AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Statystyka wielowymiarowa

Projekt

Autor:

Gabriela Bocheńska

Spis treści

1. Wprowadzenie

Bangalore, znane również jako Bengaluru, to jedno z najszybciej rozwijających się miast w Indiach, będące technologicznym i start-upowym centrum kraju. Dynamiczny rozwój miasta wiąże się z intensywnym rozwojem rynku nieruchomości, co czyni je idealnym miejscem do analizy danych dotyczących mieszkań. W niniejszym projekcie zajmiemy się analizą zbioru danych "Bangalore.csv", który zawiera informacje na temat mieszkań dostępnych na rynku nieruchomości w Bangalore.

Zbiór danych składa się z następujących zmiennych:

- **Price**: Cena mieszkania (w rupiach indyjskich)

- **Area**: Powierzchnia mieszkania (w stopach kwadratowych)
- **Location**: Lokalizacja mieszkania w Bangalore
- **Bedrooms**: Liczba sypialni
- **Resale**: Informacja, czy mieszkanie jest z rynku wtórnego (1) czy pierwotnego (0)
- **MaintenanceStaff**: Dostępność personelu konserwacyjnego (0/1)
- **Gymnasium**: Dostępność siłowni (0/1)
- **SwimmingPool**: Dostępność basenu (0/1)
- **LandscapedGardens**: Dostępność zagospodarowanych ogrodów (0/1)
- **JoggingTrack**: Dostępność bieżni (0/1)
- **RainWaterHarvesting**: Dostępność systemu zbierania deszczówki (0/1)
- **IndoorGames**: Dostępność gier wewnętrznych (0/1)
- **ShoppingMall**: Bliskość centrum handlowego (0/1)
- **Intercom**: Dostępność systemu domofonowego (0/1)
- **SportsFacility**: Dostępność obiektów sportowych (0/1)
- **ATM**: Bliskość bankomatu (0/1)
- **ClubHouse**: Dostępność domu klubowego (0/1)
- **School**: Bliskość szkoły (0/1)
- **24X7Security**: Całodobowa ochrona (0/1)
- **PowerBackup**: Dostępność zapasowego zasilania (0/1)
- **CarParking**: Dostępność parkingu (0/1)
- **StaffQuarter**: Kwatera dla personelu (0/1)
- **Cafeteria**: Dostępność kafeterii (0/1)
- **MultipurposeRoom**: Dostępność sali wielofunkcyjnej (0/1)
- **Hospital**: Bliskość szpitala (0/1)
- **WashingMachine**: Dostępność pralki (0/1)
- **Gasconnection**: Dostępność przyłącza gazowego (0/1)
- **AC**: Klimatyzacja (0/1)
- **Wifi**: Dostępność Wi-Fi (0/1)
- **Childrensplayarea**: Plac zabaw dla dzieci (0/1)
- **LiftAvailable**: Dostępność windy (0/1)
- **BED**: Liczba łóżek (0/1)
- **VaastuCompliant**: Zgodność z zasadami Vaastu (0/1)
- **Microwave**: Dostępność mikrofalówki (0/1)
- **GolfCourse**: Bliskość pola golfowego (0/1)
- **TV**: Dostępność telewizora (0/1)
- **DiningTable**: Dostępność stołu jadalnego (0/1)
- **Sofa**: Dostępność sofy (0/1)
- **Wardrobe**: Dostępność szafy (0/1)
- **Refrigerator**: Dostępność lodówki (0/1)

Celem niniejszego projektu jest zastosowanie różnych technik statystycznych i uczenia maszynowego do analizy i modelowania danych dotyczących nieruchomości w Bangalore. Projekt obejmuje następujące etapy:

1. **Regresja liniowa**: Analiza zależności między ceną mieszkania a innymi cechami, w celu stworzenia modelu predykcyjnego.
2. **Podstawowe metody klasyfikacji**: Klasyfikacja mieszkań na podstawie różnych cech, takich jak liczba sypialni, lokalizacja czy dostępność udogodnień.
3. **Walidacja krzyżowa**: Ocena dokładności i stabilności modeli przy użyciu technik walidacji krzyżowej.

4. **Regularyzacja w modelach liniowych**: Zastosowanie technik regularyzacji w celu poprawy modelu regresji liniowej i uniknięcia przeuczenia.
5. **Drzewa decyzyjne i modele pochodne**: Zastosowanie drzew decyzyjnych oraz algorytmów takich jak Random Forest i Gradient Boosting do analizy i predykcji danych.

Przed przystąpieniem do analizy danych przeprowadzono wstępną obróbkę, która polegała na usunięciu wierszy zawierających same wartości „9" w kolumnach. Wartości te powodowały problemy w analizie, stąd konieczność ich eliminacji. Dalsze kroki w projekcie pozwolą na lepsze zrozumienie czynników wpływających na ceny mieszkań w Bangalore oraz na stworzenie skutecznych modeli predykcyjnych, które mogą być użyteczne zarówno dla deweloperów, jak i potencjalnych nabywców mieszkań.

2. Regresja liniowa

Regresja liniowa to jedna z podstawowych technik statystycznych, wykorzystywana do modelowania zależności między zmienną zależną a jedną lub wieloma zmiennymi niezależnymi. W kontekście analizy danych dotyczących nieruchomości w Bangalore, regresja liniowa pozwala przewidywać ceny mieszkań na podstawie ich powierzchni (Area).

2.1. Model Regresji Liniowej

W tym projekcie zastosowano podstawowy model regresji liniowej, który przewiduje cenę mieszkania (Price) na podstawie jego powierzchni (Area). Model ten można zapisać równaniem:

$$Price = \beta_0 + \beta_1 Area + \in$$

gdzie:

- β_0 jest wyrazem wolnym (Intercept),
- β_1 jest współczynnikiem regresji (Estimate) dla zmiennej Area,
- $\in$ jest składnikiem losowym (residual).

