2 czerwca 2021

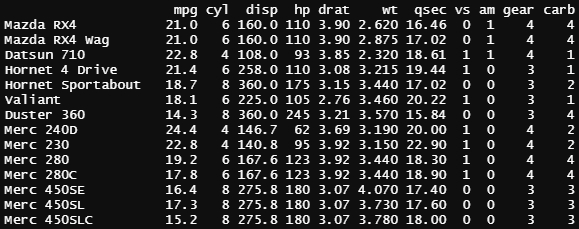
Kinga Kalisz & Marek Pałdyna

Porównanie Metod Regresji

Pakiet R

# MTCars

## Opis danych



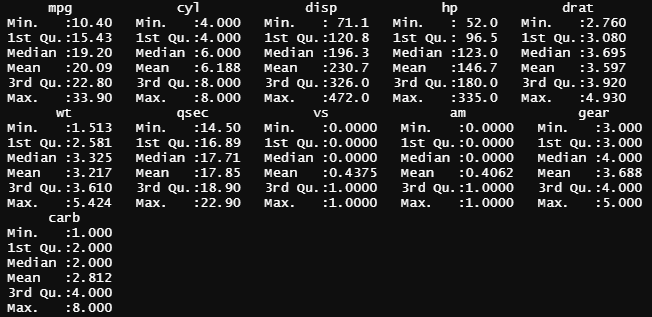
Zbiór danych MTCars zawarty jest w zbiorach danych zawartych w środowisku R.   
Składa się on z 11 kolumn i przedstawia dane na temat 32 samochodów z lat 1973-74. Zaciągnięty został z „Motor Trend US magazine”

Za cel przewidywań na tym zbiorze została obrana kolumna „qsec” która opisuje czas na ćwierć mili.

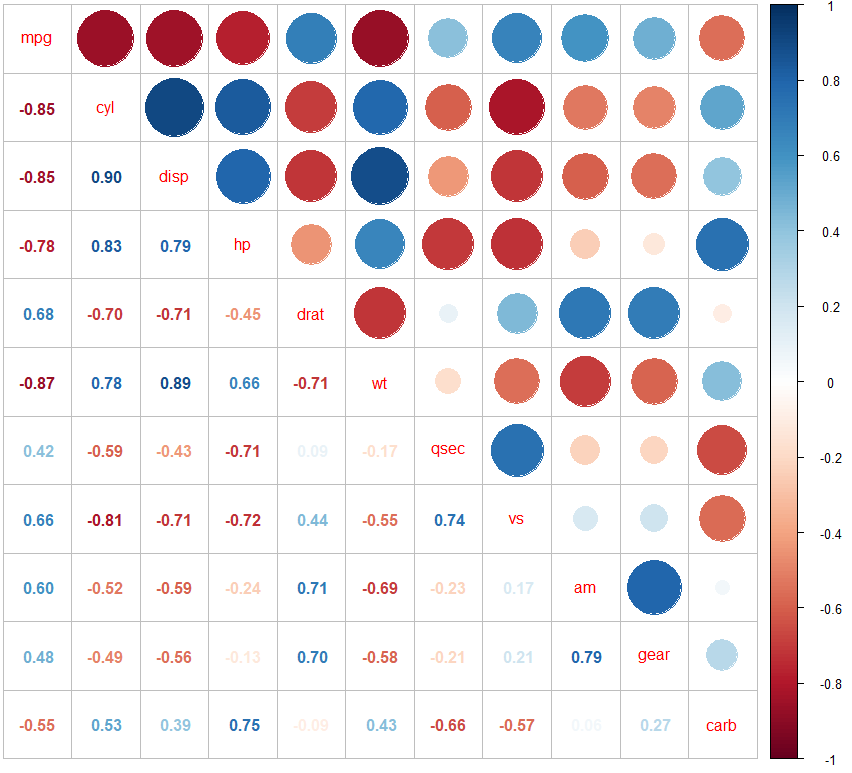
Kolumny:

* „mpg” – spalanie paliwa (mil/galon)
* „cyl” – Liczba cylindrów
* „disp” – pojemność silnika (cu.in.)
* „hp” – moc silnika (KM )
* „drat” – przełożenie tylnej osi
* „wt” – waga (1000 lbs)
* „qsec” – czas na ¼ mili
* „vs” – typ silnika (0 – w kształcie V, 1 - prosty)
* „am” – skrzynia biegów (0 – automatyczna, 1 -manualna)
* „gear” – liczba biegów do jazdy do przodu
* „carb” – liczba gaźników

## Podsumowanie danych:



## Korelacja zmiennych:



W tym zbiorze danych istnieją duże korelacje „pozytywne” i „negatywne”.

