Statistical Learning w praktyce

Zdawalność na amerykańskich uczelniach

Projekt zaliczeniowy

Bartłomiej Gibas, Szymon Duraj, Wojciech Frączek

09.02.2023r.

1. Cel analizy danych

Celem naszej analizy będzie przetestowanie metod, które poznaliśmy na zajęciach w celu analizy zmiennych, które wpływają na wynik zdawalności na wybranych amerykańskich uczelniach oraz wykonania klasyfikacji z użyciem regresji logistycznej oraz LDA czy uczelnia jest prywatna czy publiczna w oparciu o posiadane obserwacje.

2. Opis zbioru danych

Dane, z których skorzystaliśmy pochodzą z zestawu ISLR2 wbudowanego w RStudio z którego często korzystaliśmy również na zajęciach. Dane te można zaimplementować instalując bibliotekę ISLR2, a następnie wpisując funkcję data(College).

Dane zawierają statystyki dla dużej liczby amerykańskich uczelni wydane przez US News and World Report z 1995 roku. Zawierają one 777 obserwacji oraz 18 zmiennych:

1. *Private*

   Czynnik z poziomami ‘Nie’ i ‘Tak’ wskazujący czy uczelnia jest prywatna czy publiczną.

2. *Apps:*

   Liczba otrzymanych wniosków przez daną uczelnię.

3. *Accept:*

   Liczba zaakceptowanych wniosków.

4. *Enroll:*

   Liczba przyjętych studentów.

5. *Top10perc:*

   Procentowa ilość nowych studentów z najlepszych 10% studentów w szkole średniej.

6. *Top25perc:*

   Procentowa ilość nowych studentów z najlepszych 25% studentów w szkole średniej.

7. *F.Undergrad:*

   Liczba studentów studiów stacjonarnych.

8. *P.Undergrad:*

   Liczba studentów studiów niestacjonarnych.

9. *Outstate:*

   Czesne poza stanem – (def. czesne, które studenci płacą, gdy uczęszczają do publicznej szkoły wyższej lub uniwersytetu, który znajduje się poza ich stanem zamieszkania).

10. *Room.Board:*
    Koszty pokoju i wyżywienia.
11. *Books:*
    Szacowane koszty książek.
12. *Personal:*
    Szacowane wydatki osobiste.
13. *PhD:*
    Procentowa ilość wydziałów z doktorami
14. *Terminal:*
    Procentowa ilość wydziałów z dyplomem końcowym
15. *S.F.Ratio:*
    Wskaźnik student/wydział.
16. *Perc.alumni:*
    Procentowa ilość absolwentów którzy przekazują darowizny.
17. *Expend:*
    Wydatki dydaktyczne na studenta.
18. *Grad.Rate:*
    Współczynnik zdawalności.

3. Wybór narzędzi do analizy

Nasz zbiór danych będziemy analizować pod wieloma kątami, użyjemy następujących narzędzi:

- regresji logistycznej,
- analizy dyskryminacyjnej (LDA),
- selekcji krokowej,
- drzew regresyjnych,
- lasów losowych,
- bagging,
- XGboost.

