Raport 2A

Dawid Sikorski 291951

W celu przewidzenia, czy klienci banku z którymi kontaktowano się podczas kampanii telemarketingowej decydowali się na założenie lokaty. Do realizacji zadania wykorzystano dane zebrane w pliku ‘bank_marketing_training.txt’, w którym znajduje się 26874 obserwację z 20 predyktorami oraz zmienną celu ‘response’. Stosunek odpowiedzi wynosił 11:89.

```
> proporcja
  no  yes
0.89 0.11
```

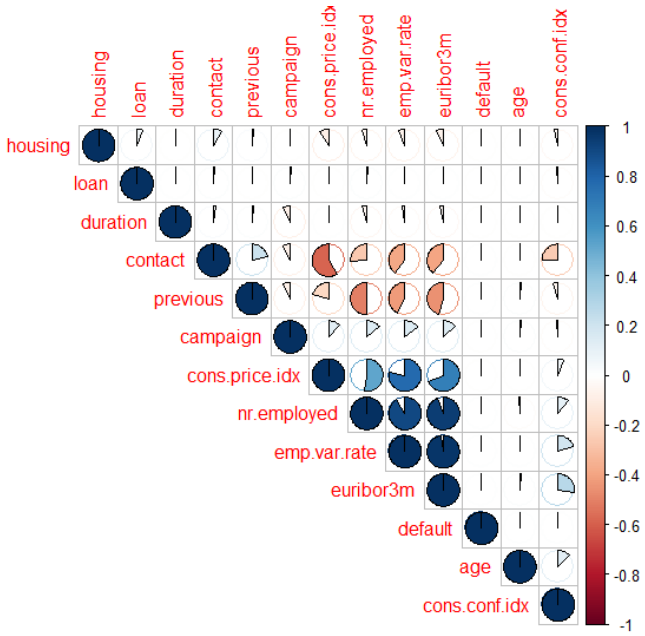
W celu klasyfikacji wykorzystałem algorytm C5.0 oraz Las losowy.

	response	
days_since_previous	no	yes
0	5	7
1	15	6
2	15	25
3	98	183
4	36	43
5	9	19
6	79	184
7	8	28
8	4	9
9	19	20
10	15	17
11	9	14
12	24	18
13	5	15
14	8	9
15	4	9
16	3	4
17	5	0
18	2	2
19	1	0
20	1	0
22	1	1
25	0	1
26	0	1
27	0	1
999	23520	2372

Zestawienie danych days_since_previous z response

Największe braki danych zostały zaobserwowane w kolumnie ‘days_since_previous’, która na podstawie macierzy korelacji została wykluczona.

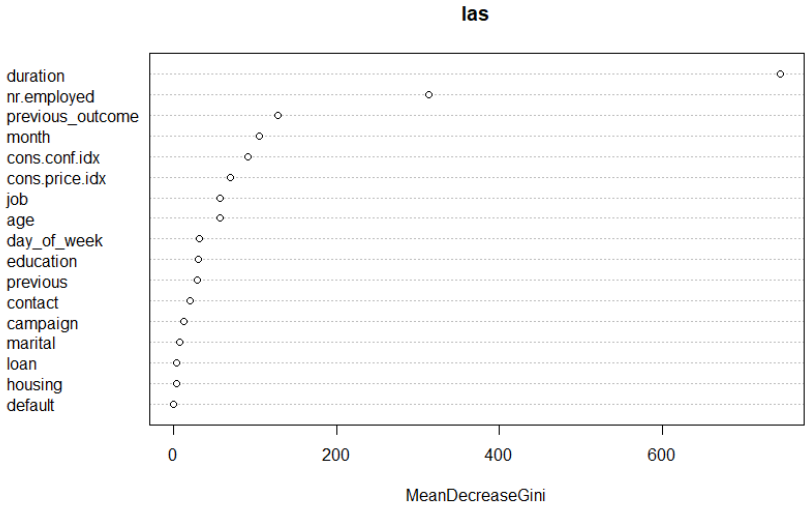
Na początku kolumny ‘default’, ‘housing’, ‘loan’, zostały zamienione z odpowiedzi 1 dla ‘yes’, a 0 dla ‘no’ oraz ‘unknown’. Kolumna ‘contact’ otrzymała 1 jeżeli był to kontakt ‘cellular’ a 0 dla ‘telephone’. Pozostałe zmienne jakościowe zostały sfaktoryzowane, tak aby można ich było użyć w algorytmie C5.0 oraz Lesie losowym.



