Drzewa decyzyjne, lasy losowe, boosting

Wojciech Kosztyła, Wojciech Węgrzynek

Regresja

W problemie regresji przewidujemy medianę dochodu gospodarstwa domowego w danym hrabstwie w USA, na podstawie danych demograficznych.
Przeprowadziliśmy regresję za pomocą wszystkich trzech modeli, w trzech wariantach

1. Za pomocą wszystkich danych.
2. Za pomocą danych bez `Poverty`.
3. Za pomocą danych bez `Poverty` i `ChildPoverty`.

Dwa wyróżnione w ten sposób predyktory, odpowiadające odpowiednio za: procent społeczeństwa żyjącego poniżej linii ubóstwa i procent społeczeństwa, które stanowią dzieci żyjące poniżej linii ubóstwa, zostały wykluczone gdyż można by sądzić, że dodanie ich do zadaia regresji znacząco trywializuje zadanie.

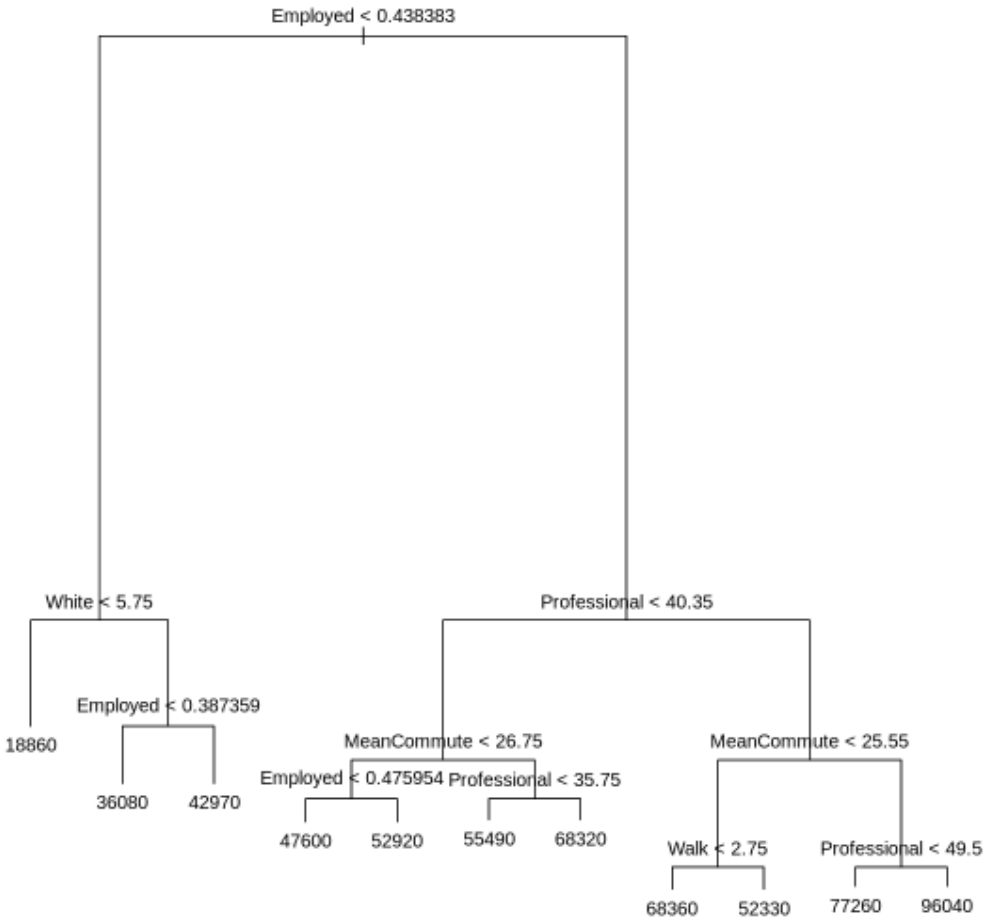
Poniżej przedstawione są rezultaty w postaci błędu średniego kwadratowego na zbiorze testowym.