4. Opis budowy modeli/ opis projektu

Nasze dane dla większości zmiennych są danymi ilościowymi. Zmienna Private jest jednak zapisana w postaci binarnej za pomocą poziomów 'Tak' oraz 'Nie' dlatego też będziemy chcieli zamienić je na postać binarną ilościowa (0-1). **Uwaga!** W przypadku analizy regresji logistycznej i LDA użyjemy zmiennej *Private* jako zmiennej jakościowej, nominalnej. Przedstawmy nasze dane za pomocą funkcji summary():

```
> summary(College)
 Private        Apps           Accept          Enroll       Top10perc       Top25perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
  F.Undergrad     P.Undergrad         Outstate       Room.Board       Books           Personal
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
      PhD            Terminal       S.F.Ratio      perc.alumni        Expend         Grad.Rate
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

Jak możemy zauważyć nie mamy żadnych wartości brakujących, lecz w kolumnach reprezentujących wartości procentowe PhD oraz Grad.Rate wartość maksymalna wykracza ponad 100. Musimy więc usunąć te obserwacje, gdzie pojawiają się wartości odstające. Po dokładniejszej analizie wiemy, że łącznie są to 2 obserwacje.

W większości przypadków będziemy dzielić nasze dane na zbiór treningowy oraz testowy w stosunku 80:20 lub 70:30 (w drugim przypadku będziemy to wyraźnie zaznaczali). Tak przygotowane dane mogą zostać wykorzystane do tworzenia naszych modeli.

5. Wyniki i podsumowania

Regresja logistyczna oraz LDA dla przewidywania etykiety (0 – uczelnia publiczna, 1 – uczelnia prywatna). UWAGA! W tej części zmienną *Private* traktujemy jako zmienną jakościową nominalną o dwóch poziomach.

Obserwacje dzielimy na zbiór uczący oraz testowy w proporcji 70% (treningowy) i 30% (testowy) wszystkich obserwacji, losując odpowiednie ilości wierszy z ramki danych. Budujemy model GLM, dla którego składowa losowa ma rozkład Bernoulliego z prawdopodobieństwem sukcesu p, zatem właściwe będzie użycie funkcji logitowej, jako wiążącej. Następnie z użyciem funkcji predict() przewidujemy wartości prawdopodobieństw dla obserwacji ze zbioru testowego i przypisujemy na ich podstawie etykiety zgodnie z warunkiem:

- Jeśli wyestymowane prawdopodobieństwo z modelu jest mniejsze bądź równe niż 0.5 oznaczamy jako 0 (uczelnia publiczna),
- Jeśli wyestymowane prawdopodobieństwo z modelu jest większe niż 0.5 oznaczamy jako 1 (uczelnia prywatna).

Wyznaczone etykiety porównujemy z etykietami ze zbioru testowego, wyniki prezentujemy w tablicy kontyngencji poniżej.

Przypisane etykiety	PUBLICZNA	PRYWATNA
PUBLICZNA	51	8
PRYWATNA	8	165

Zatem nasz model poprawnie przewiduje klasę uczelni z prawdopodobieństwem **0,931.** Jest to dobry wynik.

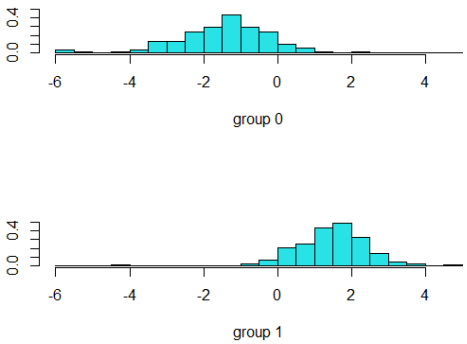
Zweryfikujmy jaki wynik otrzymamy używając analizy dyskryminacyjnej. Ponownie budujemy model jednak teraz korzystamy z wbudowanej funkcji w programie R: lda() z pakietu MASS. Używamy wciąż tych samych zbiorów: uczącego i testowego. Wynikiem przeprowadzonej procedury są m.in. wyestymowane średnie wartości w grupach dla osobnych cech:

```
Group means:
      Apps   Accept   Enroll Top10perc Top25perc F.Undergrad P.Undergrad
0 5728.776 3906.816 1632.651  22.72368  52.71053    8587.500   2017.2105
1 1924.051 1280.026  446.578  29.56010  57.66496    1799.384    429.3964
   Outstate Room.Board    Books Personal      PhD Terminal S.F.Ratio
0  6754.375   3748.276 554.8421 1654.553 76.80263 82.67105  16.88487
1 11834.355   4562.450 544.8824 1194.693 72.13043 78.55243  12.95115
  perc.alumni    Expend Grad.Rate
0    14.57237  7547.414  56.26974
1    26.23529 10308.335  69.34527
```

Z powyższego widać, że w szkołach prywatnych wskaźniki Top10perc, Top25perc, Outstate, Grad.Rate są większe. Również wydatki czy liczba absolwentów różnią się znacząco – nie jest to sprzeczne z naszą intuicją, wszak doświadczenie pokazuje, że szkoły publiczne są wybierane