Korelacje negatywne:

* pomiędzy spalaniem i ilością cylindrów silnika,
* pomiędzy spalaniem i wagą samochodu
* pomiędzy spalaniem i pojemnością silnika
* pomiędzy liczbą cylindrów i typem silnika

Korelacje pozytywne:

* pomiędzy pojemnością silnika, a wagą samochodu
* pomiędzy pojemnością silnika, a liczbą cylindrów
* pomiędzy rodzajem silnika a liczbą biegów do jazdy do przodu
* pomiędzy liczbą cylindrów, a mocą silnika

## Obróbka danych

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.8523863
* backward r.squared = 0.8453189
* **both r.squared = 0.8642928**

Wybrany został więc model “both” –> „qsec ~ cyl + disp + wt + vs + am + carb”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + …
  + …
* SVM z kernelami:
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |
| --- | --- | --- |
| Model | Typ | RMSE |
| LM |  | ??? |
| SVM | Linear | 0,7227685 |
| Polynomial | 0,8880503 |
| Radial | 0,7340955 |
| Sigmoid | 1,173061 |
| GLM | Gaussian | ??? |
| Gamma | ??? |
|  | … | … |

## Podsumowanie

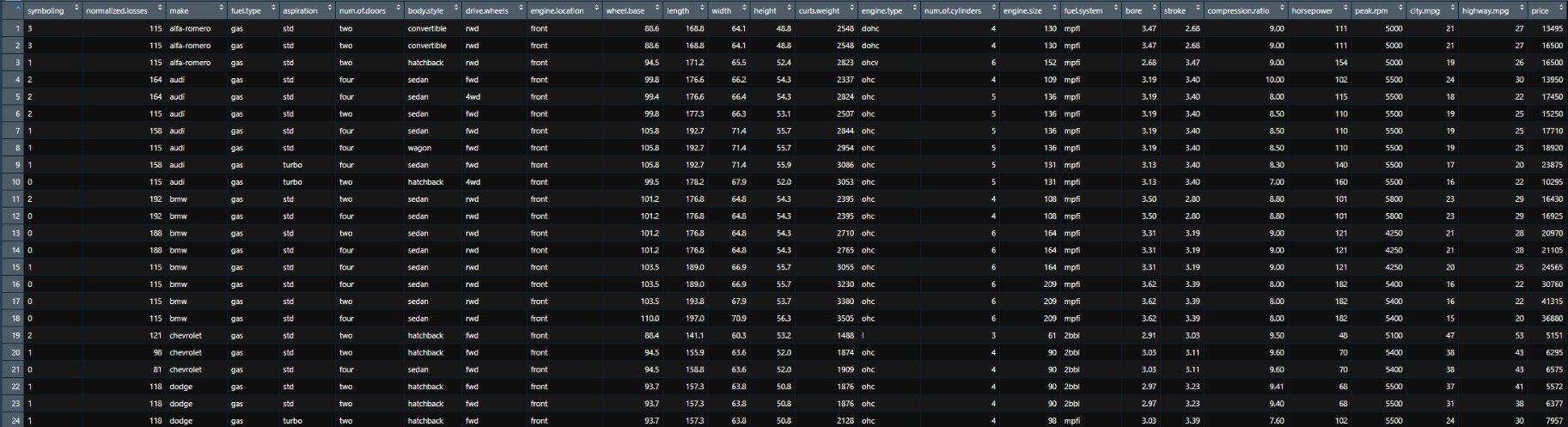
Najlepszym modelem jest : etc.

