

Dokumentacja projektu technopol.ai

Patrycja Cieplicka, Krzysztof Łapan,
Maciej Morawski, Adam Napieralski

09.04.2022

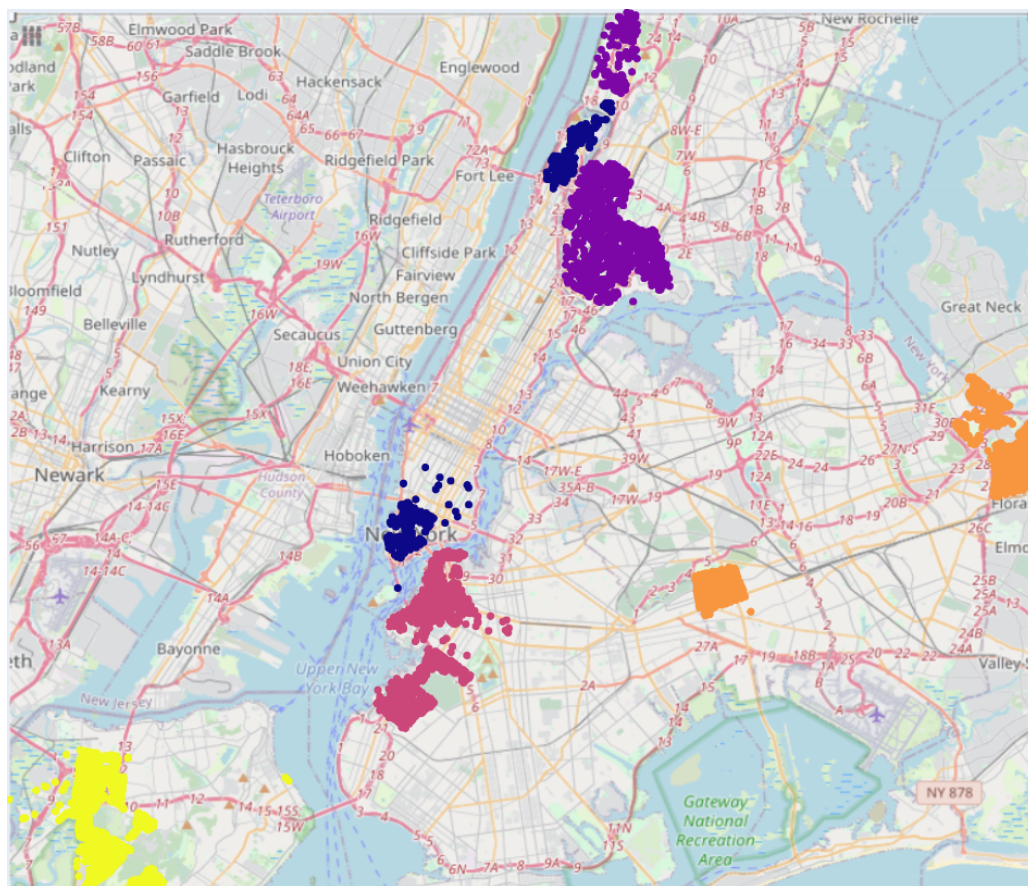
Spis treści

| | | |
|----------|-------------------------------------------------|----------|
| 1 | Analiza zbioru danych | 3 |
| 1.1 | Opis zbioru danych | 3 |
| 1.2 | Rekordy wybrakowane | 3 |
| 1.3 | Korelacja wartości kolumn | 4 |
| 1.4 | Redukcja wymiarowości danych | 4 |
| 1.5 | Dodatkowe kolumny | 4 |
| 1.6 | Podział danych na grupy | 5 |
| 2 | Opis technologii | 5 |
| 2.1 | Model uczenia maszynowego | 5 |
| 2.2 | Wizualizacja | 5 |
| 3 | Model uczenia maszynowego | 5 |
| 3.1 | Opis algorytmu | 5 |
| 3.2 | Otrzymane wyniki | 7 |
| 3.3 | Przeszukiwanie przestrzeni parametrów | 7 |
| 4 | Interpretacja wyników | 7 |

1 Analiza zbioru danych

1.1 Opis zbioru danych

Wykorzystywany zbiór danych jest fragmentem dokumentu *Property Valuation and Assessment Data*, przygotowanego przez nowojorski Department of Finance. Dokument ten dotyczący ewaluacji wartości rynkowej nieruchomości znajdujących się w tym mieście. W dostarczonym zbiorze znajdują się dane dla każdej z pięciu dzielnic miasta (*boroughs*). Plik zawiera 40 kolumn, opisujących m.in. położenie czy wymiary danej działki.



