

1 Zadania teoretyczne

Zadanie 1

- Krzywa ROC ilustruje skuteczność klasyfikatora binarnego. Jest to wykres, który przedstawia stosunek wartości TPR (czułość) do FPR przy różnych progach decyzyjnych.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

gdzie:

TP – prawdziwie pozytywne,

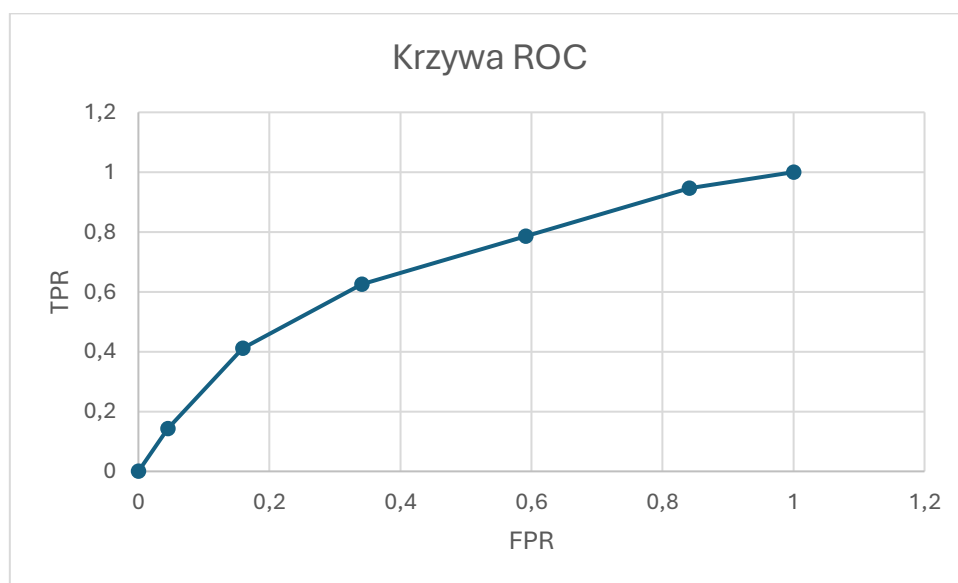
FN – fałszywie negatywne,

FP – fałszywie pozytywne,

TN – prawdziwie negatywne.

-

Próg	TP	FP	TN	FN	TPR	FPR
1.0	0	0	44	56	0	0
0.9	8	2	42	48	0,143	0,045
0.8	23	7	37	33	0,411	0,159
0.7	35	15	29	21	0,625	0,341
0.6	44	26	18	12	0,786	0,591
0.5	53	37	7	3	0,946	0,841
0.4	56	44	0	0	1	1



Zadanie 2

Root Mean Square Error (RMSE) to miara używana do oceny jakości modelu predykcyjnego w kontekście regresji. RMSE mierzy średnią wielkość błędów pomiędzy wartościami przewidywanymi przez model a rzeczywistymi wartościami.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

gdzie:

n – liczba obserwacji,

y_i – rzeczywista wartość obserwacji i

\hat{y}_i – przewidywana wartość obserwacji i .

Mean Absolute Error (MAE) to również miara używana do oceny jakości modelu predykcyjnego w kontekście regresji. Jest to średnia wartość bezwzględnych różnic między przewidywanymi a rzeczywistymi wartościami.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

gdzie:

n – liczba obserwacji,

y_i – rzeczywista wartość obserwacji i

\hat{y}_i – przewidywana wartość obserwacji i .

Relative Absolute Error (RAE) tak jak w pozostałych przypadkach to miara stosowana w ocenie jakości modeli predykcyjnych w kontekście regresji. Pozwala ona na ocenę modelu poprzez porównanie błędów modelu do błędów modelu bazowego (najczęściej średniej arytmetycznej z wartości rzeczywistych).

$$RAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}|}$$

n – liczba obserwacji,

y_i – rzeczywista wartość obserwacji i ,

\hat{y}_i – przewidywana wartość obserwacji i ,

\bar{y} – to średnia z rzeczywistych wartości y .

