Predykcja cen nieruchomości

Anna Marjankowska, Weronika Sadzik

Czerwiec 2023

1 Opis zbioru danych

Zastosowano zbiór danych, który znalazł się w artykule z 1997 roku pt. "Sparse Spatial Autoregressions" autorstwa R. Pace'a, K. Kelleya i R. Barry'ego, opublikowanym w czasopiśmie Statistics and Probability Letters. Dane te zostały pozyskane z spisu ludności Kalifornii z 1990 roku i składają się z pojedynczego wiersza na grupę bloków spisowych. Grupa blokowa stanowi najmniejszą jednostkę geograficzną, dla której Biuro Spisu Ludności Stanów Zjednoczonych dostarcza próbkowe dane. Przeciętnie grupa blokowa obejmuje populację od 600 do 3000 osób.
Cechy, które posiada zbiór danych to:

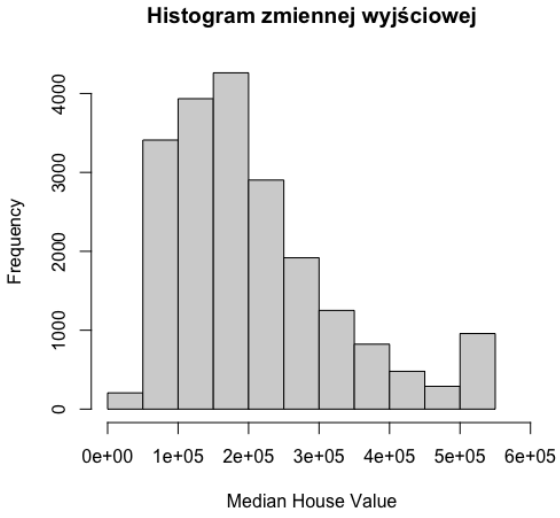
1. longitude: miara tego, jak daleko na zachód znajduje się dom; wyższa wartość oznacza dalej na zachód,
2. latitude: miara tego, jak daleko na północ znajduje się dom; wyższa wartość oznacza dom położony dalej na północ,
3. housing_median_age: średni wiek domu w bloku; niższa liczba oznacza nowszy budynek,
4. total_rooms: całkowita liczba pokoi w bloku,
5. total_bedrooms: całkowita liczba sypialni w bloku,
6. population: całkowita liczba osób mieszkających w bloku,
7. households: całkowita liczba gospodarstw domowych, grupa osób mieszkających w jednostce mieszkalnej,
8. median_income: mediana dochodu dla gospodarstw domowych w bloku domów (mierzona w dziesiątkach tysięcy dolarów amerykańskich),
9. median_house_value: mediana wartości domu dla gospodarstw domowych w bloku (mierzona w dolarach amerykańskich),
10. ocean_proximity: lokalizacja domu względem oceanu/morza.

2 Sformułowanie tematu analizy danych

Nasz projekt ma na celu przewidywanie cen mieszkań na podstawie dostępnych danych, aby dostarczyć informacje i narzędzia wspomagające podejmowanie decyzji dotyczących rynku nieruchomości.

3 Eksploracja i przygotowanie danych

Po zaimportowaniu danych do środowiska R zauważyliśmy, że mamy 20649 obserwacji. W kolumnie "total_bedrooms" stwierdziliśmy 207 brakujących wartości. Z uwagi na niewielką liczbę brakujących danych w porównaniu do wielkości próby, zdecydowaliśmy się usunąć wiersze, w których występują te braki. Dodatkowo, postanowiliśmy usunąć kolumny "latitude" i "longitude", ponieważ uważamy, że nie są one istotne dla predykcji cen mieszkań i mogą wprowadzać szum do modelu. Jedna z naszych zmiennych, "ocean_proximity", jest zmienną kategoryczną, dlatego zastosowaliśmy dla niej metodę "One-Hot Encoding" w celu przekształcenia jej w zmienną numeryczną. Następnie, aby uniknąć wpływu układu danych na proces uczenia i oceny modelu, dokonałyśmy mieszania naszych danych. W celu uzyskania powtarzalności wyników predykcji modelu ustawiłyśmy "seed" generatora liczb pseudolosowych.
Pierwszym krokiem było stworzenie histogramu zmiennej wyjściowej -
- median_house_value. Widzimy, że zmienna objaśniana ma rozkład prawostornie skośny. Najczęściej występujące wartości zmiennej znajdują się w przedziale $(100000; 200000)$

