Metody Odkrywania Wiedzy

Dokumentacja końcowa projektu

„Predykcja zużycia energii na podstawie danych czujnikowych”

Krzysztof Belewicz
Paweł Pieńczuk

26 stycznia 2020

1. Opis projektu

Celem projektu było wyznaczenie całkowitego zużycia energii dla zadanej chwili czasu, tzn. sumy poborów sprzętów AGD (kolumna *Appliances*) i oświetlenia (kolumna *lights*). Zbiór danych został pozyskany z archiwum dostępnego na stronie: `https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction`. Pojęciem docelowym jest wartość całkowitej pobieranej mocy przez gospodarstwo domowe. W ramach projektu zdecydowano się na oddzielne wykonania zadania regresji dla celu *Appliances* i celu *lights*, ze względu na hipotezę, że modele je wyznaczające mogą mieć inne właściwości.

Dokonano selekcji atrybutów za pomocą trzech algorytmów opisanych w rozdziale 3. Przeprowadzono procedurę oceny algorytmów liniowej regresji, drzew regresji oraz kawałkami liniowej regresji.

2. Opis danych

2.1. Charakterystyka danych

Dane wykorzystywane do eksperymentów zostały zebrane za pomocą sieci czujników w niewielkim domu w czasie 4.5 miesiąca. Składają się z:

- daty i godziny pomiaru,
- poboru energii sprzętów domowych [Wh],
- poboru energii oświetlenia [Wh],
- pomiarów temperatury i wilgotności dla 8 różnych pomieszczeń ([$^\circ C$], [$\%$]),
- pomiarów temperatury i wilgotności dla zewnętrznej, północnej strony budynku ([$^\circ C$], [$\%$]),
- danych z pobliskiej stacji pogodowej:
  - temperatura powietrza [$^\circ C$],
  - temperatura punktu rosy [$^\circ C$],
  - ciśnienie atmosferyczne [$mm\ Hg$],
  - wilgotność [$\%$],
  - prędkość wiatru [m/s],
  - widoczność [km].

2.2. Przygotowanie danych

Każdy pomiar został uśredniony z 3 próbek wykonanych w równych odstępach co ok. 3,3 min. W ramach przygotowania danych, data i godzina pomiaru zostały rozdzielone na cztery oddzielne kolumny, zawierające miesiąc, dzień, godzinę i minutę pomiaru.

Liczba wszystkich obserwacji, zebranych w pliku *energydata_complete.csv* wynosi 19735. Celem przyspieszenia obliczeń, algorytmy przedstawione w zadaniu zostały wykonane na danych zawierających 2000 pierwszych rekordów zmienna *testDataLength*. Wszystkie operacje dot. przygotowania danych są wykonane w funkcji *data_org*.

3. Selekcja atrybutów

Aby zapobiec nadmiernemu dopasowaniu, stosuje się selekcję atrybutów, która wybiera kilka najważniejszych atrybutów do późniejszego stworzenia modeli. Po zastosowaniu selekcji, modele oparte o ograniczoną liczbę atrybutów zwykle są lepsze od opartych o wszystkie atrybuty. Istnieje wiele metod selekcji atrybutów; w ramach projektu zostało sprawdzone kilka metod (w nawiasach umieszczono opcję type funkcji *feature_selection*):

- prosty filtr statystyczny (*„simple"*) - pomiędzy każdym z atrybutów a celem regresji stosuje się miarę statystyczną, która określa zależność celu od danego atrybutu (dalej „miara zależności"). Następnie wybiera się kilka atrybutów o największej „mierze zależności". W ramach regresji pomiędzy atrybutami ciągłymi zastosowano współczynnik korelacji (Pearsona);
- bazująca na drzewach losowych (*„rf"*) - w tym celu wykorzystano pakiet randomForest i jego wbudowaną opcję zwracają parametr IMPORTANCE (bazujący na mierze MSE), o możliwości konfiguracji ilości drzew, w badaniu wykorzystano generowanie 500 drzew;
- metoda RRELIEF (*„relief"*) - wersja algorytmu RELIEF do zastosowań w zadaniu regresji. Algorytm RELIEF, początkowo zaprojektowany dla zadania klasyfikacji binarnej, polega na losowym wybraniu obserwacji (jednego rekordu klasy+atrybuty). Następnie wyszukuje się k najbardziej podobnych obserwacji tej samej klasy, oraz k klasy przeciwnej. Dla każdego atrybutu oblicza się wagę istotności. Po wykonaniu K operacji, wykonuje się średnią wag istotności. Atrybuty segreguje się według wag istotności. W zadaniu regresji stosuje się inne funkcje obliczające wagę np. funkcję rozkładu. W projekcie $k=3$, $K=50$.