```
Prior probabilities of groups:
        0         1
0.2799263 0.7200737
```

częściej i są tańsze. Prawie 28 procent obserwacji w zestawie uczącym odpowiada szkołom publicznym, pozostałe – prywatnym. Poniższe wykresy prezentują rozkład wartości liniowej dyskryminanty dla obserwacji na zbiorze uczącym. Naturalnie rozkłady są względem siebie przesunięte.



Używając funkcji predict() wyznaczamy w oparciu o dyskryminantę etykiety dla obserwacji ze zbioru testowego i porównujemy w dwuwymiarowej tablicy kontyngencji.

Przypisane etykiety	PUBLICZNA	PRYWATNA
PUBLICZNA	53	7
PRYWATNA	6	166

Prosty rachunek pozwala wywnioskować, że model poprawnie klasyfikuje z prawdopodobieństwem 0.944. **Poprawiliśmy zatem predykcję o 0.013.**

Regresja liniowa ze zmienną Grad.Rate (współczynnik zdawalności) jako zmienną objaśnianą

Chcemy przewidzieć, wyestymować współczynnik zdawalności, w oparciu o pozostałe cechy. Tym razem również zmienną, która określa status uczelni (publiczna, prywatna) oznaczymy jako jakościową nominalną.

Z użyciem funkcji *regsubsets()* wybierzemy zestaw predyktorów dla „najlepszego modelu". Zostanie on wybrany w oparciu o RSS. Po wywołaniu komendy regsubsets() otrzymujemy

macierz, której wiersze informują o liczności zmiennych w modelu, a gwiazdki, o tym, które zmienne można wybrać, aby uzyskać najlepszy model (warunkowo dla ustalonej liczby zmiennych). I tak otrzymujemy, że najlepszy model z dwoma zmiennymi zawiera predyktory: Top25perc i Outstate. Co ciekawe Outstate i Top25perc są wybierane w każdym przypadku (od i=2 do nawet i=10).

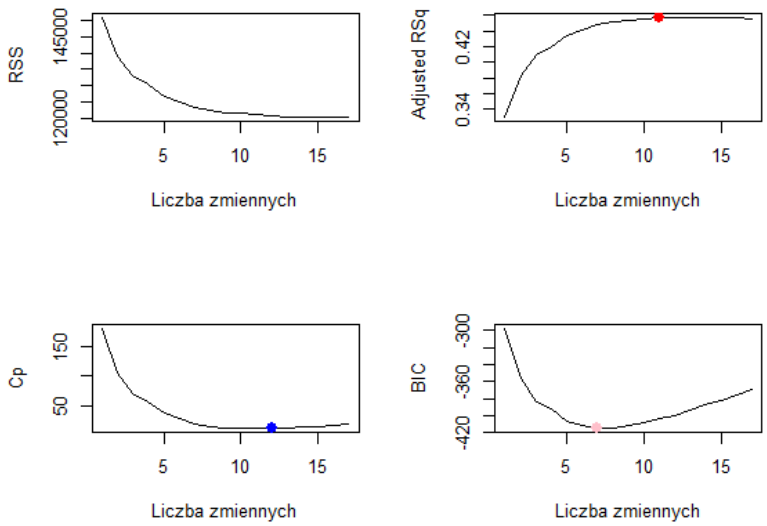
Podsumowanie *summary()* wywołane dla wyniku *regsubsets()* pozwala na przyjrzenie się innym statystykom, takim jak np. R-kwadrat.

