Modelowanie statystyczne w ZWM

Szeregowanie najlepiej rokujących spraw

Chmiela Bartosz, Melka Kamil

Uniwersytet Wrocławski

28 stycznia 2021

Contents

Parametry analizy

- Horyzont 6 miesięcy od importu,
- modele zbudowane per sprawa,
- poszukujemy "dobrych"klientów,
- dobry – dokonał wpłaty w ciągu 6M,
- szeregujemy zbiór wg prawdp. zapłaty.

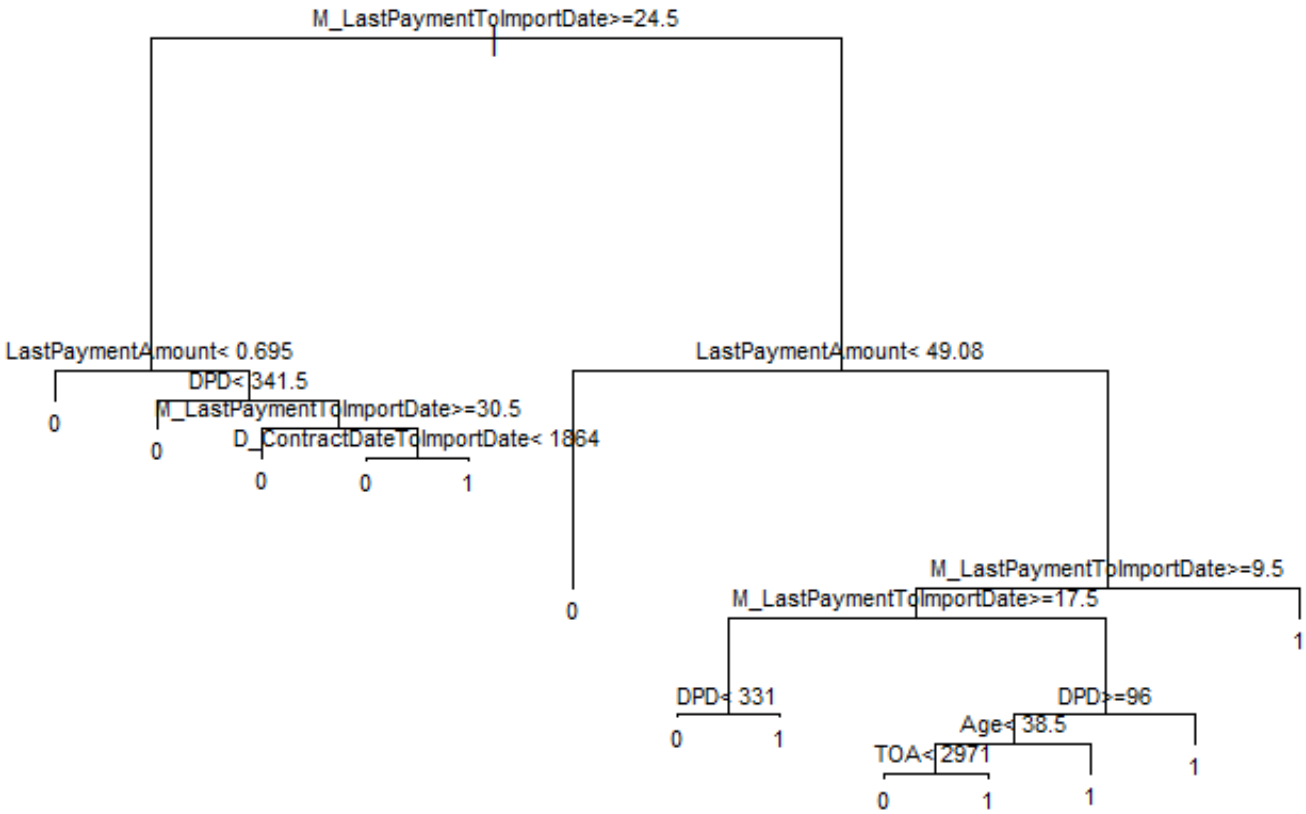
Imputacja, odstające wartości

- Uzupełnianie eksperckie Interests,
- uzupełnianie medianą wielu zmiennych,
- uzupełnianie rozkładem Land,
- zachowanie zależności między Land, a GDP i MeanSalary,
- uzupełnianie zmiennych 0/1 za pomocą rozkładu jednostajnego,
- usunięcie odstających wartości dla LoanAmount, DPD, LastPaymentAmount,
- zachowujemy 99% danych.