W ramach projektu stosuje się następujące podejście: dla każdej wymienionej metody wykonuje się selekcję połowy atrybutów (*part=0.5*) atrybutów, następnie wyznaczoną formułę aplikuje się do stworzenia modelu *rpart()*, i procedurze oceny (10-krotnej walidacji krzyżowej *model_eval()*)[1]. Następnie największy współczynnik korelacji Pearsona wyznacza najlepszą metodę selekcji atrybutów oraz formułę do stworzenia modelu.

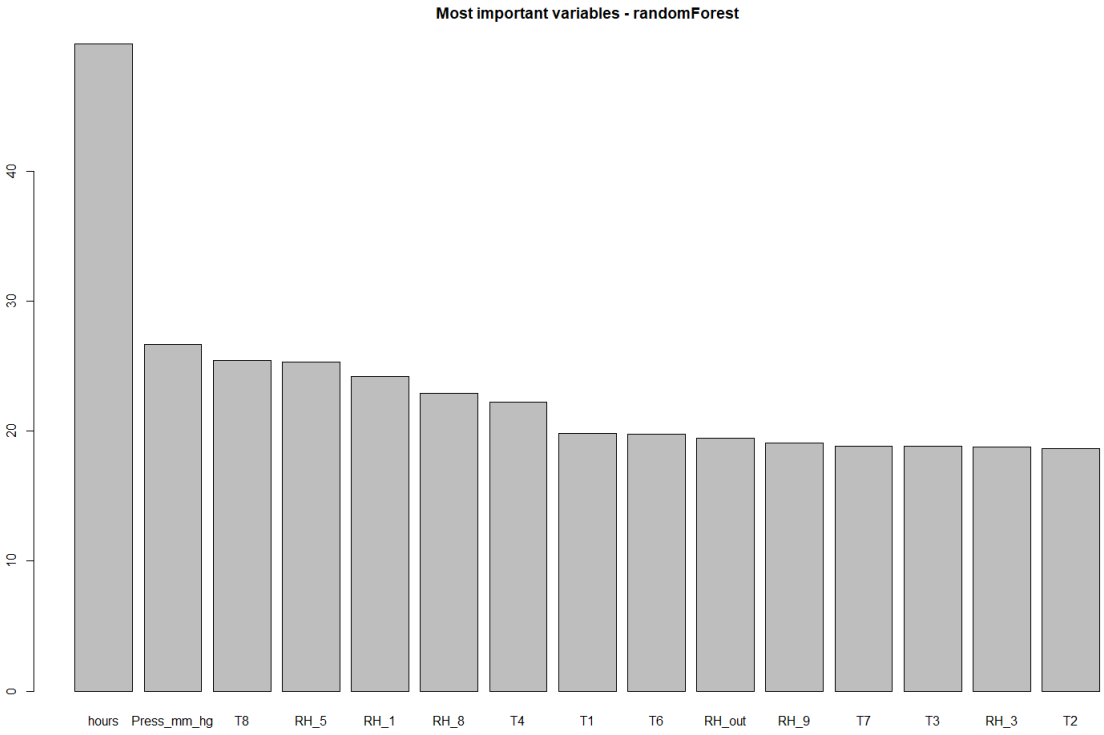
3.1. Wyniki selekcji atrybutów

Na rysunkach 3.4 – 3.6 przedstawiono wyniki każdej z selekcji. Dla *randomForest* i *simple.filter* atrybut *hours* dominuje nad pozostałymi. W tabeli 3.1 przedstawiono porównanie wyników każdej z selekcji. Wynika z tego, że atrybuty wyznaczone funkcją *simple.filter* pozwalają na najlepsze wyznaczenie modelu. Dla porównania przedstawiono też wynik walidacji krzyżowej dla modelu opartego o wszystkie atrybuty. Tylko selekcja atrybutów za pomocą *simple.filter* pozwala na poprawę dla obecnych warunków testowych (dostępne dane, ilość selekcjonowanych argumentów, algorytm do walidacji krzyżowej). W związku z wynikami, atrybuty wyznaczono za pomocą prostego filtru statystycznego posłużą w dalszej konstrukcji modelów.