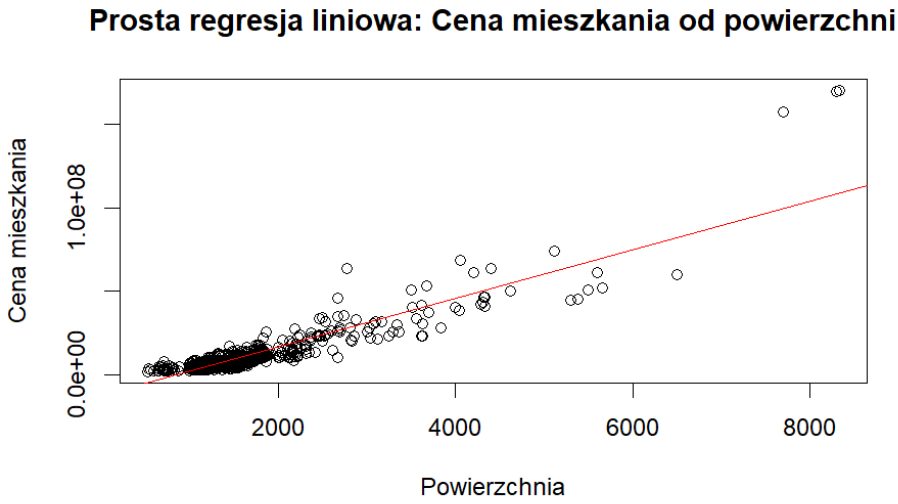
2.2. Wyniki Modelu

Po przeprowadzeniu analizy regresji liniowej, otrzymano następujące wyniki:

```
Call:
lm(formula = Price ~ Area, data = dane)

Residuals:
      Min        1Q    Median        3Q       Max
-22243755  -1888183     42220   1649342  61589065

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12324006     341585  -36.08   <2e-16 ***
Area            14564        202   72.09   <2e-16 ***
---
```

