Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Predykcja ceny majątku w San Francisco

Sprawozdanie z laboratorium

Autor

Daniil Hardzetski

nr albumu: **252463**

kierunek: **Informatyka Stosowana**

19 czerwca 2022

Streszczenie

W pracy spróbujemy na podstawie danych majątku jak liczba pokojów, pole majątku itd predykować jego cenę przy pomocy modelu liniowego. Niestety nie otrzymano dobrych rezultatów i model liniowy nie pasuje do danych. Dataset został pobrany ze strony data.sfgov.org. Pobrane dane następnie zostały napriawione zastępowaniem wartości NaN zerami.

1 Wstęp – sformułowanie problemu

Musimy zbudować model dla urządu skarbowego do wspomogania oceniania wartości mienia, aby wyliczyć na jego podstawie ilość podatków. Model musi na podstawie wybranych cech dawać przykładowy koszt

2 Opis danych

Dane pobrane ze strony:
`https://data.sfgov.org/Housing-and-Buildings/Assessor-Historical-Secured-Property-Tax-Rolls/wv5m-`

Wielkość datasetu 2,67 miliony wierszy. Wybrane do analizy atrybuty

Zmienna	Zakres	Opis
Year Property Built	1791 - 2020	Rok budowy majątku
Number of Bathrooms	0-1002	Liczba łazienek
Number of Bedrooms	0-3800	Liczba sypialń
Number of Rooms	0-3606	Liczba pokojów
Number of Stories	0-999	Liczba piętr
Number of Units	0-4000	Liczba budynków na terenie
Property Area	0-4701100	Pole budynku
Basement Area	0-55570	Pole piwnicy
Lot Area	0-58001446	Pole terytorium

Następnie musimy pobrać sam koszt majątku, ale nie jest ono podane wprost, wartość jest rozdzielona wśród pol: Assessed Fixtures Value, Assessed Improvement Value, Assessed Land Value, Assessed Personal Property Value. „Assesed value" to pewna cześć kosztu rynkowej, która jest wykorzystywana dla liczenia podatku, zwykle jest równa 80-90Formuła dla całkowitego kosztu majątku będzie:

Assessed Value=Assessed Fixtures Value + Assessed Improvement Value
+ Assessed Land Value + Assessed Personal Property Value

3 Opis rozwiązania

Dane zostały pobrany ze strony i zrealizowane na nich profilowanie przy pomocy "pandas_profiling". Odnaleziono, że dla wybranych atrybutów jest dużo wartości null, dlatego zdecydowano na wyrzucenie wszystkich rekordów z NaN. Zdecydowanie na wyrzucenie Assessed Personal Property Value,

Assessed Fixtures Value, Basement Area i Number of bedrooms z powodu braku znaczącej części wartości atrybutów.

Potem sprawdzimy korelacje atrybutów

Dalej na danych X używamy *sklearn.preprocessing.MinMaxScaler* aby przygotować ich do nauczania i zatem przeprowadzimy uczenie modelu i popatrzymy, czy uda się zrobić sensowny model

4 Rezultaty obliczeń

4.1 Plan badań

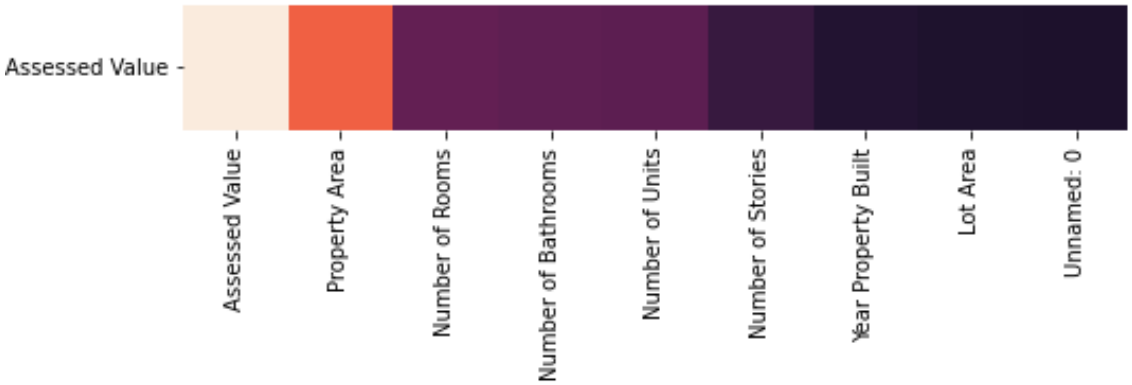
Zbiór danych zostanie podzielony na dwie części: treningową i testową w stosunku 90:10.