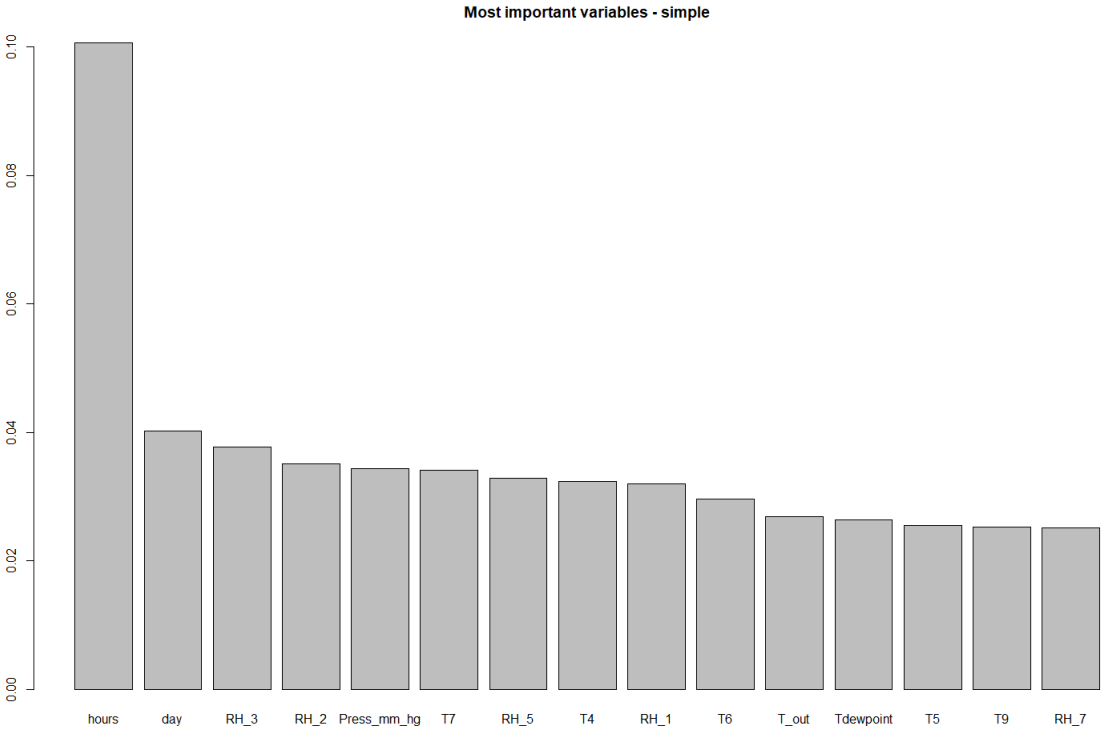
Tablica 3.1: Wyniki selekcji atrybutów — współczynniki korelacji

Parametr	***randomForest***	***simple***	***RELIEF***	**bez selekcji**
Appliances	0,555	0,616	0,553	0,585
lights	0,672	0,759	0,691	0,757