	Wariant 1	Wariant 2	Wariant 3
Drzewa decyzyjne	42655376.9478501	53270654.6447646	63137067.5822155
Lasy losowe	18903917.5371139	23472895.473636	31618907.5631645
Gradient boosting	18786473.3673427	24048559.7170808	30097156.5188249

Jak widać, drzewa decyzyjne sprawdzają się znacząco gorzej w tym zadaniu od obydwu pozostałych metod. Co ciekawe, lasy losowe dają wyniki porównywalne z boostingiem we wszystkich wariantach.
Poza tym wyniki w wariantach różnią się znacząco, co wskazuje na to że problem w wariancie trzecim jest rzeczywiście znacząco trudniejszy.

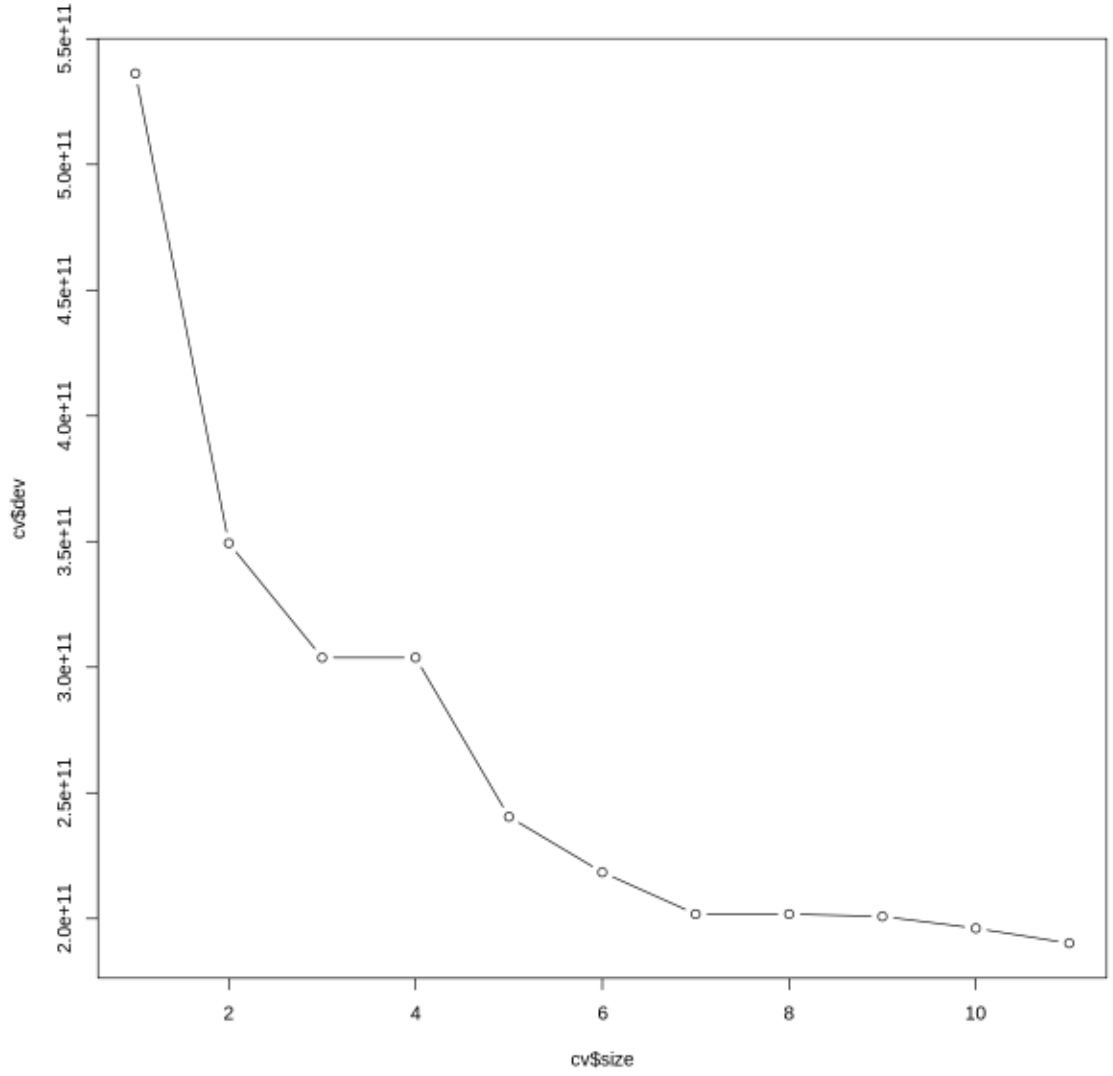
Drzewa decyzyjne

Poniżej zamieszczamy rysunek drzewa decyzyjnego uczonego na na całym zbiorze danych w wariancie 3.