```
> summary(regfit.full)$rsq
 [1] 0.3304424 0.3847760 0.4112326 0.4223091 0.4382053 0.4458334 0.4536784
 [8] 0.4576757 0.4606593 0.4623952 0.4643585 0.4660077 0.4664540 0.4669619
[15] 0.4673322 0.4676990 0.4678397
```

Widzimy, że wraz ze wzrostem liczby parametrów R^2 rośnie (co wydaje się naturalne). Dynamika wzrostu „wypłaszcza się" dla liczności zmiennych w modelu i=9 (dodawanie zmiennych nie polepsza wtedy znacznie dopasowania).

Ponadto możemy sprawdzić jak zachowują się RSS i R-kwadrat skorygowany o liczbę zmiennych tzw. R-squared adjusted. Wyznaczmy wartość maksymalną tego ostatniego. Wynosi **0.4575984**.

Poniższe wykresy prezentują wartości dodatkowo współczynnika Cp-Mallowsa oraz BIC w zależności od liczby zmiennych. Kolorowymi punktami zostały zaznaczone wartości minimalne (Cp, BIC) lub maksymalne (R^2 adj.).



W zależności którym kryterium będziemy się kierować można wybrać różne modele. Rozsądnie byłoby wybrać w tym wypadku model z 12 zmiennymi. Minimalizuje to

współczynnik Cp-Mallowsa oraz maksymalizuje dopasowanie R^2-adj. Współczynniki modelu prezentują się wówczas następująco:

```
> coef(regfit.full ,12)
  (Intercept)      Private1          Apps        Accept     Top25perc
32.8741370790  3.4769252332  0.0014317421 -0.0009598195  0.1619414515
  P.Undergrad     Outstate    Room.Board      Personal           PhD
-0.0016129362  0.0010507330  0.0017121974 -0.0017313232  0.1307432586
     Terminal   perc.alumni        Expend
-0.0954964812  0.2816975919 -0.0004570831
```

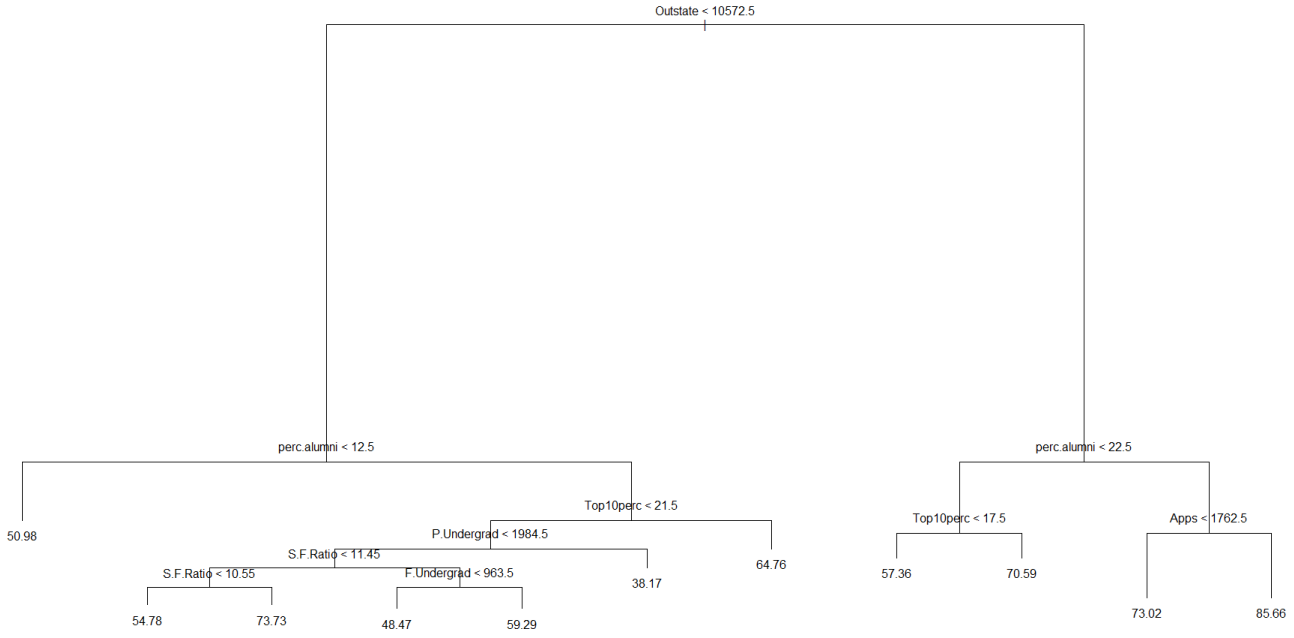
Przejrzymy jeszcze jak będzie zachowywać się model i odpowiednie współczynniki jeśli użyjemy selekcji krokowej (forward, backward selection). W funkcji *regsubset()* dodajemy specyfikację metody tzn. odpowiednio *method = „backward”* lub *method = „forward”*.Rezultaty prezentują się następująco.