[1] Więcej nt. procedury walidacji krzyżowej w rozdziale 4.3

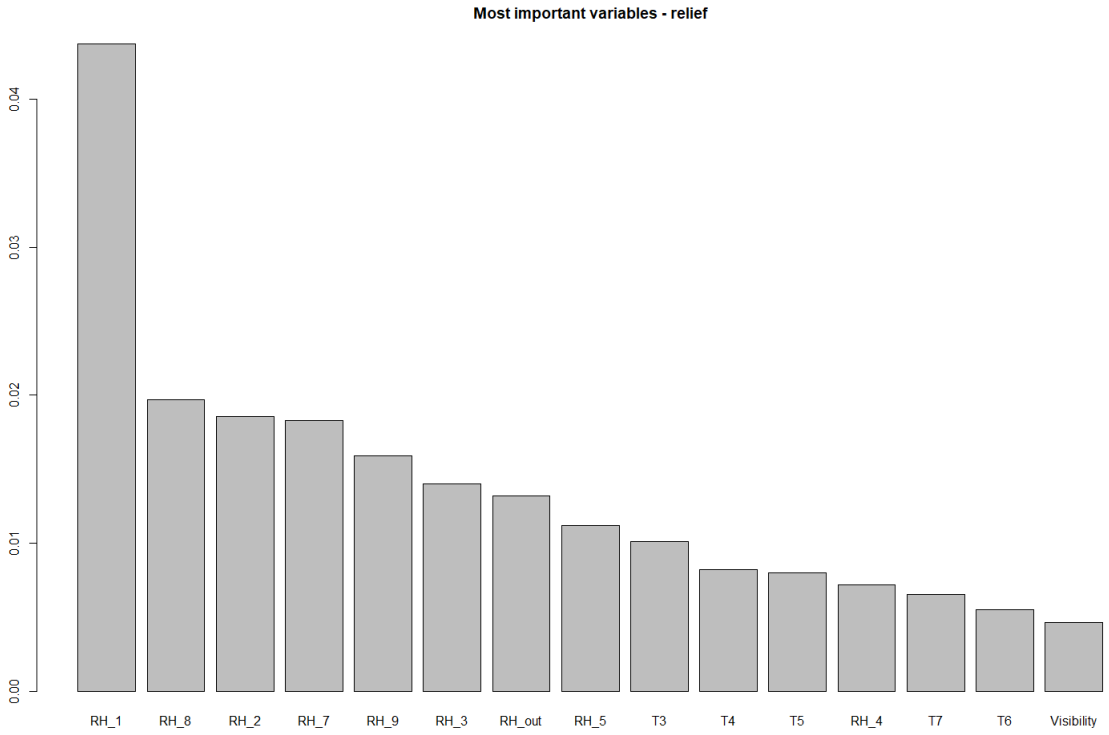


Rysunek 3.1: Wyniki selekcji z wykorzystaniem pakietu *randomForest*; cel: lights

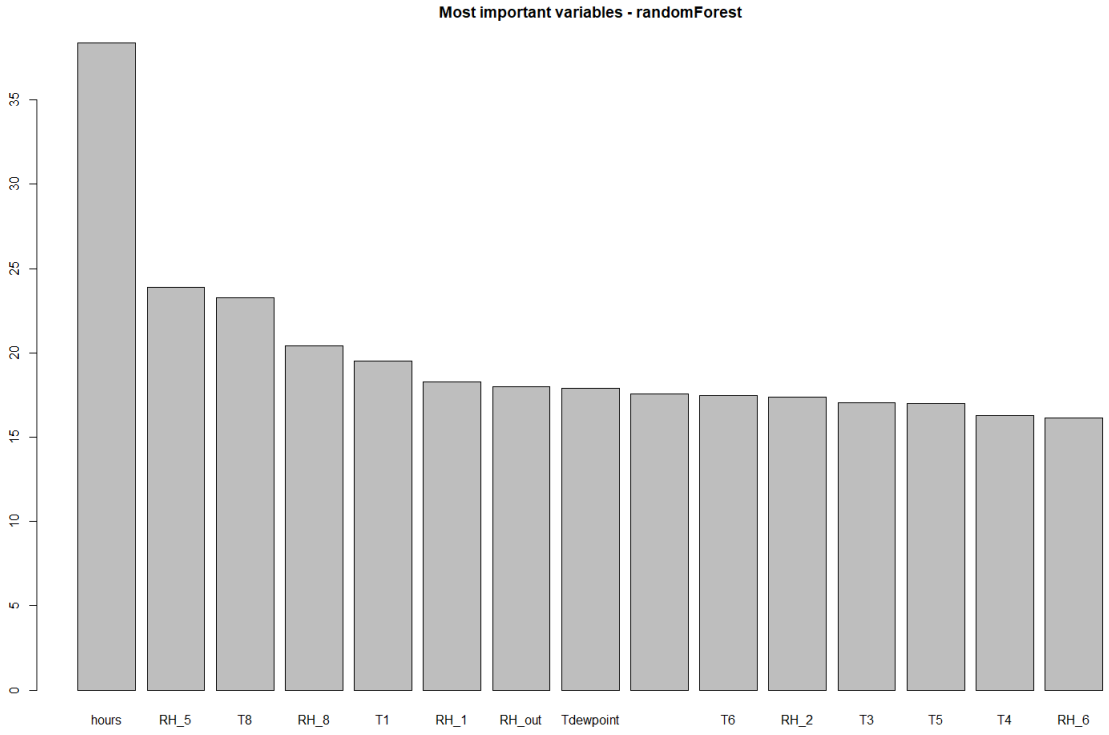


Rysunek 3.2: Wyniki selekcji z wykorzystaniem funkcji *simple.filter*; cel: lights