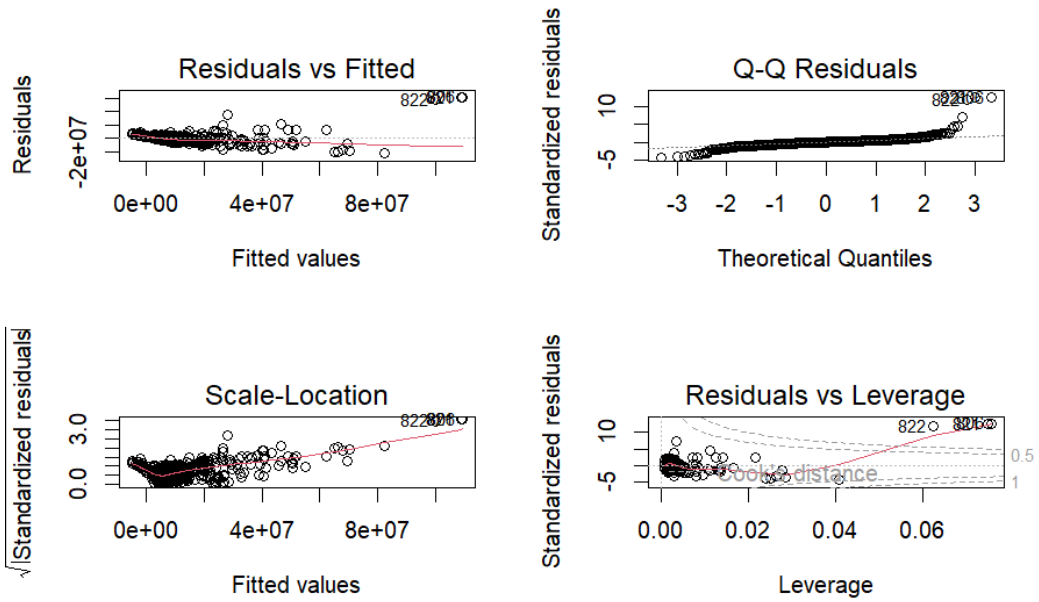
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5028000 on 1112 degrees of freedom
Multiple R-squared:  0.8237,  Adjusted R-squared:  0.8236
F-statistic:  5196 on 1 and 1112 DF,  p-value: < 2.2e-16
```

- β_0 (Intercept) wynosi -12324006, co sugeruje, że przy zerowej powierzchni (co w praktyce jest niemożliwe), model przewidywałby ujemną cenę. Wartość ta nie ma praktycznego znaczenia, ale jest częścią matematycznej formuły modelu.
- β_1 (Estimate for Area) wynosi 14564, co oznacza, że każdy dodatkowy stopień kwadratowy powierzchni mieszkania zwiększa jego cenę średnio o 14564 rupii indyjskich. Wysoka wartość współczynnika sugeruje silny wpływ powierzchni na cenę mieszkania.
- Oba współczynniki są istotne statystycznie przy poziomie istotności 0.001 (p-value < 2e-16), co oznacza, że istnieje bardzo małe prawdopodobieństwo, że uzyskane wyniki są dziełem przypadku.
- Wartość R^2 (Multiple R-squared) wynosi 0.8237, co wskazuje, że model wyjaśnia 82.37% zmienności ceny mieszkań. Jest to bardzo wysoka wartość, co sugeruje, że powierzchnia mieszkania jest bardzo silnym predyktorem jego ceny.
- Skorygowana wartość R^2 (Adjusted R-squared) wynosi 0.8236, co potwierdza, że model jest dobrze dopasowany do danych.
- Reszty (Residuals) modelu mają rozkład asymetryczny, co może sugerować, że niektóre obserwacje mają znaczne odchylenia od przewidywanych wartości. Jednakże, przy dużej liczbie danych, wpływ tych odchyleń jest minimalizowany.



Rys.1. Regresja liniowa.

Punkty leżące blisko linii sugerują, że przewidywane przez model wartości cen mieszkań są zbliżone do rzeczywistych wartości. Jednakże, istnieje kilka punktów, które znacząco odbiegają od linii regresji, co wskazuje na obecność wartości odstających.



Rys.2. Wykresy diagnostyczne dla regresji liniowej.

- Wykres reszt względem wartości dopasowanych pokazuje, że większość reszt oscyluje wokół zera, jednak widać pewne odchylenia dla większych wartości dopasowanych, co może sugerować heteroskedastyczność.
- Wykres Q-Q pokazuje, że reszty w dużej mierze podążają za linią prostą, jednak widać pewne odchylenia na końcach, co może sugerować obecność wartości odstających.
- Wykres scale-location pokazuje, że rozproszenie reszt rośnie wraz z wartością dopasowaną, co może wskazywać na niejednorodność wariancji reszt (heteroskedastyczność).
- Wykres reszt względem dźwigni (leverage) pozwala zidentyfikować obserwacje wpływowe. Wartości o wysokiej dźwigni mają znaczący wpływ na dopasowanie modelu. Kilka punktów, takich jak obserwacja numer 822, ma wysoką dźwignię i odchylenie, co może sugerować, że są to wartości odstające, które wpływają na model.

Przeprowadzona analiza regresji liniowej pokazuje, że powierzchnia mieszkania jest kluczowym czynnikiem wpływającym na jego cenę w Bangalore. Model regresji liniowej z jedną zmienną niezależną (Area) wyjaśnia znaczną część zmienności cen mieszkań, co potwierdza, że powierzchnia jest istotnym predyktorem ceny.

Jednakże analiza diagnostyczna wskazuje na pewne problemy, takie jak heteroskedastyczność oraz obecność wartości odstających. W związku z tym, w dalszych krokach analizy można rozważyć zastosowanie bardziej zaawansowanych metod regresji, takich jak regresja ważona lub regresja z transformacją zmiennych, aby poprawić dopasowanie modelu i uwzględnić zmienność reszt. Ponadto, rozszerzenie modelu o dodatkowe zmienne niezależne może również przyczynić się do lepszego przewidywania cen mieszkań.

3. Podstawowe metody klasyfikacji

Klasyfikacja jest jedną z kluczowych technik w analizie danych, której celem jest przewidywanie kategorii lub etykiet dla nowych obserwacji na podstawie modelu wyuczonego z dostępnych danych. W tym rozdziale omówimy zastosowanie różnych metod klasyfikacyjnych, takich jak regresja logistyczna, analiza dyskryminacyjna (LDA i QDA) na podstawie różnych cech nieruchomości.

3.1. Macierz korelacji:

W macierzy korelacji przedstawiono współczynniki korelacji pomiędzy różnymi zmiennymi aby zrozumieć ich zależność. Każda komórka macierzy przedstawia wartość współczynnika korelacji pomiędzy parą zmiennych. Wartości te mieszczą się w zakresie od -1 do 1, gdzie:

- 1 oznacza pełną dodatnią korelację (gdy jedna zmienna rośnie, druga również rośnie proporcjonalnie),
- 0 oznacza brak korelacji,
- -1 oznacza pełną ujemną korelację (gdy jedna zmienna rośnie, druga maleje proporcjonalnie).

Z racji wielu zmiennych (41) nie ma możliwości przedstawienia w dodgny sposób macierzy korelacji. Na jej podstawie możemy wyciągnąć kilka istotnych obserwacji dotyczących wzajemnych zależności między zmiennymi:

- **Silna dodatnia korelacja między ceną a powierzchnią (0.902)**: Oznacza to, że im większa powierzchnia nieruchomości, tym zazwyczaj wyższa cena. Jest to typowe zjawisko na rynkach nieruchomości.
- **Brak znaczącej korelacji między ceną a sprzedażą powtórną (-0.078)**: To oznacza, że cena nieruchomości nie jest istotnie związana z ich historią sprzedaży.
- **Umiarkowana dodatnia korelacja między ceną a rokiem budowy (0.426)**: Może to sugerować, że nowsze nieruchomości są zazwyczaj droższe, co może wynikać z lepszego standardu budowy lub lokalizacji.
- **Ujemna korelacja między ceną a odległością od centrum (-0.523)**: Im bliżej centrum, tym zazwyczaj wyższe ceny nieruchomości. Jest to typowe zjawisko na rynkach miejskich.

3.2. Regresja logistyczna

Regresja logistyczna jest powszechnie stosowaną metodą do binarnej klasyfikacji, gdzie zmienną zależną jest zmienna binarna. W naszym przypadku, celem jest przewidzenie obecności basenu na podstawie różnych cech nieruchomości.

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.8566     0.3746  -4.956 7.19e-07 ***
MaintenanceStaff    -1.5687     0.8679  -1.807 0.070686 .
Gymnasium            0.6636     0.3933   1.687 0.091518 .
LandscapedGardens    2.8252     0.3560   7.936 2.10e-15 ***
JoggingTrack         1.0826     0.3010   3.597 0.000322 ***
RainWaterHarvesting -1.5381     0.3208  -4.794 1.63e-06 ***
IndoorGames          1.7328     0.3571   4.853 1.22e-06 ***
ShoppingMall        10.5676  1727.1931   0.006 0.995118
Intercom             0.1571     0.3420   0.459 0.646031
```

```
SportsFacility          2.2954     0.3937   5.830 5.53e-09 ***
ATM                    15.5569  1255.8738   0.012 0.990117
ClubHouse               1.7820     0.3165   5.630 1.80e-08 ***
School                 11.5460  2393.2086   0.005 0.996151
X24X7Security          -2.0034     0.3671  -5.457 4.83e-08 ***
PowerBackup            -2.3746     0.5010  -4.740 2.14e-06 ***
CarParking             -0.2372     0.3040  -0.780 0.435296
StaffQuarter            2.9266     1.1343   2.580 0.009876 **
Cafeteria              18.6539  1083.9705   0.017 0.986270
MultipurposeRoom        2.3209     0.3998   5.806 6.41e-09 ***
Hospital               -5.6969  3471.0355  -0.002 0.998690
WashingMachine        -35.1522 17076.9340  -0.002 0.998358
Gasconnection          -1.0667     0.4675  -2.282 0.022513 *
AC                     17.5691 10008.5526   0.002 0.998599
Childrensplayarea       1.9081     0.3216   5.934 2.96e-09 ***
LiftAvailable           0.3449     0.3814   0.904 0.365778
BED                    20.2361 13836.5663   0.001 0.998833
VaastuCompliant        -0.3257     0.4317  -0.755 0.450492
GolfCourse             13.7355  2593.8435   0.005 0.995775
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1695.02  on 1950  degrees of freedom
Residual deviance:  550.03  on 1923  degrees of freedom
AIC: 606.03
```