Macierz korelacji

Na podstawie której, pozwoliłem sobie wykluczyć ‘emp.var.rate’ i ‘euribor3m’ na silną korelację z ‘nr.employed’.

Następnie na podstawie pozostawionych predyktorów powstał las losowy oraz drzewo C5.0, które posłużyły do dalszej eliminacji, aby uprościć modele.



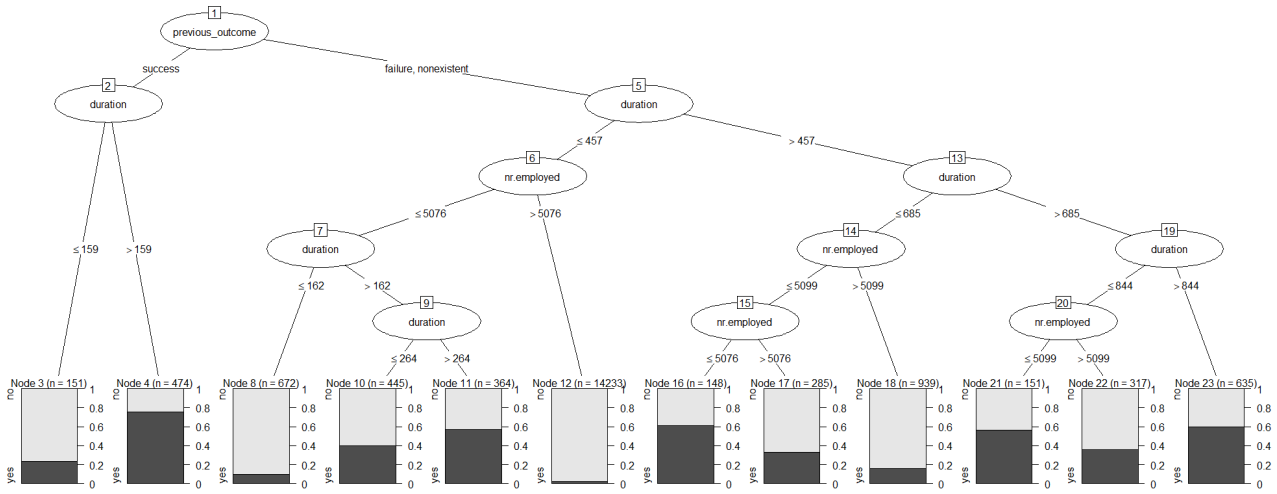

```
Attribute usage:

100.00% duration
100.00% previous_outcome
 90.82% nr.employed
  7.87% month
  5.86% contact
```

Ważność predyktorów w lesie losowym (z lewej) oraz w C5.0 (z prawej)

Na tej podstawie w prostszym modelu pozwoliłem sobie wybrać wyłącznie predyktory: ‘duration’, ‘previous_outcome’, ‘nr.employed’, ‘month’. Ponieważ, one odgrywały główną rolę w obu algorytmach.