4.2 Wyniki obliczeń

Model oceny wartości mienia można przedstawić następującym wzorem:

Y = PropertyArea*a0 + NumberOfRooms*a1 + NumberOfBath*a2 + NumberOfUnit*a3 + NumberOfStories*a4 + a5 (1)

Następnie zbadamy jak wybrane pola korelują z wartością1, wykorzystamy dla tego pandas.corr dla wyliczenia i sns.heatmap dla reprezentacji. Dodatkowo korelacji zostały posortowane dla Assessed Value. To daje następny rezultat:



Rysunek 1: Stopnie korelacji

Możemy zobaczyć, ze samym związanym z ceną parametrem jest pole majątku, potem idzie liczba pokojów, łazienek i budynków. Liczba pięter i pole piwnicy mają małe powiazanie z ceną, a pozostałe wartości prawie niezwiązane. Dlatego do dalszej pracy zostaną wybrane wymienione atrybuty oprócz tych, które zostały wyrzucone na etapie profilowania danych.

Dla oceny modelu wykorzystujemy odnośny RSME, która wylicza się według wzoru:

$$RelRMSE = RMSE/AVG(Y_{pred}) * 100\% \quad (2)$$

4.3 Wyniki obliczeń

Wykorzystując wzór (4.2) teraz będziemy patrzeć na nasze predykcje. Niestety, stopień jakości rezulatów odpowiada stopniu korelacji. Chociaż średnia wartości testowej i predykowanej są podobne, błąd RMSE jest większy od średniej wartości o 1200%2. Wygląda, że eksperement nie udał się. Patrząc na wykres scatter możemy zobaczyć dlaczego 3. Chociaż wykres modelu i znajduje się w okresie wartości, one są zbyt rozbieżne, aby było możliwa ich predykcja.

Predykcja dla wszystkich rekordów:
AVG value: 897468.7362065997
AVG traint value: 873213.9037849721
AVG pred value: 1808530.4179682531
RMSE: 10045459.56617556
Error % from avg: 1119.0%

Jednak sprobujemy jeszcze zrobić testy na podzbiorach danych i zobaczymy, czy dla jakiegoś regionu predykcja jest lepsza.

I tak, podział danych według regionów polepszył predykcje, chociaż i nie dla wszystkich. Top trzy najlepszych predykcji:

Location: Pine Lake Park
AVG value: 533025.119992477
AVG traint value: 535508.102443609
AVG pred value: 435954.139866001
RMSE: 366167.554903196
RelRMSE: 69.0%

Location: Merced Manor
AVG value: 649360.0775988287
AVG traint value: 647568.6121951219
AVG pred value: 666816.8197166435
RMSE: 448442.1589665158
Error % from avg: 69.0%

Location: Monterey Heights
AVG value: 845737.0466536395
AVG traint value: 832940.9609279609
AVG pred value: 818438.3733129352
RMSE: 620339.622633226
Error % from avg: 73.0%

Wizualizacja danych dla Pine Lake Park przedstawiona na rysunkach z błędem 4 i wartościami 5. Możemy zobaczyć, że sukces tutaj też jest ograniczony.