```
> coef(regfit.fwd,12)
  (Intercept)      Private1          Apps        Accept     Top25perc
32.8741370790  3.4769252332  0.0014317421 -0.0009598195  0.1619414515
  P.Undergrad     Outstate    Room.Board      Personal           PhD
-0.0016129362  0.0010507330  0.0017121974 -0.0017313232  0.1307432586
     Terminal   perc.alumni        Expend
-0.0954964812  0.2816975919 -0.0004570831
> coef(regfit.bwd,12)
  (Intercept)      Private1          Apps        Accept     Top25perc
32.8741370790  3.4769252332  0.0014317421 -0.0009598195  0.1619414515
  P.Undergrad     Outstate    Room.Board      Personal           PhD
-0.0016129362  0.0010507330  0.0017121974 -0.0017313232  0.1307432586
     Terminal   perc.alumni        Expend
-0.0954964812  0.2816975919 -0.0004570831
> coef(regfit.full,12)
  (Intercept)      Private1          Apps        Accept     Top25perc
32.8741370790  3.4769252332  0.0014317421 -0.0009598195  0.1619414515
  P.Undergrad     Outstate    Room.Board      Personal           PhD
-0.0016129362  0.0010507330  0.0017121974 -0.0017313232  0.1307432586
     Terminal   perc.alumni        Expend
-0.0954964812  0.2816975919 -0.0004570831
>
```

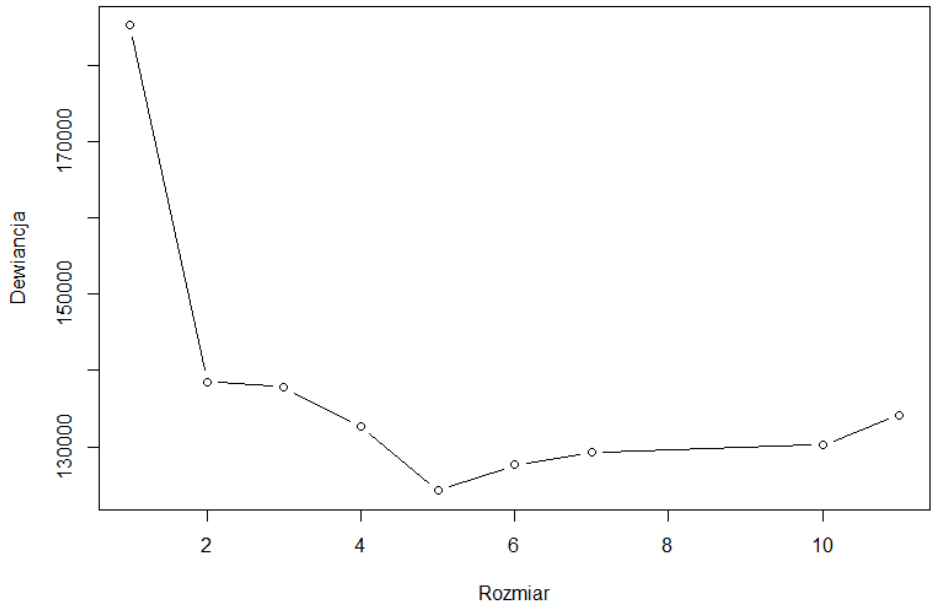
Zauważmy, że współczynnik przy zmiennej Private1 (odp. uczelniom prywatnym) jest wysoce dodatni w porównaniu do reszty zmiennych. Oznacza to, że w grupie uczelni prywatnych ten współczynnik zdawalności jest średnio wyższy niż w