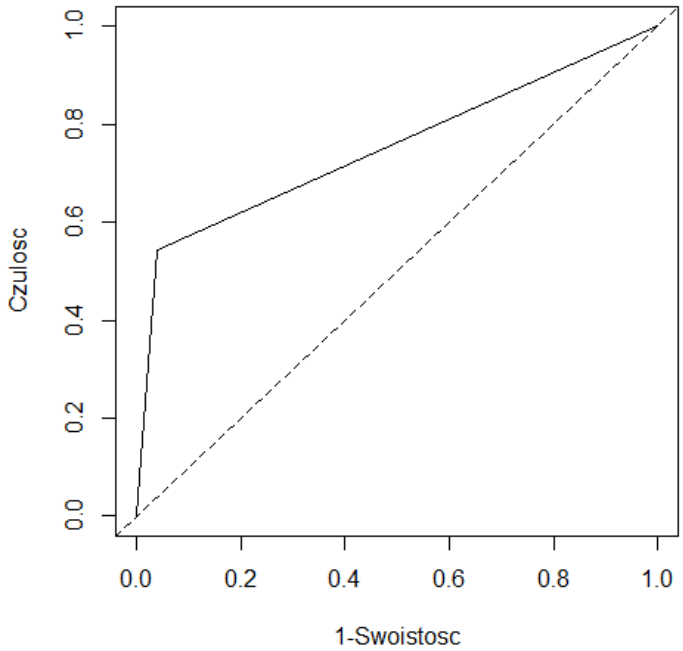
Drzewo C5.0 zostało zbudowane w oparci o predyktory: ‘duration’, ‘previous_outcome’, ‘nr.employed’, ‘month’ oraz zmienną celu ‘response’. Dodatkowo ustawiony został parametr minCase na 75.



Drzewo C5.0

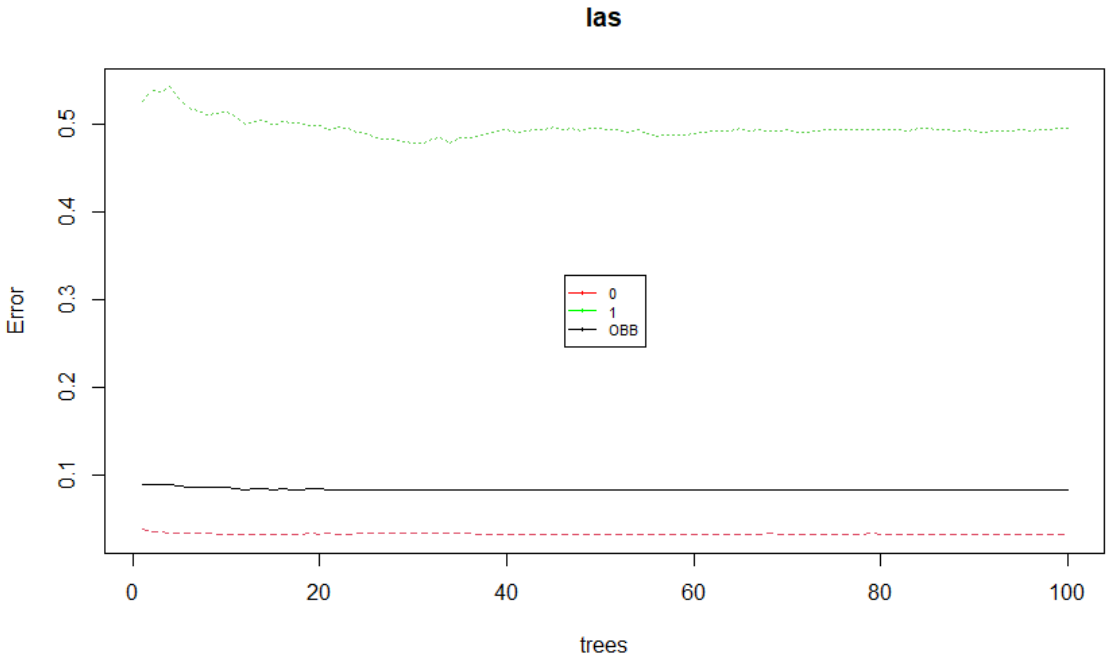
Najważniejszym predyktorem podziału został ‘previous_outcome’. Następnie drzewo w pierwszych liściach skupiło się na podziale względem ‘duration’. Dalszy podział zależał już wyłącznie od ‘duration’ oraz ‘nr.employed’.

Z otrzymanego drzewa, możemy zaobserwować iż mamy jeden czysty liść (numer 12) zawierający aż 14233 obserwacji. Liście 3, 4, 8 oraz 18 odpowiadają w mniej więcej 80% poprawnie. Ogólny błąd klasyfikacji podczas nauki wynosił 8,4%.