Podsumowując, symulacje ujawniają że wg. (4.2) ocena wartości majątku jest bardziej złożona, niż było zapropowane tutaj. Jednak dla pewnych regionów udało sbudować odnośnie dobry projekt

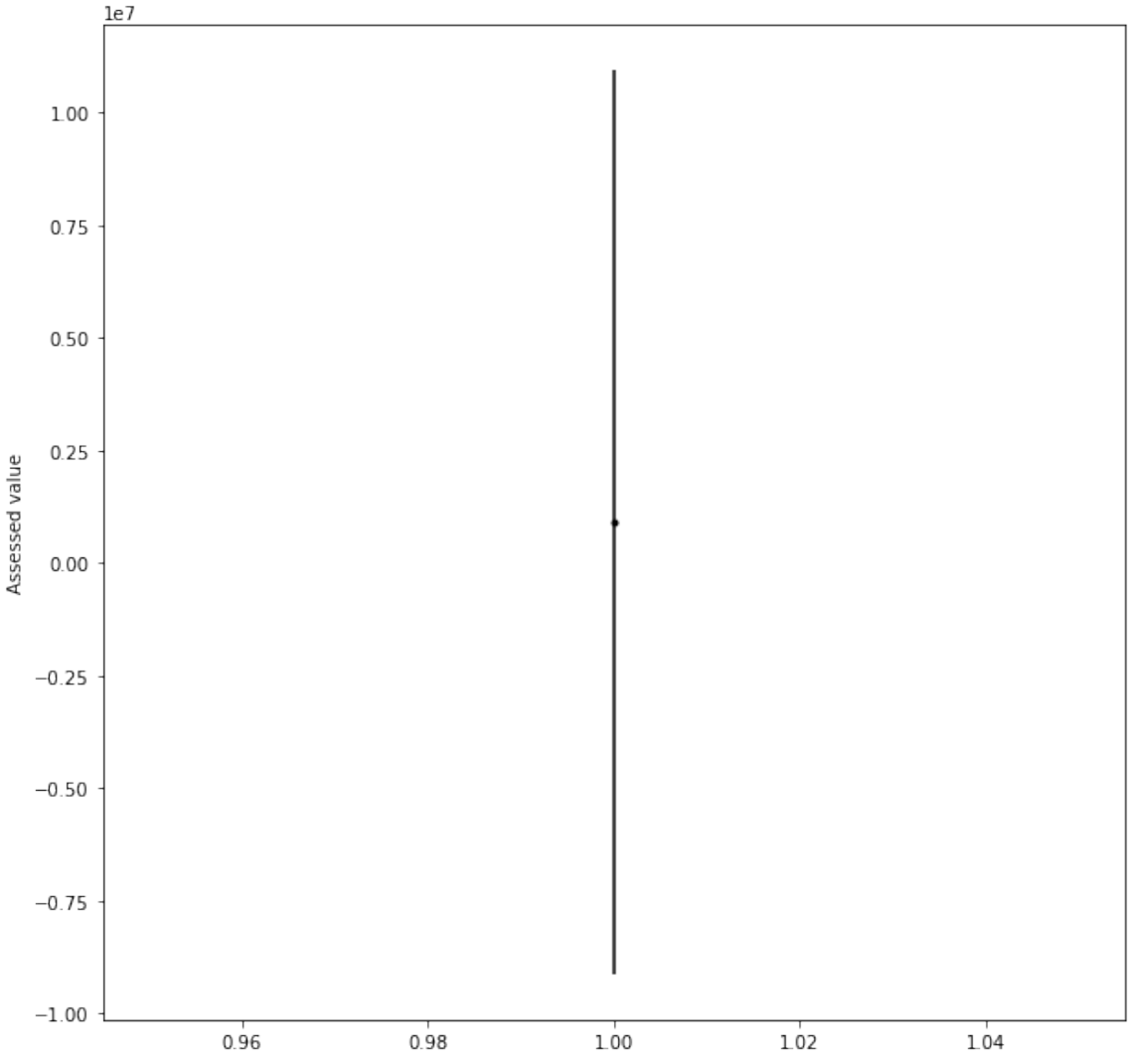
5 Wnioski

Nie udało się stworzyć modelu danych kosztu mienia przy pomocy regresji liniowej dla wszystkich danych. Dane źródłowe są zbyt rozbierzne. Lepsze rezultaty mamy na podzbiorach danych, ale one też są daleko od dobrych. Dla dobrej predykcji potrzebujemy bardziej szczególowe dane.