Rysunek 1: Dostępne dane na mapie Nowego Jorku

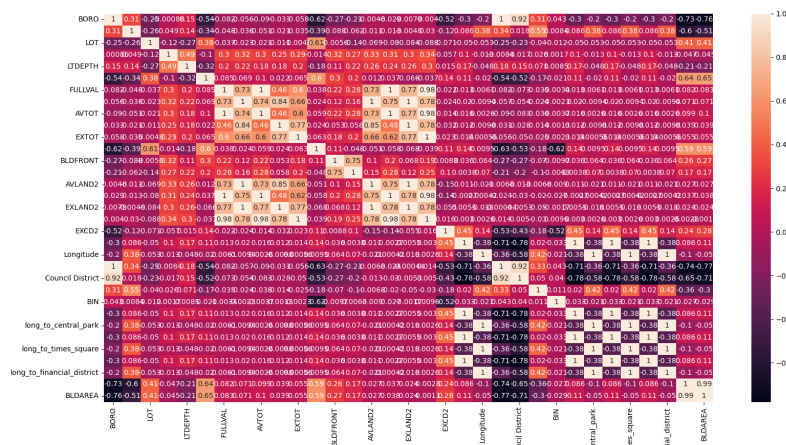
1.2 Rekordy wybrakowane

Pomimo 40 zdefiniowanych kolumn, wiele wierszy jest wybrakowanych, tj. nie posiada wartości w danej kolumnie. Ze względu na małą liczbę kompletnych wierszy, zdecydowaliśmy

się na uzupełnianie niektórych kolumn zerem lub średnią wartością dla dzielnicy.

1.3 Korelacja wartości kolumn

Aby zapoznać się z danymi oraz powiązaniem pomiędzy poszczególnymi kolumnami, obliczyliśmy macierz korelacji dla zbioru danych. Widoczna jest ona na rysunku 2.



Rysunek 2: Macierz korelacji

1.4 Redukcja wymiarowości danych

W celu uproszczenia zadania i zoptymalizowania procesu uczenia, przystąpiliśmy do redukcji wymiarowości zbioru danych. Na początek usunęliśmy kolumny, które były redundantne, a także kolumny tekstowe, które nie przydałyby się do trenowania modelu. Były to następujące kolumny:

- BBLE
- OWNER
- PERIOD
- Borough
- New Georeferenced Column

1.5 Dodatkowe kolumny

Istniejące dane wykorzystaliśmy do stworzenia nowych kolumn, które wyrażały kluczowe wartości określające charakter działek w mieście. Poniżej znajduje się lista dodanych kolumn oraz opis ich pozyskania.

- POSTCODE, Latitude i Longitude - jeśli dany wiersz posiadał dane nt. adresu działki, ale miał puste wartości w 3 wymienionych kolumnach, uzupełnialiśmy brakujące dane poprzez *geocoding* przy pomocy biblioteki Geocoder w języku Python

1.6 Podział danych na grupy

Dodatkowo podzieliliśmy kolumny ze względu na to, czy miały postać danych numerycznych (np. wielkość działki) czy kategoriycznych (np. dzielnica). Na podstawie wiedzy eksperckiej postanowiliśmy podzielić dane na grupy w zależności od TAXCLASS. Okazało się, że wartość rynkowa posiadłości o TAXCLASS równej "2", "3", "4", są liniowo skorelowane z aktualną wartością całkowitą (AVTOT). Dodatkowo obiekty należące do klasy "3" miały wartość rynkową równą 0, co jest spodziewane na podstawie szczegółowego opisu w kontekście wyznaczania wartości rynkowej nieruchomości znajdującej się na rysunku 3.

Klasa "1" zawierała bardzo wiele danych, które nie były wyraźnie skorelowane z wartością, dlatego dodatkowo w ramach tej klasy zastosowaliśmy podział w zależności od dzielnicy.

2 Opis technologii

2.1 Model uczenia maszynowego

Stworzone modele uczenia napisane zostały w języku Python. W celu wytrenowania samych modeli wykorzystane zostały powszechnie stosowane biblioteki jak Keras oraz Tensorflow.

2.2 Wizualizacja

W celu wizualizacji danych, nad którymi pracowaliśmy, opracowaliśmy prostą aplikację w języku Python, w której wyświetlane są wykresy dwu- i trójwymiarowe przy pomocy biblioteki Plotly. Biblioteka Dash umożliwiła nam natomiast stworzenie interfejsu, w którym użytkownik może wybrać interesujące go parametry i wyświetlić je na interaktywnej mapie miasta.





