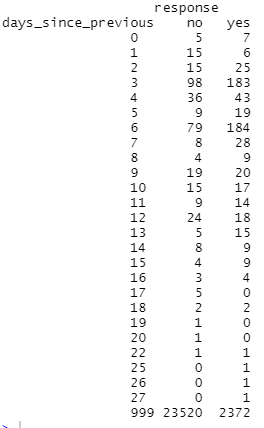
Raport 2A

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.



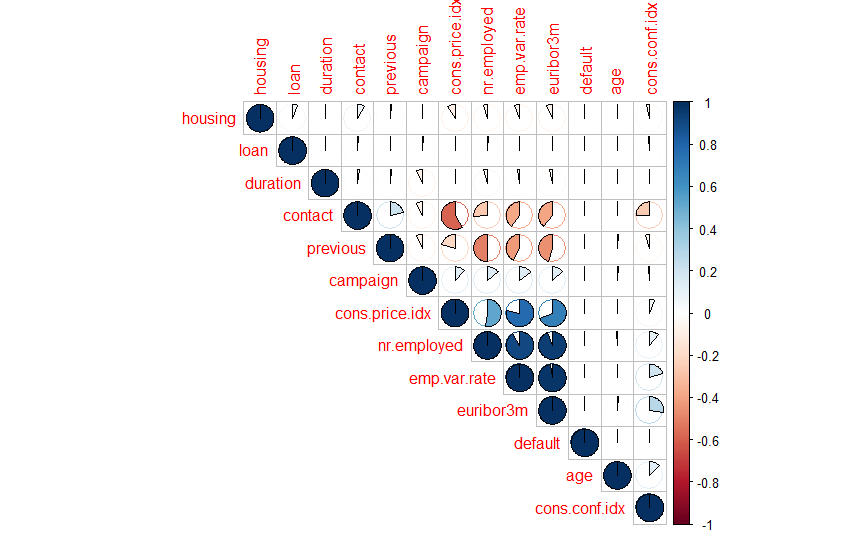
W celu klasyfikacji wykorzystałem algorytm C5.0 oraz Las losowy.



*Zestawienie danych days\_since\_previous z response*

Największe braki danych zostały zaobserwowane w kolumnie ‘days\_since\_previous’, która na podstawie macierzy korelacji została wykluczona.

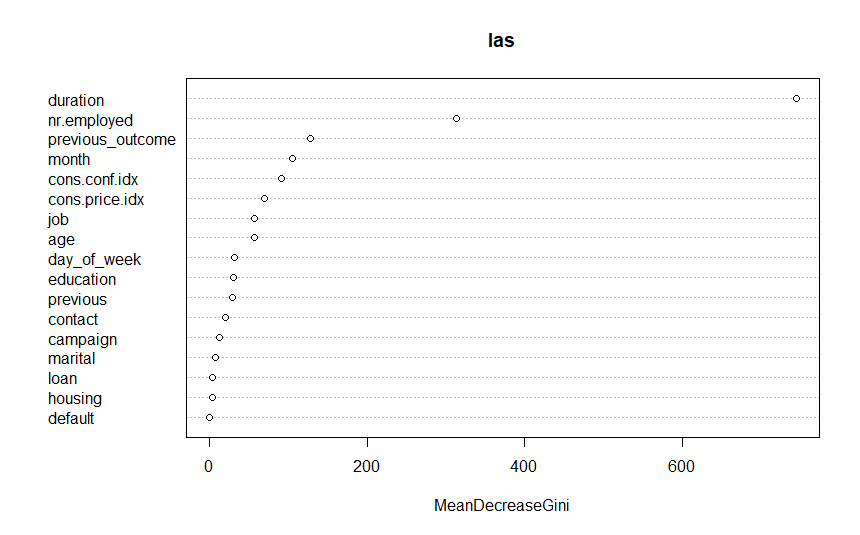
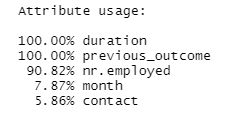
Na początku kolumny ‘default’, ‘housing’, ‘loan’, zostały zamienione z odpowiedzi 1 dla ‘yes’, a 0 dla ‘no’ oraz ‘unknown’. Kolumna ‘contact’ otrzymała 1 jeżeli był to kontakt ‘cellular’ a 0 dla ‘telephone’. Pozostałe zmienne jakościowe zostały sfaktoryzowane, tak aby można ich było użyć w algorytmie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