4. Konstrukcja i ocena modeli

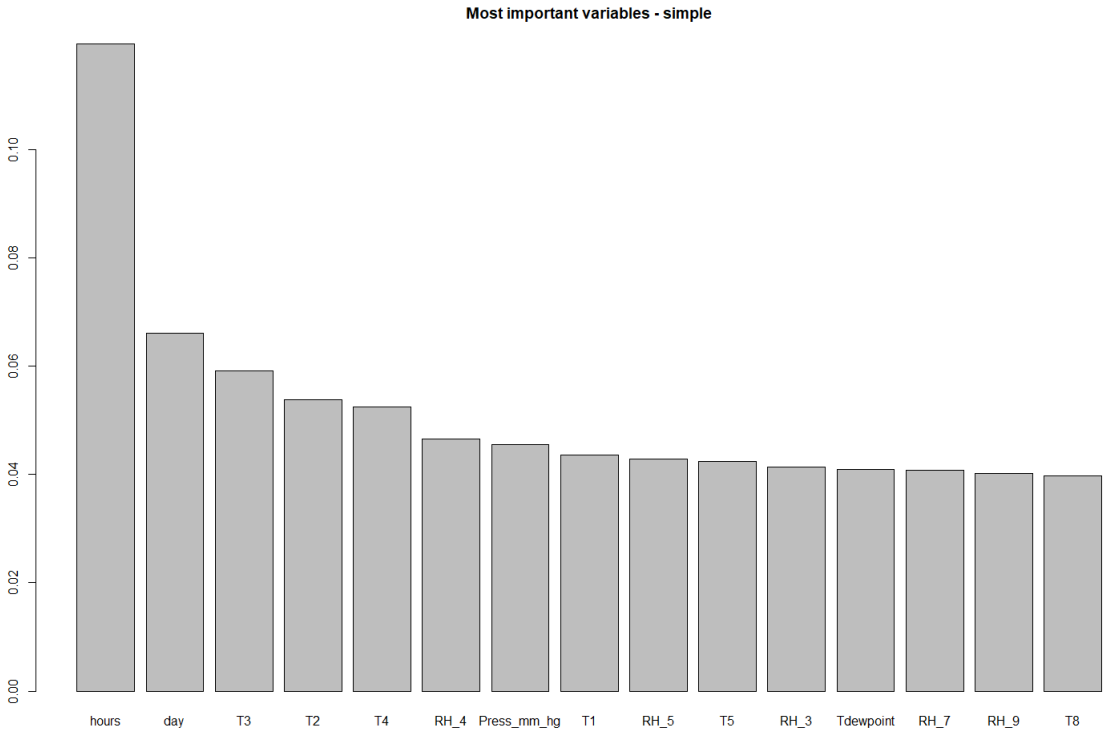


Rysunek 3.3: Wyniki selekcji z wykorzystaniem funkcji *rrelief.filter*; cel: lights