Najważniejszym predyktorem jest `Employed`, czyli procent populacji pracującej - występuje ona w korzeniu i w dwóch innych węzłach. Wśród ważnych predyktorów mamy również: `White`, `Professional`, `MeanCommute` i `Walk`.

Sporządziliśmy również wykres wyniku walidacji krzyżowej w zależności od rozmiaru drzewa.



Widać na nim, że zależność jest malejąca, a więc model raczej nie przeucza.

Boosting

Poniżej zamieszczamy tabelę z summary modelu w wariancie 3

	var	rel.inf
	<chr>	<dbl>
Employed	Employed	35.2605050
Professional	Professional	10.7944436
MeanCommute	MeanCommute	8.7551421
Citizen	Citizen	3.9347048
Asian	Asian	3.4598327
Hispanic	Hispanic	3.3164477
TotalPop	TotalPop	3.0757102
White	White	2.9158485
Walk	Walk	2.8302426
Service	Service	2.2796937
Production	Production	2.0341382
Unemployment	Unemployment	1.9868791
Men	Men	1.7998582
Construction	Construction	1.7567978
SelfEmployed	SelfEmployed	1.7061039
Black	Black	1.4851645
Women	Women	1.4652731
Carpool	Carpool	1.2540739
Office	Office	1.2297672
PrivateWork	PrivateWork	1.2056622
Drive	Drive	1.1569529
Transit	Transit	1.1474548
Native	Native	1.1383017
PublicWork	PublicWork	1.1062669
OtherTransp	OtherTransp	1.0513406
WorkAtHome	WorkAtHome	1.0141583
Pacific	Pacific	0.4557977
FamilyWork	FamilyWork	0.3834379

.

Widzimy, że w boostingu najważniejsze 3 predyktory pojawiły się również w przypadku drzew decyzyjnych, niosą one ze sobą znacząco większą ilość informacji niż dwa pozostałe.

Klasyfikacja

W problemie klasyfikacji przewidujemy zwycięzcę wyborów prezydenckich w USA w danym hrabstwie w USA, na podstawie danych demograficznych.
Przeprowadziliśmy regresję za pomocą wszystkich trzech modeli, tym razem w jednym wariancie (wszystkie predyktory).

