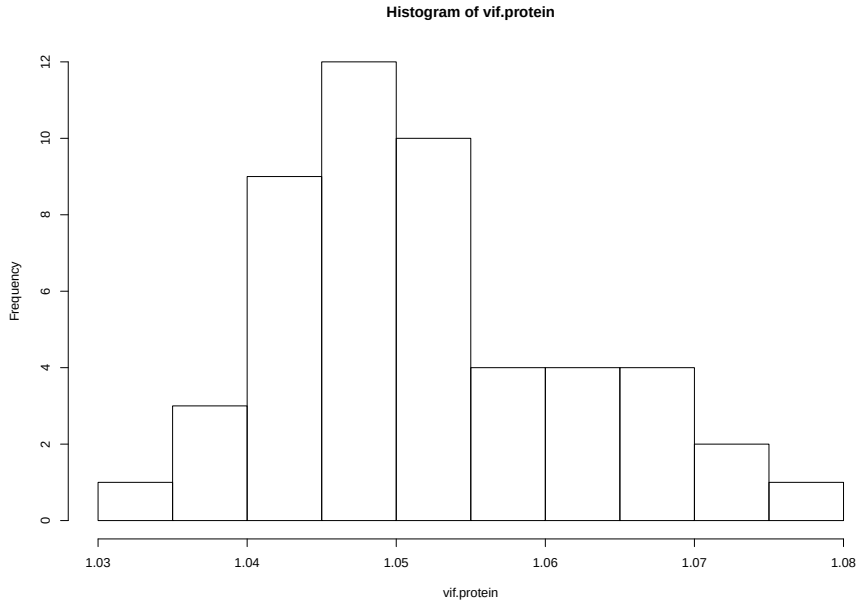
ANALIZA ZMIENNYCH OBJAŚNIAJĄCYCH ORAZ UTWORZENIE MODELU DLA DANYCH PROTEIN I CANCER

Jan Malinowski
Wydział Fizyki
372202

DANE PROTEIN

a) W segmencie data.train pliku protein.RData znalazło się 600 obserwacji dla 2000 numerycznych zmiennych objaśniających i odpowiadającej im numerycznej zmiennej objaśnianej. Wśród danych nie brakowało żadnych wartości (pola typu "NA"). Wartości wszystkich predyktorów, z wyjątkiem x1288, znalazły się pomiędzy -5, a 4,5, natomiast zakres cechy x1288 był w przybliżeniu równy (-19, 10). Zmienna objaśniająca przyjmowała w przybliżeniu od -83 do 46.



Rys. 1. Histogram VIF dla podzbioru najlepszych zmiennych objaśniających danych protein.

b) Do danych protein wybrano metodę LMForward z kryterium AIC oraz metodę lasów losowych. Metoda LMForward polega na zachłannym przeszukiwaniu przestrzeni predyktorów w celu znalezienia tych, które powodują największy spadek statystyki RSS. W każdej iteracji dodawany jest predyktor, który taki spadek powoduje jeśli uwzględniając karę wynikającą z kryterium AIC jest powoduje on wciąż najniższy spadek tej statystyki. Metoda ta jest dobrą metodą do znalezienia cech istotnie wpływających na wartość zmiennej objaśnianej. W metodzie Random Forest budowany jest las złożony z drzew, które budowane są poprzez wylosowanie pewnego podzbioru predyktorów, co pozwala często uwolnić się od wpływu najistotniejszych predyktorów. Za pomocą funkcji varImp można następnie odczytać, które predyktory mogą dobrze służyć w celu predykcji zmiennej objaśnianej, co również spełnia warunki zadania.

c)

Tabela 1. Znaleziony błąd CV_5 cross-walidacji dla danych protein.

metoda	LM Forward (AIC)	RF
CV_5	25.1	69.7

Model stworzony za pomocą metody LMForward z kryterium AIC wydaje się dokładniejszy, ponieważ ma prawie 3 krotnie mniejszy błąd cross-walidacyjny i to on został wykorzystany do znalezienia predykcji na danych z data.test z pliku protein. Wybrano model posiadający niezerowe współczynniki przy 50 zmiennych objaśniających.