*Macierz korelacji*

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

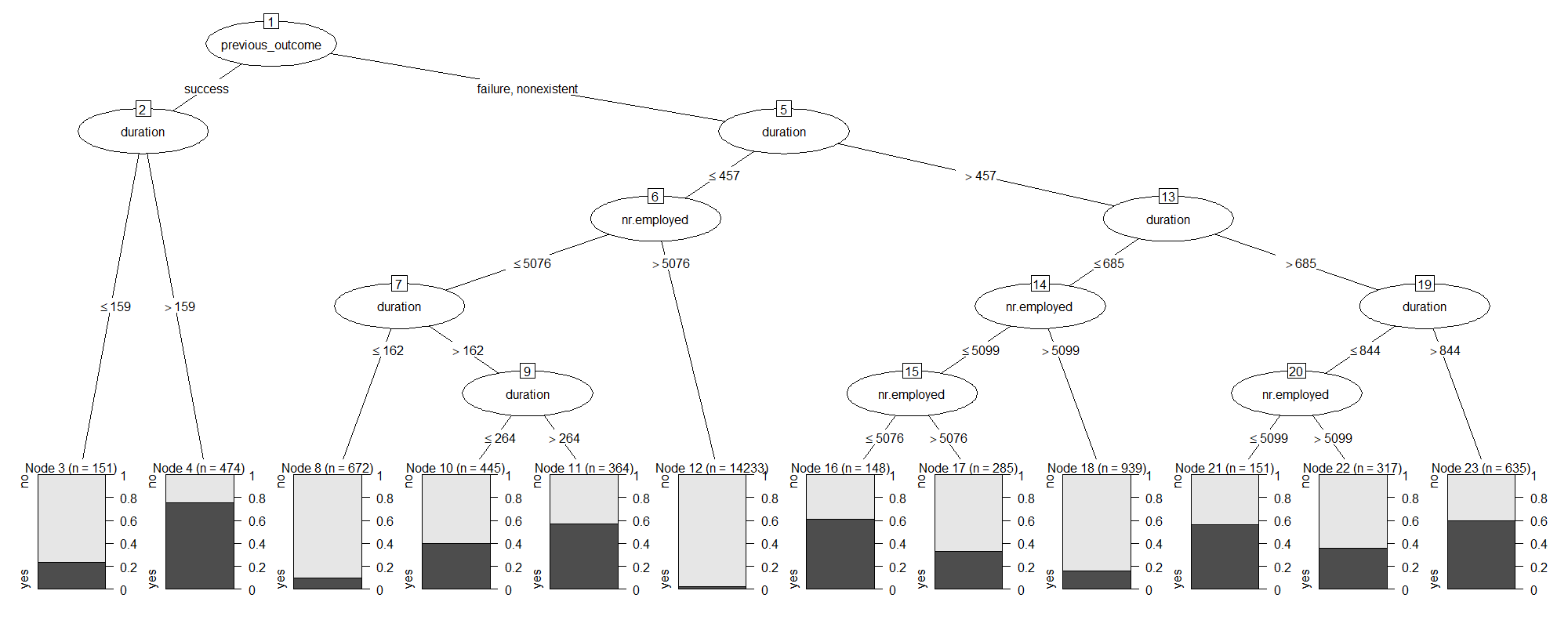
Następnie na podstawie pozostawionych predyktorów powstał las losowy oraz drzewo C5.0, które posłużyły do dalszej eliminacji, aby uprościć modele.



*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.