2 Model regresji

Zadanie 1

b) 731 rekordów, 16 atrybutów,

instant – liczba porządkowa,

dteday – data wypożyczenia roweru,

season – pora roku,

yr – rok,

mnth – miesiąc,

holiday – czy dzień jest świętem,

weekday – dzień tygodnia,

workingday – czy dzień jest dniem roboczym,

weathersit – sytuacja pogodowa,

temp – znormalizowana temperatura,

atemp – znormalizowana odczuwalna temperatura,

hum – znormalizowana wilgotność,

windspeed – znormalizowana prędkość wiatru,

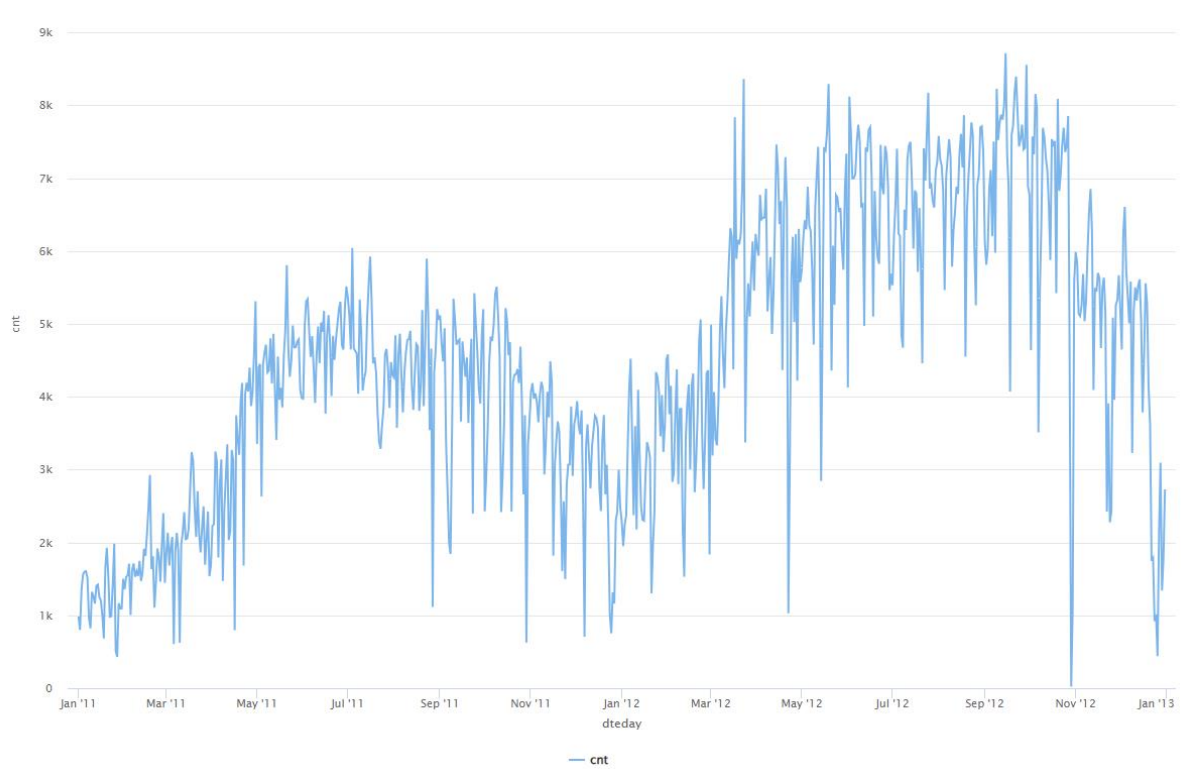
casual – liczba wypożyczonych rowerów przez użytkowników niezarejestrowanych,

registered – liczba wypożyczonych rowerów przez użytkowników zarejestrowanych,

cnt – łączna liczba wypożyczeń (suma casual i registered).