# Imports-85

## Dane

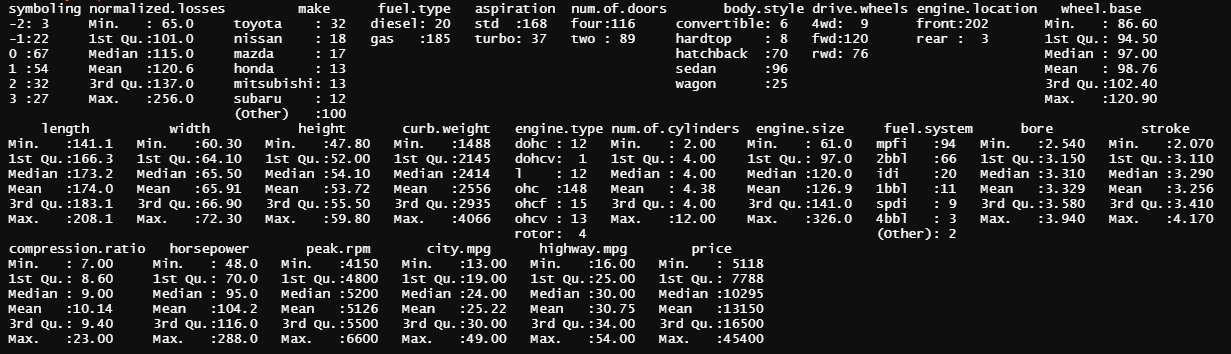
Plik imports-85.csv to plik zawierający dane o samochodach. Przechowuje 205 obserwacji.

Kolumny:

* symbolling – poziom ryzyka samochodu
* normalized-losses – względna średnia wysokość odszkodowania za rok ubezpieczenia pojazdu
* make – marka
* fuel-type – typ paliwa
* aspiration – rodzaj ssania silnika
* num-of-doors – liczba drzwi
* body-style - typ nadwozia
* drive-wheels - rodzaj napędu
* engine-location - położenie silnika
* wheel-base - rozstaw osi
* length - dłogość
* width - szerokość
* height - wysokość
* curb-weight – masa własna
* engine-type – typ silnika
* num-of-cylinders – ilość cylindrów
* engine-size: rodzaj silnika
* fuel-system: system paliwowy
* bore: prześwit
* stroke: suw silnika
* compression-ratio: stopień sprężenia
* horsepower: moc silnika w KM
* peak-rpm: maksymalna ilość obrotów/min silnika
* city-mpg: zużycie benzyny w mieście
* highway-mpg: zużycie benzyny na drodze szybkiego ruchu
* price - cena

Na podstawie danych będziemy przeprowadzać przewidywanie „poziomu ryzyka ubezpieczenia samochodu” (symbolling) na podstawie innych parametrów.

## Podsumowanie danych



## Korelacja zmiennych

W tym zbiorze danych istnieją duże korelacje „pozytywne” i „negatywne”.

Korelacje negatywne:

* pomiędzy spalaniem w mieście a ilością KM
* pomiędzy spalaniem w mieście/na szosie a wagą własną
* pomiędzy współczynnikiem kompresji a rodzajem paliwa
* pomiędzy spalaniem a rodzajem paliwa

Korelacje pozytywne:

* pomiędzy rozstawem osi a długością
* pomiędzy długością a szerokością
* pomiędzy rozmiarem silnika a masą własną
* pomiędzy zużyciem na szosie a zużyciem w mieście

## Obróbka danych

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.7512137
* **backward r.squared = 0.8202633**
* both r.squared = 0.7014353

Wybrany został więc model “backward” –> „symboling ~ normalized.losses + make + num.of.doors + drive.wheels + length + width + engine.type + num.of.cylinders + engine.size + fuel.system + bore + city.mpg + highway.mpg”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + …
  + …
* SVM z kernelami:
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |
| --- | --- | --- |
| Model | Typ | RMSE |
| LM |  | ??? |
| SVM | Linear | 0.7931797 |
| Polynomial | 0.6681694 |
| Radial | 0.5222752 |
| Sigmoid | 2.188139 |
| GLM | Gaussian | ??? |
| Gamma | ??? |
|  | … | … |

## Podsumowanie

Najlepszym modelem jest : etc.

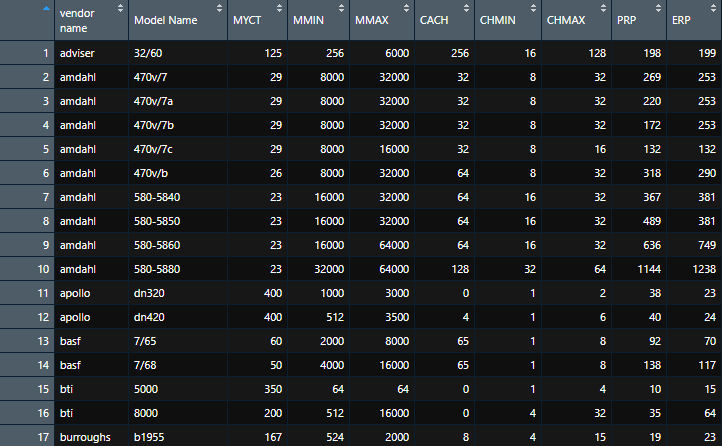
# Machine

## Dane

Plik machine.data zawiera dane na temat Komputerów PC. Posiada 209 wierszy danych.

