przeanalizuj zbiór danych pod kątem rekordów wybrakowanych i korelacji wartości w kolumnach w celu zredukowania rozmiaru danych i zoptymalizowania procesu uczenia.

2. Wybór i wytrenowanie modelu uczenia maszynowego (30 pkt)

Wybierz algorytm uczenia maszynowego i za jego pomocą stwórz model, który po wytrenowaniu na dostarczonych danych będzie jak najlepiej przewidywał cenę nieruchomości w Nowym Jorku (wartość w kolumnie "FULLVAL") na podstawie innych parametrów zawartych w zbiorze danych.

3. Zrozumienie i interpretacja wyników (20 pkt)

Wyciągnij wnioski na podstawie efektywności i wyników uzyskiwanych przez model. Zastanów się, które kolumny mają największy wpływ na kształtowanie się cen nieruchomości, a także które kolumny zdają się nie wpływać na ich wartość. Postaraj się sformułować jak najdokładniejsze wnioski dotyczące procesu kształtowania się cen nieruchomości.

4. Prezentacja i zastosowanie (40 pkt)

Wymyśli i zaproponuj w jaki sposób zbudowany model oraz wiedza pozyskana podczas pracy ze zbiorem danych może zostać zastosowana do rozwiązania jakiegoś problemu obecnego w dzisiejszym świecie lub który może pojawić się w przyszłości. Forma prezentacji jest dowolna, ale oceniana będzie także jej kreatywność - może to być na przykład prototyp aplikacji czy też systemu wspomagania decyzji biznesowych.

Dodatkowe informacje i zasady oceniania:

- **Zbiór treningowy** dostępny jest [pod tym linkiem](#).
- Rozwiązanie może zostać stworzone w dowolnej technologii.
- **Wymagane jest utworzenie repozytorium w serwisie github** i przesłanie linka do niego na adres kacper.klassa@best.pw.edu.pl do godziny od rozpoczęcia kodowania.
- Wymagane jest utworzenie podstawowej **dokumentacji stworzonego projektu**.
- Dokładność modelu będzie weryfikowana na zbiorze testowym posiadanym jedynie przez Jury i udostępnionym uczestnikom tuż przed rozpoczęciem prezentacji podczas preeliminacji - ze względu na dowolność w używanej technologii to **zadaniem uczestników jest przygotowanie fragmentu kodu, który sprawdzi dokładność modelu na zbiorze testowym**.
- W celu polepszenia jakości prezentacji poza zadaniem opisanym w punkcie 4 należy przygotować prezentację multimedialną i przesłać ją na adres kacper.klassa@best.pw.edu.pl do godziny **16:00**.
- O godzinie **16:00** rozpoczynają się **preeliminacje**.
- Punkty wypisane w nawiasach będą przyznawane podczas prezentacji finałowych, do których przejdą **trzy drużyny**. Podczas preeliminacji celem drużyny jest zwięźle przedstawienie działania oraz zalet swojego rozwiązania oraz przekonanie Jury że to właśnie ono jest warte udziału w finałowych prezentacjach.
- W trakcie preeliminacji drużyny będą miały **6 min** na prezentację oraz **4 minuty** na pytania do Jury.
- Prezentacje finałowe rozpoczynają się o **18:00**.
- Czas na prezentacje finałowe wynosi **8 minut**, po których nastąpią **4 minuty** na pytania od Jury.
- Planowo zwycięzcy zostaną ogłoszeni o godzinie **20:35**.

Dostarczony zbiór danych jest fragmentem zbioru 'Property Valuation and Assessment Data' zebranego przez agencję NYC Department of Finance (DOF).