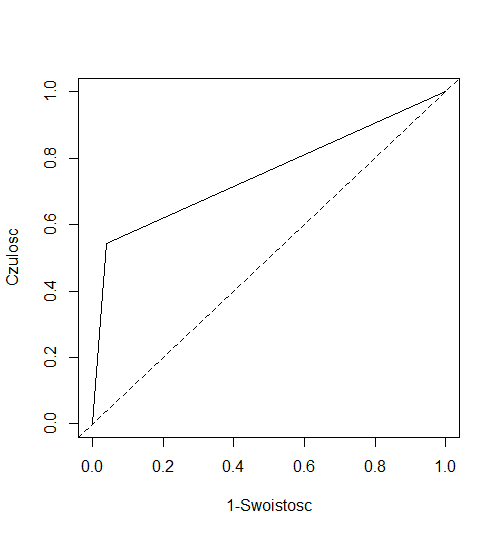
Drzewo C5.0 zostało zbudowane w oparci o predyktory: ‘duration’, ‘previous\_outcome’, ‘nr.employed’, ‘month’ oraz zmienną celu ‘response’. Dodatkowo ustawiony został parametr minCase na 75.



*Drzewo C5.0*

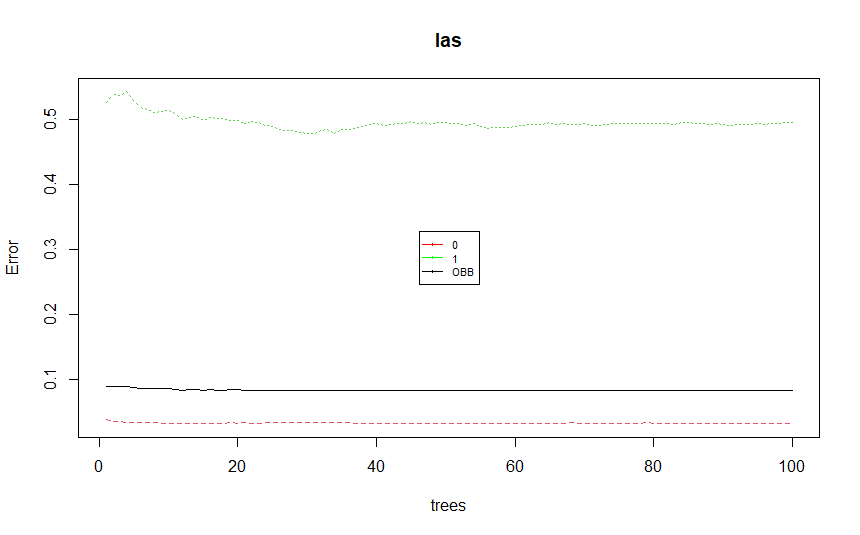
Najważniejszym predyktorem podziału został ‘previous\_outcome’. Następnie drzewo w pierwszych liściach skupiło się na podziale względem ‘duration’. Dalszy podział zależał już wyłącznie od ‘duration’ oraz ‘nr.employed’.

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%.



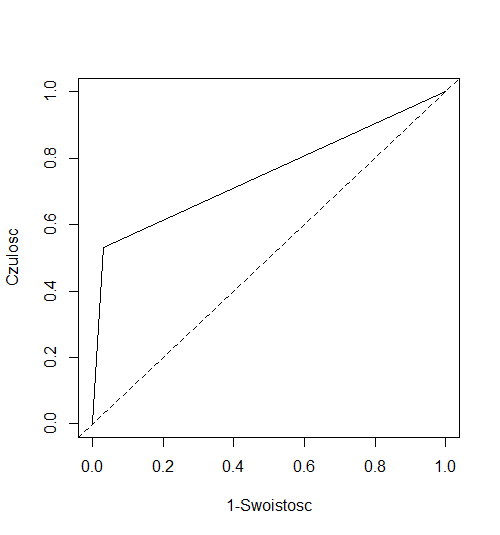
*Krzywa ROC dla drzewa C5.0 dla danych treningowych*

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.



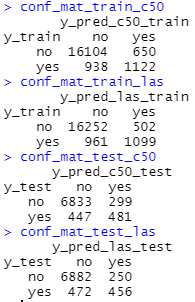
*Wykres błędy klasyfikacji*

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:

**

*Macierze klasyfikacji*

** \_ 

*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.