Kolumny:

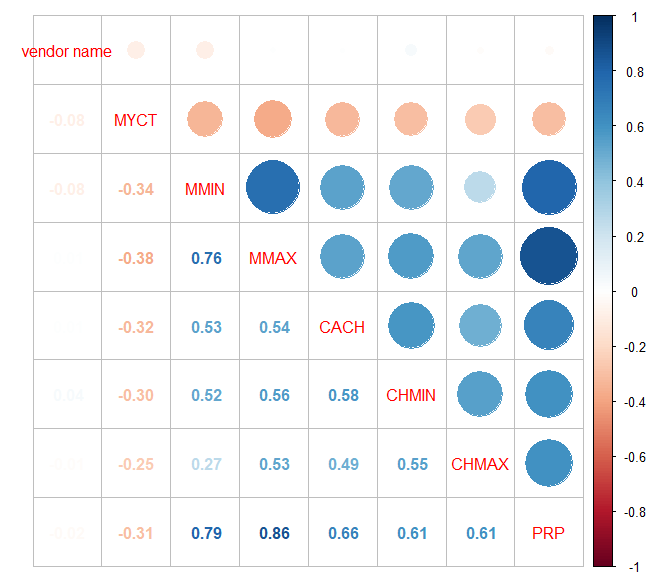
1. vendor name: nazwa dostawcy
2. Model Name: nazwa modelu
3. MYCT: czas cyklu maszyny in nanoseconds
4. MMIN: minimalna główna pamięć w kilobajtach
5. MMAX: maksymalna główna pamięć w kilobajtach
6. CACH: pamięć cache w kilobajtach
7. CHMIN: minimalna liczba kanałów w jednostkach
8. CHMAX: maksymalna liczba kanałów w jednostkach
9. PRP: opublikowana wydajność względna
10. ERP: przewidziana wydajność względna z oryginalnego artykułu



Na podstawie danych będziemy przeprowadzać przewidywanie „mocy względna” (PRP) komputera.

## Podsumowanie danych

## Korelacja zmiennych



W tym zbiorze danych istnieją duże korelacje „pozytywne” i słabe „negatywne”.

Korelacje pozytywne:

* pomiędzy maksymalną/minimalną główną pamięcią, a opublikowana wydajność względna
* pomiędzy minimalną, a maksymalną główną pamięcią

Korelacje negatywne:

* pomiędzy czasem cyklu maszyny, a maksymalną/minimalną główną pamięcią
* pomiędzy czasem cyklu maszyny, a pamięcią cache

Zmienna vendor name jest bardzo słabo skorelowana. Dodatkowo z danych zostały usunięte dwie zmienne:

* Model name – usunięta ze względu, że jest to zmienna unikalna dla każdego rekordu w danych
* ERP – wartość przewidywana i bardzo podobna do ERP więc będzie zaburzać pracę algorytmów

## Obróbka danych

Do najlepszego dopasowania danych do zadanego problemu użyto funkcji step. Dobrała ona kolumny tak aby regresja wychodziła jak najlepsza -> najlepsze R-squared. Funkcja ta ma 3 „metody” – forward, backward i both.

Wartości squared dla danych:

* forward r.squared = 0.8648239
* backward r.squared = 0.8648239
* both r.squared = 0.8648239

Dla tych danych wszystkie metody doboru zmiennych wybrały ten sam model -> „PRP ~ MYCT + MMIN + MMAX + CACH + CHMAX”

## Modele regresji

Zostały zaimplementowane 3 modele regresji:

* LM
* GLM
  + Gaussian
  + Poisson
  + …
  + …
* SVM z kernelami:
  + Linear
  + Polynomial
  + Radial
  + Sigmoid

## Porównanie poszczególnych modeli i ich błędów RMSE

|  |  |  |
| --- | --- | --- |
| Model | Typ | RMSE |
| LM |  | ??? |
| SVM | Linear | 70.25375 |
| Polynomial | 39.50634 |
| Radial | 66.4254 |
| Sigmoid | 525.4628 |
| GLM | Gaussian | ??? |
| Gamma | ??? |
|  | … | … |

## Podsumowanie

Najlepszym modelem jest : etc.