



## Zadanie dla kandydata na stanowisko Data Science

### Wstęp

Przygotowane zadanie ma na celu sprawdzenie umiejętności kandydatów aplikujących do na stanowisko Data Science w dziedzinie budowania modeli Machine Learning. Zadanie to ma przede wszystkim pokazać pewną świadomość działań podejmowanych przez kandydata w procesie analizy danych i budowania modelu statystycznego.

### Opis zadania

W pliku „**house.csv**” znajdują się dane dotyczące cen nieruchomości o danych cechach. Należy wykonać analizę danych prowadzącą do skonstruowania najlepszych modeli, które na podstawie cech danej nieruchomości dokona predykcji jej ceny, zarówno ciągłej (price) jak i binarnej (price\_bin). Powinno zostać zbudowane 2 niezależne modele do ciągłej oraz do binarnej predykcji.

Preferowanym językiem do rozwiązania tego zadania jest Python. Można korzystać z dowolnych bibliotek do analizy danych statystycznych i ML. Można również sięgać do zewnętrznych baz danych. Rozwiązanie problemu powinno być przedstawione w postaci analizy dojścia do najlepszego modelu, z opisami poszczególnych kroków.

### Dane

W pliku „house.csv” znajdują się nieruchomości z ich atrybutami:

- 'id':str - ID wpisu w bazie,
- 'date':str - data wpisu,
- 'price':float - cena
- 'price\_bin':int - binarna zmienna, gdzie '1' występuje wtedy, kiedy cena jest wyższa od 1mln
- 'bedrooms':float - ilość sypialni,
- 'bathrooms':float - ilość łazienek,
- 'sqft\_living':float - powierzchnia użytkowa,
- 'sqft\_lot':int - powierzchnia działki,
- 'floors':float - ilość pięter,
- 'waterfront':int - położenie na nabrzeżu {0,1},
- 'view':int - widok (0:4),
- 'condition':int - stan nieruchomości (1:5),
- 'grade':int - nachylenie działki (1:13),
- 'sqft\_above':int - powierzchnia poddasza,
- 'sqft\_basement':int – powierzchnia piwnicy,
- 'yr\_built':int - rok budowy
- 'yr\_renovated':int - rok renowacji
- 'zipcode':int - kod pocztowy
- 'lat': float - latitude

- 'long':float - longitude

### **Sposób prezentacji wyników**

Wyniki powinny zostać przesłane w postaci raportu wraz z elementami kodu Python i komentarzami wykonywanych czynności (np. jupyter notebook). Dodatkowo należy dodać krótki opis zrozumienia i analizy danego problemu. Także warto wzbogacić raport o przemyślenia na temat potencjalnego rozwoju danego rozwiązania lub zupełnie innych podejść (wystarczy na koniec umieścić w komentarzach).

Wyniki należy wysłać na adres: [hr@addepto.com](mailto:hr@addepto.com) w jednym ciągu mailowym, w którym była prowadzona korespondencja. Z kandydatami, którzy prześlą najlepsze rozwiązania będziemy się kontaktować w celu przeprowadzania dyskusji nad opracowaną analizą.

W razie jakichkolwiek pytań lub problemów z danymi prosimy o kontakt mailowy na powyżej podany adres.