- Kluczowe zmienne istotnie wpływające na obecność basenu to: `LandscapedGardens`, `JoggingTrack`, `RainWaterHarvesting`, `IndoorGames`, `SportsFacility`, `ClubHouse`, `X24X7Security`, `PowerBackup`, `StaffQuarter`, `MultipurposeRoom`, oraz `Childrensplayarea`.
- Zmienna `ShoppingMall` oraz inne zmienne mające wysokie oszacowania, ale z dużymi błędami standardowymi, mogą wskazywać na problem z kolineacją lub brakiem istotności.
- Wskaźnik błędu klasyfikacji: 5.95%
- Model dobrze przewiduje obecność basenu, ale ma problemy z dokładnym przewidzeniem braku basenu.

Aby poprawić model, zmniejszono liczbę zmiennych. Wyniki są następujące:

- Kluczowe zmienne istotnie wpływające na posiadanie basenu obejmują `Price`, `Resale`, `Gymnasium`, `LandscapedGardens`, `JoggingTrack`, `RainWaterHarvesting`, `IndoorGames`, `Intercom`, `SportsFacility`, `ClubHouse`, `X24X7Security`, `PowerBackup`, `StaffQuarter`, `MultipurposeRoom`, oraz `Childrensplayarea`.
- Wskaźnik błędu klasyfikacji: 5.17%

- **Regresja logistyczna** wykazała się dobrymi wynikami, zwłaszcza w drugim modelu, z wskaźnikiem błędu klasyfikacji wynoszącym 5.17%.
- **LDA** i **QDA** miały nieco wyższe wskaźniki błędu klasyfikacji, odpowiednio 9.60% i 9.42%.
- **k-NN** z k=5 osiągnęło najniższy wskaźnik błędu klasyfikacji wynoszący 8.05%, ale miało problemy z dokładnym przewidzeniem braków basenów.

Wybór najlepszej metody zależy od specyficznych wymagań i priorytetów danego zadania. Regresja logistyczna i k-NN z k=5 były najbardziej obiecującymi modelami w tym przypadku.

4. Walidacja krzyżowa i bootstrap

W analizie danych i modelowaniu statystycznym ważne jest, aby oszacować i zweryfikować zdolność modelu do generalizacji na nieznane dane. Dwa popularne podejścia do tego celu to walidacja krzyżowa i metoda bootstrap. W tym rozdziale przedstawimy te techniki i zilustrujemy je na podstawie konkretnego kodu w języku R, który został zastosowany do zbioru danych dotyczących cen nieruchomości w Bangalore.

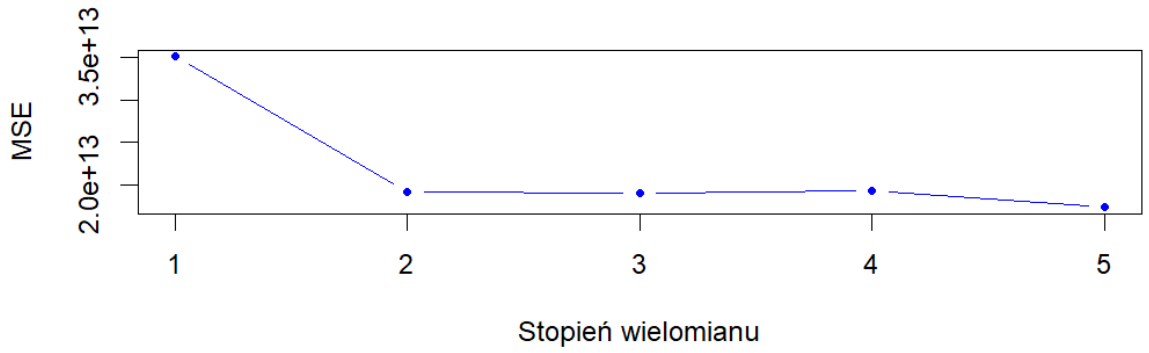
4.1. Walidacja Krzyżowa

Walidacja krzyżowa to technika, która pozwala ocenić, jak wyniki modelu będą generalizować na niezależny zestaw danych. Istnieje kilka wariantów walidacji krzyżowej, z których najczęściej używane to k-krotna walidacja krzyżowa oraz Leave-One-Out Cross-Validation (LOOCV).

4.1.1. K-krotna Walidacja Krzyżowa

W k-krotnej walidacji krzyżowej dane są dzielone na k równych części. Model jest trenowany k razy, za każdym razem używając $k - 1$ części jako zbioru treningowego, a pozostałej części jako zbioru walidacyjnego. Następnie średnia z k oszacowań błędu walidacyjnego jest używana jako miara błędu modelu.

W przykładzie zastosowana została 10-krotna walidacja krzyżowa dla regresji wielomianowej o różnych stopniach, a następnie wykreślone wyniki:

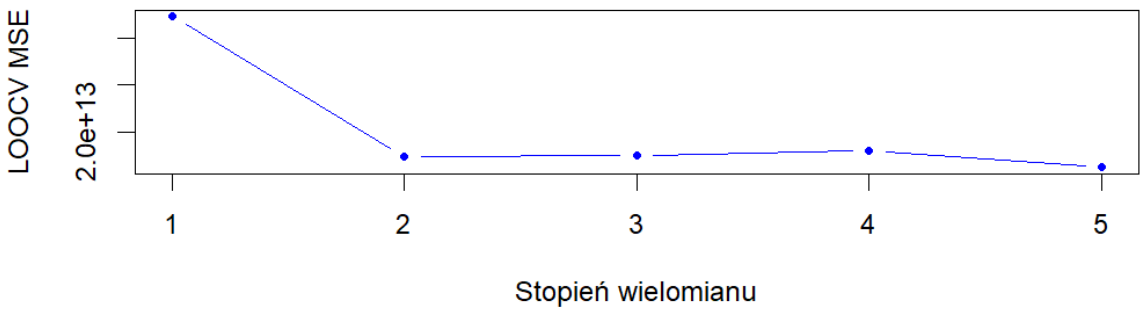


Rys.3. Błąd MSE regresji wielomianowej w zależności od stopnia wielomianu dla 10-krotnej walidacji krzyżowej.