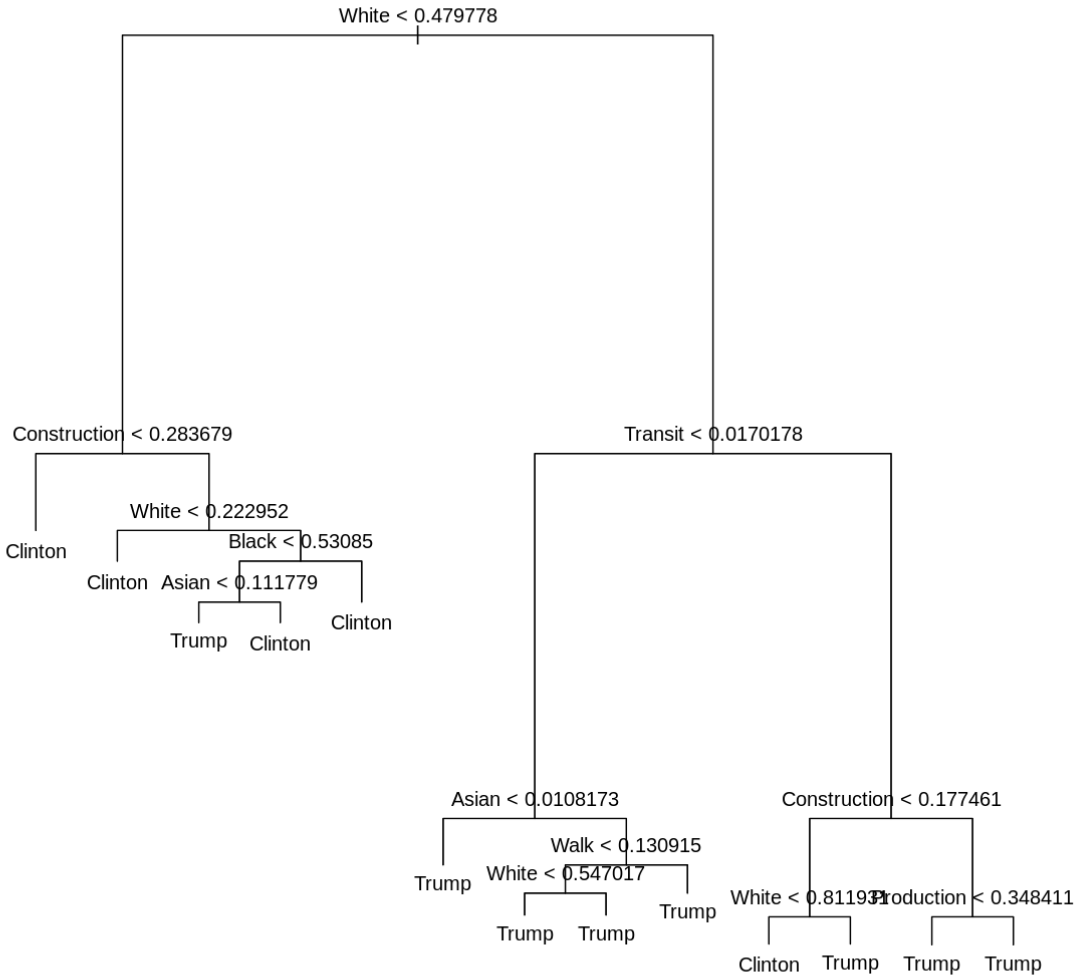
Do porównania wyników również wykorzystaliśmy błąd średni kwadratowy na zbiorze testowym, jednak w tym wypadku wartościom `"Trump"` i `"Clinton"` przypisane zostały wartości liczbowe `1` i `0`. Z tego powodu błędy będą z zakresu `[0, 1]`.

Drzewa decyzyjne	`0.070247934892373`
Lasy losowe	`0.0526992287917738`
Gradient boosting	`0.0491526356577975`

Łatwo można zauważyć, że drzewa decyzyjne wypadły tutaj gorzej od pozostałych modeli. Lasy losowe i Gradient boosting mają wyniki bardzo zbliżone do siebie, co poparte jest wybranymi przez nie predyktorami.