d)

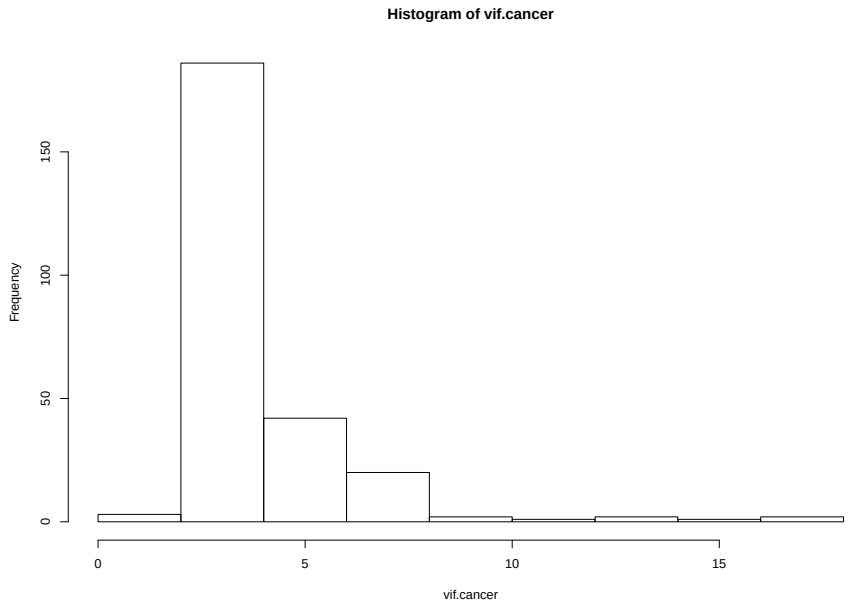
Tabela 2. Najważniejsze predyktory dla danych protein

id predyktora	x966	x603	x1678	x1288	x31

Powyższe predyktory zostały wybrane, ponieważ były 5-cioma najbardziej istotnymi cechami w modelu stworzonym za pomocą metody LMForward z kryterium AIC, a znajdowały się one również wśród 5 najczystszych węzłów modelu Random Forest.

DANE CANCER

a) W segmencie data.train pliku cancer.RData znalazły się 643 obserwacje dla 17737 numerycznych zmiennych objaśniających i odpowiadającej im numerycznej zmiennej objaśnianej. Wśród danych nie brakowało żadnych wartości (pola typu "NA"). Wartości wszystkich predyktorów znalazły się w przedziale między 2 a 14, natomiast zakres zmiennej objaśnianej wyniósł w przybliżeniu (0,2, 1,0).



Rys. 2. Histogram VIF dla podzbioru najlepszych zmiennych objaśniających danych cancer.

c) Dla danych cancer postanowiono skorzystać z metod: Ridge/Lasso, Bagging oraz boostingu na drzewach. Ridge/Lasso to podejście łączące w sobie regresję grzbietową oraz regresję typu "lasso". Obie te metody szukając współczynników regresji nakładają na nie dodatkową karę, która powoduje ich efektywne zmniejszenie. Odpowiada za to dodatkowy człon pojawiający się w funkcji minimalizacyjnej dla regresji grzbietowej związany z kwadratem estymowanego współczynnika, natomiast w regresji lasso z jego wartością bezwzględną. Podejście, które zastosowano w tej pracy korzysta z obydwu metod regresji, dla α równej 0 stosując regresję grzbietową, natomiast dla $\alpha = 1$ lasso. Metody te z powodzeniem mogą służyć do znalezienia istotnych współczynników poprzez ściąganie tych mniej istotnych do 0. Bagging jest szczególnym rodzajem metody Random Forest, wymienionej już w tej pracy, dla którego liczba cech losowanych przy budowie każdego drzewa jest równa po prostu wszystkim dostępnym predyktorom. Z tego powodu Bagging tak samo jak metoda Random Forest jest dobrą metodą do szukania istotnych zmiennych objaśniających. Boosting jest metodą, w której w kolejnych krokach budowane są drzewa korzystające z wyników z poprzednich iteracji, ważny przy tym jest parametr douczający λ. Drzewa wytworzone za pomocą tej metody mogą tak samo jak w wypadku metody Random Forest posłużyć do znalezienia najważniejszych predyktorów, poprzez znalezienie tych predyktorów, które znajdują się wysoko w ich węzłach.

