WTUM11 - Raport 1

Aleksandra Mach, Jędrzej Wicha, Piotr Duperas, Szymon Pawłowski

March 2022

1 Przedstawienie tematu

Tematem projektu jest przewidywanie ceny domu na podstawie jego cech. Okazuje się, że na łączną kwotę do zapłaty za mieszkanie może się składać znacznie więcej czynników niż liczba sypialni czy powierzchnia lokalu. Mimo, że wydaje się to nieintuicyjne, to na cenę mogą wpływać czynniki takie jak np. wysokość sufitu w piwnicy.

Mając do dyspozycji dane o domach mieszkalnych z miasta Ames ze stanu Iowa, naszym zadaniem będzie przewidzieć cenę danego domu na podstawie jego 79 cech, miedzy innymi:

- data budowy
- powierzchnia
- rodzaj fundamentu
- wielkość basenu
- rodzaj ogrzewania domu
- jakość płotu
- lokalizacja garażu

2 Krótki opis projektu

Na początku zajmiemy się szczegółową obróbką danych - na pewno znajdą się wartości brakujące czy outliery, którymi trzeba będzie się odpowiednio zająć. Ponadto dane zawierają 79 kolumn - zaczynając od takich typowych cech domów jak powierzchnia, liczba pokoi czy okolica, a kończąc na mocno wyszukanych - np. wysokość sufitu, liczba kominków, rodzaj dachu, z których trzeba będzie wybrać te, który mają największy wpływ na wynik. Prawdopodobnie zbudujemy kilka modeli (na pewno będą wśród nich las losowy czy gradient boosting zaproponowane w temacie, ale też sprawdzimy inne, które według nas będą pasować), do których będziemy próbować dopasowywać najlepsze parametry. W celu porównania modeli będziemy sprawdzać różne miary takie jak dokładność, AUC czy krzywe ROC.

3 Podział zadań

- Wczytanie i początkowa obróbka danych – Wicha
- Przygotowanie danych do modeli – Duperas
- Zbudowanie modeli – Wszyscy
- Ewaluacja modeli – Pawłowski
- Przedstawienie wyników projektu – Mach

4 Narzędzia przetwarzania danych

Język programowania: PYTHON 3.9.
Skorzystamy z przetwarzania danych w chmurze w środowisku Google Colab.

Wykorzystane biblioteki:

- Pandas – biblioteka do obróbki danych
- Numpy – biblioteka do obróbki danych
- sklearn – biblioteka do budowania modeli regresji
- matplotlib – biblioteka do wizualizacji danych

Rozważamy skorzystanie z kilku modeli regresji:

- las losowy
- gradient boosting
- Support Vectors Regression
- regresja bayesowska