Picipolo - dokumentacja

April 9, 2022

1 Czyszczenie danych i redukcja wymiarowości

W celu wyczyszczenia danych:

- usuneliśmy zmienne identyfikujace (numer budynku, numer ulicy itp.)
- usuneliśmy zmienne skorelowane, zostawiliśmy najbardziej predykcyjne, użyliśmy korelacji spearmana i pearsona
- usunelismy zmienne, które mogliśmy uzyskać z innych zmiennych
- usuneliśmy zmienne, w których wystepuje jedna wartość oraz te które w wiekszości były brakami
- w niektórych kolumnach NaN oznaczał 0, wiec wystarczyło zrobić One Hot Encoding, aby poradzić sobie z brakami.

2 Wybór i wytrenowanie modelu uczenia maszynowego

Wykonaliśmy preprocessing:

- usuneliśmy zmienne, które były mało znaczace w naszym modelu
- imputowaliśmy rekordy, w których powierzchnia działki i powierzchnia budynku były równe 0 (KNNImputer), nie mogliśmy ich usunać, bo stanowiły 20 % ramki
- wykonaliśmy OneHot Encoding zmiennych
- zmieniliśmy koumne YEAR z przedziałów na rok
- zmieniliśmy wymiary budynków i działek na ich powierzchnie
- skalowanie zmiennych za pomoca MinMax na przedział [0,1]
- usuneliśmy ouliery
- wykonaliśmy transformacje logarytmiczna, kolumn o wysokiej skośności

- w celu oszacowania oceny wytrenowaliśmy RandomForestRegressor (ze wszystkich modeli, które testowaliśmy, ten radził sobie najlepiej na domyślnych ustawieniach), próbowaliśmy także dobrać hiperparametry, jednak te domyślne okazały sie być najlepsze

3 Zrozumienie i interpretacja wyników.

Otrzymaliśmy średni bład na poziomie 100 000, jednak model mylił sie także nawet o 22 000 000. Ramka danych nie była najlepsza do predykcji, ponieważ jedna zmienna miała udział w modelu równy 0.95. Dodatkowo ta zmienna też oznaczała cene i była skorelowana z etykieta na poziomie 1. Pozostałe zmienne nie miały dużego udziału w predykcji, przez co model mylił sie znaczaco w niektórych przypadkach. Dodatowo wiele ważnych zmiennych takich jak wymiary budynku, czy działki w wielu przypadkach były równe 0, a z pewnościa ta zmienna miałaby duży wpływ na cene.

4 Prezentacja i zastosowanie

Uważamy, że przygotowany przez nas model mógłby zostać wykorzystany w obydwu podejściach biznesowych - do oszacowania cen nieruchomości w Technopolis takiej jak Nowy Jork zarówno dla potencjalnych nabywców, jak i sprzedajacych. W naszej aplikacji użytkownik mógłby podać parametry nieruchomości, a nastepnie otrzymać jej wycene. Mógłby on także zaznaczyć miejsce na mapie, a nasza aplikacja ściagnełaby informacje na temat pobliskich nieruchomości na sprzedaż ze strony aukcyjnej, a także sprawdziłaby na podstawie pobranych parametrów, czy nieruchomości te sa dobrze wycenione. Pozwoliłoby to kupujacym na nieprzepłacanie i uczciwym sprzedajacym na wyznaczanie odpowiedniej ceny.