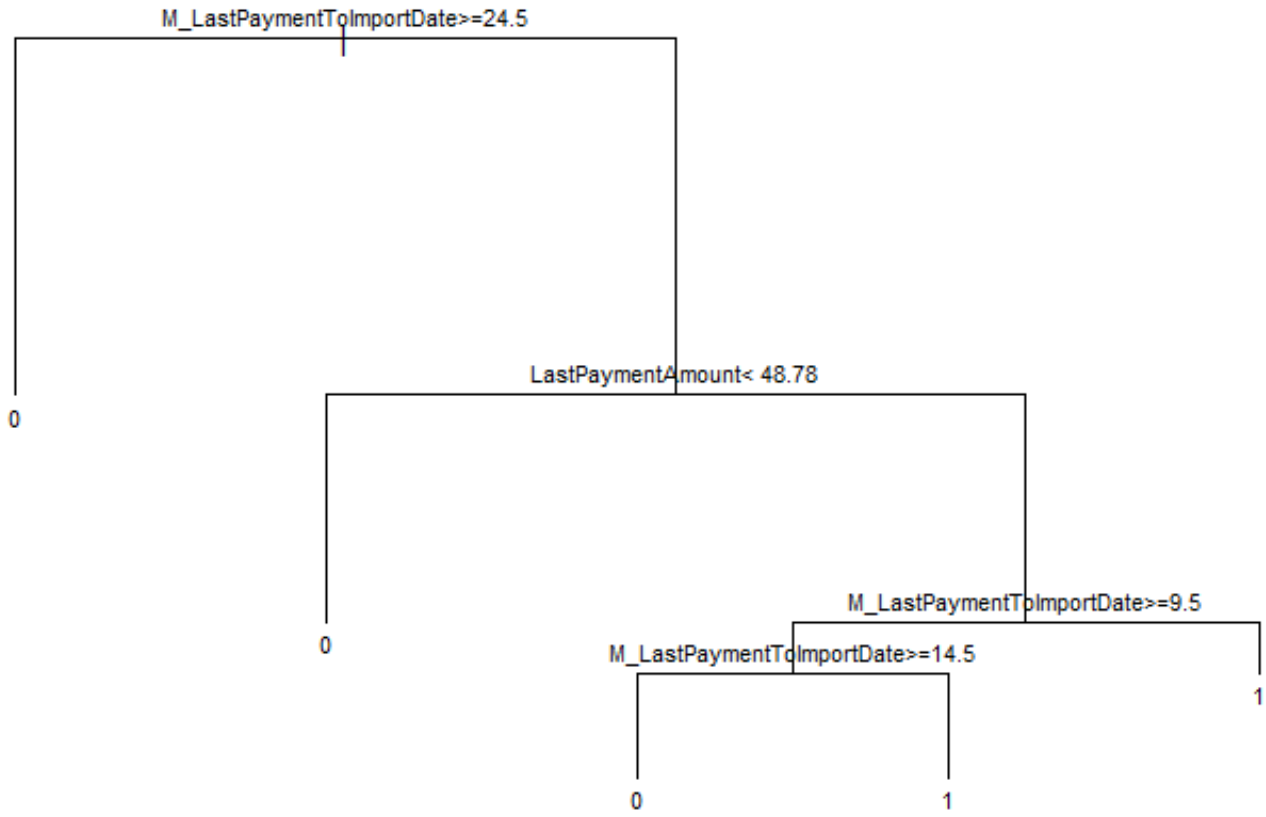
Modele statystyczne

- Drzewo binarne oraz lasy losowe,
- modele regresji GAM,
- logloss jako miara jakości modeli,
- 500 sztucznych porfteli, losowanych ze zwracaniem,
- $|w - p|/p$ miara jakości porfeli,
- zbiór treningowy, walidacyjny, testowy.