3 Model uczenia maszynowego

3.1 Opis algorytmu

Po preprocessingu, polegającym na usunięciu pewnych kolumn i dodaniu agregacji innych, przeszliśmy do tworzenia właściwego modelu uczenia maszynowego. Dla każdego parametru zdefiniowanego jako kategoriyczne (np. dzielnica), trenowany jest osobny model. Przy przewidywaniu wartości dla nowych danych, wybierany jest model dla odpowiedniej kombinacji danych kategoriycznych. Przy trenowaniu model sprawdzany jest za pomocą 10-krotnej krosvalidacji. Wypróbowane zostały 3 warianty regresji:

Determining Your Market Value

The Department of Finance assigns market values to all properties in New York City. Market Value is the worth of your property determined by the Department of Finance based on your property's tax class and the New York State Law requirements for determining market value.

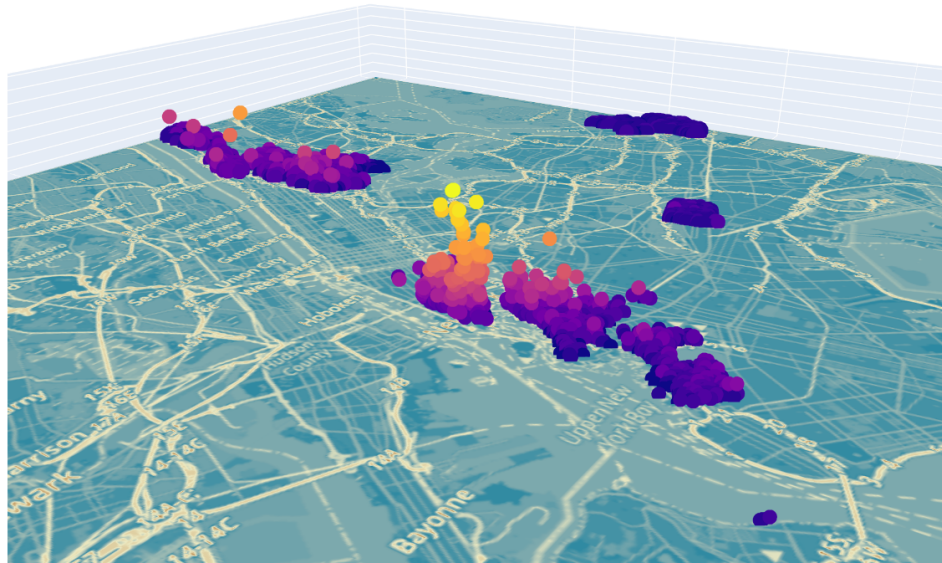
| | |
|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | One, two or three unit residential properties. The Department of Finance uses statistical modeling to analyze prices of similar properties (based on factors such as size and location) that sold in your neighborhood in the prior 3 years. |
|  | Residential property with more than 3 units including cooperatives and condominiums. NY State Law mandates that we value all class 2 properties as income producing, based on their income and expenses. We use a statistical model as a tool to find typical income and expenses for similar properties to yours (in terms of size, location, number of units and age). Then we apply a formula to the income data to get to your Market Value. The law requires that we value co-ops and condos as if they were a rental buildings, even though they are not income producing. |
|  | Utility company equipment and special franchise property - The Department of Finance uses the cost of constructing, reproducing or replacing the building added to the land value. |
|  | Most other real property, such as office buildings, factories, stores, hotels, and lofts - The Department of Finance uses your property's income earning potential and expenses. Estimated annual income is based in part on information you provide on the annual Real Property Income and Expense (RPIE) Filing . Statistical modeling is also used as a tool in this process. |

Rysunek 3: Sposób wyznaczania wartości rynkowej

- regresja liniowa
- regresja z *Gradient boosting*
- regresja z lasem losowym

Dla każdego modelu regresji obliczane są następujące miary błędu:

- średni błąd bezwzględny



Rysunek 4: Heatmapa wysokości budynków

- współczynnik determinacji R^2

3.2 Otrzymane wyniki

3.3 Przeszukiwanie przestrzeni parametrów

W celu lepszego dobrania kolumn pozytywnie wpływających na szacowanie ceny posiadłości, wykonane zostało przeszukiwanie przestrzeni parametrów:

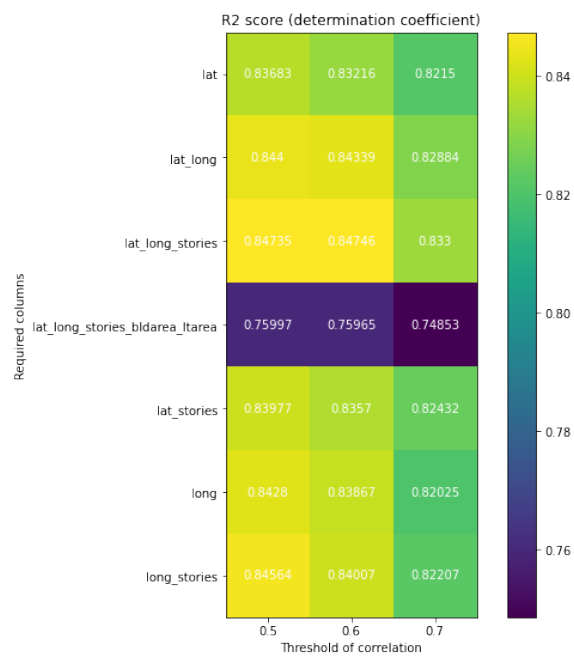
- kolumn wymaganych do uwzględnienia w każdym modelu regresji;
- próg wartości korelacji od której kolumny dobierane są dodatkowo.

W przeprowadzonej analizie najlepsze wyniki zostały uzyskane dla kombinacji kolumn wymaganych: Longitude, Latitude, STORES, z progiem równym 0.6.

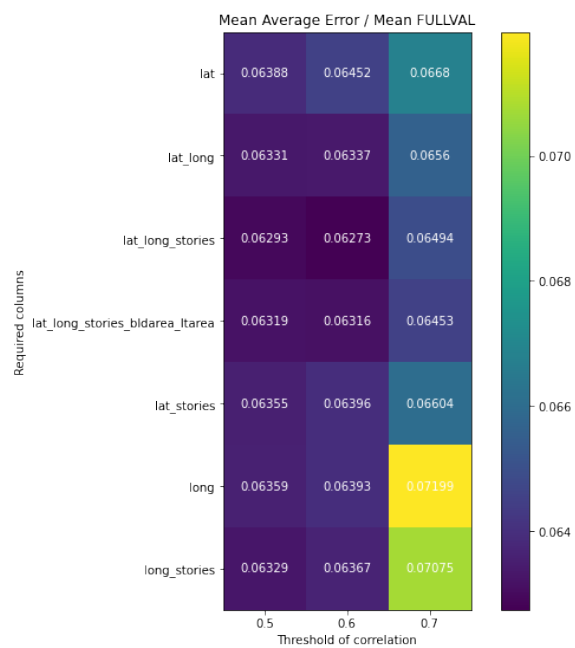
4 Interpretacja wyników

Na cenę rynkową posiadłości najbardziej i wprost wpływają ich wartości uzyskane na drodze wyceniania w celach podatkowych (kolumny *AVTOT*, *AVLAND*, *EXLAND*, *EXTOT*). Ponadto, zarówno na bazie wiedzy eksperckiej, jak i tej pozyskanej z danych, można stwierdzić, że przydzielona klasa podatkowa jest silnie związana z poszukiwaną wartością rynkową.

| | | MAE | R2 | FULLVAL | MAE / FULLVAL |
|--------------|---------------|--------|-------|---------|---------------|
| BORO1 | TAX1 | 241418 | 0,963 | 2478000 | 0,097 |
| BORO2 | TAX1 | 23331 | 0,975 | 587952 | 0,04 |
| BORO3 | TAX1 | 145189 | 0,97 | 2003044 | 0,072 |
| BORO4 | TAX1 | 16066 | 0,981 | 624182 | 0,026 |
| BORO5 | TAX1 | 8242 | 0,989 | 501691 | 0,016 |
| | TAX2 | 3876 | 0,98 | 305420 | 0,013 |
| | TAX3 | 74934 | 0,92 | 307716 | 0,244 |
| | TAX4 | 110924 | 0,93 | 1040774 | 0,107 |
| | TAX5 | 170563 | 0,948 | 1554425 | 0,110 |
| | TAX6 | 141084 | 0,949 | 962369 | 0,146 |
| | TAX7 | 1806 | 1 | 1109507 | 0,002 |
| | Ogółem | 24773 | 0,996 | 837883 | 0,022 |



Rysunek 5: Wartości miary R2 dla przeszukiwanej przestrzeni parametrów



Rysunek 6: Wartości miary MAE / FULLVAL dla przeszukiwanej przestrzeni parametrów