Zaobserwowano, że błąd modelu ulega redukcji, jednakże z uwagi na brak normalizacji danych, wartości błędów wydają się być znacząco zawyżone.

4.2. Leave-One-Out Cross-Validation (LOOCV)

LOOCV jest specjalnym przypadkiem k-krotnej walidacji krzyżowej, gdzie *k* jest równy liczbie obserwacji w zbiorze danych. Każda obserwacja jest używana jako zbiór walidacyjny dokładnie raz, a pozostałe dane jako zbiór treningowy.



Rys.4. Błąd MSE regresji wielomianowej w zależności od stopnia wielomianu dla LOOCV walidacji krzyżowej.

4.3. Metoda Bootstrap

Bootstrap jest metodą statystyczną, która polega na wielokrotnym losowaniu próbek ze zbioru danych z zwracaniem (sampling with replacement) i stosowaniu tych próbek do oszacowania rozkładu statystyki lub błędu standardowego.

W przykładzie, użyta została metoda bootstrap do oszacowania błędów standardowych współczynników regresji liniowej:

```
Call:
boot(data = dane, statistic = lm_coefs, R = 1000)


Bootstrap Statistics :
        original      bias     std. error
t1* -13190012.01 219626.2545 1419504.7329
t2*     15269.84   -154.4749     996.2801
```

Wyniki pokazują, że dla współczynnika t1* (intercept) średni błąd standardowy wynosi około 1,419,504.73, a dla współczynnika t2* (Area) błąd standardowy wynosi około 996.28. Te wyniki są ważne, ponieważ dostarczają informacji o stabilności oszacowań współczynników regresji.

Wszystkie uzyskane wyniki wykazywały zbliżone wartości. Ze względu na znaczne wartości błędów, zdecydowano się na normalizację danych. Po przeprowadzeniu standaryzacji, wartość błędu zmniejszyła się do poziomu 0.01.

Walidacja krzyżowa i metoda bootstrap są nieocenionymi narzędziami w ocenie modelu statystycznego. Walidacja krzyżowa pomaga w wyborze modelu o najlepszej zdolności generalizacji, podczas gdy metoda bootstrap dostarcza informacji na temat niepewności oszacowań modelu. W tym rozdziale pokazaliśmy, jak zastosować te techniki w praktyce na przykładzie modelowania cen nieruchomości w Bangalore, co pozwala na lepsze zrozumienie i interpretację wyników modeli statystycznych.

5. Regularyzacja w modelach liniowych

Regularyzacja to technika stosowana w statystyce i uczeniu maszynowym w celu przeciwdziałania nadmiernemu dopasowaniu (overfitting) modeli do danych treningowych. W modelach liniowych regularyzacja wprowadza karę za duże wartości współczynników, co pomaga w uzyskaniu bardziej stabilnych i uogólniających się modeli. W tym rozdziale omówwione zostaną dwie popularne techniki regularyzacyjne: regresja grzbietowa (ridge regression) i regresja Lasso, ilustrując je na przykładzie analizy cen nieruchomości w Bangalore.

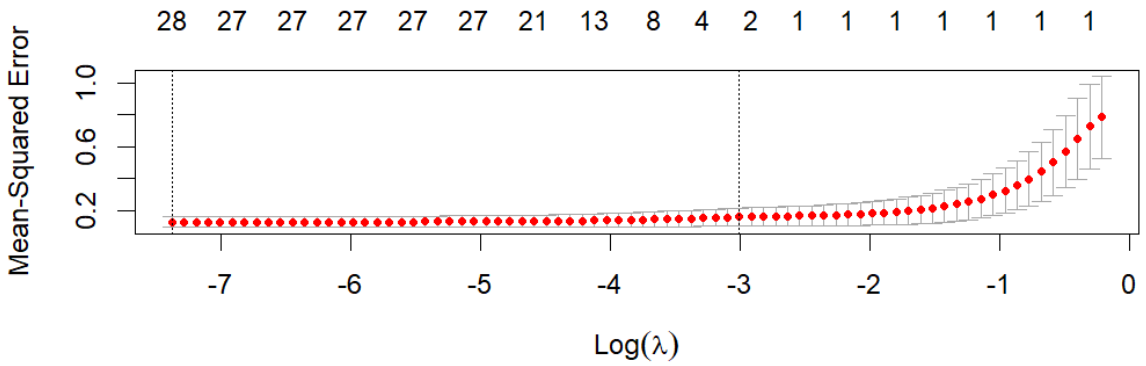
5.1. Regresja grzbietowa

Regresja grzbietowa (ridge regression) wprowadza karę w postaci sumy kwadratów współczynników regresji. Karę tę kontroluje parametr λ, który określa siłę regularyzacji.

Dla regresji grzbietowej optymalną wartość λ uzyskaliśmy przy pomocy krzyżowej walidacji, która wyniosła 0.08099967. Po zastosowaniu tej wartości λ, średni błąd kwadratowy (MSE) wyniósł 0.2183579.

5.2. Regresja Lasso

Regresja Lasso (Least Absolute Shrinkage and Selection Operator) wprowadza karę w postaci sumy modułów współczynników regresji, co może prowadzić do wyzerowania niektórych współczynników, a tym samym selekcji zmiennych.



Rys.5. Błąd MSE w zależności od dobranego parametru regularyzacji λ.