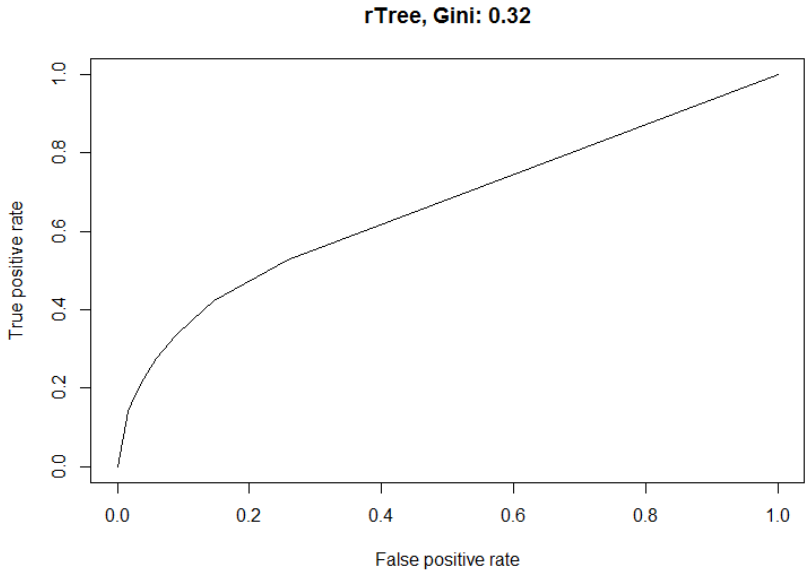
Drzewo binarne

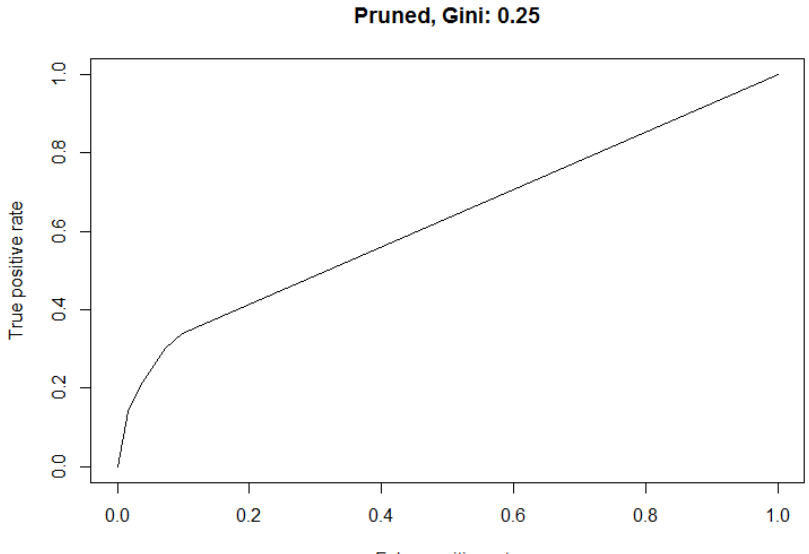


Przycięte drzewo binarne



Gini


rTree, Gini: 0.32
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


Pruned, Gini: 0.25
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Lasy losowe

- 80 modeli, 4 zmienne parametry,
- ntree $\in \{500, 1000\}$ - ilość zbudowanych drzew w lesie,
- mtry $\in \{13, 10, 5, 3\}$ - ilość zmiennych losowanych do drzewa,
- nodesize $\in \{1, 5, 20, 100, 200\}$ - minimalna ilość liści, (?)
- cutoff $\in \{0.5, 0.26\}$ - odcięcie od którego klasyfikujemy 1.