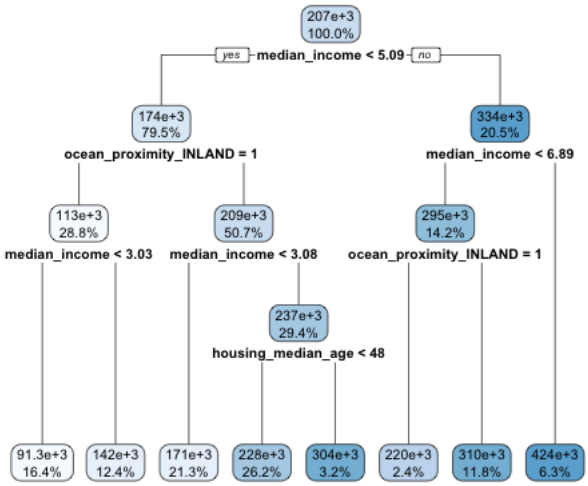


Rysunek 1: Histogram zmiennej wyjściowej

Kolejnym krokiem było wykrycie wartości odstających. Pierwszą metodą, którą zastosowałyśmy, było obliczenie punktów odstających przy użyciu metody "z-score".
Następnie postanowiłyśmy stworzyć wykresy pudełkowe (boxploty) dla każdej z cech. Kolejno przeprowadziłyśmy analizę obecności odstępów za pomocą metody rozstępu międzykwartylowego (IQR). Ponieważ w przypadku rozdzielenia danych na dane jednowiarowe liczba odstępów dla każdej cechy była stosunkowo duża w porównaniu do wielkości próbki, zdecydowałyśmy się zastosować wielowymiarową odległość Mahalanobisa jako wielowymiarowy odpowednik współczynnika "z-score", a następnie usunęłyśmy 10% obserwacji z największą odległością Mahalanobisa.
Ostatnim krokiem przygotowania danych bīło podzielenie zbioru na zbiór treningowy oraz testowy w stosunku 20% do 80%.
Skalowanie zmiennych w naszym projekcie nie jest potrzebne, ponieważ używamy wyłącznie modeli opartych na strukturach drzewiastych, które nie wymagają w takiej formie przeskalowanych zmiennych, gdyż nie wykorzystują w żadnym sensie pojęcia metryki.

4 Budowa modelu

Z pomocą biblioteki "rpart" zbudowałyśmy model drzewa regresyjnego.
Następnie zaimplementowałyśmy trzy metryki do oceny naszego modelu - Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), które kolejno zwróciły następujące wartości: $75740,56$, $55964,85$ i 5736632288.
W celu weryfikacji czy model zachowuje się lepiej od naiwnego podejścia - stałego modelowania wartości objaśnionej wartością średnią ze zmiennej wyjściowe wyznaczyłyśmy RMSE dla wartości objaśnianej w zbiorze testowym oraz średniej wartości zmiennej objaśnianej ze zbioru testowego. Otrzymałyśmy wartość $114896,7$, która jest wyższa niż RMSE z modelu drzewa regresyjnego, zatem możemy uznać, że model działa lepiej od naiwnego podejścia.
Kolejnym krokiem było wyznaczenie współczynnika korelacji między prognozami modelu a zmienną objaśnianą ze zbioru testowego. Otrzymana wartość - $0,75$ jest satysfakcjonująca.



Rysunek 2: Drzewo regresyjne

Następnie narysowałyśmy powyższe drzewo regresyjne z któego możemy odczytać strukturę modelu.
Zą pomocą biblioteki "Cubist" zbudowałyśmy drzewo modeli regresyjnych.
Następnie używając poprzednio zaimplementowanej metryki do oceny naszego modelu - Root Mean Squared Error (RMSE) otrzymałyśmy następującą wartośś: RMSE - $64095,03$.
Jak widać wartość ta jest niższa od wartości RMSE w przypadku pojedynczego drzewa regresyjnego, a zatem ulepszony model działa lepiej.
Tak jak poprzendio wyznaczyłyśmy współczynik korelacji między prognozami modelu a zmienną objaśniającą ze bioru testowego, otrzymana wartość - $0,83$ jest znacznie wyższa od współczynnika korelacji w przypadku modelu drzewa regresyjnego.
Ostatnim krokiem było wyświetlenie podstawowych statystyk prognoz.

5 Podsumowanie

Najmniejszy błąd na zbiorze testowym otrzymaliśmy dla drzewa modeli regresyjnych. Postawiony problem został rozwiązany, zbudowany przez nas model przewiduje ceny nieruchomości na satysfakcjonującym poziomie, co może być pomocne w podejmowaniu decyzji biznesowych.