publicznych (kontrasty proste). Drugi współczynnik, który zwraca naszą uwagę stoi przy zmiennej procent darowizn od absolwentów. Być może wysoka zdawalność przekłada się na to, że studenci, a właściwie absolwenci dobrze wspominają uczelnie, dlatego wspierają ją finansowo. Wszystkie te spekulacje wymagają dokładniejszej analizy, m.in. względem istotności współczynników w problemie testowania.

Metody drzewiaste (drzewa regresyjne, las losowy, bagging, XGBoost)

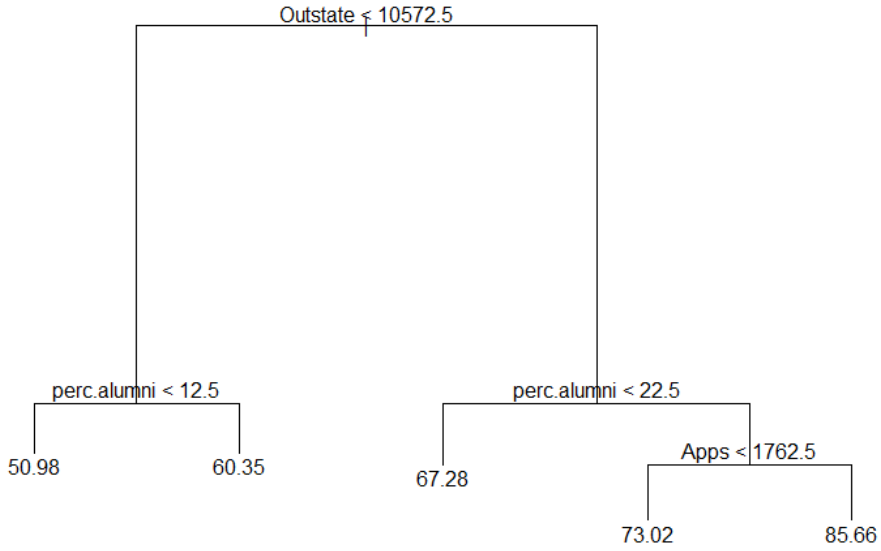
Na początek stwórzmy drzewo regresyjne, gdzie zmienną objaśnianą będzie współczynnik zdawalności (Grad.Rate).



Przy pomocy walidacji krzyżowej sprawdzamy czy warto przyciąć otrzymane drzewo.



Wykres mówi nam że najmniejsza wartość dewiancji będzie dla drzewa o 5 liściach.



Drzewo można więc interpretować w następujący sposób: zmienna Outstate jest najważniejszym czynnikiem, który decyduje o poziomie zdawalności na danej uczelni. Oznacza to, że uczelnie, gdzie studenci muszą zapłacić większe czesne, mają większy procent zdawalności. Bez względu na to ile studenci płacą za czesne procent absolwentów, którzy przekazują darowizny również ma znaczenie w kontekście zdawalności. Jeśli natomiast czesne za studia oraz procent absolwentów, którzy wpłacają darowizny są wysokie to równie wysokie znaczenie ma również liczba wniosków do przyjęcia na daną uczelnię.