Tabela 3. Znaleziony błąd CV_5 cross-walidacji dla danych cancer.

metoda	Ridge/Lasso	Bagging	Tree Boosting
$\hat{MSE}_{test}$	0.00412	0.00585	0.00573

Miarą wyboru najlepszego modelu był ponownie estymowany błąd cross-walidacji, który najmniejszy posiadał model uzyskany metodą Ridge/Lasso z parametrem $\alpha = 0.004$ oraz $\lambda = 0.296$.

d)

Najważniejsze predyktory dla danych cancer.

[1] "ENSG00000135077" "ENSG00000198092" "ENSG00000153802" "ENSG00000152672" [5]
"ENSG00000103569" "ENSG00000177300" "ENSG00000177575" "ENSG00000177483" [9]
"ENSG00000111536" "ENSG00000189252" "ENSG00000100055" "ENSG00000184557" [13]
"ENSG00000144290" "ENSG00000171094" "ENSG00000167850" "ENSG00000155875" [17]
"ENSG00000003987" "ENSG00000130584" "ENSG00000100453" "ENSG00000167207" [21]
"ENSG00000112293" "ENSG00000189419" "ENSG00000127507" "ENSG00000147481" [25]
"ENSG00000112116" "ENSG00000158714" "ENSG00000147174" "ENSG00000143365" [29]
"ENSG00000249111" "ENSG00000163959" "ENSG00000180644" "ENSG00000134460" [33]
"ENSG00000153495" "ENSG00000149635" "ENSG00000070269" "ENSG00000204351" [37]
"ENSG00000087157" "ENSG00000091137" "ENSG00000159733" "ENSG00000139572" [41]
"ENSG00000064932" "ENSG00000142698" "ENSG00000133937" "ENSG00000170525" [45]
"ENSG00000198327" "ENSG00000214900" "ENSG00000183323" "ENSG00000140368" [49]
"ENSG00000068024" "ENSG00000187672" "ENSG00000115138" "ENSG00000105374" [53]
"ENSG00000099985" "ENSG00000115607" "ENSG00000113263" "ENSG00000197587" [57]
"ENSG00000125810" "ENSG00000086967" "ENSG00000165181" "ENSG00000170801" [61]
"ENSG00000131068" "ENSG00000213171" "ENSG00000203923" "ENSG00000160679" [65]
"ENSG00000215114" "ENSG00000166435" "ENSG00000189430" "ENSG00000151023" [69]
"ENSG00000144410" "ENSG00000234560" "ENSG00000174946" "ENSG00000140678" [73]
"ENSG00000165195" "ENSG00000168405" "ENSG00000131126" "ENSG00000137090" [77]
"ENSG00000127318" "ENSG00000174606" "ENSG00000027869" "ENSG00000152463" [81]
"ENSG00000167470" "ENSG00000112685" "ENSG00000116984" "ENSG00000164509" [85]
"ENSG00000198183" "ENSG00000165566" "ENSG00000213071" "ENSG00000166507" [89]

"ENSG00000128578" "ENSG00000136630" "ENSG00000146809" "ENSG00000150361" [93]
"ENSG00000177047" "ENSG00000042781" "ENSG00000162594" "ENSG00000145649" [97]
"ENSG00000137473" "ENSG00000232040" "ENSG00000172348" "ENSG00000170832"

Najważniejsze predyktory zostały wybrane poprzez znalezienie 100 największych wartości bezwzględnych spośród współczynników przy poszczególnych zmiennych objaśniających, które zostały obliczone dla metody z najlepszym modelem (Ridge/Lasso).