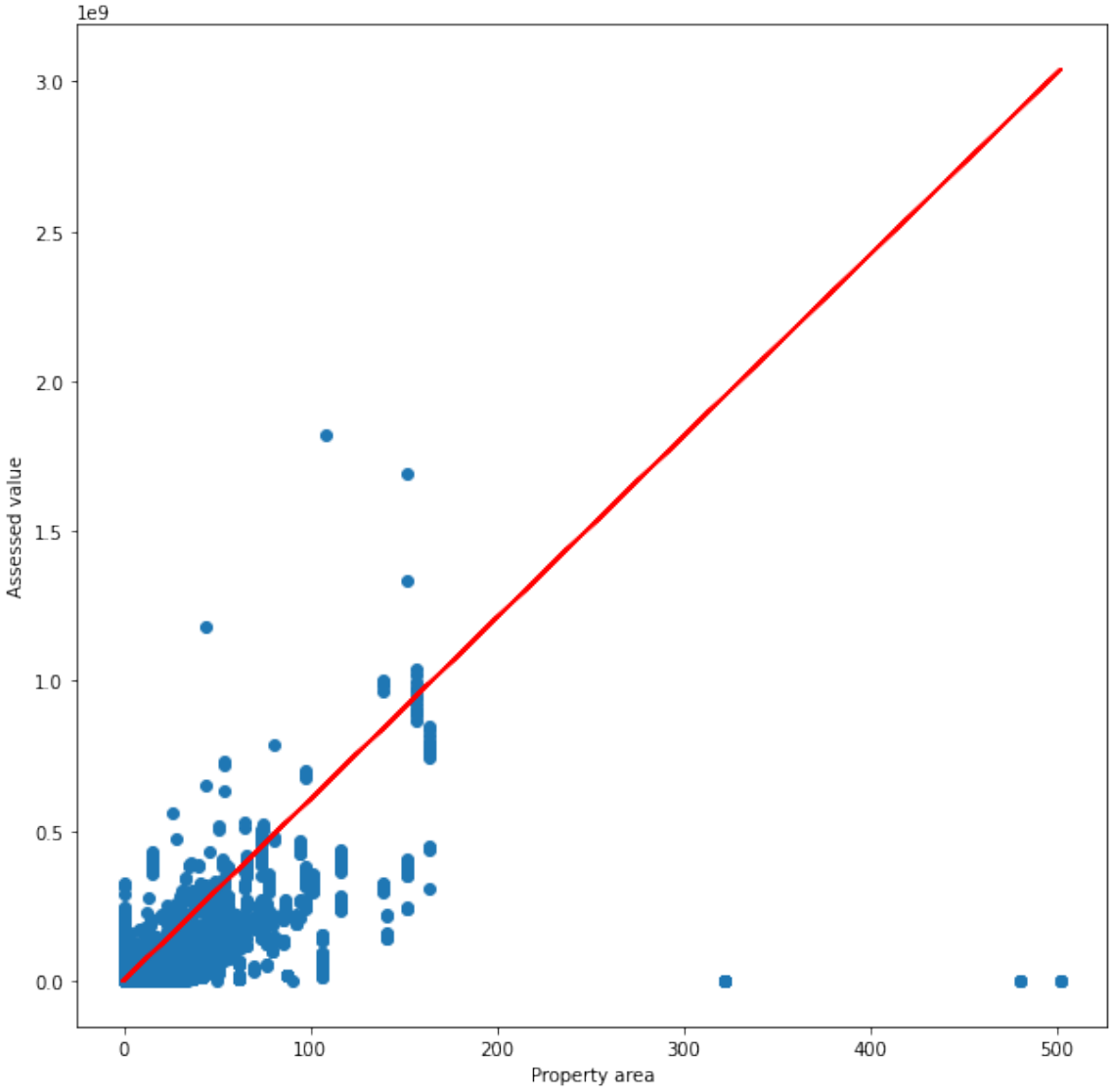
A Dodatek

Kody źródłowe(utrzymane w konwencji języka Python wraz z instrukcjmi uruchomienia) umieszczone zostały w repozytorium github:
`https://github.com/DanH4rd/MSID-Miniprojekt2`.