Następnie użyjmy drzewa regresyjnego do naszego zbioru testowego w celu predykcji. Otrzymujemy błąd średniokwadratowy (MSE) na poziomie 159.26. Sprawdźmy zatem czy jesteśmy w stanie ten błąd zmniejszyć.

Pierwszym narzędziem, które użyjemy będzie bagging. Jest to przypadek szczególny lasy losowego, gdzie w funkcji *randomForest()* parametr mtry ustawiamy na liczbę zmiennych w danych.

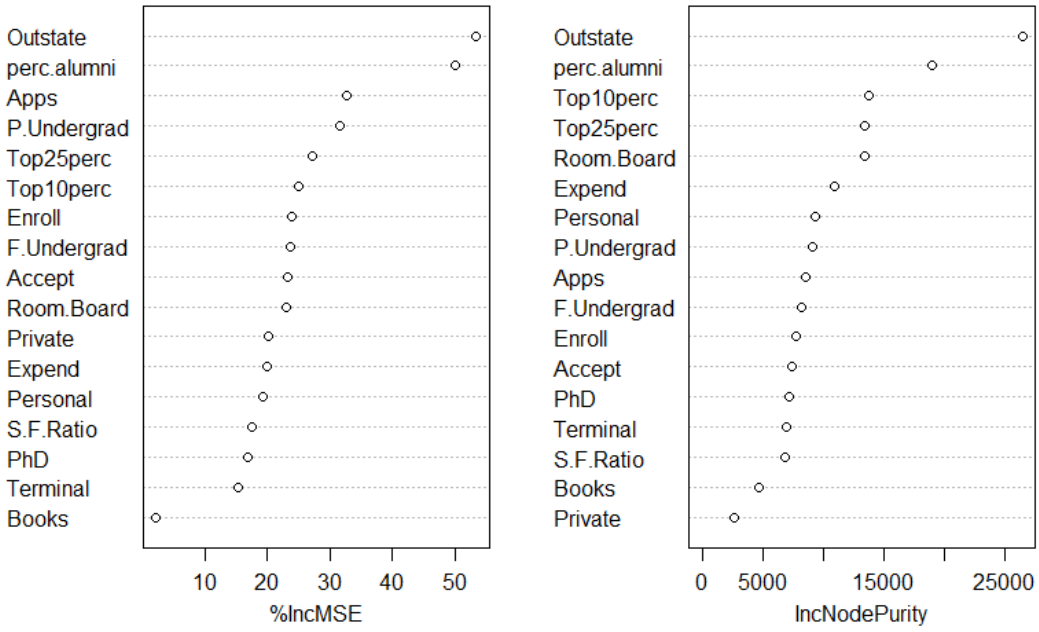
```
> bag.grad

Call:
 randomForest(formula = Grad.Rate ~ ., data = New_College, mtry = 18,      importance = TRUE, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 17

          Mean of squared residuals: 172.8129
                    % Var explained: 41.8
```

Dla baggingu nasz błąd średniokwadratowy wynosi **131.82**. Jest zatem znacznie mniejszy niż błąd przy drzewie regresyjnym.

Następnie sprawdźmy jaki błąd wyświetli nam las losowy. Po sprawdzeniu kilku różnych parametrów funkcji, najmniejszy błąd zwraca ona dla mtry=3 oraz ntry=2000.

Jest on równy 127.51.

Dla lasów losowych możemy także wyświetlić wykresy istotności czynników w naszym modelu.



Jak widzimy one również wskazują na Outstate, perc.alumni oraz Apps jako te najbardziej istotne.

Ostatnim sposobem na zmniejszenie naszego błędu średniokwadratowego jest zastosowanie XGBoostu do naszych danych. Po sprawdzeniu różnych parametrów w funkcji

xgboost udało się nam maksymalnie zmniejszyć nasz błąd do poziomu 124.25 i jest to najmniejszy błąd jaki udało się nam osiągnąć w metodach drzewiastych.