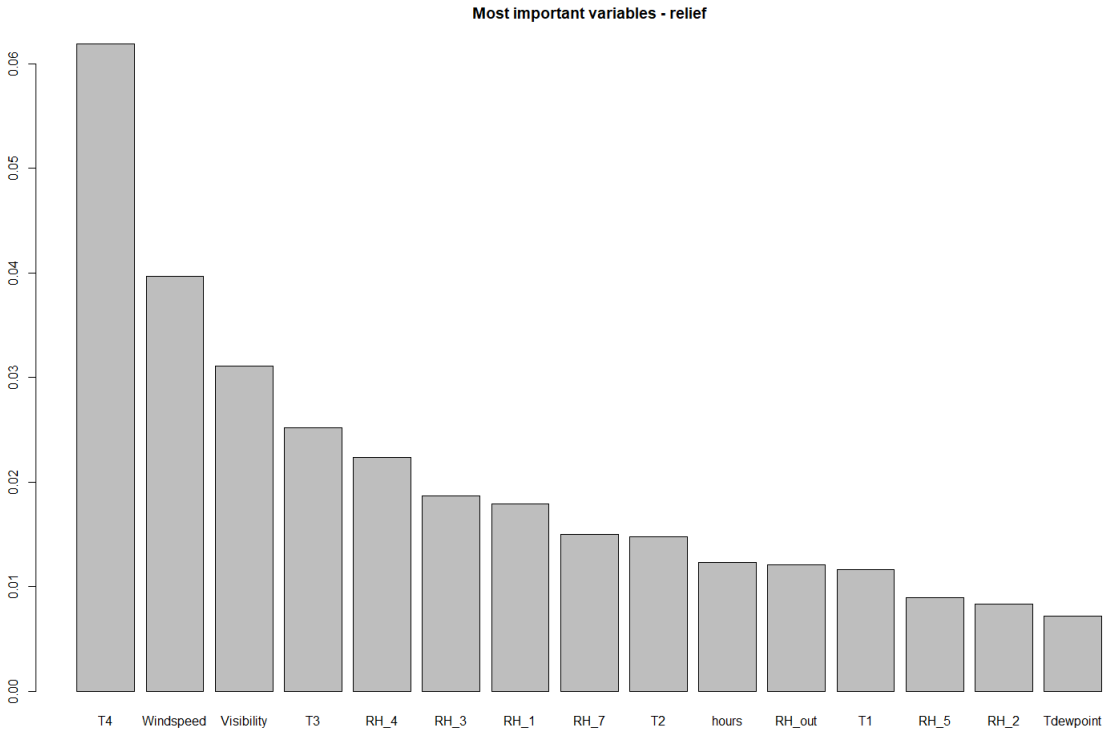


Rysunek 3.4: Wyniki selekcji z wykorzystaniem pakietu *randomForest*; cel: Appliances

data.frame. Algorytm na podstawie średniej lub mediany wyznacza współczynniki funkcji liniowej tj. współczynnik kierunkowy i wyraz wolny. Następnie od każdego atrybuty wyznacza wagę współczynniku kierunkowego oraz jeden wyraz wolny. Zaletą tego algorytmu jest jego prostota i szybkość



Rysunek 3.5: Wyniki selekcji z wykorzystaniem funkcji *simple.filter*; cel: Appliances



Rysunek 3.6: Wyniki selekcji z wykorzystaniem funkcji *rrelief.filter*; cel: Appliances

działania, niewątpliwą wadą jest fakt, że większość procesów zachodzących w świecie nie da się opisać za pomocą liniowych zależności.

Następnie dokonano modelowania za pomocą drzew regresji (algorytm *rpart()*). Funkcja buduje model rekurencyjnie dzieląc zbiór na mniejsze i wyznaczając dla nich średnią. Dla metody „anova" dedykowanej do zadania regresji kryterium podziału wyznaczone jest na podstawie sum kwadratów dla danego węzła i dla jego potomnych. Algorytm może przyjąć także argumenty na minimalną liczbę podziałów *minsplit* oraz maksymalną głębokość drzewa *maxdepth* (czyli długość pomiędzy korzeniem a liśćmi).

Pojedyncze modele drzew regresji zazwyczaj cierpią z powodu wysokiej wariancji – jedną z metod jej redukcji jest tzw. Bagging (**B**ootstrap **agg**regat**ing**). W ramach tej metody tworzonych jest pewna liczba zbiorów "bootstrapowych". Dla każdego z tych zbiorów tworzy się nieprzycięte drzewo regresji. Następnie uśrednia sie każdy z tych modeli, zmniejszając wariancję i redukując zbytnie dopasowanie. Bagging może zostać zrealizowany za pomocą pakietu *ipred* lub *caret*. Zasada działania modelowania przy pomocy tych pakietów jest podobna, pakiet *caret* pozwala natomiast na łatwą analizę istotności atrybutów. Celem porównania wyników realizowanych przez oba algorytmy, zdecydowano się na użycie ich obu.

Modele oparte o drzewo regresji cierpią z powodu faktu, że w liściach, które reprezentują pewny podzbiór przestrzeni (na której buduje się model), wynikiem jest pojedyncza liczba. Przykładowo dla zależności celu od jednego atrybutu można to funkcja reprezentująca model może być nieciągła i stworzona z odcinków o zerowym współczynniku kierunkowym. Dużo lepiej byłoby aproksymować tę zależność funkcją kawałkami liniową. W tym celu wykorzystuje się regresję kawałkami liniową (*grow.modtree* z pakietu *dmr.regtree*). Za pomocą listy *plr_args* ograniczono głębokość drzewa do 10 oraz wymuszono co najmniej 2 podziały.