Rezultaty

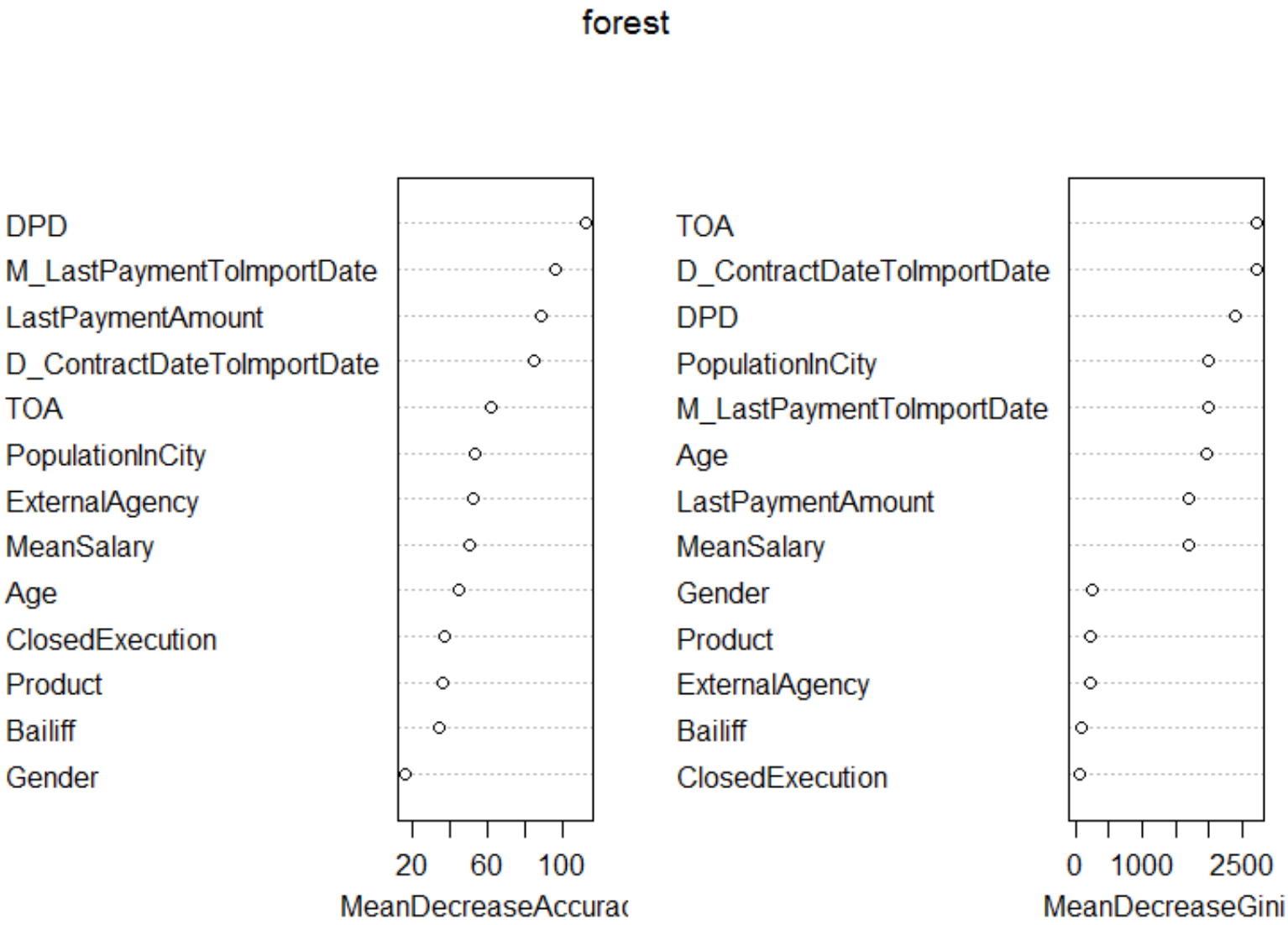
	ntree	mtry	n	cutoff	tr. size	accuracy	gini ind	logloss	|w-p|/p
54	1000	10	5	0.26	6098.377	0.7542	0.3732	1.56195	0.55798
53	1000	10	5	0.50	6097.522	0.7646	0.3736	1.56255	0.55809
43	1000	13	5	0.50	5938.813	0.7646	0.3705	1.56265	0.55847
44	1000	13	5	0.26	5939.298	0.7539	0.3706	1.56328	0.55836
13	500	10	5	0.50	6097.598	0.7648	0.3743	1.56331	0.55812
14	500	10	5	0.26	6097.536	0.7543	0.3729	1.56427	0.55864