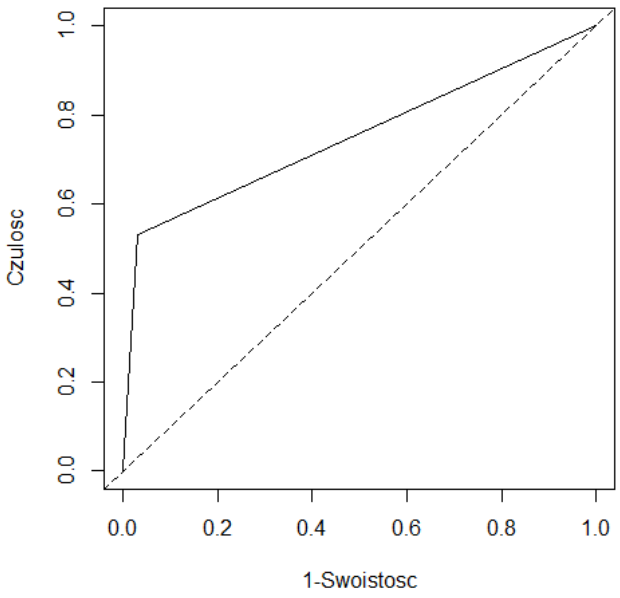
Krzywa ROC dla drzewa C5.0 dla danych treningowych

Las losowy został zbudowane w oparci o predyktory: ‘duration’, ‘previous_outcome’, ‘nr.employed’, ‘month’ oraz zmienną celu ‘response’. Dodatkowo ustawiony został parametr ‘nodesize’ na 75 oraz ‘ntree’ na 100.



Wykres błędy klasyfikacji

Warto tu wspomnieć, iż pomimo ustawionego ziarna losowości podczas tworzenia lasu losowego można otrzymać różne wyniki. Ogólny błąd podczas uczenia wynosił od 8.3% do 8.5%.



Krzywa ROC dla lasu losowego dla danych treningowych

W celu sprawdzenia poprawności przetestowałem model dla zbioru treningowego oraz testowego otrzymując następujące wyniki:

```
> conf_mat_train_c50
       y_pred_c50_train
y_train    no   yes
     no  16104   650
     yes   938  1122
> conf_mat_train_las
       y_pred_las_train
y_train    no   yes
     no  16252   502
     yes   961  1099
> conf_mat_test_c50
      y_pred_c50_test
y_test   no  yes
   no  6833  299
   yes  447  481
> conf_mat_test_las
      y_pred_las_test
y_test   no  yes
   no  6882  250
   yes  472  456
```

Macierze klasyfikacji

Name	Czulosc	Trafnosc	Swoistosc
C5.0 Train	0.5446602	0.9155948	0.9612033
C5.0 Test	0.5183190	0.9074442	0.9580763
Las Train	0.5223301	0.9221856	0.9713501
Las Test	0.4816810	0.9104218	0.9662086

Name	Czulosc	Trafnosc	Swoistosc
C5.0 Train	0.5446602	0.9155948	0.9612033
C5.0 Test	0.5183190	0.9074442	0.9580763
Las Train	0.5334951	0.9222388	0.9700370
Las Test	0.4913793	0.9104218	0.9649467

Czułość, Trafności i Swoistość dla C5.0 oraz Lasu

Możemy zaobserwować, iż modele się nie przeuczyły ze względu na podobne rezultaty. Porównując obrazki możemy zaobserwować, iż ponowne uruchomienie dało identyczne wyniki na tym samym ziarnie losowości dla drzewa C5.0 a odrobinę różniące się dla Lasu losowego.

Podsumowując, obydwa modele klasyfikują poprawnie. Jeżeli chodziłoby nam o swoistość lepszym wyborem było by las losowy, lecz w przypadku klasyfikacji potrzebujemy czułości, a algorytm C5.0 daje nam odrobinę lepszy wynik.