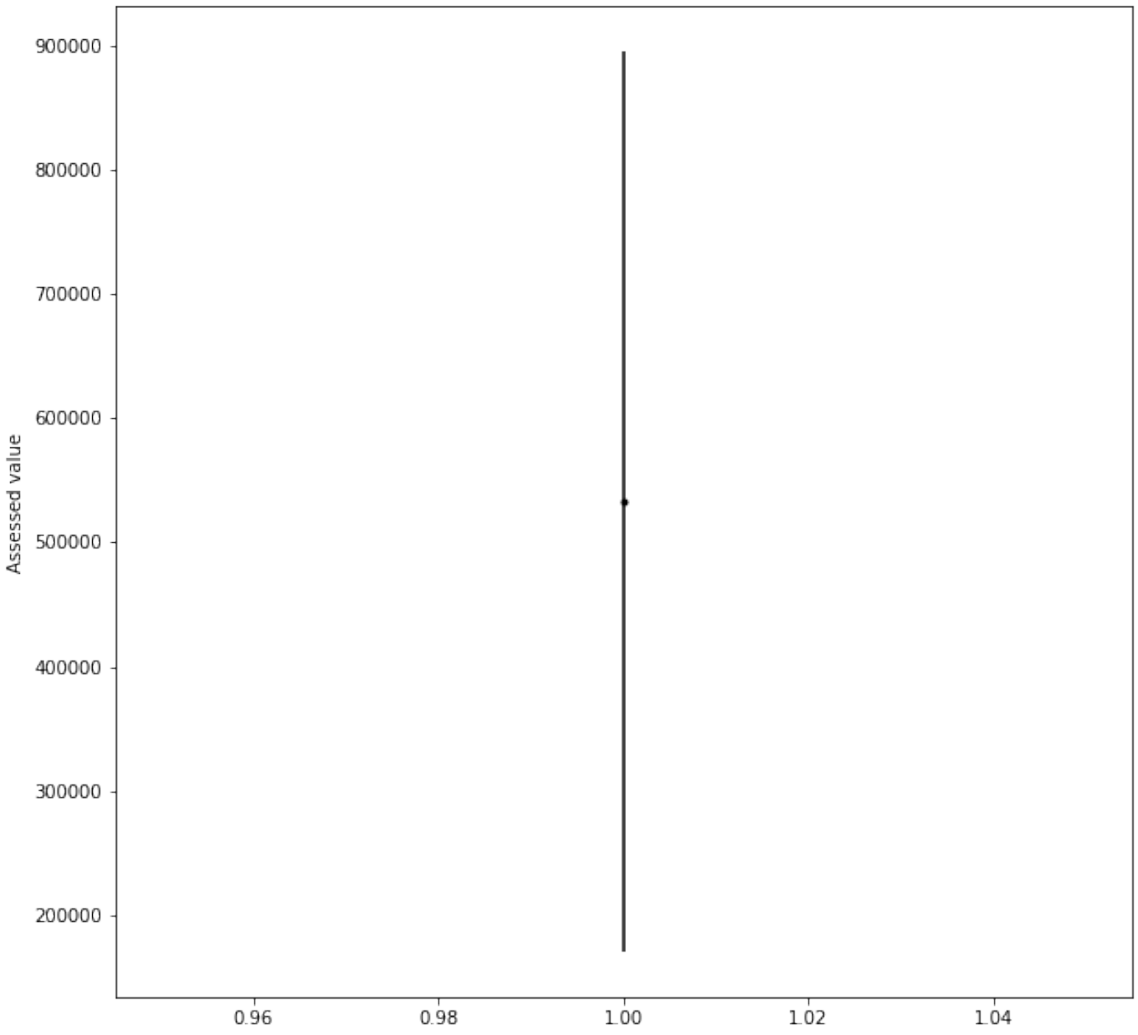
A Wykresy



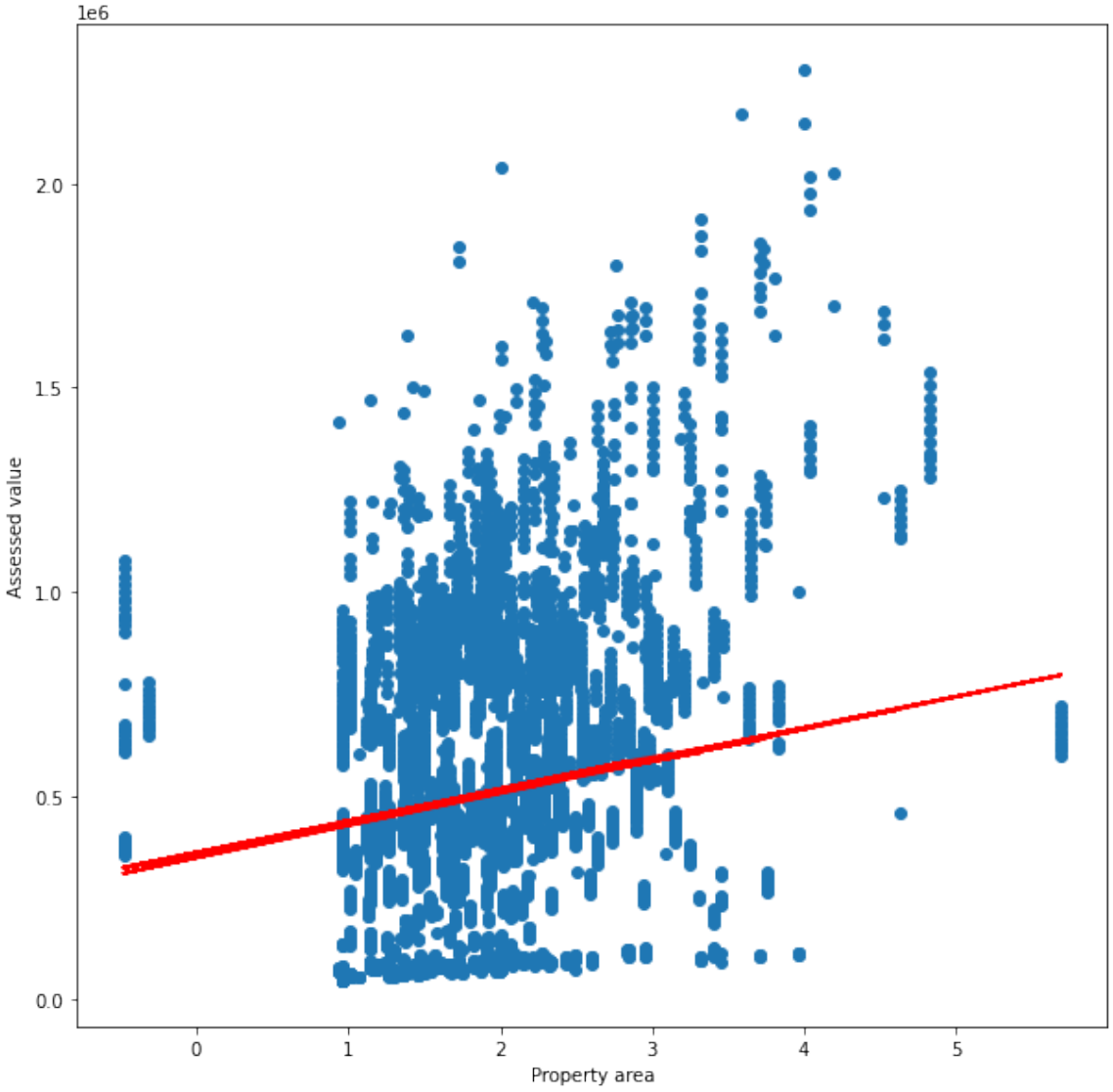
Rysunek 2: Bląd dla średniej wartości



Rysunek 3: Wykres wartości mienia od Property area i wykres modelu



Rysunek 4: Błąd dla średniej wartości najlepszego regionu



Rysunek 5: Wykres wartości mienia od Property area i wykres modelu najlepszego regionu