Modele opisane w tym podrozdziale zostały ocenione za pomocą procedury k-krotnej walidacji krzyżowej z wykorzystaniem miar jakości opisanych dalej.

4.2. Miary jakości

Dla zbudowanych modeli oblicza się następujące miary jakości:

1. CC - współczynnik korelacji liniowej Pearsona

$$CC = \frac{cov(P,A)}{var(P) \cdot var(A)}$$

2. MSE - błąd średniokwadratowy

$$MSE = \frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{n}$$

3. RMSE - pierwiastek z błędu średniokwadratowego

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{n}}$$

4. MAE - średni błąd względny

$$MAE = \frac{|p_1 - a_1| + ... + |p_n - a_n|}{n}$$

5. RSE - względny błąd kwadratowy

$$RSE = \frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + ... + (a_n - \overline{a})^2}$$

6. RRSE - pierwiastek ze względnego błędu kwadratowego

$$RRSE = \sqrt{\frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + ... + (a_n - \overline{a})^2}}$$

7. RAE - błąd względny

$$RAE = \frac{|p_1 - a_1| + ... + |p_n - a_n|}{|a_1 - \overline{a}| + ... + |a_n - \overline{a}|}$$

TODO tutaj ładnie się wpasuje co po kolei w kodzie poszło

4.3. Procedury oceny

Aby móc ocenić model pod względem przydatności zastosowano metodę k-krotnej walidacji krzyżowej. Kod zawarto w funkcji *model_eval()*, która stanowi . Zbiór testowy jest dzielony losowo na k podzbiorów równej wielkości. W kolejnych iteracjach każdy ze zbiorów jest traktowany jako zbiór testowy, podczas gdy na reszcie danych buduje się model. Następnie modele są uśredniane i następuje predykcja. Po predykcji modelu na zbiorze testowym wyznacza się miary jakości opisane w 4.2.

4.4. Wyniki walidacji krzyżowej

Wyniki 10-krotnej walidacji krzyżowej zostały przedstawione w tabeli 4.1. Przedstawia ona miary jakości wyznaczone za pomocą tej procedury dla każdego algorytmu. Zauważa się przewagę metod *bootstrapowych* nad innymi. Błędy i współczynniki korelacji dla tych metod osiągają najmniejsze wartości. Ewentualne różnice pomiędzy wynikami algorytmów z pakietów *ipred* i *caret* można tłumaczyć losowością procesu tworzenia ich modelu i/lub procedury oceny. Zauważa się też różnicę pomiędzy predykcją celu *Appliances* a *lights*. Wydaje się to byc zgodne z intuicją — zwykle światła włącza się w dzień, a wyłącza w nocy. Z kolei różne urządzenia AGD stosuje się w różnych okresach, więc znalezienie szczególnej zależności może być skomplikowanym zadaniem.

Tablica 4.1: Walidacja krzyżowa — porównanie parametrów

Appliances	**MSE**	**RRSE**	**MAE**	**RMSE**	**RAE**	**CC**	**RSE**
lm()	14914.800	2.5702027	75.20398	122.12616	1.9553172	0.3425775	6.6059421
rpart()	11646.595	1.2051757	58.13124	107.91939	1.0320860	0.5698647	1.4524484
ipred	9016.412	1.3234627	53.02221	94.95479	1.0308015	0.6983742	1.7515534
PLR	266453.137	0.9929399	78.68274	516.19099	0.8137909	0.1532704	0.9859296
caret	8958.720	1.2985949	52.82693	94.65051	1.0113670	0.6991588	1.6863487
lights	**MSE**	**RRSE**	**MAE**	**RMSE**	**RAE**	**CC**	**RSE**
lm()	63.61679	1.6605096	5.640188	7.976013	1.4319283	0.50520939	2.7572920
rpart()	38.27473	0.8563671	3.815467	6.186657	0.7194938	0.74393671	0.7333646
ipred	30.00860	0.8274880	3.578404	5.478011	0.7473722	0.81096864	0.6847364
PLR	7264.81589	0.9960550	9.868371	85.233889	0.7595684	0.09086541	0.9921255
caret	30.92202	0.8363734	3.593142	5.560757	0.7441449	0.80317142	0.6995204

5. Wnioski

TODO