Wykres pokazuje, że istnieje wartość λ, która minimalizuje średni błąd kwadratowy (MSE). Jest to wartość $\lambda_{min} = -7$,zaznaczona na wykresie przerywaną linią. Ta wartość λ zapewnia najlepsze dopasowanie modelu do danych walidacyjnych. Dla wartości λ mniejszych od λ_{min} , MSE pozostaje stosunkowo niskie i stabilne, co sugeruje, że model nie jest nadmiernie skomplikowany i jest w stanie dobrze generalizować na danych testowych. Wartości λ większe od λ_{min} powodują wzrost MSE, co sugeruje nadmierną regularyzację i niedopasowanie modelu (underfitting). Wartość λ, również zaznaczona przerywaną linią, jest największą wartością λ, której MSE mieści się w granicach jednego błędu standardowego od minimalnego MSE. Ta wartość może być używana, jeśli chcemy bardziej konserwatywnie regularyzować model, kosztem nieco większego błędu predykcji. Szare linie wokół czerwonych kropek reprezentują przedziały błędu standardowego.

Wynik MSE dla Ridge: Wartość MSE dla optymalnej λ wynosiła 0.218, co wskazuje na dobry poziom dopasowania modelu. Dla porównania, MSE dla modelu bez regularyzacji było wyższe, co potwierdza skuteczność regularyzacji Ridge w redukcji błędu.

Wynik MSE dla Lasso: Optymalne λ dla Lasso dawało MSE wynoszące 0.181, co jest lepszym wynikiem niż dla Ridge i wskazuje na skuteczność Lasso w redukcji błędu oraz uproszczeniu modelu przez eliminację mniej istotnych zmiennych.

6. Drzewa decyzyjne i modele pochodne

Drzewa decyzyjne są jednymi z popularnych algorytmów uczenia maszynowego, które znajdują zastosowanie zarówno w zadaniach klasyfikacji, jak i regresji. Ich struktura przypomina drzewo, gdzie każdy wierzchołek reprezentuje test na jednej z cech danych, każda gałąź reprezentuje wynik tego testu, a liście drzewa zawierają wartości prognozowane dla zbiorów danych. W kontekście analizy danych nieruchomości w Bangalore drzewa decyzyjne mogą być użyte do przewidywania różnych cech, takich jak dostępność udogodnień czy kategorie cenowe mieszkań.

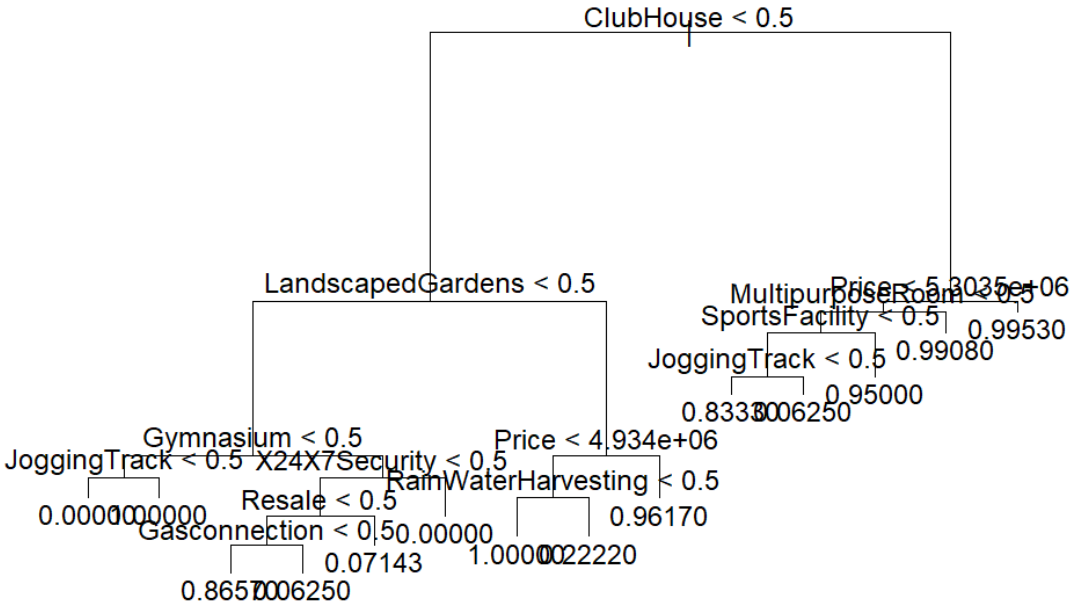
6.1. Drzewa Decyzyjne

Drzewa decyzyjne są wszechstronnymi modelami, które segmentują dane na hierarchiczne struktury oparte na podziale cech w celu przewidywania wyników.

6.1.1. Drzewo Klasyfikacyjne: sales_high_tree

Model `sales_high_tree` został zbudowany przy użyciu funkcji `tree` w R i przewiduje obecność basenu (`SwimmingPool`) na podstawie cech takich jak `ClubHouse`, `LandscapedGardens`, `Gymnasium` oraz inne. Oto kluczowe obserwacje z tego modelu:

- **Węzły i Podziały:** Drzewo składa się z węzłów, gdzie dane są dzielone na podstawie warunków (`ClubHouse`, `LandscapedGardens`, itp.).
- **Węzły Końcowe:** Istnieje 14 węzłów końcowych, gdzie dokonywane są końcowe predykcje.
- **Wykonanie:** Odchylenie średnie reszt (`Residual mean deviance: 0.02114`) sugeruje dobrą zgodność modelu.



Rys.5. Podstawowe drzewo klasyfikacyjne dla danych Bangalore.