Rysunek: 5 najlepszych lasow wg logloss

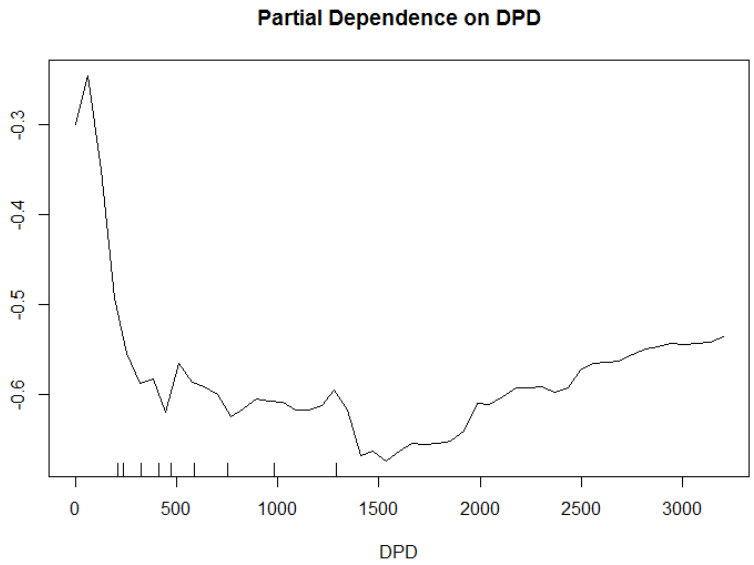
	ntree	mtry	n	cutoff	tr. size	accuracy	gini ind	logloss	|w-p|/p
71	1000	3	1	0.50	8495.513	0.7698	0.3906	1.87935	0.55705
33	500	3	5	0.50	5985.434	0.7696	0.3889	1.82809	0.55706
32	500	3	1	0.26	8451.252	0.7516	0.3895	1.89236	0.55708
73	1000	3	5	0.50	5988.79	0.7697	0.3912	1.82580	0.55717
22	500	5	1	0.26	9377.174	0.7519	0.3812	1.64998	0.55727
31	500	3	1	0.50	8438.452	0.7695	0.3905	1.89523	0.55734

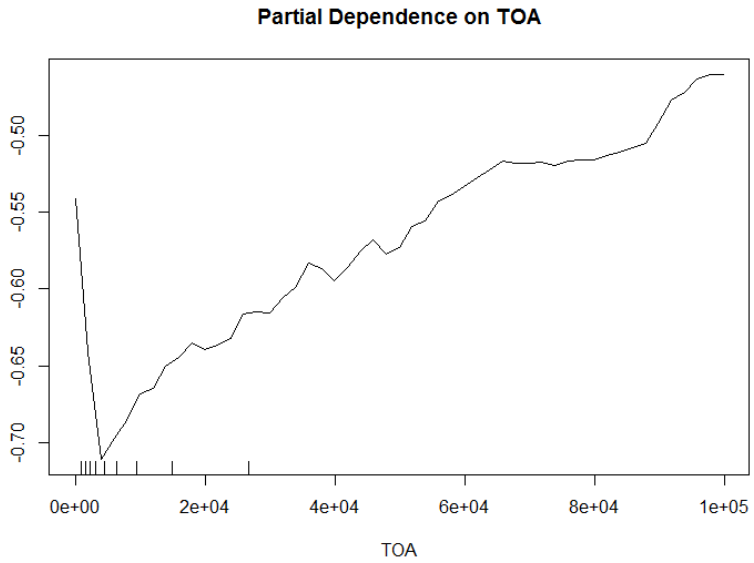
Rysunek: 5 najlepszych lasow wg odchylenia

Zmienne w najlepszym lesie



Partial dependance plots





Testowy błąd

logloss	|w-p|/w	|w-p|/p
1.871848	2.136542	0.6012786