Projekt 1

Projekt 1 - Modelowanie Predykcyjne

Dane

Dane to 50 zmiennych niezależnych o nazwach od A1 do X2 i jedna zmienna zależna y - odpowiedź binarna, zmienna przyjmująca wartość 'klasa -' lub 'klasa +'. ilościowe (rozkład jednostajny) i jakościowe (4 poziomy) zmienne

```
head(zbior_uczacy, n = 5)
```

```
##         y A1 B1 C1   D1 E1 F1   G1   H1   I1   J1 K1 L1   M1   N1   O1 P1
## 1 klasa -  A  B  A  7.2  B  B 17.3 96.8  8.9 62.7  A  A 98.2 48.1  5.3  C
## 2 klasa -  B  A  B 99.6  D  C 26.2 64.1 17.8  4.5  D  A 40.6 11.9 30.3  A
## 3 klasa +  B  A  B 79.7  C  D  2.5 47.2 62.8 40.0  C  C 17.9 83.3 61.1  C
## 4 klasa +  B  B  B 86.3  C  C 25.8 48.1 80.6 89.2  B  D 59.6 98.2 31.0  A
## 5 klasa +  A  A  A 67.6  C  C 67.3 41.0 10.9 57.5  C  C 45.8 85.1 10.6  C
##     Q1 R1   S1 T1   U1   V1   W1   X1 Y1 Z1   A2   B2 C2 D2   E2   F2   G2
## 1  2.3  B 17.0  B 37.6 61.9 20.9 54.3  B  C 43.8 18.7  C  D 84.3 33.9 86.0
## 2 20.5  B  4.0  A 18.6 86.5 11.4 45.8  B  A 35.7  8.6  C  A 72.2  8.1 61.8
## 3 24.4  B 38.3  D 14.7 51.2 84.7 79.0  B  C 37.5 94.2  C  C 62.5 63.4 88.5
## 4 13.0  A 34.0  B 17.2  8.0 24.1 16.4  B  C 72.5 40.1  B  D 26.4 14.6 64.8
## 5 67.5  B 48.9  D 62.1 42.1 46.1 47.7  A  A 89.0 54.5  C  A 59.3 67.6 88.9
##     H2 I2 J2   K2   L2   M2   N2 O2 P2 Q2 R2   S2   T2   U2   V2   W2   X2
## 1 92.8  B  B 37.3  8.4 97.8 70.2  B  C  A  B 81.4 53.9 60.9 51.0 42.1 39.1
## 2 72.9  B  A  4.6 54.4 63.9 11.5  A  A  A  B 37.5 27.5 60.9 80.1 45.6 21.6
## 3 53.1  B  B  9.6 38.6  4.7 48.5  B  B  A  B 99.9  7.8 43.8 68.7 38.4 24.1
## 4 72.6  A  B 98.8 88.0  7.9 75.6  A  D  B  B 77.8 90.7 55.5  4.5 65.2 78.9
## 5  2.6  A  A 96.7 22.9 78.4 97.3  B  C  B  A 37.4 40.5 18.3 41.8 71.5 13.1
```

W zbiorze znajdują się zarówno zmienne ilościowe (około 30) oraz jakościowe (około 20). Dodatkowo przy użyciu funkcji summary możemy dokładniej przyjrzeć sie danym w zbiorze.

```
#summary(zbior_uczacy)
#zbior_uczacy_numeric <- model.matrix(y~., zbior_uczacy)
zbior_uczacy_numeric <- as.character(zbior_uczacy)
```

Zmienne ilościowe mają rozkład jednostajny

Selekcja zmiennych

Etap selekcjowania zmiennych wykonujemy aby pozbyć się możliwych szumów i zbyt wysokiej wariancji, dzięki temu możemy otrzymać bardziej stabilne rozwiązanie.