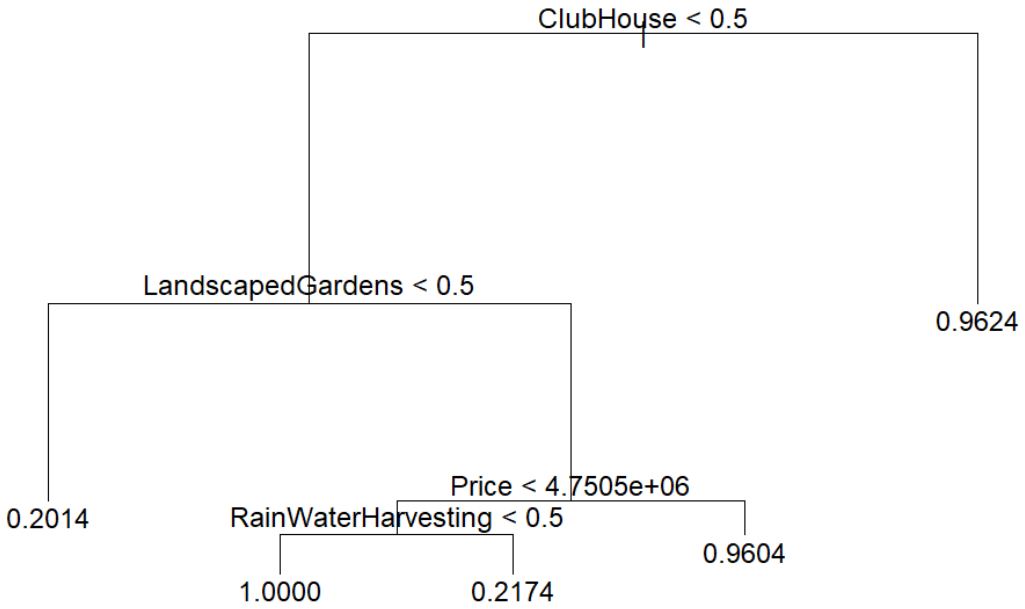
6.1.2. Drzewo Regresyjne: medv_tree

Podobnie jak `sales_high_tree`, `medv_tree` przewiduje `SwimmingPool` przy użyciu technik regresyjnych w ramach drzewa decyzyjnego. Wykorzystuje ten sam zestaw

predyktorów do oszacowania prawdopodobieństwa występowania basenów w nieruchomościach.

Główne obserwacje:

- **Resztki i Dopasowanie:** `medv_tree` pokazuje odchylenie średnie reszt wynoszące `0.02114`, co wskazuje na stabilne wykonanie w porównaniu do drzewa klasyfikacyjnego.
- **Węzły Końcowe:** Składa się z 5 węzłów końcowych, odzwierciedlających segmentację danych na podstawie cech nieruchomości.

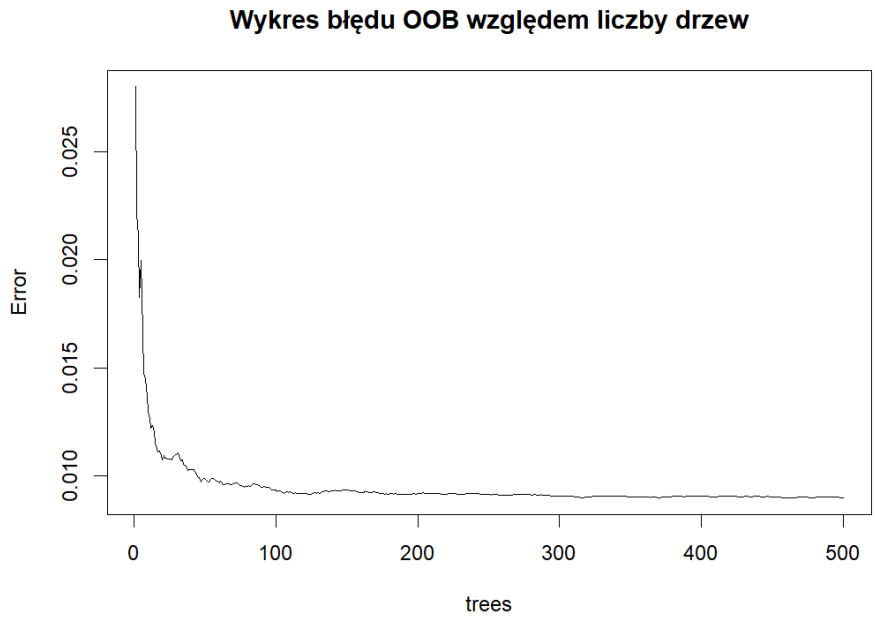


Rys.6. Drzewo regresyjne wyznaczone metodą przycinania sterowanego złożonością.

6.2. Metody Ensemble

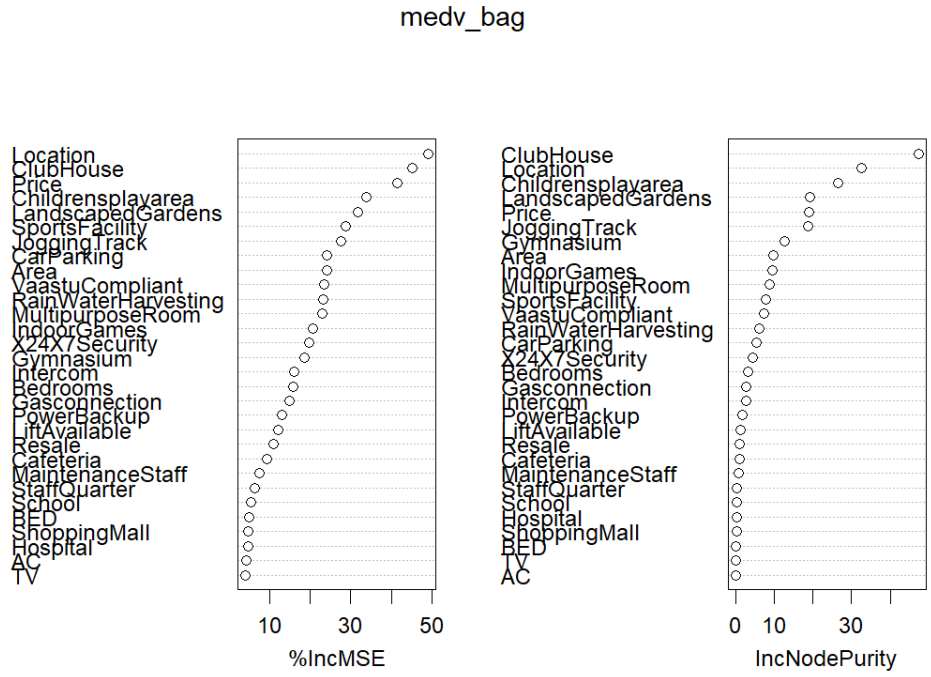
Metody ensemble polegają na łączeniu wyników wielu modeli bazowych w celu uzyskania bardziej dokładnych i stabilnych prognoz. Przykłady takich metod to bagging (np. Random Forest) i boosting (np. Gradient Boosting), które redukują wariancję i błąd systematyczny, poprawiając ogólną wydajność modelu. W praktyce, metody ensemble są szczególnie skuteczne w złożonych problemach, gdzie pojedynczy model może nie być wystarczająco elastyczny lub dokładny.