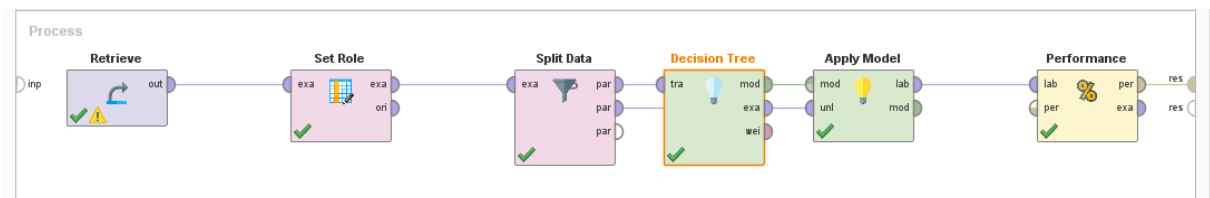
c) Integer, 22-8714

d)



Największy popyt jest w okresie od wiosny (koniec kwietnia w 2011 i połowa marca w 2012) do końca października.

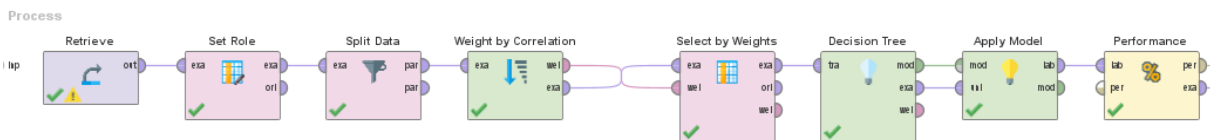
Zadanie 2



root_mean_squared_error: 233.953 +/- 0.000
relative_error: 4.49% +/- 5.79%

Zadanie 3

-

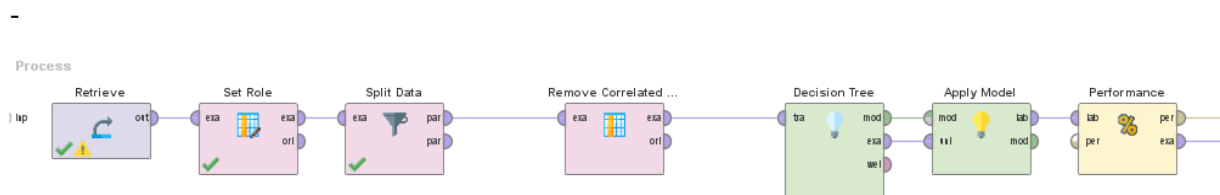


Select by Weights

weight relation:

k:

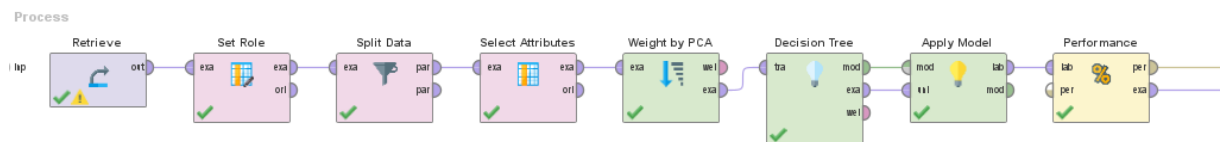
root_mean_squared_error: 226.368 +/- 0.000
 relative_error: 17.52% +/- 205.75%



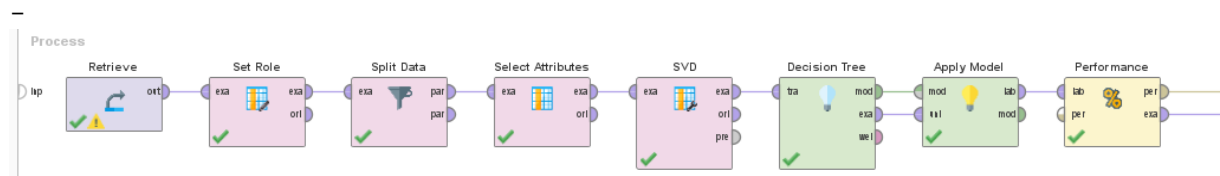
Remove Correlated Attributes

correlation:

root_mean_squared_error: 198.963 +/- 0.000
 relative_error: 4.39% +/- 6.18%



root_mean_squared_error: 162.833 +/- 0.000
 relative_error: 3.03% +/- 3.10%



dimensionality reduction

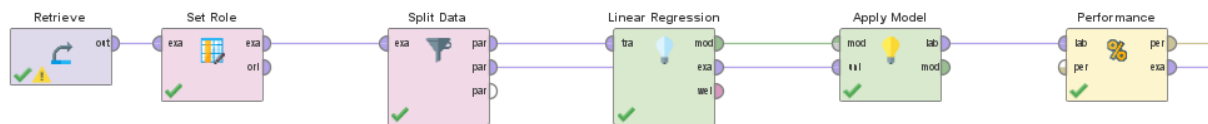
dimensionality reduction:

dimensions:

root_mean_squared_error: 165.854 +/- 0.000
 relative_error: 18.88% +/- 230.44%

Zadanie 4

a)



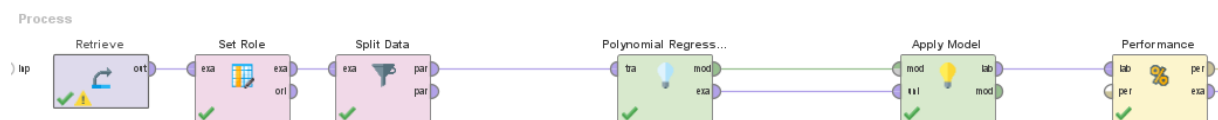
Linear Regression

min tolerance

ridge

root_mean_squared_error: 732.768 +/- 0.000
 relative_error: 99.11% +/- 1,174.47%

b)



Polynomial Regression

max iterations

replication factor

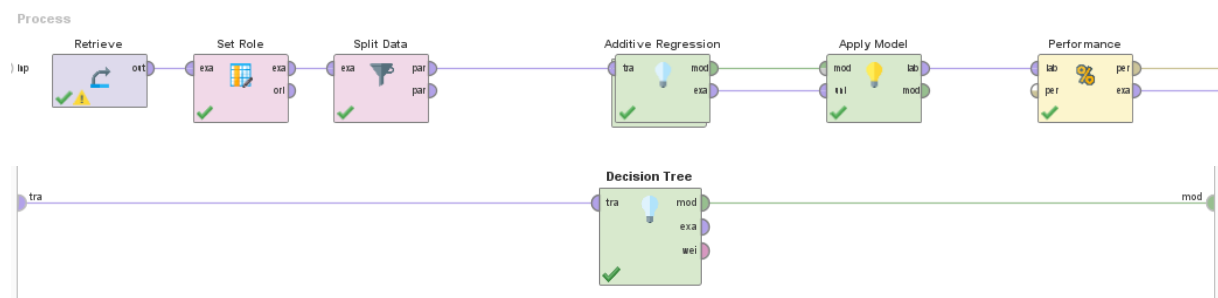
max degree

min coefficient


max coefficient



root_mean_squared_error: 1.807 +/- 0.000
 relative_error: 0.06% +/- 0.09%

c)



Parameters ×

 **Additive Regression**

iterations	<input type="text" value="10"/>	
shrinkage	<input type="text" value="0.5"/>	

root_mean_squared_error: 10.684 +/- 0.000
relative_error: 1.88% +/- 39.88%

Zadanie 5

- Predykcje można przeprowadzić bez selekcji cech, ale selekcja cech powinna być użyta w takim zadaniu ze względu na to, że wyniki dzięki niej są lepsze.
- Jeśli chodzi o selekcję cech to najlepsza okazała się selekcja na podstawie korelacji między atrybutami. Mniejsze błędy RMSE i RAE względem zwykłego DT i korelacji z decyzją. Jeśli uwzględnimy także redukcję wymiarów to wtedy najlepsze okazuje się PCA – najmniejsze błędy.
- Najmniejsze wartości błędów były w przypadku Polynomial, jednak mam wątpliwości czy nie wynikają one z przetrenowania lub jakiegoś błędu. Jeśli tak to wtedy najlepsze jest Additive Regression (z Decision Tree).