Test niezależności Chi Square

Test niezależności (Pearsona) chi-kwadrat jest jednym z najpopularniejszych testów statystycznych. Mała liczba założeń oraz prostota przeprowadzenia powoduje jego częste wykorzystanie w analizie danych. Zmienne wybrane spośród wszystkich to takie których p-value < 0.05, czyli takie które wykazują istotną zależność między daną zmienną a zmienną y.

```
##             y            A1            B1            E1            F1
##  0.000000e+00  1.166366e-02  1.356103e-02 5.659211e-285  8.234509e-03
##            J1            L1            M1            Q1            W1
##  1.107462e-26  4.656045e-02  0.000000e+00  6.923376e-25  0.000000e+00
##            F2            L2            N2            P2            T2
##  2.572146e-28  4.665200e-02  2.149561e-03 4.603660e-260  3.478023e-21
##            U2
##  3.796093e-40
```

Selekcja zmiennych przy użyciu lasów losowych

Dzięki bibliotece 'caret', można również wyznaczyć "ważność" zmiennej.

```
##      Overall
## A1  49.75434
## B1  49.74931
## C1  50.30031
## D1 582.29508
## E1 677.60082
## F1 223.76957
```

Kolejne wagi to: 1985.18258, 1912.84356, 913.79699, 842.06419, 839.62006, 779.46947, 771.00936, 681.57905, 623.21855, 590.72945 a odpowiadające im zmienne to: M1, W1, U2, Q1, T2, F2, J1, E1, P2, H2.

Selekcja zmiennych metodą BIC

Jedną z metod stosowanych do predykcji zmiennych jest regresja logistyczna. Budując model regresji decydujemy, ile i które zmienne o charakterze jakościowym i ilościowym do niego włączymy. Wiadomo, że dopasowanie modelu na ogół poprawia się wraz z kolejnym dodanym regresorem, jednak zbyt rozbudowane modele mogą być przeuczone. Aby ograniczyć liczbę zmiennych, możemy korzystać z kryteriów wyboru modelu AIC lub BIC, które porównując modele między sobą biorą pod uwagę również ich wymiar.

Kryterium BIC

Model regresji liniowej biorący pod uwagę wszystkie dostępne zmienne objaśniające.

Istotne zmienne zostały wskazane przez dopasowany model (p-value jest mniejsze od 0.05 co wskazuje istotną zależność między zmienną objaśnianą a zmienną objaśnianą) to : A1, B1, E1, F1, M1, P1, Q1, W1, Z1, O2, P2, T2, U2.

Używając kryterium BIC funkcja step wskazała 7 naistotniejszych dla modelu zmiennych czyli E1, M1, Q1, W1, P2, T2 i U2.

Step: AIC=63225.39 y ~ E1 + M1 + Q1 + W1 + P2 + T2 + U2

```
   Df Deviance   AIC
```

63096 63225 - M1 1 63156 63275 - T2 1 63739 63858 - Q1 1 63743 63862 - U2 1 63812 63931 - P2 3 64481 64578 - E1 3 64605 64703 - W1 1 64785 64904

Klasyfikacja

Klasyfikator glm

W poprzednim kroku wybrane zostały istotne dla modelu zmienne. Mając odpowiednio dobrany model można przystąpić do klasyfikacji. Zbiór danych został podzielony w stosunku 1:5, gdzie zbiór treningowy posiada 40000 obserwacji a zbiór testowy 10000. Dobrany został model ze zmiennymi wyznaczonymi w poprzednim punkcie czyli, E1, M1, Q1, W1, P2, T2 i U2. Na jego podstawie przewidziane zostały prawdopodobieństwa przynależności do klasy "+" dla 10000 testowych obserwacji. Wyniki zostały posortowane od największego prawdopodobieństwa do najmniejszego. Następnie próbka 1000 najwyższych wyników została porównana z autentyczną klasą obserwacji. Klasyfikator glm osiągnął skuteczność 0.756%.

sortowanie indexów najlepszych wyników najwyzszych prawdopodobieństw 1000 wybieram i sprawdzam klase+

```
## [1] 0.756
```

Klasyfikator random forest

W tym wypadku zastosowano klasyfikator random forest i postępowano jak wyżej. Klasyfikator random forest również osiągnął skuteczność 0.756%.

```
## [1] 0.756
```

Wyniki - zbiór testowy

Ostateczny zbiór testowy został poddany klasyfikacji glm a wyniki można znaleźć w pliku magdalena_baracz.txt.

zbior_testowy