6.2.1. Bagging z Random Forests



Rys.7. Wykres błędu OOB względem liczy drzew dla modelu Random Forest.

- **Model Random Forest (medv_bag)**: Uśrednia predykcje z wielu drzew decyzyjnych (w tym przypadku `500` drzew) w celu poprawy dokładności i kontrolowania nadmiernej dopasowania.



Rys.8. Wykres ważności predyktorów.

- **Istotność Zmiennych**: Identyfikuje kluczowe predyktory (`Price`, `Area`, `ClubHouse`, itp.), wpływające na obecność basenów.

6.2.2. Boosting z Gradient Boosting Machines (GBM)

- **Model Gradient Boosting (`medv_boost`)**: Konstruuje zespół słabych modeli predykcyjnych (drzew decyzyjnych) iteracyjnie, aby zminimalizować błędy resztowe i poprawić dokładność predykcji.

Porównanie Modeli i Ocena Wydajności

- **Walidacja Krzyżowa i Przycinanie**: Techniki stosowane do optymalizacji drzew decyzyjnych przez wybór odpowiedniej złożoności drzewa (`cv.tree`, `prune.tree`).
- **Ocena Wydajności**: Metryki takie jak średni błąd kwadratowy (MSE) służą do oceny dokładności modelu (`0.0325374` dla przyciętego `medv_tree`).

Drzewa decyzyjne oraz pochodne modele takie jak Random Forests i Gradient Boosting stanowią solidne ramy do przewidywania obecności basenów w nieruchomościach. Modele te wykorzystują uczenie zespołowe w celu poprawy dokładności predykcji, co czyni je nieocenionymi narzędziami w analizie rynku nieruchomości oraz w procesach podejmowania decyzji.

7. Wnioski

Analiza danych dotyczących rynku nieruchomości w Bangalore dostarcza szeregu istotnych wniosków.

1. **Powierzchnia mieszkania jako główny czynnik wpływający na cenę:** Regresja liniowa wykazała, że powierzchnia mieszkania jest kluczowym predyktorem ceny nieruchomości. Wysoka wartość współczynnika regresji (14564 rupii indyjskich za dodatkowy metr kwadratowy) potwierdza, że większe mieszkania zazwyczaj mają wyższe ceny. Znacząca część zmienności cen mieszkań (82.37%) jest wyjaśniana przez tę jedną zmienną. Dla deweloperów i nabywców istotne jest zatem rozważenie tej cechy przy ocenie wartości nieruchomości.
2. **Znaczenie udogodnień i lokalizacji:** Analiza klasyfikacyjna pokazała, że dostępność różnych udogodnień, takich jak ogrody krajobrazowe, trasy joggingowe czy baseny, istotnie wpływa na atrakcyjność nieruchomości. Udogodnienia te mogą nie tylko podnosić komfort życia mieszkańców, ale także zwiększać wartość rynkową nieruchomości.
3. **Wpływ lokalizacji na cenę nieruchomości:** Korelacja ujemna między ceną a odległością od centrum miasta wskazuje, że nieruchomości położone bliżej centrum zazwyczaj są droższe. Jest to zgodne z ogólnymi trendami na rynkach miejskich, gdzie dostępność do kluczowych punktów usługowych i komunikacyjnych ma duże znaczenie dla cen nieruchomości.
4. **Potrzeba dalszej analizy reszt i poprawy modeli:** Diagnostyka przeprowadzona na modelu regresji liniowej wykazała pewne problemy, takie jak heteroskedastyczność i

obecność wartości odstających. Zastosowano do tego różne techniki regularyzacji oraz transformacji danych. Dodatkowo, rozszerzenie modelu o dodatkowe zmienne niezależne mogłoby pomóc w lepszym przewidywaniu cen nieruchomości.

5. **Zastosowanie różnych metod klasyfikacji:** Porównanie różnych metod klasyfikacji, takich jak regresja logistyczna, LDA, QDA i k-NN, pozwoliło wyłonić regresję logistyczną jako najbardziej obiecującą pod względem dokładności predykcji. Mimo że każda z metod miała swoje zalety i ograniczenia, to regresja logistyczna osiągnęła najniższy wskaźnik błędu klasyfikacji (5.17%), co potwierdza jej przydatność w przewidywaniu obecności różnych udogodnień w nieruchomościach.
6. **Walidacja krzyżowa jako narzędzie oceny modeli:** Zastosowanie walidacji krzyżowej potwierdziło zdolność modeli do generalizacji na nieznane dane. Dla lepszej oceny modeli zaleca się stosowanie tej techniki, aby uniknąć zjawiska przeuczenia i zapewnić ich skuteczność w praktycznych zastosowaniach.
7. **Drzewa decyzyjne:** Oprócz szczegółowej analizy danych nieruchomości w Bangalore za pomocą drzew decyzyjnych, istotnym wnioskiem jest możliwość uzyskania intuicyjnych reguł decyzyjnych, które mogą być łatwo interpretowane przez decydentów i analityków. Dzięki strukturze drzewa, której wizualizacja jest klarowna, możliwe jest zrozumienie kluczowych czynników wpływających na ceny nieruchomości oraz preferencje rynkowe. Ta interpretowalność czyni drzewa decyzyjne cennym narzędziem nie tylko do prognozowania cen, ale również do analizy rynku i identyfikacji kluczowych trendów w branży nieruchomości.

Wnioski te podkreślają kompleksowość rynku nieruchomości w Bangalore oraz znaczenie dokładnej analizy danych przy podejmowaniu decyzji inwestycyjnych i zakupowych. Dalsze badania w tym obszarze mogą przyczynić się do jeszcze lepszego zrozumienia dynamiki cenowej oraz preferencji konsumentów na tym rozwijającym się rynku.