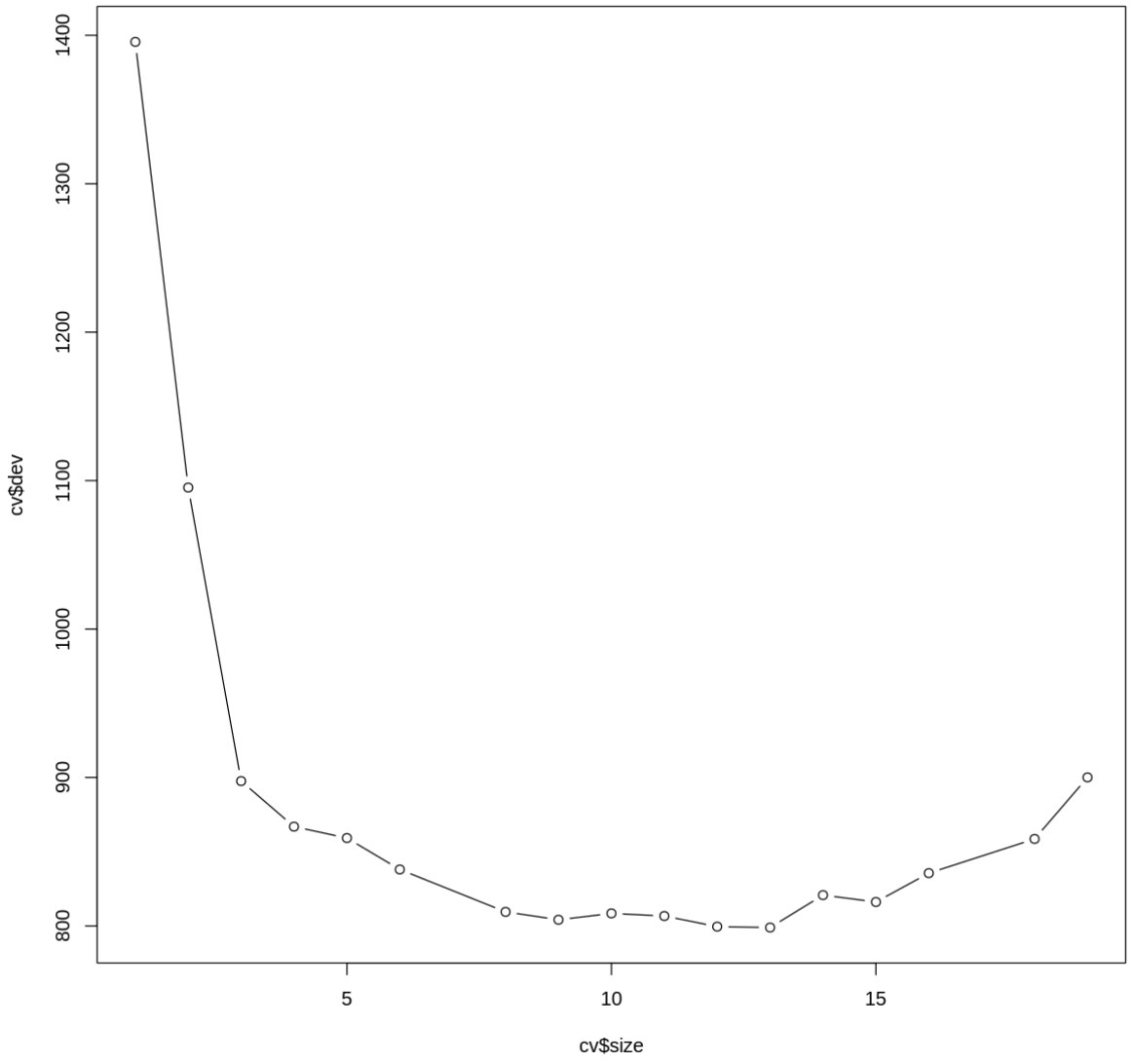
Drzewa decyzyjne

Poniżej zamieszczamy rysunek drzewa decyzyjnego.



Najważniejszym predyktorem jest `White`, czyli procent populacji rasy białej - występuje ona w korzeniu i w trzech innych węzłach. Wśród ważnych predyktorów mamy również: `Transit`, `Construction`, `Asian` i `Black`.

Sporządziliśmy również wykres wyniku walidacji krzyżowej w zależności od rozmiaru drzewa.



Zależność jest w większości malejąca, lecz przy 14 zmienia się w rosnącą, więc model może się przeuczac.

Boosting

Poniżej zamieszczamy tabelę z summary modelu

White	White	18.9343372
Transit	Transit	12.5511979
Construction	Construction	7.6364801
Asian	Asian	7.1355293
Black	Black	5.2820932
TotalPop	TotalPop	4.9057761
Service	Service	3.6075291
Employed	Employed	3.3909899
Citizen	Citizen	3.0248177
Hispanic	Hispanic	2.7309096
Unemployment	Unemployment	2.5530055
Professional	Professional	2.1431625
Production	Production	2.0076973
Walk	Walk	1.9901130
SelfEmployed	SelfEmployed	1.9439380
WorkAtHome	WorkAtHome	1.7152637
IncomePerCap	IncomePerCap	1.7123590
MeanCommute	MeanCommute	1.6688233
Drive	Drive	1.5947604
Income	Income	1.5685536
Poverty	Poverty	1.4032245
Office	Office	1.3560840
OtherTransp	OtherTransp	1.2771196
PublicWork	PublicWork	1.2669334
PrivateWork	PrivateWork	1.1372929
Native	Native	1.1081729
Carpool	Carpool	0.9927589
Men	Men	0.9706534
Women	Women	0.9186637

Widzimy, że w boostingu najważniejsze 4 predyktorów pojawiło się również w przypadku drzew decyzyjnych. Przy następnych predyktorach, ilość niesionych informacji znacząco spada.